Language and Ontology (LangOnto2) & Terminology and Knowledge Structures (TermiKS)

Workshop Programme

Session 1
09:00 – 09:30 – Introductory Talk
Pamela Faber, *The Cultural Dimension of Knowledge Structures*

09:30 – 10:10 – Introductory Talk
John McCrae, *Putting ontologies to work in NLP: The lemon model and its applications*

10:10 – 10:30
Gregory Grefenstette, Karima Rafes, *Transforming Wikipedia into an Ontology-based Information Retrieval Search Engine for Local Experts using a Third-Party Taxonomy*

10:30 – 11:00 Coffee break

Session 2
11:00 – 11:30
Juan Carlos Gil-Berrozpe, Pamela Faber, *Refining Hyponymy in a Terminological Knowledge Base*

11:30 – 12:00
Pilar León-Araúz, Arianne Reimerink, *Evaluation of EcoLexicon Images*

12:00 – 12:30
Špela Vintar, Larisa Grčić Simeunović, *The Language-Dependence of Conceptual Structures in Karstology*

12:30 – 13:00
Takuma Asaishi, Kyo Kageura, *Growth of the Terminological Networks in Junior-high and High School Textbooks*

13:30 – 14:00 Lunch break

Session 3
14:00 – 14:20
Gabor Melli, *Semantically Annotated Concepts in KDD's 2009-2015 Abstracts*

14:20 – 14:40
Bhaskar Sinha, Somnath Chandra, *Neural Network Based Approach for Relational Classification for Ontology Development in Low Resourced Indian Language*

14:40 – 15:00
Livy Real, Valeria de Paiva, Fabricio Chalub, Alexandre Rademaker, *Gentle with the Gentilics*

15:00 – 15:30
Dante Degl'Innocenti, Dario De Nart and Carlo Tasso, *The Importance of Being Referenced: Introducing Referential Semantic Spaces*

15:30 – 16:00
Haizhou Qu, Marcelo Sardelich, Nunung Nurul Qomariyah and Dimitar Kazakov, *Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events*

16:00 – 16:30 Coffee break

Session 4
16:30 – 16:50
Christian Willms, Hans-Ulrich Krieger, Bernd Kiefer, ×-*Protégé An Ontology Editor for Defining Cartesian Types to Represent n-ary Relations*

16:50 – 17:10
Alexsandro Fonseca, Fatiha Sadat and François Lareau, *A Lexical Ontology to Represent Lexical Functions*

17:10 – 17:40
Ayla Rigouts Terryn, Lieve Macken and Els Lefever, *Dutch Hypernym Detection: Does Decompounding Help?*

17:40 – 18:30 Discussion and Closing

Editors

Fahad Khan	Istituto di Linguistica Computazionale “A. Zampolli” - CNR, Italy
Špela Vintar	University of Ljubljana, Slovenia
Pilar León Araúz	University of Granada, Spain
Pamela Faber	University of Granada, Spain
Francesca Frontini	Istituto di Linguistica Computazionale “A. Zampolli” - CNR, Italy
Artemis Parvizi	Oxford University Press
Larisa Grčić Simeunović	University of Zadar, Croatia
Christina Unger	Bielefeld University, Germany

Workshop Organizers/Organizing Committee

Fahad Khan	Istituto di Linguistica Computazionale “A. Zampolli” - CNR, Italy
Špela Vintar	University of Ljubljana, Slovenia
Pilar León Araúz	University of Granada, Spain
Pamela Faber	University of Granada, Spain
Francesca Frontini	Istituto di Linguistica Computazionale “A. Zampolli” - CNR, Italy
Artemis Parvizi	Oxford University Press
Larisa Grčić Simeunović	University of Zadar, Croatia
Christina Unger	Bielefeld University, Germany

Workshop Programme Committee

Guadalupe Aguado-de-Cea	Universidad Politécnica de Madrid, Spain
Amparo Alcina	Universitat Jaume I, Spain
Nathalie Aussenac-Gilles	IRIT, France
Caroline Barrière	CRIM, Canada
Maja Bratanić	Institute of Croatian Language and Linguistics, Croatia
Paul Buitelaar	Insight Centre for Data Analytics, Ireland
Federico Cerutti	Cardiff University
Béatrice Daille	University of Nantes, France
Aldo Gangemi	LIPN University, ISTC-CNR Rome
Eric Gaussier	University of Grenoble, France
Emiliano Giovannetti	ILC-CNR
Ulrich Heid	University of Hildesheim, Germany
Caroline Jay	University of Manchester
Kyo Kageura	University of Tokio, Japan
Hans-Ulrich Krieger	DFKI GmbH
Roman Kutlak	Oxford University Press
Marie-Claude L'Homme	OLST, Université de Montréal, Canada
Monica Monachini	ILC-CNR
Mojca Pecman	University of Paris Diderot, France
Silvia Piccini	ILC-CNR
Yuan Ren	Microsoft China
Fabio Rinaldi	Universität Zürich, Switzerland
Irena Spasic	University of Cardiff
Markel Vigo	University of Manchester

Table of contents

Author Index

Introduction

This joint workshop brings together two different but closely related strands of research. On the one hand it looks at the overlap between ontologies and computational linguistics and on the other it explores the relationship between knowledge modelling and terminologies. In particular the workshop aims to create a forum for discussion in which the different relationships and commonalities between these two areas can be explored in detail, as well as presenting cutting edge research in each of the two individual areas.

A significant amount of human knowledge can be found in texts. It is not surprising that languages such as OWL, which allow us to formally represent this knowledge, have become more and more popular both in linguistics and in automated language processing. For instance ontologies are now of core interest to many NLP fields including Machine Translation, Question Answering, Text Summarization, Information Retrieval, and Word Sense Disambiguation. At a more abstract level, however, ontologies can also help us to model and reason about phenomena in natural language semantics. In addition, ontologies and taxonomies can also be used in the organisation and formalisation of linguistically relevant categories such as those used in tagsets for corpus annotation. Notably also, the fact that formal ontologies are being increasingly accessed by users with limited to no background in formal logic has led to a growing interest in developing accessible front ends that allow for easy querying and summarisation of ontologies. It has also led to work in developing natural language interfaces for authoring ontologies and evaluating their design.

Additionally in recent years there has been a renewed interest in the linguistic aspects of accessing, extracting, representing, modelling and transferring knowledge. Numerous tools for the automatic extraction of terms, term variants, knowledge-rich contexts, definitions, semantic relations and taxonomies from specialized corpora have been developed for a number of languages, and new theoretical approaches have emerged as potential frameworks for the study of specialized communication. However, the building of adequate knowledge models for practitioners (e.g. experts, researchers, translators, teachers etc.), on the one hand, and NLP applications (including cross-language, cross-domain, cross-device, multi-modal, multi-platform applications), on the other hand, still remains a challenge.

The papers included in the workshop range across a wide variety of different areas and reflect the strong inter-disciplinary approach, which characterises both areas of research. In addition we are very happy to include two invited talks in the program presented by authorities in their respective fields: Pamela Faber from the field of terminology, and John McCrae, an expert on linguistic linked data and the interface between NLP and ontologies.

Thanks:
The organising committee of the workshop would like to thank the authors for contributing to the success of the workshop, and the LREC organisers for their advice and help.

The LangOnto2 and TermiKS Organising Committee

The Cultural Dimension of Knowledge Structures

Pamela Faber
University of Granada, Spain
Email: pfaber@ugr.es

Frame-Based Terminology (FBT) is a cognitive approach to terminology, which directly links specialized knowledge representation to cognitive linguistics and cognitive semantics (Faber 2011, 2012, 2014). More specifically, the FBT approach applies the notion of *frame* as a schematization of a knowledge structure, which is represented at the conceptual level and held in long-term memory and which emphasizes both hierarchical and non-hierarchical conceptual relations. Frames also link elements and entities associated with general human experience or with a particular culturally embedded scene or situation. Culture is thus a key element in knowledge structures.

Cultural frames are directly connected to what has been called 'design principle' (O'Meara and Bohnemeyer 2008), 'template', 'model', 'schema' or 'frame '(Brown 2008; Burenhult 2008, Cablitz 2008, Levinson 2008). In EcoLexicon, a frame is a representation that integrates various ways of combining semantic generalizations about one category or a group of categories. In contrast, a *template* is a representational pattern for individual members of the same category. Burenhult and Levinson (2008: 144) even propose the term, *semplate*, which refers to the cultural themes or linguistic patterns that are imposed on the environment to create, coordinate, subcategorize, or contrast categories.

Although rarely explored, cultural situatedness has an impact on semantic networks, which reflect differences between terms used in closely related language cultures. Nevertheless, the addition of a cultural component to term meaning is considerably more complicated than the inclusion of terms that designate new concepts specific to other cultures. One reason for this is that certain conceptual categories are linked, for example, to the habitat of the speakers of a language and derive their meaning from the geographic and meteorological characteristics of a given geographic area or region. This paper explains the need for typology of cultural frames or profiles linked to the most prominent semantic categories. As an example, the terms for different types of local wind are analyzed and a set of meaning parameters are established that structure and enrich the cultural schemas defining meteorological concepts. These parameters highlight the cultural dimension of wind as a meteorological force.

Bio:
Pamela Faber lectures and works in terminology, translation, lexical semantics, and cognitive linguistics. She holds degrees from the University of North Carolina at Chapel Hill, the University of Paris IV, and the University of Granada where she has been a full professor in Translation and Interpreting since 2001. She is the director of the LexiCon research group, with whom she has carried out various research projects on terminological knowledge bases, conceptual modeling, and ontologies, funded by the Spanish government. One of the results of these projects and the practical application of her Frame-based Terminology Theory is EcoLexicon (ecolexicon.ugr.es), a terminological knowledge base on environmental science. She has published more than 100 articles, book chapters, and books, and has been invited to present her research in universities in Madrid, Barcelona, Leipzig, Brussels, Zagreb, Mexico D.F., Lodz, and Strasbourg, among other places. She serves on the editorial and scientific boards of several journals, such as Fachsprache, Language Design,Terminology, and the International Journal of Lexicography. She is also a member of the AENOR standardization committee.

References:
Brown, P. 2008. Up, Down, and Across the Land: Landscape Terms, Place Names, and Spatial Language in Tzeltal. *Language Sciences* 30: 151–181.
Burenhult, N. 2008. Streams of Words: Hydrological Lexicon in Jahail. *Language Sciences* 30: 182–199.
Burenhult, N. and S. C. Levinson. 2008. Language and Landscape: A Crosslinguistic Perspective. *Language Sciences* 30: 135–150.
Cablitz, G. 2008. When 'What' is 'Where': A Linguistic Analysis of Landscape Terms, Place Names and Body Part Terms in Marquesan (Oceanic, French Polynesia). *Language Sciences* 30: 200–226.
Faber, P. 2011. The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception–Action Interface. *Terminology* 17 (1): 9–29.
Faber, P. (ed.) 2012. *A Cognitive Linguistics View of Terminology and Specialized Language.* Berlin/New York: De Gruyter.
Faber, P. 2014. Frames as a Framework for Terminology. In *Handbook of Terminology, Vol. 1* edited by H. J. Kockaert and F. Steurs, 14-33. Amsterdam/Philadelphia: John Benjamins.
Levinson, S. C. 2008. Landscape, Seascape and the Ontology of Places on Rossel Island, Papua New Guinea. *Language Sciences* 30 (2/3): 256–290.
O'Meara, C. and J. Bohnemeyer. 2008. Complex Landscape Terms in Seri. *Language Sciences* 30: 316–339.

Putting Ontologies to Work in NLP: The *lemon* Model and its Applications

John McCrae
Insight Centre for Data Analytics, National University of Ireland
Email: john@mccr.ae

From the early development of lexicons such as WordNet it has been a goal to record rich information about the meanings of words and how they relate. In particular, there has been an ambition to provide full and formal definitions of concepts so that they can be clearly disambiguated and understood. Moreover, it is important to be able to represent the meaning of ontological concepts relative to how they are expressed in natural language and this syntax-ontology mapping is still poorly understood. To this end, we developed the *lemon* model, firstly in the Monnet project and secondly in the context of the W3C OntoLex Community Group and I will discuss how this model can express the mapping of words into ontological contexts. This model has found application is several practical areas, which I will describe in detail, but has so not yet achieved the goal of unifying ontological reasoning with natural language processing and I will describe the next steps I envision to achieve this goal.

Bio:
John McCrae completed his PhD at the National Institute of Informatics, Tokyo, while contributing to the BioCaster system for detecting disease outbreaks by reading texts in East Asian languages. From 2009-2015 he was at Bielefeld University, where he played a leading role in the development of the lemon (Lexicon Model for Ontologies), a major contribution to the representation of semantics relative to natural language, which is now being used by most relevant research groups, notably in recent efforts by the Global WordNet Association to standardize an interlingual index of concepts. Secondly, out of the work on this topic he has been instrumental in creating the topic of linguistic linked open data as a major research theme, which has been supported by over a dozen workshops and events and was a major theme of the previous Language Resource and Evaluation Conference (LREC). This topic led to the Lider project, which used linguistic linked open data as an enabler for content analytics in enterprise. Since August 2015, he has been working at the Insight Centre for Data Analytics at the National University of Ireland, Galway, where his work has focussed on linguistic linked data, reproducible research, ontologies and NLP for minority languages.

Transforming Wikipedia into an Ontology-based Information Retrieval Search Engine for Local Experts using a Third-Party Taxonomy

Gregory Grefenstette, Karima Rafes
Inria Saclay/TAO and BorderCloud, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
{gregory.grefenstette,karima.rafes}@inria.fr, karima.rafes@bordercloud.com

Abstract

Wikipedia is widely used for finding general information about a wide variety of topicss. Its vocation is not to provide local information. For example, it provides plot, cast, and production information about a given movie, but not showing times in your local movie theatre. Here we describe how we can connect local information to Wikipedia, without altering its content. The case study we present involves finding local scientific experts. Using a third-party taxonomy, independent from Wikipedia's category hierarchy, we index information connected to our local experts, present in their activity reports, and we re-index Wikipedia content using the same taxonomy. The connections between Wikipedia pages and local expert reports are stored in a relational database, accessible through as public SPARQL endpoint. A Wikipedia gadget (or plugin) activated by the interested user, accesses the endpoint as each Wikipedia page is accessed. An additional tab on the Wikipedia page allows the user to open up a list of teams of local experts associated with the subject matter in the Wikipedia page. The technique, though presented here as a way to identify local experts, is generic, in that any third party taxonomy, can be used in this to connect Wikipedia to any non-Wikipedia data source.

Keywords: ontology, pivot ontology, Wikipedia, local search

1. Introduction

Wikipedia is a large multilingual crowd-sourced encyclopedia, covering millions of topics. Its goal is to provide a "compendium of knowledge" incorporating "elements of general and specialized encyclopedias, almanacs, and gazetteers."[1] Possessing over 5 million pages in the English version, Wikipedia involves two principal parties: hundreds of thousands of human editors who contribute new information and edit existing pages, and anonymous internet users who use Wikipedia as a resource for finding facts involving general knowledge. To help end users find information, Wikipedia contributors can also add general categories "to group together pages on similar subjects."[2] The categories added to Wikipedia is a graph and can contain cycles (Zesch and Gurevych, 2007) but it can be transformed into a taxonomy Ponzetto and Navigli (2009).

But one taxonomy is not good for all purposes (McDermott, 1999, Veltman, 2001), and different domains produce different taxonomies covering some of the same topics. For example, the Association for Computing Machinery publishes every four years or so its taxonomy (Santos and Rodrigues, 2009) that is used to annotate computer science articles for many conferences. This taxonomy or other available taxonomies are difficult to merge with Wikipedia category set "because the concepts of source taxonomies are in different granularity, different structure, ambiguous, and partly incompatible" (Amini, *et al.*, 2015). Though merging taxonomies is difficult (Swartout and Tate, 1999), taxonomies are useful for providing a hierarchic, faceted view on data (Hearst, 2006), and because they limit the conceptual space to the principal terms that interest the group for which the taxonomy was created.

The problem we address here is how to combine the

(i) domain-directed usefulness of a taxonomy, created for one group, with the

(ii) familiarity of use and search that Wikipedia offers, without altering the generality of Wikipedia's content, with a

(iii) local source of data independent from Wikipedia.

The use case we describe involves finding local experts for mathematical or computer science problem, the domain areas of the ACM classification, using Wikipedia as a search engine. Demartini (2007) proposed using Wikipedia to find experts by extracting expertise from Wikipedia describing people, or from Wiki editors that edited page corresponding to a given topic. West et al. (2012) tried to characterize what makes Wikipedia editors be considered as experts in a given field. Our approach differs from this previous work since it allows us to connect Wikipedia content and subject matter to people who do not appear in Wikipedia either as subjects or editors.

2. Finding Experts with Wikipedia

In large organizations, such as multinational corporations, universities, and even large research centers, it can be difficult to know who is an expert about a given subject. A common response to this problem is to create an ontology of expertise and manually or automatically assign, to experts, labels from this ontology. Beyond the cost or effort needed to produce

[1] en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_an_encyclopedia
[2] en.wikipedia.org/wiki/Help:Category

the ontology, this solution creates an additional problem. Once such a knowledge base of experts exists, the searcher still has to internalize the ontology labels and their meaning in order to find the expert. The difficulty the user faces explains why some expert knowledge bases are found useful for one division in a large organization but be useless for another division which does not share the same terminology or perspective (Hahn & Subrami, 2000). We propose a method for finding experts that exploits a pivot ontology, hidden from the user, and which allows the searcher to browse Wikipedia, a resource that he or she is probably familiar with, in order to find his or her local expert.

We suppose that the user of our system is familiar with Wikipedia, but would like to find some local experts to solve a given problem. We also suppose that the user is able to find the Wikipedia page concerning the problem they have in mind, i.e., the user knows how to navigate in the semantic space implicit in Wikipedia.

Since Wikipedia contains information about historical figures and entities of general and not necessarily local interest, we need to connect our set of local experts to the pages of Wikipedia. It would run counter to the philosophy of Wikipedia to alter its content to cover local events and local unhistorical information. One solution, then, would be to copy all of Wikipedia into a local wiki, and then we could modify the subject pages as we wished to include names of local experts, laboratories, companies that are concerned with every page of interest. Although this would solve the problem of connecting Wikipedia pages to local experts, this solution has a few drawbacks: (i) users would no longer be inside the "real Wikipedia" but in a copy, which would have to be kept up to date regularly, and accessed with a different URL, (ii) since Wikipedia contains over 4 million pages in its English version, one would have to make as many decisions as to which pages to modify.

Figure 1 An extra tab appears while browsing Wikipedia once the expert finder module is activated by a logged-in user. Here the tab is labeled 'Inria' since we are searching for experts inside the Inria Institute

We have developed an automated technique that uses the same URL for Wikipedia, and which automatically establishes the connection between local expert and Wikipedia pages, using the ACM subject classification as a pivot ontology that is used to index the local experts and Wikipedia pages into the same semantic space, and a Wikipedia gadget (plugin) that adds a discrete tab to Wikipedia pages. To find a local expert for a given problem, the user searches the subject of interest in Wikipedia. The extra tab on each Wikipedia page can be clicked to reveal a list of experts concerning the some topic mentioned in the page.

2.1 Example

Before explaining the mechanisms and ontological resources involved, let us see an example. In this example, our local experts are any of the research teams in the French nation-wide computer science public research institute, Inria[3]. The Inria Institute employs 3600 scientists spread over 200 research teams, each specializing in some branch of computer science and mathematics. In our example, finding an expert will mean finding a team who is competent to answer questions about a given subject.

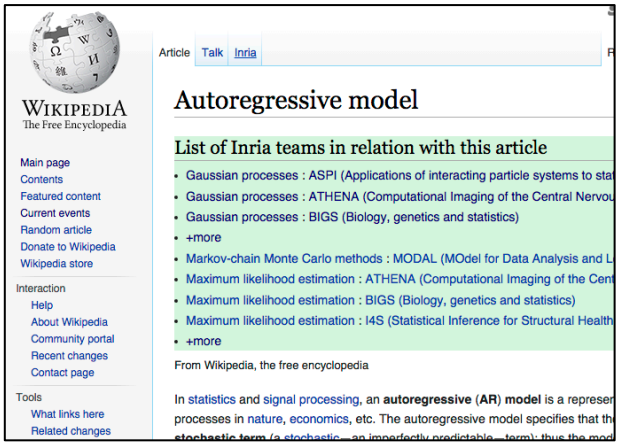

Figure 2. Clicking on the tab expands a box that shows the Inria teams that are associated with subjects on the Wikipedia page. For example, one subject mentioned on the page is "Gaussian processes" and 3 Inria teams that work in this domain are listed ASPI, ATHENA, and BIGS, with their expanded team names. Clicking on "Gaussian processes" goes to an ACM 2012 ontology page for this concept. Clicking on a team name goes to a web page from their 2014 annual report where a project involves "Gaussian processes". On this page, the user can find team members who are experts in the area. (See Figures 3 and 4)

In our solution, when someone is looking for an expert in a given subject inside the Inria institute, an additional tab appears to the identified user on the Wikipedia interface[4]. In Figure 1, the tab appears with the label "Inria" to the right of the "Article" and "Talk" tabs above the article title. Clicking on the tab expands a box listing the ACM subjects found in the articles and the Inria research teams treating those subjects, as seen in Figure 2.

[3] https://en.wikipedia.org/wiki/Inria

[4] This tab appears when the user is logged in, and has the resource module for expert finding, this tab appears while the user browses http://en.wikipedia.org. Wiki resource modules are additions that anyone can develop and activate. See further explanations at https://www.mediawiki.org/wiki/ResourceLoader/Developing_with_ResourceLoader

Both subjects and teams (see Figure 2) are linked to pages outside Wikipedia: if the user clicks on the subject name (e.g. “Maximum Likelihood Method”) in the expanded box, they are sent to the corresponding page in the ACM classification; if they click on the Inria team name (e.g. ATHENA), they are sent to the page in the team’s annual report which mentions the subject.

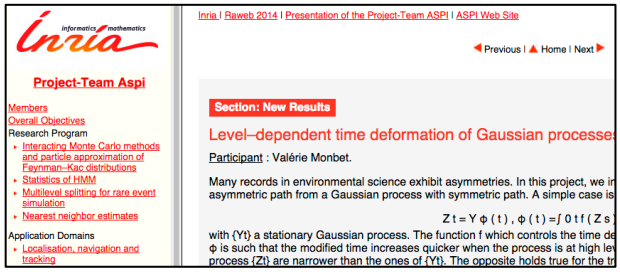

Figure 3. The Inria team annual report page found by following the link in the expert finding module. There the user sees that “Valerie Moribet” is involved in a project using Gaussian processes, and would probably be a good expert contact for finding out about Auto-regressive models.

Thus, Wikipedia has become a search engine with the user browsing towards their query (here “Autoregressive Models”, with the pull down expert box corresponding the search engine results page, leading to outside content. The user can find the Wikipedia article closest to his or her concern, and use the expert finding tab to find local experts who know about the subjects on the page.

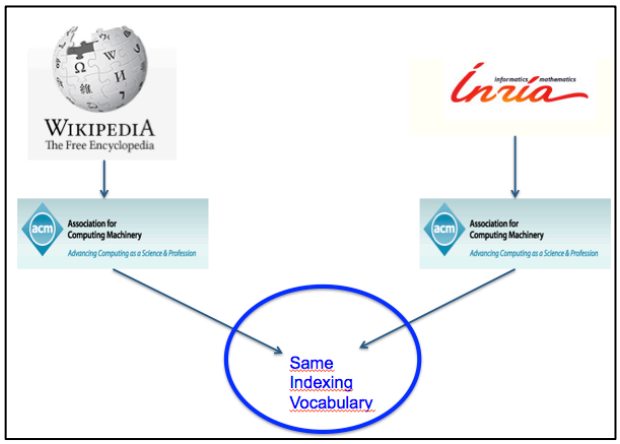

Figure 4. The ACM 2012 is the latest version of a computer science ontology created by the Association for Computing Machinery. "Gaussian processes" is a hyponym of “Theory of Computation”. This specific concepts provides a link between the Wikipedia web page on “Autoregressive models” and the Inria team experts, but the searcher need not understand the ACM hierarchy nor know its contents in order to make the connection.

This seems to us a natural and intuitive method for finding experts that obviates the need for learning the underlying, pivot ontology by which the experts are connected to the topic page. Even if the connecting ontology terms are explicitly displayed in the results, the user need not ever use them in an explicit query.

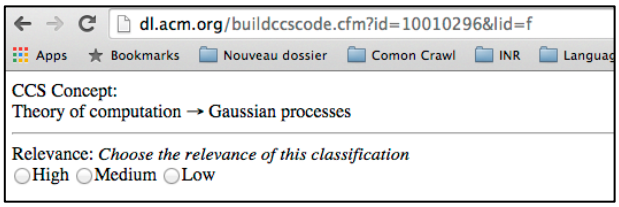

Figure 5. The content of all Wikipedia pages, and all information associated with the local expert teams are indexed using the same controlled vocabulary, here the ACM classification scheme.

2.2 Language Resources and Processing

Both Wikipedia page content and the expert profiles are mapped into the same ontology categories, which provide a pivot, or link between them. In our implemented example, we used the ACM 2012 Classification schema[5] as the shared ontological space. Here is an entry in this ontology:

<skos:Concept rdf:about="#10010296" xml:lang="en">
<skos:prefLabel xml:lang="en">Gaussian processes
</skos:prefLabel>
<skos:altLabel xml:lang="en">Gaussian process
</skos:altLabel>
<skos:inScheme
rdf:resource="http://totem.semedica.com/taxonomy/The ACM Computing Classification System (CCS)"/>
<skos:broader rdf:resource="#10010075"/>

This entry gives a synonym for “Gaussian processes”, an internal ACM code for this concept (10010296), and a link to a hypernym (10010075), “Theory of computation”. This SKOS taxonomy, augmented by any Wikipedia redirects as additional synonyms, was converted into a natural language processing program that associates the internal ACM code to raw English text (delivered as resource with this paper).

Wikipedia text was extracted from a recent dump.[6] The text was extracted from each article, tokenized, lowercased, and transformed into sentences. If any preferred or alternative term from the ACM classification scheme was found, then that ACM code was associated to the Wikipedia article[7]. Any ACM concept appearing in more than the adhoc threshold of 10,000 articles was eliminated as being too general[8].

The source data for expert profiles in our implementation are the public web pages of Inria teams 2014 activity

[5] http://www.acm.org/about/class/class/2012
[6] http://dumps.wikimedia.org/biwiki/latest
[7] The programs for turning English text into sentences, and for recognizing ACM codes in text can be found athttp://pages.saclay.inria.fr/gregory.grefenstette/LRECpack.gz
[8] This threshold eliminates concepts ‘Women’ found under ‘Gender’.

reports [9]. Each page was downloaded, boilerplate removed, text extracted, tokenized lowercased and split into sentences, as with the Wikipedia text.

In all 3123 Inria web pages were associated 129,499 Wikipedia articles were tagged one or more of with 1049 different ACM codes.

For example, the Wikipedia page "Artificial Intelligence" contains the ACM classification phrase "supervised learning". Many of the Inria research team annual report web pages also mention this topic, for example, the page uid70.html of the Inria team TAO. A link between the Wikipedia page "Artificial Intelligence", the term "supervised learning" and the TAO web page is created and stored in a publicly accessible SPARQL endpoint.

Figure 6. The link between each concept found in each Wikipedia page, and each outside data source, is rewritten as RDF and stored in a publicly accessible SPARQL database.

This SPARQL endpoint is accessed by the program "DisplayInriaTeam2.js" which is embedded in a Wikipedia gadget. In order to install this gadget in your Wikipedia;

1) visit wikipedia.org and register on the site to create an account
2) log in, and go to the following page: https://meta.wikimedia.org/wiki/Special:MyPage/global.js
3) create the page, and add the following line to it: *mw.loader.load("www.wikidata.org/w/index.php?title=User:ggrefen/DisplayInriaTeam2.js&action=raw&ctype=text/javascript");*
4) refresh your browser and visit a page with a computing topic
5) you can de-activate this gadget by adding two slashes (//) at the beginning of the line (before the word *mw.loader.load*).

Once you have followed the above steps, every time you log in, this gadget is activated and the "Inria" tab is inserted to every Wikipedia page, accessing the SPARQL data base when clicked on and producing the expanded box (see Figure 2) of the ACM concepts present on the page, and the Inria teams concerned with these concepts, providing the explicit links between the Wikipedia page content and the local external data.

2.3 Variations

We have currently implemented this search solution linking Wikipedia to Inria research team annual reports. Instead of using annual reports to create the "expert profiles", one could instead use the publications of researchers from a given team or research center. For example, each Inria team is obliged to publish their papers in the open access repository hal.inria.fr and the title of the papers, or their abstracts, or their full content could also be used to link Wikipedia page content to individual researchers. In this case, a different SPARQL endpoint could instantiated in this gadget, or a different gadget could be used for this local data. In place the ACM hierarchy as a pivot ontology[10], one could extract a taxonomy of Wikipedia categories and subcategories[11], or use another existing ontology, to index the Wikipedia content and the outside expert profiles. For example, one could use MeSH as the anchor ontology and publications of doctors at local hospital to transform Wikipedia into a search engine of specialists for medical problems.

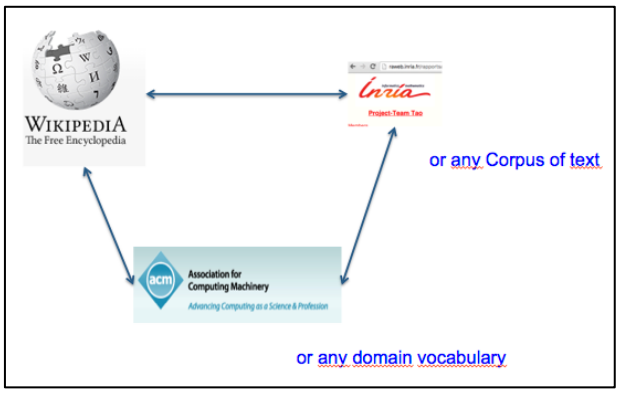

Figure 7. Our approach allows us to substitute and textual data source as the local expert repository, and any domain vocabulary as the pivot ontology. Thus Wikipedia, wihout altering its source text, can be used to search over any local data repository.

3. Conclusion

We have presented an Ontology-based information retrieval system for finding local experts using Wikipedia. It is constructed using a pivot ontology, indexing Wikipedia pages, and some textual representation of the experts (web pages, reports, or publications). The pivot ontology must be represented as a natural language processing resource, i.e. a resource that a program can be used to index to natural language text, that is applied to both Wikipedia pages and the textual representation of experts, moving them to the

[9] https://raweb.inria.fr/rapportsactivite/RA2014/index.html

[10] As a side effect of our work, Wikidata has decided to include the class system of the ACM in its hierarchy of concepts.

[11] Starting from, for example, https://en.wikipedia.org/wiki/Category:Computer_science

same space defined by the vocabulary of the resource. Once this mapping is done, and once the expert finding tab is opened, a Wikimedia resource loader dynamically makes the connection between the Wikipedia subject matter and the local experts by accessing a SPARQL endpoint, and constructs the list of local experts which is inserted into the displayed Wikipedia text. Neither the text of Wikipedia, not the expert profile text are permanently altered. An additional advantage of this system is that the user seeking an expert does not interact explicitly with the pivot ontology, but only with Wikipedia and the original textual representations of the experts. One plan to improve this system is to produce a more contextual markup, rather than the current string matching, by first categorizing the Wikipeida page as belonging to the subject of interest (in the case of Inria, this would be mathematics and computing).

As the general public becomes more used to using Wikipedia to find information, they are able to find the best page that characterizes their need. With this gadget, Wikipedia is transformed into a local expert search engine[12].

3.1 Acknowledgments

This work is supported by the Center for Data Science, funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02, and by an Advanced Researcher grant from Inria.

4. Bibliographical References

Amini, Bahram, Roliana Ibrahim, Mohd Shahizan Othman, and Mohammad Ali Nematbakhsh. (2015). A Reference Ontology For Profiling Scholar's Background Knowledge In *Recommender Systems. Expert Systems with Applications* 42, no. 2, pp. 913-928.

Demartini, Gianluca. (2007). Finding Experts Using Wikipedia. *FEWS* 290: 33-41.

Hristoskova, Anna, Elena Tsiporkova, Tom Tourw, Simon Buelens, Mattias Putman, and Filip De Turck. (2012). Identifying Experts through a Framework For Knowledge Extraction from Public Online Sources. In *12th Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, Ghent, Belgium, pp. 19-22.

Hahn, Jungpil, and Mani R. Subramani. (2000). A Framework of Knowledge Management Systems: Issues and Challenges for Theory and Practice. In *Proceedings of the 21st international conference on Information systems*, pp. 302-312.

Hearst, Marti. (2006). Design Recommendations For Hierarchical Faceted Search Interfaces. In *ACM SIGIR Workshop on Faceted Search*.

McDermott, Richard. (1999). Why Information Technology Inspired but Cannot Deliver Knowledge Management. *California Management Review* 41, no. 4: 103-117.

Ponzetto, Simone Paolo, and Roberto Navigli. (2009) Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *IJCAI*, vol. 9, pp. 2083-2088.

Rafes, Karima, and Cécile Germain. (2015). A Platform For Scientific Data Sharing. In *BDA2015-Bases de Données Avancées*. France.

Santos, António Paulo, and Fátima Rodrigues. (2009). Multi-Label Hierarchical Text Classification Using The ACM Taxonomy. In *14th Portuguese Conference on Artificial Intelligence (EPIA)*.

Swartout, W. and Tate, A. (1999). Guest editors' Introduction: Ontologies. *IEEE Intelligent Systems*, (1), pp.18-19.

Veltman, Kim H. (2001). Syntactic And Semantic Interoperability: New Approaches To Knowledge And The Semantic Web. *New Review of Information Networking* 7.1: 159-183.

West, Robert, Ingmar Weber, and Carlos Castillo. A Data-Driven Sketch Of Wikipedia Editors. In *Proceedings of the 21st International Conference Companion on World Wide Web*, pp. 631-632. ACM, 2012.

Zesch, Torsten, and Iryna Gurevych (2007). Analysis Of The Wikipedia Category Graph For NLP Applications. *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007).*

[12] 3 other Wikipedia gadgets for local data sets have also been developed using the LinkedWiki platform (Rafes & Germain, 2015):
https://io.datascience-paris-saclay.fr/appDisplayDatasets.php
https://io.datascience-paris-saclay.fr/appDisplayDevices.php
https://io.datascience-paris-saclay.fr/appDisplayScientists.php

Refining Hyponymy in a Terminological Knowledge Base

Juan Carlos Gil-Berrozpe, Pamela Faber
University of Granada
Calle Buensuceso, 11. 18002. Granada, Spain
juancarlosgb@correo.ugr.es, pfaber@ugr.es

Abstract

Hyponymy or *type_of* relation is the backbone of all hierarchical semantic configurations. Although recent work has focused on other relations such as meronymy and causality, hyponymy maintains its special status since it implies property inheritance. As reflected in EcoLexicon, a multilingual terminological knowledge base on the environment, conceptual relations are a key factor in the design of an internally and externally coherent concept system. Terminological knowledge bases can strengthen their coherence and dynamicity when the set of conceptual relations is wider than the typical generic-specific and part-whole relations, which entails refining both the hyponymy and meronymy relations. This paper analyzes how hyponymy is built in the EcoLexicon knowledge base and discusses the problems that can ensue when the *type_of* relation is too simplistically defined or systematically represented. As a solution, this paper proposes the following: (i) the correction of property inheritance; (ii) the specification of different subtypes of hyponymy; (iii) the creation of 'umbrella concepts'. This paper focuses on the first two solutions and proposes a set of parameters that can be used to decompose hyponymy.

Keywords: terminological knowledge bases, conceptual modelling, generic-specific relations

1. Introduction

In recent years, research in specialized language has begun to acknowledge the need for an interdisciplinary approach and for a set of theoretical premises that will make conceptual modelling more objective (León-Araúz et al., 2012). In fact, the study of terminology and specialized communication is currently experiencing a 'cognitive shift' (Faber, 2009), which is granting greater importance to conceptual organization as reflected in neurological processes (Faber et al., 2014). Terms are specialized knowledge units used to designate the objects, events and processes characteristic of a specialized domain. In the same way as language mirrors the mind, terminological structure can be regarded as a reflection of conceptual structure.

However, the specification of conceptual structure must be grounded on a set of theoretical assumptions regarding categorization, more specifically, whether and to what extent sensory information is part of semantic representation and processing (Meteyard et al., 2012). In this sense, Patterson et al. (2007), propose a supramodal format for semantic representations, which is modality-invariant though derived from mappings across sensory and motor input. In Terminology, the correlate of this supramodal representation is a category schema or template as posited by various authors (Faber et al., 2014; Roche et al., 2009; Leonardi, 2010). This top-level schema constrains perceptual input though, at the same time, it is also derived from sensorimotor mappings. This type of schema facilitates the retrieval of all the information stored, and is the frame for any semantic network.

Not surprisingly, the configuration of specialized concepts in networks with both hierarchical and non-hierarchical or associative relations has proven to be one of the most important aspects of terminology work (León-Araúz et al., 2012). Nevertheless, this task is far from simple because, in certain cases, the semantics of the relations are too vague, as can be observed in many thesauri, conceptual maps, and semantic networks (Jouis, 2006). That is the reason why a wide range of methods for structuring knowledge have been considered in Terminology. These include extending non-hierarchical relations, specifying the properties of the relations, and integrating innovative theories from linguistics and artificial intelligence. In order to guarantee high-quality terminological work, it is thus necessary to establish a methodology based on logical properties that will facilitate the accurate organization of conceptual relations.

2. Terminological Knowledge Bases and Conceptual Relations

Regarded by Meyer et al. (1992) as a hybrid between term banks and knowledge bases, terminological knowledge bases (TKBs) represent the specialized knowledge of a certain field through related concepts and the terms that designate them in one or various languages. A TKB is thus a product that reflects both linguistic and cognitive processes. Optimally, TKBs should reflect how conceptual networks are established and structured in our minds. They must also be designed to meet the needs of a specific group of users, whether they are experts or lay public.

According to León-Araúz et al. (2013), TKBs should account for the representation of natural and contextual knowledge dynamism. Various issues must thus be considered when designing and creating a TKB. On the one hand, the organization of the knowledge field should accurately represent the concepts and the semantic relations linking them. On the other hand, access to information and its retrieval should facilitate knowledge acquisition.

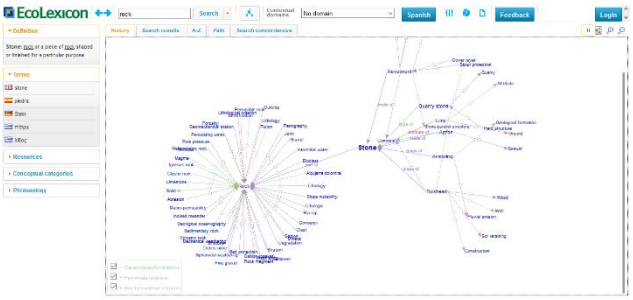

Figure 1: Visual interface of EcoLexicon

One difficulty in concept representation stems from the fact that the characteristics of a concept may vary depending on the perspective taken. Such conceptual multidimensionality can affect a wide range of properties from shape to function (Kageura, 1997). The representation of multidimensionality is thus a major challenge in TKB design since extracting a few concepts and establishing simple relations between them results in monodimensional systems, which are unrealistic and only permit *in vitro* knowledge acquisition (Dubuc & Lauriston, 1997; Cabré, 1999).

Nevertheless, the representation of multidimensionality must also follow rules. In this sense, conceptual (semantic) relations cannot be created on demand, but should systematically be derived from a set inventory (León-Araúz et al., 2012).

EcoLexicon[1] is one example of a multidimensional TKB. It is a multilingual knowledge resource on the environment whose content can be accessed through a user-friendly visual interface with different modules for conceptual, linguistic and graphical information (Faber et al., 2014; Reimerink et al., 2010) (see Figure 1). EcoLexicon targets different user groups interested in expanding their knowledge of the environment for text comprehension and generation, such as environmental experts, technical writers, and translators. This resource is available in English and Spanish, though five more languages (German, Modern Greek, Russian, French and Dutch) are being progressively implemented. It currently contains a total of 3,599 concepts and 20,070 terms.

In EcoLexicon, conceptual relations are classified in three main groups: generic-specific relations, part-whole relations and non-hierarchical relations (see Figure 2). As can be observed, hierarchical relations have been divided into two groups to distinguish between hyponymic relations and meronymic relations. The set of generic-specific relations only comprises *type_of*. In contrast, the set of part-whole relations contains *part_of*, *made_of*, *delimited_by*, *located_at*, *takes_place_in*, and *phase_of*. In the last place, the set of non-hierarchical relations includes *affects*, *causes*, *attribute_of*, *opposite_of*, *studies*, *measures*, *represents*, *result_of*, *effected_by*, and *has_function*. The set of all conceptual relations in EcoLexicon comes to a total of 17. In some cases, these relations are domain-specific (e.g. *measures*), which means that the set of conceptual relations of a TKB may vary from one field of knowledge to another.

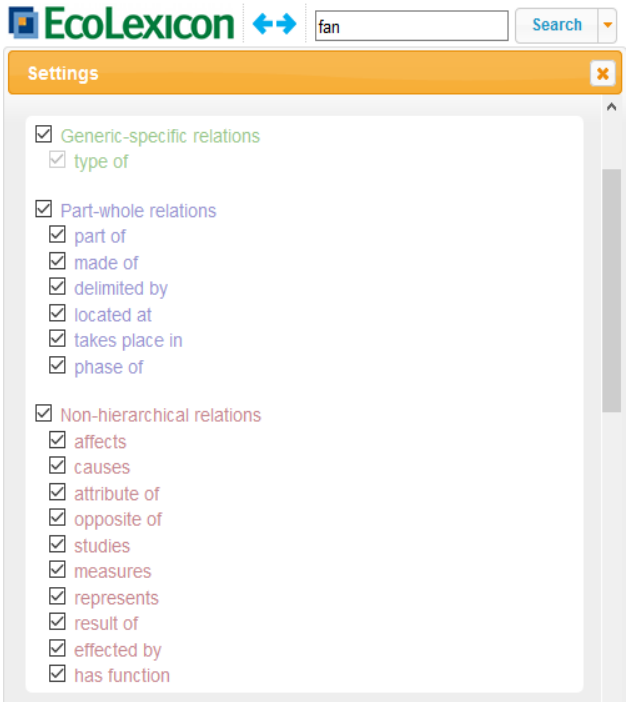

Figure 2: Semantic relations in EcoLexicon

3. Refining Hyponymy in EcoLexicon

TKBs can acquire greater coherence and dynamicity when the range of conceptual relations is wider than the traditional generic-specific and part-whole relations (León-Araúz et al., 2012), which entails taking into consideration non-hierarchic relations and, in addition, expanding the original sense of both hyponymy and meronymy. In EcoLexicon, the meronymic relation *part_of* has already been divided into subtypes, as shown in Figure 2. For example, even though CONDENSATION is *part_of* the HYDROLOGIC CYCLE, it is more accurate to say that CONDENSATION is a *phase_of* the HYDROLOGIC CYCLE. This distinction was made in EcoLexicon because of the following factors: (i) domain-specific needs, (ii) ontological reasoning, and (iii) transitivity-related consistency (León-Araúz & Faber, 2010).

Nevertheless, the *type_of* relation still has not been subdivided. This is the source of a wide range of problems in EcoLexicon, such as different cohyponyms at the same level (see Figure 3), which produces noise as well as information overload and redundancy. Still another problem lies in transitivity and property inheritance. For example: LIMESTONE is currently represented as a hyponym to both ROCK and SEDIMENTARY ROCK. A possible solution would be to refine the *type_of* relation.

[1] http://ecolexicon.ugr.es/

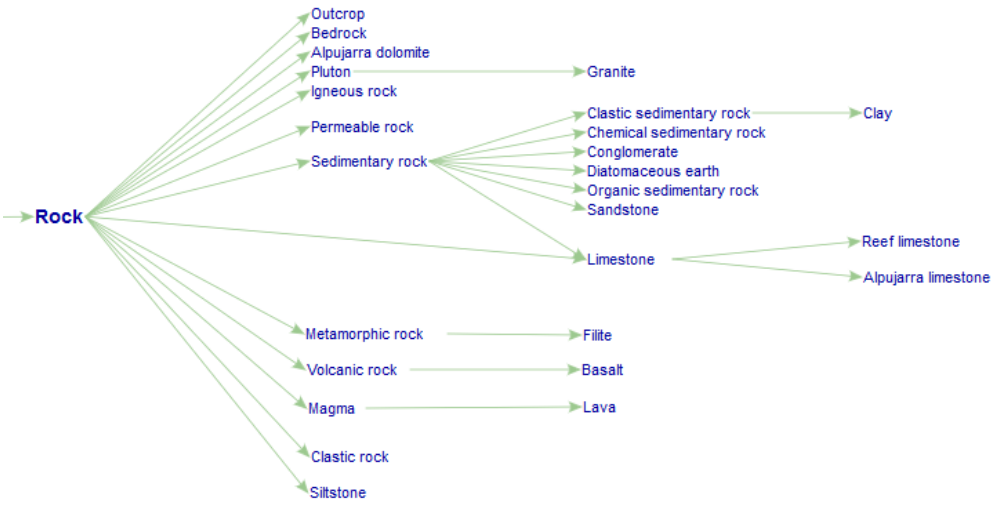

Figure 3: Presence of different dimensions of cohyponyms at the same level, and transitivity problems

Murphy (2003, 2006) states that hyponymy is a relation of inclusion whose converse is hyperonymy. Due to its inference-invoking nature, its importance in definitions, and its relevance to selectional restrictions in grammar, hyponymy is central in many models of the lexicon. Similarly to meronymy, hyponymy can be divided into subtypes (Murphy, 2003), but those subtypes should provide a valid taxonomy of generic-specific relations. According to Murphy (*ibid*: 219-220), the most commonly established distinction is among taxonomic hyponymy ('is-a-kind-of' relation) and functional hyponymy ('is-used-as-a-kind' relation). In some way, this dichotomy is related to what Cruse (2002) calls 'facets'.

On the other hand, another important phenomenon in the specification of hyponymic relations is the existence of 'microsenses' (Cruse, 2002), which are only activated in a certain context. In EcoLexicon, microsenses can already be made explicit by entering different concepts as hyperonyms of the same hyponym. It is even possible to filter the query by restricting the represented contextual domain (see Figure 4). More precisely, conceptual propositions (concept-relation-concept) in EcoLexicon are triggered or constrained based on their salience in different discipline-based subdomains.

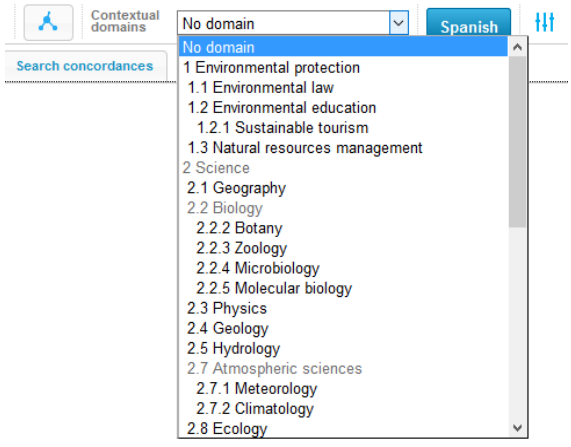

Figure 4: Contextual domains in EcoLexicon

This shows that the same concept may have different context-dependent hyperonyms, reflecting only a selection of its 'microsenses'. This phenomenon can be observed in Figures 5 and 6, which illustrate the different conceptual networks of SAND in the Soil Sciences domain (see Figure 5) and in the Geology domain (see Figure 6).

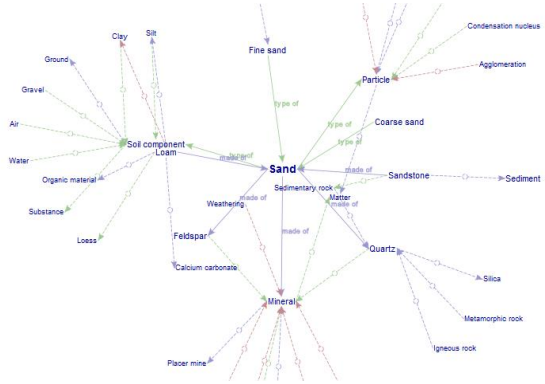

Figure 5: SAND network in Soil Sciences

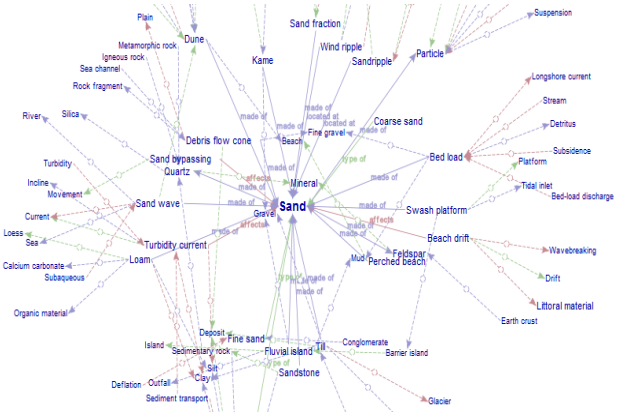

Figure 6: SAND network in Geology

Furthermore, Gheorghita & Pierrel (2012) state that the meaning of an input in a TKB can be disambiguated just by adding a domain to the definition. In the case of EcoLexicon, domains are not applied to definitions, but to conceptual relations. However, according to its database, only 2624 (50%) of all relations have been classified using a domain, and this invalidates the possibility of being completely accurate using this method.

Moreover, there is still the need to further refine the *type_of* relation. Possible solutions include: (i) the correction of property inheritance; (ii) the specification of different subtypes of hyponymy; and (iii) the creation of 'umbrella concepts'. This paper focuses on the first two, since the correction of property inheritance is the first step towards dividing the *type_of* relation into subtypes.

3.1 Correcting Property Inheritance

As previously mentioned, hyponymy is a unidirectional relation where child concepts inherit the properties of their parent concepts, though they also have differentiating properties that make their meaning more specific. In a TKB, property inheritance between hyperonyms and hyponyms can be represented through genus–differentia definitions, based on the explicitation of the *genus* (hyperonym or superordinate) and one or many *differentiae* (characteristics that vary between cohyponyms) (Temmerman, 2000).

Despite the fact that most of the concepts in EcoLexicon are defined in this way, transitivity problems still arise (see Figure 3). This paper proposes a solution, as exemplified in the analysis of two types of concept: an entity (ROCK) and a process (EROSION).

3.1.1. Property Inheritance in the Conceptual Network of an Entity: ROCK

The original *type_of* network of ROCK was initially not accurately defined (see Figure 3). For example, LIMESTONE appeared as a direct hyponym of both ROCK and SEDIMENTARY ROCK, and there were two similar entities that designated 'clastic rock' at two different levels (CLASTIC ROCK and CLASTIC SEDIMENTARY ROCK). In order to solve such problems and related issues, the conceptual network of ROCK was enhanced with the addition of new concepts (e.g. SOLID ROCK, MOLTEN ROCK or DOLOMITE) and the property inheritance relations were restructured.

Table 1 shows an example of property inheritance in the original conceptual network. BASALT is defined as a 'rock of igneous origin', but its hyperonym (VOLCANIC ROCK) is also defined as an 'igneous rock'. Furthermore, the hyperonym of VOLCANIC ROCK is assumed to be ROCK, regardless of the fact that the only types of rock mentioned in its definition are 'igneous, sedimentary and metamorphic'.

ROCK: consolidated or unconsolidated aggregate or mass of minerals or organic materials. The three types of rock are igneous, sedimentary, and metamorphic.
VOLCANIC ROCK: extrusive igneous rock solidified near or on the surface of the Earth, resulting from volcanic activity.
BASALT: very hard rock of igneous origin, consisting of augite and triclinic feldspar, with grains of magnetic or titanic iron, and also bottle-green particles of olivine. It is formed by decompression melting of the Earth's mantle.

Table 1: ROCK – BASALT in the former conceptual network (original definitions)

Table 2 shows how property inheritance has been improved in the enhanced conceptual network. In this case, it is respected in all senses: BASALT is a *type_of* VOLCANIC ROCK, which is a *type_of* IGNEOUS ROCK, which is a *type_of* SOLID ROCK, which is a *type_of* ROCK. In other words, BASALT in the end reflects the inheritance of the characteristics possessed by all of its hyperonyms.

ROCK: consolidated or unconsolidated aggregate or mass of minerals or organic materials.
SOLID ROCK: rock in solid state, formed by the compression of sediments or the solidification of molten material.
IGNEOUS ROCK: solid rock formed by solidification of molten magma either beneath or at the Earth's surface.
VOLCANIC ROCK: extrusive igneous rock solidified near or on the surface of the Earth, resulting from volcanic activity.
BASALT: very hard volcanic rock, consisting of augite and triclinic feldspar, with grains of magnetic or titanic iron, and also bottle-green particles of olivine. It is formed by decompression melting of the Earth's mantle.

Table 2: ROCK – BASALT in the new conceptual network (enhanced definitions)

Finally, as a result of modifications in the remaining conceptual relations, improved terminological definitions, and the addition of new concepts to fill semantic gaps, the conceptual network of ROCK was enhanced (see Table 3).

LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
ROCK < *type_of* -	solid rock < *type_of* -------	sedimentary rock < *type_of* ----	limestone < *type_of* ----------	reef limestone | Alpujarra limestone |	
			clastic rock < *type_of* ----------	clay |	
			chemical sedimentary rock < *type_of*	dolomite < *type_of* -	Alpujarra dolomite |
			organic sedimentary rock |		
			conglomerate |		
			diatomaceous earth |		
			sandstone |		
			siltstone |		
		igneous rock < *type_of* ---------	plutonic rock < *type_of* ---------	granite |	
			volcanic rock < *type_of* ---------	basalt |	
		metamorphic rock < *type_of* ---	filite |		
		permeable rock |			
		bedrock |			
		outcrop |			
	molten rock < *type_of* ----	magma < *type_of* ----------	lava |		

Table 3: Enhanced conceptual network of ROCK (generic-specific relations)

3.1.2. Property Inheritance in the Conceptual Network of a Process: EROSION

Property inheritance is also manifested in 'process' type of concepts. In this case, the original conceptual network of EROSION was also analyzed to examine if the inheritance of characteristics between parent and child concepts was accurate. As a result, certain concepts had to be relocated, and the definitions of some hyponyms needed to be enhanced to correct property inheritance.

To portray how property inheritance has been corrected in the definitions of concepts, another comparison has been made to show the differences in the *type_of* relation established from EROSION to CHANNEL SCOUR. As can be observed in the original conceptual network (see Table 4), CHANNEL SCOUR, located at the third level with respect to EROSION, is defined as 'erosion' when it should inherit the traits of its direct hyperonym, SCOUR.

EROSION: process by which materials of the Earth's crust are worn away, loosened, or dissolved while being transported from their place of origin by different agents, such as wind, water, bacteria, etc.
FLUVIAL EROSION: erosion of bedrock on the sides and bottom of the river; the erosion of channel banks; and the breaking down of rock fragments into smaller fragments by the flow of water in the channel.
SCOUR: localized erosive action of water in streams, excavating and carrying away material from the bed and banks.
CHANNEL SCOUR: erosion of a stream bed.

Table 4: EROSION – CHANNEL SCOUR in the former conceptual network (original definitions)

In contrast, in the enhanced conceptual network (see Table 5), property inheritance is well expressed, since each hyponym adopts the characteristics of its hyperonym: CHANNEL SCOUR is a *type_of* SCOUR, which is a *type_of* FLUVIAL EROSION, which is a *type_of* WATER EROSION, which is a *type_of* EROSION. Therefore, CHANNEL SCOUR (at the fourth level of hyponymy), is now defined as a type of SCOUR rather than as a type of EROSION.

EROSION: process by which materials of the Earth's crust are worn away, loosened, or dissolved while being transported from their place of origin by different agents, such as wind, water, bacteria, etc.
WATER EROSION: erosion of rocks and sediment by water, involving detachment, transport and deposition.
FLUVIAL EROSION: water erosion of bedrock on the sides and bottom of the river; the erosion of channel banks; and the breaking down of rock fragments into smaller fragments by the flow of water in the channel.
SCOUR: localized fluvial erosion in streams, excavating and carrying away material from the bed and banks.
CHANNEL SCOUR: scour of a stream bed.

Table 5: EROSION – CHANNEL SCOUR in the new conceptual network (enhanced definitions)

Nevertheless, the previously mentioned modifications were not the only changes made to refine the conceptual network, as new concepts (e.g. WATER EROSION, RILL EROSION or STREAMBANK EROSION) were also added. In the end, an enhanced version was obtained (see Table 6). The correction of property inheritance not only enhances content, but also indicates how hyponymy can be decomposed into subtypes as discussed in the following section.

LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
EROSION **< *type_of* -**	water erosion **< *type_of* -**	fluvial erosion **< *type_of***	scour < ***type_of* ------**	channel scour | outflanking |
			sheet erosion | rill erosion | gully erosion | streambank erosion |	
		sea erosion |		
	wind erosion **< *type_of* --**	deflation |		
	abrasion < ***type_of* -------**	glacier abrasion |		
	anthropic erosion | glacier erosion | internal erosion | potential erosion | differential erosion | attrition | denudation |			

Table 6: Enhanced conceptual network of EROSION (generic-specific relations)

3.2. Specifying Subtypes of Hyponymy

According to Murphy (2003, 2006), hyponymy can be divided into subtypes, such as taxonomic hyponymy and functional hyponymy. In this case, the conceptual networks in EcoLexicon show a more fine-grained set of subtypes, which are initially based on whether the concept is an entity (ROCK) or a process (EROSION).

3.2.1. Subtypes of Hyponymy in the Conceptual Network of an Entity: ROCK

Based on the improved conceptual network of ROCK (see Table 3) and the enhanced concept definitions, up to five different subtypes of hyponymy related to entities can be established:

- **Formation-based hyponymy:** a *type_of* relation dependent on the formation process or the origin of the hyponyms.
- **Composition-based hyponymy:** a *type_of* relation dependent on the components or the constituents of the hyponyms.
- **Location-based hyponymy:** a *type_of* relation dependent on the physical situation or location of the hyponyms.
- **State-based hyponymy:** a *type_of* relation dependent on the state of matter of the hyponyms.
- **Attribute-based hyponymy:** a *type_of* relation dependent on the traits or features of the hyponyms.

Table 7 offers examples of these subtypes in the conceptual network of ROCK. For instance, IGNEOUS ROCK is considered to be a *formation-based_type_of* SOLID ROCK because it is 'formed by solidification of molten magma'; REEF LIMESTONE is presented as a *composition-based_type_of* LIMESTONE since it is 'composed of the remains of sedentary organisms'; and VOLCANIC ROCK is represented as a *location-based_type_of* IGNEOUS ROCK because it is 'solidified near or on the surface of the Earth'.

Formation-based hyponymy: (X *formation-based_type_of* Y)
• SEDIMENTARY ROCK < SOLID ROCK • IGNEOUS ROCK < SOLID ROCK • CLASTIC ROCK < SEDIMENTARY ROCK
Composition-based hyponymy: (X *composition-based_type_of* Y)
• DOLOMITE < CHEMICAL SEDIMENTARY ROCK • ORGANIC SEDIMENTARY ROCK < SEDIMENTARY ROCK • REEF LIMESTONE < LIMESTONE
Location-based hyponymy: (X *location-based_type_of* Y)
• BEDROCK < SOLID ROCK • VOLCANIC ROCK < IGNEOUS ROCK • ALPUJARRA LIMESTONE < LIMESTONE
State-based hyponymy: (X *state-based_type_of* Y)
• SOLID ROCK < ROCK • MOLTEN ROCK < ROCK
Attribute-based hyponymy: (X *attribute-based_type_of* Y)
• PERMEABLE ROCK < SOLID ROCK

Table 7: Examples of the subtypes of hyponymy found in the conceptual network of ROCK

Nevertheless, not all hyponymic relations can be classified using a subtype. There are certain child concepts whose differentiating features make it impossible to determine one subtype of hyponymy. For example, GRANITE is a *type_of* PLUTONIC ROCK based on its attributes ('coarse-grained, light-colored, hard'), its composition ('consisting chiefly of quartz, orthoclase or microline, and mica') and its function ('used as a building material'). Such cases will remain classified as general taxonomic hyponymy, or as a non-specific *type_of* relation.

Nonetheless, this list of subtypes is not a closed inventory of hyponymic relations, but only those which have been distinguished so far in the conceptual network of ROCK and similar entities. In fact, in regards to WATER, two more subtypes of hyponymy have been noticed:

- **Function-based hyponymy:** a *type_of* relation dependent on the function or the purpose of the hyponyms.
  e.g. DRINKING WATER *function-based_type_of* WATER
- **Shape-based hyponymy:** a *type_of* relation dependent on the shape or the physical aspect of the hyponyms.
  e.g. AMORPHOUS FROST *shape-based_type_of* FROST

A minimum number of coincidences will eventually be established to confirm the validity (and usefulness) of a subtype of hyponymy.

3.2.2. Subtypes of Hyponymy in the Conceptual Network of a Process: EROSION

In reference to EROSION (see Table 6), up to four subtypes of hyponymy, typical of processes, were established:

- **Agent-based hyponymy:** a *type_of* relation dependent on the agent or the promoter that causes the hyponyms.
- **Patient-based hyponymy:** a *type_of* relation dependent on the entity or location affected by the hyponyms.
- **Result-based hyponymy:** a *type_of* relation dependent on the results and effects of the hyponyms.
- **Attribute-based hyponymy:** a *type_of* relation dependent on the traits or features of the hyponyms.

Table 8 contains some examples of these subtypes of hyponymy found in the conceptual network of EROSION. For example, ANTHROPIC EROSION is considered to be an *agent-based_type_of* EROSION because it is 'caused by human activities'; GLACIER ABRASION is regarded as a *patient-based_type_of* ABRASION since it is the abrasion 'of a glacier bed'; and RILL EROSION is a *result-based_type_of* FLUVIAL EROSION because it 'forms small channels'.

As shown in Table 8, the subtypes of process hyponymy are different from those of an entity (except for attribute-based hyponymy, which is common to both). A process is generally a nominalization of a verb, and thus it often involves an agent, a patient, and a result. This differs from formation, composition, and state, which are typical of entities. Moreover, in the case of processes, patient-based hyponymy sometimes overrides location-based hyponymy, as the patient can be a physical location (e.g. CHANNEL SCOUR affects a stream bed, and therefore takes place in a stream bed).

Furthermore, the general taxonomic hyponymy (*type_of*) is also present in processes. In fact, various examples of it can be found in the conceptual network of EROSION. For instance, DENUDATION is a *type_of* EROSION based on its agents ('caused by the action of water, ice, wind and waves'), its patient ('the Earth's surface') and its result ('redistribution of Earth surface material').

Agent-based hyponymy (X *agent-based_type_of* Y)
• SEA EROSION < EROSION • ANTHROPIC EROSION < EROSION • FLUVIAL EROSION < WATER EROSION
Patient-based hyponymy (X *patient-based_type_of* Y)
• STREAMBANK EROSION < FLUVIAL EROSION • GLACIER ABRASION < ABRASION • CHANNEL SCOUR < SCOUR
Result-based hyponymy (X *result-based_type_of* Y)
• SHEET EROSION < FLUVIAL EROSION • RILL EROSION < FLUVIAL EROSION • GULLY EROSION < FLUVIAL EROSION
Attribute-based hyponymy: (X *attribute-based_type_of* Y)
• POTENTIAL EROSION < EROSION • DIFFERENTIAL EROSION < EROSION

Table 8: Examples of the subtypes of hyponymy found in the conceptual network of EROSION

In the same way as for entities, this set of subtypes of hyponymy is not a closed set since further research is needed to determine the extension and scope of subtypes of hyponymy applied to processes.

4. Conclusion

In this paper we have analyzed how to refine hyponymy in EcoLexicon, a multilingual terminological knowledge base on the environment. We have revised the theoretical background behind the fundamental characteristics of TKBs, their representation of multidimensionality and their reflection of conceptual relations. We have also studied how hyponymy is built in EcoLexicon and how different facets and microsenses can be expressed. The correction of property inheritance is a preliminary though essential phase in the refinement of the *type_of* relation.

This preliminary study has shown how to refine generic-specific relations and establish subtypes of hyponymy through the analysis of the concepts in a network and their definitions. In this way, several subtypes of hyponymy have been distinguished for entities (e.g. formation-based hyponymy), for processes (e.g. agent-based hyponymy) and for both types (e.g. attribute-based hyponymy).

We have also demonstrated how this type of refined hyponymy can be implemented in EcoLexicon, thus increasing its informativity for users. Another important issue is the formal modelling of relations in ontologies. Although restrictions on length oblige us to reduce the scope of our discussion, future work will focus on this topic in greater depth.

Moreover, the different subtypes of hyponymy and the new conceptual hierarchy must be corrected and validated by domain experts. Further research is also required to verify the existence of these subtypes of hyponymy in other fields of knowledge, to establish systematic parameters for the creation of new subtypes, and to explore how semantic relations are expressed in nominal clauses and compound nouns (Downing, 1977; Nastase & Szpakowicz, 2003).

This research opens the door to enhancing conceptual networks in TKBs and making them more informative.

5. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

6. Bibliographical References

Cabré, M. T. (1999). *La Terminología: Representación y Comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cruse, D. A. (2002). Hyponymy and its Varieties. In Rebecca Green, Carol A. Bean, and Sung Hyon Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective*, 3–22. Dordrecht, Boston, London: Kluwer Academic Publishers.

Downing, P. (1977). On the Creation and Use of English Compound Nouns. *Language*, 53 (4): 810–842. Linguistic Society of America.

Dubuc, R. and Lauriston, A. (1997). Terms and Contexts. In Sue Ellen Wright and Gerhard Budin, editors, *Handbook of Terminology Management*, 1: 80–88. Amsterdam, Philadelphia: John Benjamins.

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. *MonTI (Monografías de Traducción e Interpretación)*, 1: 107–134. Valencia: Universitat de València.

Faber, P., León-Araúz, P., Reimerink, A. (2014). Representing Environmental Knowledge in EcoLexicon. In Elena Bárcena, Timothy Read, and Jorge Arús, editors, *Languages for Specific Purposes in the Digital Era*, 19: 267–301. Berlin, Heidelberg: Springer.

Gheorghita, I., and Pierrel, J. (2012). Towards a Methodology for Automatic Identification of Hypernyms in the Definitions of Large-Scale Dictionary. In Nicoletta Calzolari (Conference Chair) et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*, 2614–2618. Istanbul, Turkey: ELRA.

Jouis, C. (2006). Hierarchical Relationships "is-a": Distinguishing Belonging, Inclusion and Part/of Relationships. In Nicoletta Calzolari (Conference Chair) et al., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, 571–574. Genoa, Italy: ELRA.

Kageura, K. (1997). Multifaceted/Multidimensional Concept Systems. In Sue Ellen Wright and Gerhard Budin, editors, *Handbook of Terminology Management*, 1: 119–132. Amsterdam, Philadelphia: John Benjamins.

León-Araúz, P., and Faber, P. (2010). Natural and Contextual Constraints for Domain-Specific Relations. In Verginica Barbu Mititelu, Viktor Pekar, and Eduard Barbu, editors, *Proceedings of the Workshop Semantic Relations, Theory and Applications*, 12–17. Valletta, Malta.

León-Araúz, P., Faber, P., and Montero Martínez, S. (2012). Specialized Language Semantics. In Pamela Faber, editor, *A Cognitive Linguistics View of Terminology and Specialized Language*, 95–175. Berlin, Boston: De Gruyter Mouton.

León-Araúz, P., Reimerink, A., and García Aragón, A. (2013). Dynamism and Context in Specialized Knowledge. *Terminology*, 19 (1): 31–61. Amsterdam, Philadelphia: John Benjamins.

Leonardi, P. M. (2010). Digital Materiality? How Artifacts Without Matter, Matter. *First Monday,* 15 (6). Available from: http://firstmonday.org/article/view/3036/2567

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of Age: A Review of Embodiment and the Neuroscience of Semantics. *Cortex*, 48: 788–804. doi: 10.1016/j.cortex.2010.11.002

Meyer, I., Bowker, L. and Eck, K. (1992). COGNITERM: An Experiment in Building a Knowledge-Based Term Bank. In Hannu Tommola, Krista Varantola, Tarja Salmi-Tolonen, and Jürgen Schopp, editors, *Proceedings of the Fifth EURALEX International Congress (EURALEX '92)*, 159–172. Tampere, Finland: Tampereen Yliopisto.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge: Cambridge University Press.

Murphy, M. L. (2006) Hyponymy and Hyperonymy. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, 1: 446–448. New York: Elsevier.

Nastase, V., and Szpakowicz, S. (2003). Exploring Noun-Modifier Semantic Relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*. Tilburg, The Netherlands, pp. 285–301.

Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where Do You Know What You Know? The Representation of Semantic Knowledge in the Human Brain. *Nature Reviews Neuroscience*, 8: 976–988.

Reimerink, A., León-Araúz, P., Magaña Redondo, P. J. (2010). EcoLexicon: An Environmental TKB. In Nicoletta Calzolari (Conference Chair) et al., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC2010)*, 2322–2329. Valletta, Malta.

Roche, C., Calberg-Challot, M., Damas, L., Rouard P. (2009). Ontoterminology: A New Paradigm for Terminology. In Jan L. G. Dietz, editor, *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2009)*, 2626–2630. Madeira, Portugal.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam, Philadelphia: John Benjamins.

Evaluation of EcoLexicon Images

Pilar León-Araúz, Arianne Reimerink

Department of Translation and Interpreting, Universidad de Granada

Buensuceso 11, 18071 Granada (Spain)

pleon@ugr.es, arianne@ugr.es

Abstract

The multimodal knowledge base EcoLexicon includes images to enrich conceptual description and enhance knowledge acquisition. These images have been selected according to the conceptual propositions contained in the definitional templates of concepts. Although this ensures coherence and systematic selection, the images are related to each specific concept and are not annotated according to other possible conceptual propositions contained in the image. Our aim is to create a separate repository for images, annotate all knowledge contained in each one of them and then link them to all concept entries that contain one or more of these propositions in their definitional template. This would not only improve the internal coherence of EcoLexicon but it would also improve the reusability of the selected images and avoid duplicating workload. The first step in this process and the objective of the research here described is to evaluate the images already contained in EcoLexicon to see if they were adequately selected in the first place, how knowledge is conveyed through the morphological features of the image and if they can be reused for other concept entries. This analysis has provided preliminary data to further explore how concept type, conceptual relations, and propositions affect the relation between morphological features and image types chosen for visual knowledge representation.

Keywords: multimodality, knowledge representation, visual knowledge patterns, resource evaluation

1. Introduction

EcoLexicon (ecolexicon.ugr.es) is a multimodal, multilingual terminological knowledge base (TKB) on the environment in which concepts are interrelated, based on the information extracted from a specialized domain corpus created for EcoLexicon (Faber et al., 2014; León Araúz et al., 2009). However, the way concepts are related can also be reflected in the graphical images depicting these concepts. For this reason, a visual corpus was also compiled to enrich conceptual description in EcoLexicon.

Currently, images are stored in association with concept entries according to the semantic content described in their definition. The definitions in EcoLexicon are based on templates that define category membership and describe the basic conceptual propositions. In this way, definitions have a uniform structure that directly refers to and evokes the underlying conceptual structure of the domain. These templates can be considered a conceptual grammar that thus ensures a high degree of systematisation (Montero and García 2004; Faber et al. 2007; Faber 2012). For example, for the definition of EROSION, the template includes the four basic relations of all natural processes: *is_a*, *has_agent*, *affects* and *has_result*. It also has an additional relation because it is a complex procedural concept, which can be divided into a sequence of steps: *has_phase*. For the selection of images, the basic conceptual propositions in the definitional template are used to select images which contain the same information to reinforce knowledge acquisition (Faber et al., 2007).

The definitional template provides a systematic means to select images for each concept entry in the TKB. It has been applied by our researchers to compile the visual corpus for over ten years (Faber et al., 2007). Images are thus regarded as a whole and are only linked to the concept itself. So far, we have shown how the same concept can be (and most often should be) represented through different images, depending on perspective, or the semantic content highlighted (Faber et al., 2007; Reimerink et al., 2010). However, the same image can also work for the representation of other related concept entries (e.g. an entity and the process through which it was formed, a concept and its parts, etc.). Images should thus be further dissected according to the features they possess (i.e. image type and other morphological characteristics) and the knowledge they convey (i.e. semantic content). For example, many images show several concepts in a specific background where they establish different relations that can be explicitly labeled or inferred from previous knowledge. In this sense, we believe that images should not be stored in the TKB as the representation of a concept, but as the representation of a set of conceptual propositions. They should be annotated according to semantic and morphological information and stored in a separate repository. Since each image activates several propositions and each proposition can be activated by different concepts, one image would then be linked to several concept entries. This would enhance the reusability of images, improving the consistency of the TKB and avoiding duplicating workload.

The aim of the research here described is to take the first step in this process of improving image selection and annotation: evaluate the images already contained in EcoLexicon to see i) if they were adequately selected in the first place; ii) how conceptual knowledge is conveyed through the morphological features of the image and iii) if they can be reused for other concept entries. The results of the evaluation will shed light on the relationship between the conceptual relations conveyed and the morphological features, or visual knowledge patterns (VKPs) used in images. In the future, the data obtained will help us to create the necessary annotation criteria for the existing and newly selected images in EcoLexicon.

2. Methodology

The images were analyzed in terms of (1) their adequateness in representing one or more conceptual propositions linked to the concept entry in which there are currently included, (2) the reusability of the image for other concept entries, (3) the semantic relations expressed in the image, (4) the concept types involved, and (5) the morphological features or visual knowledge patterns (VKPs) such as colour coding, referential background, arrows, labels, etc. used to convey the information. Two of our researchers studied all entries en EcoLexicon and then discussed the fulfilment of criteria and image description. An image was considered reusable if it contained any other concept or conceptual proposition that is or could be included in the TKB.

The images and the concepts they are linked to were exported to a spread sheet, where they were manually assigned a number according to the level of adequateness: 0, not at all; 1, partially; and 2, completely. Then another number was assigned according to the reusability of the image (0, no; 1, yes). Other columns included information on the image type (photograph, drawing – including maps and diagrams – or flow chart), semantic content (concepts and relations) and VKPs (labels, arrows, colours and their specific use). The assessment outcome is explained with the example in Figure 1 and Table 1.

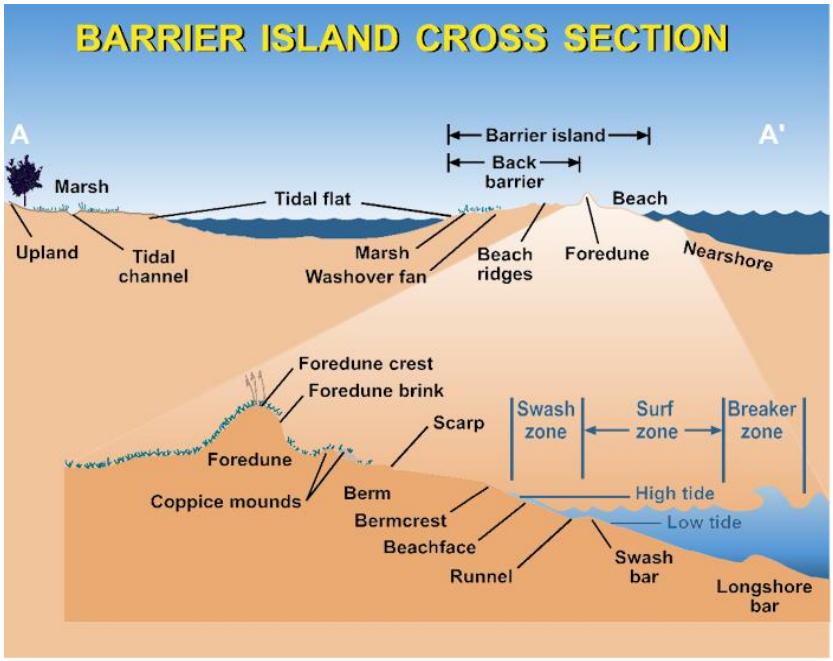

Figure 1: Drawing in concept entry BARRIER ISLAND

Concept	Image type	Conceptual propositions	VKPs	Adequate	Reusability
Barrier island	Drawing	Barrier island *has_location* nearshore Barrier island *has_part* back barrier/ foredune/ beach Barrier island *delimited_by* tidal flat/nearshore etc.	Static image Colours (referential similarity and distinction) Arrows (denomination of parts)	2	1 (foredune, tidal flat, surfzone, etc.)

Table 1: Assessment outcome for drawing in concept entry BARRIER ISLAND

The image in Figure 1 is found in the entry for BARRIER ISLAND and it represents several conceptual propositions. It has therefore been assigned a 2 (see Table 1), for it is fully adequate. The combination of image type (drawing) and VKPs (colours that provide a high level of referential similarity and arrows with labels) makes it especially adequate for the representation of the *part_of* relation. In this specific case, the image represents more conceptual propositions (*has_location* and *delimited_by*) because of the larger context in which the concept is shown. It can also be easily reused in entries of the parts of the concept and, again because of the larger context, in geographically related concepts such as TIDAL FLAT and SURF ZONE.

3. Results and Discussion

The results of our evaluation will be described and discussed in two subsections. The first one will show the data related to the adequateness of the images for the concept entry in which they are included and the possibility to use them in other concept entries. In this section examples of EcoLexicon will be included to

underline our discussion of the results. The second subsection explains how the different types of image, relation, concept and VKP are combined to convey knowledge in EcoLexicon images. Correlation graphs are used to visualize the data.

3.1 Adequateness and Reusability

Currently, EcoLexicon includes 3599 concepts and 1113 concepts have one or more images linked to their entry (31%). The total number of images amounts to 1698 of which 90.8% are adequate, 8.0% are partially adequate and 1.2% are not. Some are only partially adequate because the images are too small or unclear in the sense that you cannot see them properly, which is probably due to changes in format when introduced in EcoLexicon from other sources. The cases of complete inadequateness, normally due to misinterpretation of the conceptual information contained, must be discarded and new adequate images selected.

Although most images are adequate in the sense that they represent at least one conceptual proposition, sometimes several very similar images that do not add distinguishable conceptual knowledge are selected for the same entry in EcoLexicon. AQUIFER is a good example with four drawings with exactly the same information. Of these, the best one should be selected and linked to all the related concepts. The rest should be discarded. On the other hand, in the concepts CONFINED and UNCONFINED AQUIFER one image (Figure 2) that perfectly distinguishes these closely related concepts has been used.

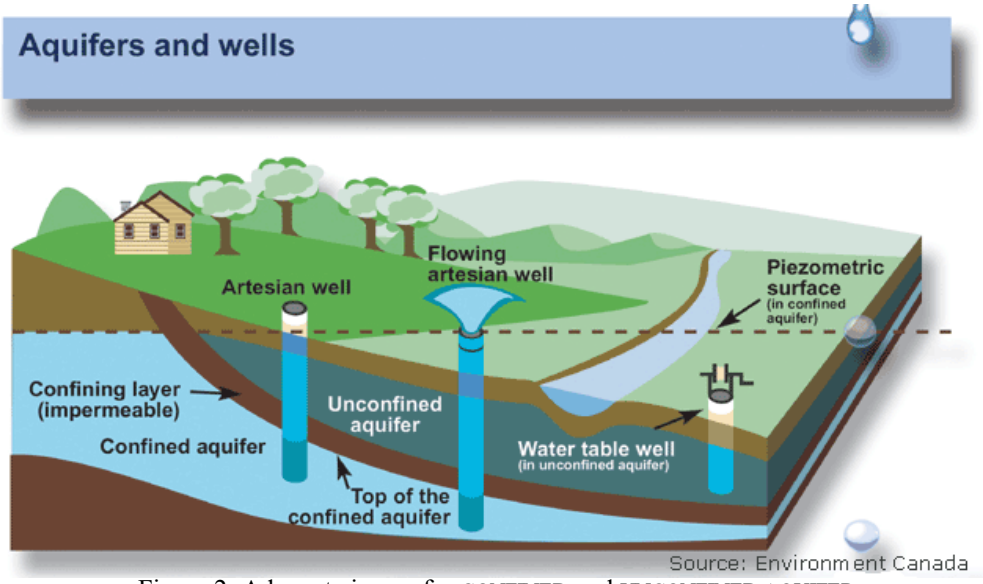

Figure 2: Adequate image for CONFINED and UNCONFINED AQUIFER

Most images are colour photographs representing *type_of* and *has_location* relations. These images cannot be reused often, because they represent the real world object. Some of them could be reused if we take into account that these real world objects are the result of certain processes, for example the concept BAR in Figure 3 is a colour photograph that represents the conceptual propositions BAR *has_location* RIVER MOUTH and BAR *result_of* LITTORAL SEDIMENTATION.

Through the annotation of all conceptual propositions in these images, they can be linked to the concept entries of the related processes – and not only to those of the entities resulting from them –, which would be a way to enhance reusability.

Instruments are often depicted by colour photographs as well. This is a problem because the most important conceptual relation for instruments is *has_function*, which cannot be easily represented by a photograph. A good example is DREDGING in Figure 4, where the *takes_place_in* and *has_instrument* relations are quite clear, but the *has_function* relation is not.

Figure 3: Adequate image for BAR *has_location* RIVER MOUTH and BAR *result_of* LITTORAL SEDIMENTATION

Figure 4: Adequate image for DREDGING *takes_place_in* SEA and *has_instrument* DREDGE

These images should therefore be combined with images that represent the process in which they participate, so that the *has_function* relation is made explicit. Another possibility to make the *has_function* relation explicit is to combine instruments with the output they provide, for example the concept entry METEOROGRAPH includes a photograph of the instrument and the concept METEOROGRAM includes a photograph of the output of the instrument. If we explicitly link both images to both concept entries, the *has_function* relation will be much clearer.

The number of photographs (51.7%) largely exceeds the number of drawings (28.8%) and flow charts (19.4%). There are several reasons to explain this.

Firstly, there are more objects than events in the knowledge base. Secondly, when the implementation of EcoLexicon began, quick deployment was considered more important than a coherent view towards image selection. Flow charts and drawings are more reusable than photographs and, if all conceptual propositions in the images are carefully annotated, they can be shown according to the specific perspective of the end-user providing more pertinent and coherent information. The entry for REEF is an example of how this can be done. The three drawings in Figure 5, for example, explain the phases of reef formation as well as the name of the subtype of reef in each phase. The image can be reused for each of the subtypes and in combination with a colour photograph of the real world entities, will provide the end-user with all the necessary information for comprehension: how each subtype is formed, how it relates to the other subtypes, and what they look like in reality.

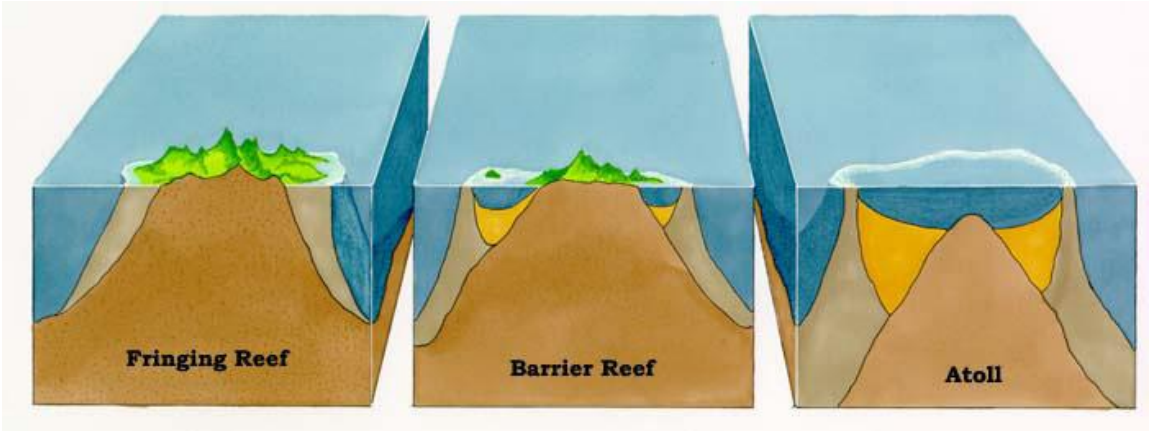

Figure 5: Adequate drawing for REEF *has_phase* FRINGING REEF/BARRIER REEF/ATOLL and FRINGING REEF/BARRIER REEF/ATOLL *type_of* REEF

Flow charts provide much conceptual information and can therefore be used to add new conceptual entries to EcoLexicon. However, this has not been done often.

Figure 6: Adequate image for conceptual propositions combining UPWELLING, COASTLINE, WIND and SURFACE WATER

For example, Figure 6 is only found in the entry for UPWELLING, but could have been reused in concept entries such as SURFACE WATER, WIND and CONTINENTAL SHELF. That way, an end-user would be able to understand the interaction among these entities no matter with which one the search started.

With the results of our study and its description of the images, all flow charts can be revised to search for new conceptual information. Moreover, annotating all conceptual propositions in the flow charts in detail as well as their VKPs will boost reusability.

3.2 Correlation between Image Types, Semantic Relations and VKPs

According to the results in Figure 7, processes are represented mostly with flow charts, whereas photographs and drawings are used to describe entities. A large number

of entities (over 40%) is also represented by flow charts. This is due to the fact that processes can affect physical entities and the latter can cause processes, thus their interaction is better conveyed in flow charts. Processes can also be depicted in a combination of drawings were different phases are shown in each one of them (e.g. Figure 5) and in photographs when the focus is on the result of the process (e.g. Figure 3). Properties (e.g. SOIL PERMEABILITY or ISOTROPIC) appear more often in drawings because you need labels to explicitly convey them, and labels are the most prototypical VKP in drawings (see Figure 10).

Not surprisingly, as flow charts are clearly preferred for representing processes, they are also the image type mostly used to convey procedural relations such as *result_of* and *causes* (Figure 8; only the most representative relations are shown). In turn, photographs are clearly more adequate for *type_of* and *has_location*, whereas drawings are used evenly for *type_of* and *has_location*, as well as *part_of* and *delimited_by*, all typical relations for the description of physical entities. Actually, *part_of* and *delimited_by* are specific to drawings.

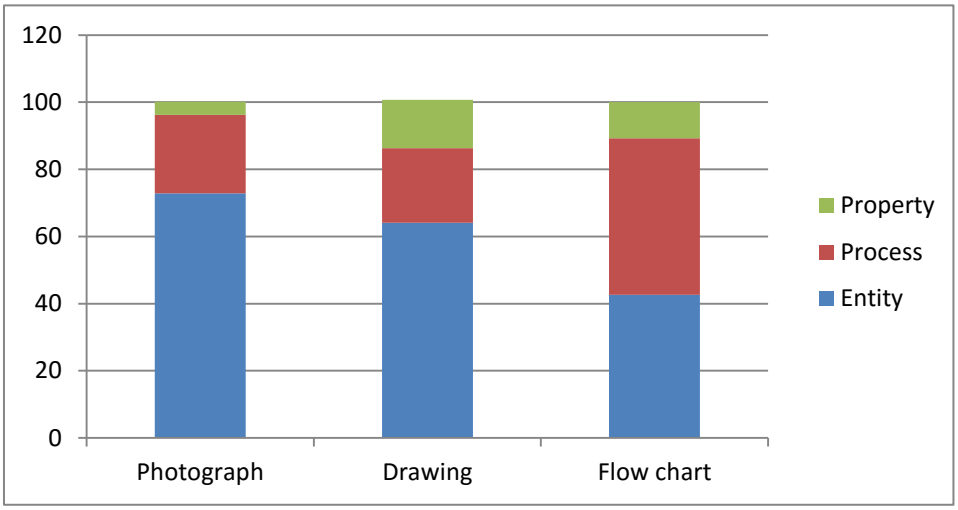

Figure 7: Correlation between image types and concept types

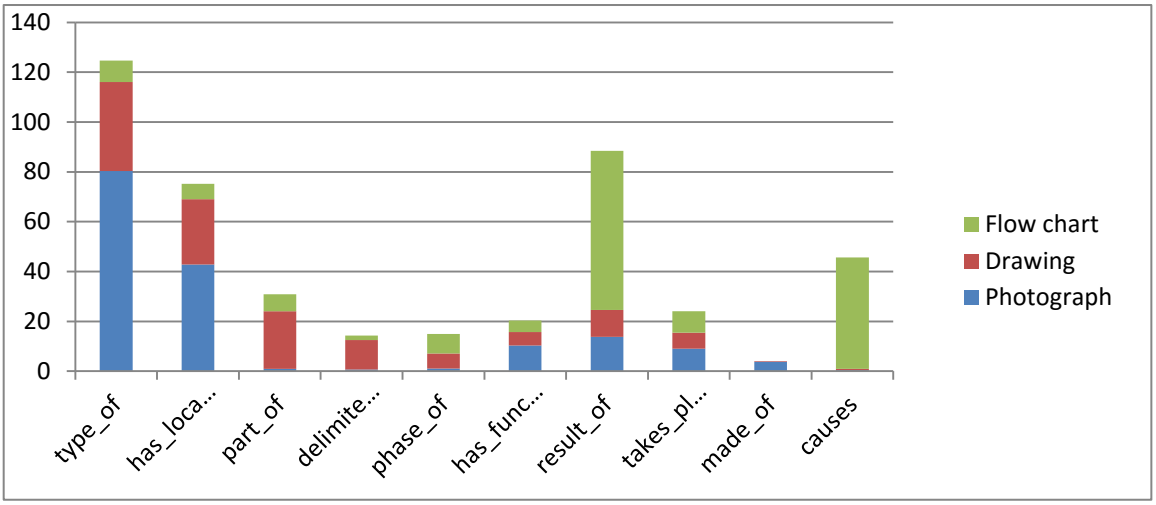

Figure 8: Correlation between conceptual relations and image types

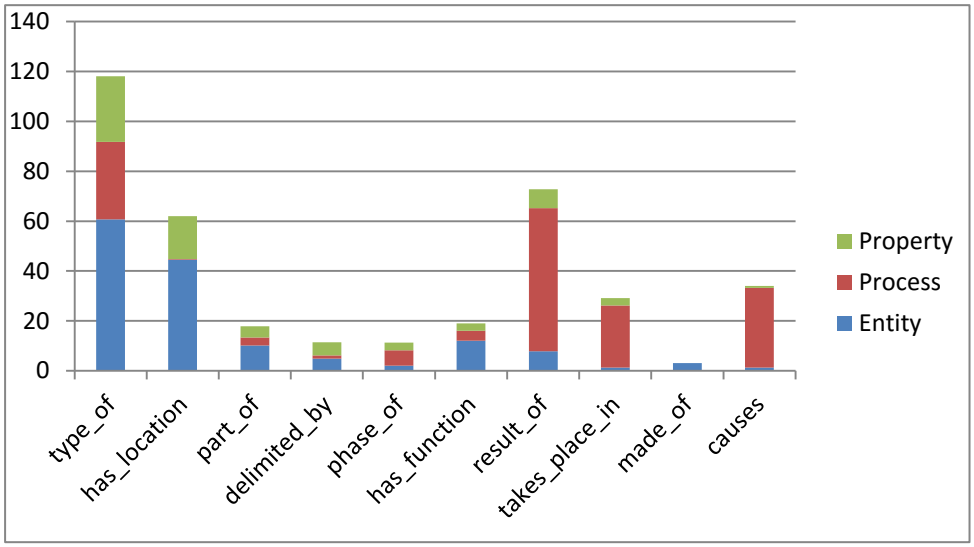

Figure 9: Correlation between conceptual relations and concept types

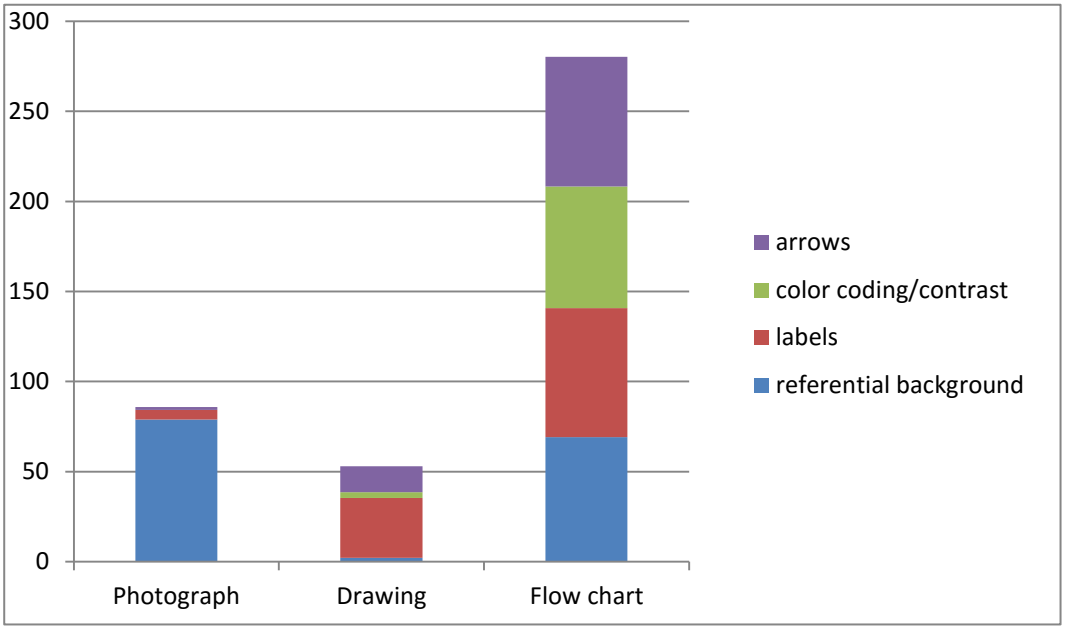

Figure 10: Correlation between image types and VKPs

Processes are mostly represented in images where relations such as *result_of*, *takes_place_in* and *causes* are present (see Figure 9). Entities, however, are found in images in combination with *type_of*, *has_location* and, to a lesser extent, *part_of* and *has_function*. *Has_function* and *part_of* are less represented because not all entities are functional nor do they have differentiated parts. Moreover, we have found that the *has_function* relation is not easy to convey in one image (see Figure 4). Properties are mostly present in images that convey *type_of* and *has_location*. They are also present, approximately in the same percentage, in combination with the other relations, except for *made_of*, which may be due to the surprising lack of data on this relation.

In Figure 10 only the most representative VKPs have been included. All of them are equally present in flow charts in EcoLexicon, they even appear all together in many of them. Photographs have clear referential backgrounds in over 75% of all cases and occasionally include labels. As previously mentioned, drawings are mostly characterized by labels and then arrows. The obvious conclusion we can draw from these results is that flow charts are the most interesting image type for the study of VKPs.

4. Conclusion

Much has been written regarding the importance of combining visual and textual information to enhance knowledge acquisition (Paivio, 1971, 1986; Mayer & Anderson, 1992). However, the combination of images and text still needs further analysis (Faber, 2012; Prieto Velasco, 2008; Prieto Velasco & Faber, 2012). An in-depth analysis of the features of images provides the means to develop selection criteria for specific representation purposes. The combination of conceptual content and image type based on morphological characteristics can be used to enhance the selection and annotation of images that explicitly focus on the conceptual propositions that best define concepts in a knowledge base. The analysis of EcoLexicon images has

provided the preliminary data to further explore how concept type, conceptual relations, and propositions affect the relation between VKPs and image types chosen for visual knowledge representation. Depending on how the annotation process evolves, new data will provide the clues for future content-based image description and retrieval.

5. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

6. Bibliographical References

Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/ Boston: De Gruyter Mouton.

Faber, P., León-Araúz, P. and Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era, Educational Linguistics* 19, ed. E. Bárcena, T. Read and J. Arhus. Berlin, Heidelberg: Springer.

Faber, P., León Araúz, P., Prieto Velasco, J. A., and Reimerink, A. (2007). Linking images and words: the description of specialized concepts. *International Journal of Lexicography* 20(1), pp. 39–65, doi:10.1093/ijl/ecl038.

León Araúz, P., Reimerink, A., and Faber, P. (2009). Knowledge extraction on multidimensional concepts: corpus pattern analysis (CPA) and concordances. In *8ème Conférence Internationale Terminologie et Intelligence Artificielle*. Toulouse.

Mayer, R. E. and Anderson, R. B. (1992). The instructive animation: helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology* 84(4), pp. 715–726.

Montero-Martínez, S. and García de Quesada, M. (2004). Designing a corpus-based grammar for pragmatic terminographic definitions. *Journal of Pragmatics* 36(2), pp. 265–291.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental Representations: A Dual-Coding Approach*. New York: Oxford University Press.

Prieto Velasco, J. A. (2008). *Información Gráfica y Grados de Especialidad en el Discurso Científico-Técnico: Un Estudio de Corpus*. PhD Thesis, University of Granada.

Prieto Velasco, J. A. and Faber, P. (2012). Graphical Information. In P. Faber (Ed.) *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton, pp. 225–248.

Reimerink, A., García de Quesada, M. and Montero Martínez, S. (2010) Contextual information in terminological knowledge bases: A multimodal approach. *Journal of pragmatics* 42(7), pp. 1928-1950.

6.1 Sources of Images

Figure 1:
http://www.beg.utexas.edu/UTopia/contentpg_images/gloss_barrier_island2.jpg

Figure 2:
http://water.usgs.gov/gotita/earthgwaquifer.html

Figure 3:
http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=464

Figure 4: http://www.dragadoshidraulicos.com/

Figure 5:
http://www.biosbcc.net/ocean/marinesci/04benthon/crform.htm

Figure 6:
http://cordellbank.noaa.gov/images/environment/upwelling_470.jpg

The Language-Dependence of Conceptual Structures in Karstology

Špela Vintar, Larisa Grčić Simeunović
University of Ljubljana, Faculty of Arts, University of Zadar, Institute of French and Iberoromance Languages
Aškerčeva 2, SI–1000 Ljubljana, Obala kralja Petra Krešimira IV, 2, HR–23000 Zadar
spela.vintar@ff.uni-lj.si, lgrcic@unizd.hr

Abstract

We explore definitions in the domain of karstology from a cross-language perspective with the aim of comparing the cognitive frames underlying defining strategies in Croatian and English. The experiment involved the semi-automatic extraction of definition candidates from our corpora, manual selection of valid examples, identification of functional units and semantic annotation with conceptual categories and relations. Our results comply with related frame-based approaches in that they clearly demonstrate the multidimensionality of concepts and the key factors affecting the choice of defining strategy, e.g. concept category, its place in the conceptual system of the domain and the communicative setting. Our approach extends related work by applying the frame-based view on a new language pair and a new domain, and by performing a more detailed semantic analysis. The most interesting finding, however, regards the cross-language comparison; it seems that definition frames are language- and/or culture-specific in that certain conceptual structures may exist in one language but not the other. These results imply that a cross-linguistic analysis of conceptual structures is an essential step in the construction of knowledge bases, ontologies and other domain representations.

Keywords: definitions, conceptual structures, frame-based terminology, karstology, definition types, English-Croatian

1. Introduction

Definitions represent the core of conceptual structuring in a domain. According to the Aristotelian scholastic principle concepts must be categorized into classes and definitions represent the formal link between the concept (*definiendum*), its parent concept (*genus*) and the differentiating set of properties, allowing its assignment into a particular class (*differentia*). In the general theory of terminology (GTT; Wüster, 1974), the definition plays an equally central role of concept delineation, in other words, the pinning down of meaning before the assignment of the (linguistic) designation.

Recent decades have brought dramatic shifts in the understanding of concepts, definitions, and the linguistic reality of (intercultural) communication. Rather than being universal, concepts seem to be transient and dynamic in nature (Temmerman, 1997, 2000; Kageura, 2002), and insights from cognitive science provide evidence that concepts are in fact layers of associative networks which are "reloaded" each time they are evoked (Faber 2009, 2012). Definitions as found in specialized texts deviate from classical typologies (Seppälä, 2007), seem fuzzy and indeterminate (Leitchik & Shelov, 2007), and the traditionally advocated onomasiological approach proves inadequate in tackling variation, register, style (Pecman, 2014; De Santiago, 2014) or the complexity of cross-language and cross-cultural differences (Faber, 2012).

In the present paper we explore the multiple formal and semantic dimensions of definitions through a corpus-based analysis. The empirical part of our study is based on the cognitive model of terminology as proposed by Faber (2009), because we find that definition frames provide a helpful tool in exploring the multidimensionality of concepts, especially if we seek to demonstrate that the cognitive-semantic components chosen to define a concept in a particular context vary.

We examine definitions in English and Croatian didactic and scientific texts in the domain of karstology, and by comparing different definitions of the same concept, we show that the choice of the defining strategy is influenced by a number of factors, including the perspective from which the concept is described in an interdisciplinary domain, register (didactic vs. scientific) and language (English vs. Croatian).

2. Finding and "Framing" Definitions

Empirical analyses of authentic texts confirm that defining strategies can be multifarious and highly dependent on register, domain, and style of writing (Pollak, 2014). Pollak (2014) explored definition types in an English and Slovene corpus of language technologies as a step preceding the design of a definition extraction algorithm. Not only did she identify over 20 definition types, but she also arrived at the conclusion that almost 40% of the definition candidates were borderline cases which could be regarded as definitions or not. A validation experiment with 20 students who were required to mark sentences as either definitions or non-definitions resulted in inter-annotator agreement of 0.36 (kappa), which is very low.

Automatic extraction of definitions from text is a well-researched topic within Natural Language Processing. Many early approaches to definition extraction relied on morphosyntactic patterns presupposing the analytical definition type (Klavans & Muresan, 2001), later extended with more sophisticated grammars or lattices (Navigli & Velardi, 2010). Several approaches use machine learning techniques to distinguish between definitions and non-definitions (Fišer et al., 2010), and the combination of a base grammar and a classifier proved most successful than either of these techniques used alone (Degórski et al., 2008; Westerhout 2010). A common problem to all these attempts is low recall and/or low accuracy when extracting definitions from highly unstructured noisy corpora. Our own approach is semi-automatic in that we use lexicogrammatical patterns to extract definition candidates, but the distinguishing between definitions and non-definitions was performed manually.

Frame-based Terminology is a relatively recent attempt to reconcile the conceptual/cognitive layers of specialised knowledge and the textual reality. It uses a modified and

adapted version of Fillmore's Frames (Fillmore, 1976) coupled with premises from Cognitive Linguistics to configure specialized domains on the basis of definitional templates and to create situated representations of specialized knowledge concepts (Faber, 2002; Faber et al., 2006; Faber, 2012). The definition templates are based on corpus evidence from which typical concept features and relations are extracted and subsequently mapped to a framework of categories.

The definition patterns of individual conceptual categories are represented by combining dynamic semantic roles such as AGENT, PATIENT, INSTRUMENT, LOCATION etc. on the one hand with concept classes such as ENTITY, EVENT, PROPERTY or PHYSICAL OBJECT on the other. The conceptual structure of the domain is described via events or situations governed by non-hierarchical semantic relations between the concept classes, e.g. *causes, measures, has_function, has_form*. Such semantic frames represent possible cognitive structures used to define the meaning of a terminological unit. For our cross-language analysis of definitions in the karstology domain, we adapted the model proposed by Faber (2012) for EcoLexicon (ecolexicon.ugr.es) and introduced several additional concept categories and semantic relations. A prototypical event in karstology could be modelled with the following frame:

Natural AGENT: *erosion, tectonics* → causes process: *dissolution, sedimentation* → affects PATIENT: *rock, limestone* → results in: *uvalas, dolines, caves*.

3. Identifying Definition Frames in Two Languages

Our corpus-based analysis of definitions was performed on a comparable English-Croatian corpus of karstology, where for each language the corpus consisted of two subcorpora, one containing scientific texts (doctoral dissertations, scientific papers, conference proceedings) and the other, didactic texts (textbooks and lecture notes). Both corpora are comparable in size: the Croatian corpus contains 881,174 tokens, whereas the English corpus has 913,416 tokens (see Table 1). The English and the Croatian corpora can be considered comparable in terms of domain and text types included, but the portion of scientific texts is slightly larger in Croatian.

		English	Croatian
Scientific	Number of texts	23	9
	Tokens	499.422	628.138
Didactic	Number of texts	17	9
	Tokens	413.974	253.036
Total	Number of texts	40	18
	Tokens	913.416	881.174

Table 1: Basic corpus data

Both corpora received standard pre-processing including tokenization, PoS-tagging and lemmatization. For Croatian, pre-processing was performed with a recently developed tagger (Agić et al., 2013). For the pre-processing of English and for corpus querying we used the SketchEngine facilities (Kilgarriff et al., 2014).

Our analysis involved the following steps:

- extraction of definition candidates using lexico-syntactic patterns,
- validation of definition candidates, and
- annotation of definitions with semantic categories and relations.

In the following subsections these steps are described in more detail.

3.1 Extraction and Validation of Definition Candidates

Definition candidates were extracted using a set of lexicosyntactic patterns, designed specifically for each language on the basis of previous research into definition extraction (Fišer & Pollak & Vintar, 2010; Pollak, 2014). Some of these patterns assume the traditional analytical definition (*[NP]-is-a-[NP]*), while others may contain only a trigger word or phrase (*term, be-defined-as*), and will therefore frequently capture definitions of an entirely different format. Croatian and English patterns are similar but not completely parallel. For example, the trigger word *term* has two near-synonyms in Croatian, and we used all three (*termin / naziv / izraz*).

Clearly these pattern lists are not exhaustive and other potentially fruitful expressions could also be used, but since we were not aiming for total recall, their yield was deemed satisfactory. Table 2 lists the patterns for Croatian and English, the number of candidates yielded by each pattern, and the number of definitions retained after manual validation.

The manual validation was not an easy task, especially considering the variability of definitions discussed in Section 2. We retained sentences which contained an explanation of the definiendum in any form by giving at least one distinguishing feature. In this way, several sentences were retained although they contained no genus. As can be observed in Table 2, the majority of candidate sentences were still discarded, and several cases, not listed above, were either marked as borderline or as KRC (knowledge-rich context). In the end we limited our analysis only to true definitions and ignored semi-definitions and KRCs, even though they also contained important conceptual relations. Some definitions were extracted via several patterns. After removing duplicates, the final data set consisted of 191 examples for Croatian and 142 for English.

3.2 Annotating Definitions with Conceptual Categories and Relations

For this step we first needed to define the domain-specific categories and relations to be used in annotation. A preliminary classification of karstology concepts into semantic classes had been previously performed by Grčić Simeunović (2014), which was a useful starting point. For pragmatic reasons semantic classes were added during annotation in case the need arose. As a result, the final inventory consisted of 30 classes, including*: limestone area, landform, water cycle, opening, process, measure, method, layer, minerals, rock characteristics, substance, territory, physical phenomenon, information system, situation, geographical boundary* etc.

Croatian	# candidates extracted	# definitions	% definitions	English	# candidates extracted	# definitions	% definitions
naziv	455	84	18.46	term	444	64	14.41
izraz	169	5	2.96				
termin	25	14	56				
N-biti-N	345	28	8.12	N-is-a-N	98	21	21.43
				N-be-used	92	4	4.35
N-predstavljati	219	31	14.15	N-represent	81	3	3.70
nazivati-se	24	17	70.83	be-called	74	20	27.03
N-biti-A-N	134	12	8.95	N-is-a-A-N	71	9	12.68
definirati-se-kao	1	1	100	be-defined-as	45	32	71.11
sadržati	61	10	16.39	N-contain	40	4	10.00
N- značiti	133	3	2.25	N-mean	27	2	7.41
zvati-se	12	2	16.67	N-refer-to	15	3	20.00
N-sastojati-se	18	2	11.11				
možemo-podijeliti-na	6	0	0				
proces-Ng	106	11	10.38				
Total	1708	220			987	162	

Table 2: Definition extraction patterns and yield

In each definition we first identified the *definiendum* (the concept being defined) and the *genus* (superordinate concept), when present. Those two concepts were then assigned to a semantic class in accordance with the information contained in the definition. For the remaining part of the definition, which in most cases represents the *differentia*, no further semantic classes were assigned. Instead, we identified the conceptual relations activated by the context. The following example illustrates this procedure:

Definition sentence:
Less permeable rock below an aquifer that keeps groundwater from draining away is called a confining bed (also known as aquitard or aquiclude).

Definiendum:	confining bed / aquitard / aquiclude
Definiendum class:	**hydrological form**
Genus:	rock
Genus class:	**mineral**
Differentia:	less permeable
Relation:	**has_attribute**
Differentia:	below an aquifer
Relation:	**has_location**
Differentia:	keeps groundwater from draining away
Relation:	**has_function**

Table 3: Example of semantic categories and relations found in a definition

Thus, a *hydrological form* is defined by specifying the *attribute*, *location* and *function* of its genus.

We described our dataset with a total of 23 relations. The total number of relation instances in the dataset was 509. Determining the semantic relation governing the relationship between the concept and its specific properties is not always straightforward, and in many cases, the distinctions between categories are difficult to draw. Our annotation preserved the AGENT - PATIENT and CAUSE - EFFECT dimensions of the karstological event, which is why we differentiate between the *causes* and *has_result* relations. This is illustrated by the examples below. In (1) *rainfall excess* is the natural agent causing *flooding*, while in (2) the *exposure of the river to the surface* happens as an effect of the *underground cavern collapsing*. In the first definition, we thus identified the relations of *causes* and *defined_as*, while the second definition contains the relations *caused_by, has_form* and *has_result.*

(1) *Threshold runoff has been defined as the amount of rainfall excess of a given duration necessary to cause flooding on small streams.*
(2) *Short steep-sided valleys caused by collapse of an underground cavern and exposing the river to the surface are called karst windows.*

Looking at the frequencies of individual relations occurring in the Croatian versus the English corpus, there are many differences, especially as some relations occur in one language but not the other. However, there are also a number of similarities. Both languages share the two most frequently observed relations *has_form* and *has_location*. For Croatian, the list continues with *caused_by, has_function* and *has_attribute*, and for English, with *has_attribute, defined_as* and *has_function*. This would indicate that LOCATION, CAUSE, FUNCTION and ATTRIBUTE (physical or other) represent the key semantic properties of concepts in the domain of karstology regardless of language or register. In the Croatian dataset,

154 out of 191 definitions contain at least one of the above relations, and in the English set, 117 out of 142.

4. Analysis of Cross-Language Aspects of Definition Frames

The frequencies of semantic categories for the concepts defined in our karstology corpus revealed some thematic differences between our subcorpora. Apparently the Croatian texts contain a larger proportion of definitions for landforms (e.g. *hum, klanac, škrip, čučevac*), while the English texts seem to place a slightly greater emphasis on hydrological phenomena and forms as well as on different types of limestone areas, mainly karst itself or karst types (e.g. *karst, epikarst, bradikarst, fluviokarst*). These differences point to the irregularities of the term formation process where some realities are given a stable name or term in one language but not in the other.

We were also interested in the distribution of semantic relations across the concept categories. The assumption that a certain conceptual category will be more likely defined via a specific set of relations, thus constituting typical definition frames for each category, led us to formulate a cognitive model of the selected domain. While we might expect such frames to be universal (e.g. a landform may be described by its form regardless of language or register), we were particularly interested in verifying this assumption with our bilingual dataset.

Table 4 shows the relations occurring in a particular concept category typical of each language. For *landform* it seems that definition frames are universal at least in the top three relations. As might be expected, a landform is typically defined by specifying its form, location, and the natural process that contributed to its formation. The lower part of the list seems less aligned though, and there seems to be little correspondence between languages. A similar impression is conveyed by the list for *process*. Processes are usually not described in terms of their form or their composition, which explains the absence of relations such as *has_form, similar_to, made_of* and *has_part*. On the other hand, a process may be defined or even computationally modelled, and may exhibit a time pattern. The category *limestone area* was more frequent in the English subcorpus, but apart from this difference, we were surprised to find the *has_result* relation in English but not in Croatian. This agent-like relation is usually expected to occur with processes and not concepts that we consider static, such as territories or areas. We thus decided to take a closer look at definitions of *karst* and related terms in both languages in order to see whether the resulting cognitive models of the domain overlapped.

The Croatian corpus contains 13 sentences defining either *karst* (*krš*, 3) or types of karst (*klastokrš* 2, *tektokrš*, *škrapavi krš, linearni krš, fluviokrš, boginjavi krš, hidrotermokarst, obalni krš*). While all of these *definienda* belong to the same category of *limestone area*, their genus concepts fall into two groups. More specifically, eight of the examples define *karst* or karst type as a kind of terrain, area or relief form, while four definitions choose the genus *pojava* (phenomenon). The differentia of the definitions contain the following relations: has_location (9), made_of (5), caused_by (3), result_of (2), has_part (2), develops_from (1). The three Croatian definitions of *karst* (examples 3-5) illustrate the context-dependence and multidimensionality of the concept *karst*. In example (3), *karst* is defined as a relief form developing on soluble rock (limestone, dolomite etc.). This is not surprising since this definition belongs to the didactic part of our corpus, which is more specifically composed of textbooks.

Example (4) is a less typical definition in that it focuses on the processes and agents contributing to the formation of karst. On the other hand, example 5 defines *karst* as a group of morphological and hydrological phenomena found on soluble rock.

landform	CRO	EN	process	CRO	EN	limestone area	CRO	EN
has_form	31	14	caused_by	3	1	has_location	4	10
has_location	21	8	has_location	2	0	result of	0	10
caused_by	13	6	has_attribute	1	1	caused_by	3	7
made_of	9	1	defined_as	1	3	has_attribute	3	6
has_attribute	7	2	computed_as	1	0	has_form	3	5
similar_to	4	0	has_result	1	4	has_result	0	4
contains	0	5	has_time_pattern	0	2	contains	1	4
result_of	0	4	causes	0	1	made of	2	4
has_function	3	2				has_function	2	2
has_part	1	2						
causes	1	0						

Table 4: : Cross-language comparison of relations occurring with selected concept categories

(3) Krš je specifičan oblik reljefa koji se razvija na topivim stijenama (vapnenac , dolomit, sol, gips).
[Karst is a specific relief form which develops on soluble rock (limestone, dolomite, salt, gypsum).]

(4) Krš kao reljef na topivim stijenama predstavlja rezultat raznolikih i međusobno uvjetovanih čimbenika kao npr. litološkog sastava, kemijskih procesa, pukotinske cirkulacije vode, tektonskih pokreta, klimatsko-bioloških čimbenika, a u novije vrijeme sve više dolazi do izražaja i utjecaj čovjeka .
[Karst as relief on soluble rocks represents the result of various and mutually interactive factors, such as the lithological composition, chemical processes, water

circulation in crevasses, tectonic movements, weather- and biology-related factors, and in recent times increasingly human interventions.]
(5) Krš je specifičan skup morfoloških i hidroloških pojava u topivim stijenama, prije svih vapnenačkim i dolomitskim [...].
[Karst is a specific set of morphological and hydrological phenomena occurring on soluble rocks, mostly limestone and dolomite [...].]

The English subcorpus has as many as 25 definitions for *karst* (10) or its subtypes: *hydrothermal karst, hypogene karst* (2), *endokarst, epikarst* (3), *contact karst, bradikarst, ore-bearing karst, anomalous hydrothermal karst, heterogeneous karst, fluviokarst, doline karst, thermal karst*. The majority of the genus concepts used to define these terms belong to the categories limestone area, territory or relief form, just as in Croatian. A surprising observation, however, was the fact that four definitions describe *karst* (or its subtype) as a process (examples 6-9), and as a consequence, has_result is one of its relations.

(6) In the broadest sense, hydrothermal karst is defined as the ***process*** *of dissolution and possible subsequent infilling of cavities in the rock by the action of thermal water.*
(7) Here we introduce the working term "anomalous hydrothermal karst" to describe the hydrothermal ***process*** *developing in zones where the steady-state thermal field of the hydrosphere is disturbed.*
(8) In the most general terms, karst may be defined as a ***process*** *of interaction between soluble rocks and different waters, as a result of which characteristic features develop on the Earth's surface and underground.*
(9) Hypogene karst is defined as ***the formation of caves by water*** *that recharges the soluble formation from below, driven by hydrostatic pressure or other sources of energy, independent of the recharge from the overlying or immediately adjacent surface.*

This observation supports the view that the cognitive structures governing knowledge presentation in a specialized text are not universal and depend not only on context, register, or the author's beliefs, but also on the language in which the definition is formulated. Our corpus-based evidence shows that the definition frame [limestone area] is_a [process] has_result [result] is possible in English, but not in Croatian. Quite possibly, the concept of karst activates slightly different layers of meaning for a speaker of Croatian (or Slovene), because the term originates from the geographical area Kras and thus bears a strong associative link to a (static and physically identifiable) landscape.

5. Conclusions

In a previous monolingual study (Grčić Simeunović & Vintar, 2015), we explored the multidimensionality of karstology concepts and the effects of register, context, and style on the range of concept properties chosen for the definition. This study extends those findings into the space of cross-language comparison. The results obtained seem to indicate that cognitive structures underlying knowledge transfer, of which specialised texts are a surface representation, may be influenced by language and culture. While the concept of *karst* can only be defined as a type of terrain in Croatian, in English within certain contexts, it is described as a process.
This observation is interesting for a number of reasons. Firstly, it challenges the efforts to build language-independent domain representations, such as ontologies or semantic networks of the WordNet type. Secondly, it could have important implications for multilingual terminography, which for the most part remains rooted in the traditional concept-oriented approach and has so far included language or translation-specific information mostly in the form of collocations and phraseology. Finally, it would be worthwhile to fully understand the reasons why such profound differences in cognitive frames come to exist, even in the realm of specialised discourse. In the case of our experiment, we suspect that the relation between the "donor" and "receiver" language regarding the origin of terms may play a certain role, in the sense that karstology concepts might have initially evolved in a close relationship with the geographical (and cultural and linguistic) reality represented by Karst as a region. Given the dynamic nature of concepts, the layers constituting the cognitive boundaries of a concept may be restructured or modified through the transfer and expansion of knowledge to other languages and cultures, as well as through interdisciplinarity.
Further research in underway to explore cross-language conceptual relations between Croatian, Slovene and English.

6. Acknowledgements

An extended version of this paper has been accepted for publication in Fachsprache, ISSN 1017-3285, to appear in 2017.

7. Bibliographical References

Agić, Ž., Ljubešić, N.,/Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. *Proceedings of BSNLP 2013*. Sofia, Bulgaria.

Degórski, L., Marcinczuk, M., Przepiórkowski, A. (2008). Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. *Proceedings of the Sixth international conference on Language Resources and Evaluation* (LREC 2008). ELRA.

De Santiago, P. (2014). De la forma al contenido, del contenido a la definición. *Normas: Revista de Estudios Lingüísticos Hispanicos* 6: 28-44.

Faber, P. (2002). Terminographic definition and roncept Representation. *Training the language services provider for the new millennium*. Ed. Belinda Maia/ Johann Haller/Margherita Ulyrich. Porto: Universidade do Porto. 343-354.

Faber, P., Montero Martínez, S., Castro Prieto, M. R., Senso Ruiz, J., Prieto Velasco, J. A., León Arauz, P., Márquez L., C., Vega Expósito, M. (2006). Process oriented terminology management in the domain of coastal engineering. *Terminology* 12 (2): 189-213.

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografias de*

Traducción e Interpretación 1: 107-134.
Faber, P., Ed. (2012). *A cognitive linguistics view of terminology and specialized language.* Berlin, Boston: De Gruyter Mouton.
Fillmore, Ch. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*: Origins and evolution of language and speech. Vol. 280/1: 20-32.
Fišer, D., Pollak, S., Vintar, Š. (2010). Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. *Proceedings of LREC*. Ed. Nicoletta Calzolari et al. Malta, Valletta: ELRA: ELDA: ILC. 2932-2936.
Grčić Simeunović, L. (2014). *Methodology of terminological description for the purposes of specialized translation.* Unpublished PhD thesis. (in Croatian) Zadar: University of Zadar.
Grčić Simeunović, L., Vintar, Š. (2015). Domain modelling: Comparative analysis of definition styles. *Od Šuleka do Schengena.* Ed. Maja Bratanić et al. (in Croatian) Zagreb: Institut za hrvatski jezik i jezikoslovlje. 251 - 266.
Kageura, K. (2002). *The dynamics of terminology: A descriptive theory of term formation and terminological growth*. Amsterdam/Philadelphia: John Benjamins.
Klavans, J. L./Muresan, S. (2001). Evaluation of DEFINDER: A system to mine definitions from consumer-oriented medical text. *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*. ACM. 201-202.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography* 1: 7-36.
Leitchik, V. M., Shelov, S. D. (2007). Commensurability of scientific theories and indeterminacy of terminological concepts. *Indeterminacy in terminology and LSP: Studies in honour of Heribert Picht*. Ed. Bassey E. Antia. Amsterdam/Philadelphia: John Benjamins. 93-106.
Navigli, R., Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics.
Pecman, M. (2014). Variation as a cognitive device: How scientists construct knowledge through term formation. *Terminology* 20(1): 1-24.
Pollak, S. (2014). *Semi-automatic domain modeling from multilingual corpora*. Unpublished PhD thesis. Ljubljana: Department of Translation Studies, Faculty of Arts.
Seppälä, S. (2007). La définition en terminologie: typologies et critères définitoires. *TOTh Terminologie et Ontologie: théories et applications.* Annecy: Institut Porphyre. 23-44.
Shelov, S. D. (1990). Typology of term definitions (comparison of normative and non-normative terminological dictionaries). *Nauchno-Tekhnicheskaya Terminologiya 4* (in Russian): 16-27.
Temmerman, R. (1997). Questioning the univocity ideal. The difference between SocioCognitive Terminology and traditional Terminology. *Hermes. Journal of Linguistics* 18: 51-91.
Temmerman, R. (2000). *Towards new ways of terminological description. The sociocognitive approach*. Amsterdam/Philadelphia: John Benjamins.
Westerhout, E. (2010). *Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch*. University of Utrecht: Lot Dissertation Series.
Wüster, Eugen (1974). Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. *Linguistics* 119: 61-106.

Growth of the Terminological Networks in Junior-high and High School Textbooks

Takuma Asaishi, Kyo Kageura
Department of Human Informatics, Aichi Shukutoku University
2-9, Katahira Nagakute City, Japan
tasaishi@asu.aasa.ac.jp, kyo@p.u-tokyo.ac.jp

Abstract

In this paper, we analyze the mode of deployment of the concepts when reading through textbooks. In order to do so, we construct terminological networks guided by the discourse, with vertices representing index terms and edges representing the co-occurrence of index terms in a paragraph. We observe the growth of the terminological networks in junior-high and high school Japanese textbooks in four domains (physics, chemistry, biology and earth science). Our primary results are as follows: (1) The rate of occurrence of new terms in high school textbooks is faster than in junior-high school textbooks regardless of the domain. (2) While several connected components grow independently in junior-high school textbooks, the largest component remains dominant in high school textbooks, regardless of the domain. (3) The average number of terms that are directly connected to a term increases, and many more terms are directly connected in high school textbooks than in junior-high school textbooks in all domains except physics. In addition, terms are indirectly connected through a few terms on average and are strongly connected in partial groups throughout the text, regardless of the domain and school level. (4) The degree of centralization (i.e., few terms are connected to many terms directly or facilitate indirect connection between terms) deteriorates regardless of the domain and school level.

Keywords: complex network, terminology, textbook, knowledge, network analysis

1. Introduction

In order to analyze the mode of deployment of concepts when reading through textbooks, we construct terminological networks guided by the discourse in text, with vertices representing index terms and edges representing the co-occurrence of index terms in a paragraph. We observe the growth of terminological networks by observing the transitions of network statistics. Our target texts are Japanese junior-high and high school textbooks in four domains (physics, chemistry, biology and earth science). We analyze textbooks because the reader and domain are clear, and therefore the result will be used as a point of reference to characterize various texts.

The structural nature of terminology has been analyzed from the viewpoint of network structure. The target texts are not only literary works (Masucci and Rodgers, 2009; Amancio et al., 2012) but also scientific articles (Grabska-Gradzinska et al., 2012) and writing by students (Ke et al., 2014). These revealed that the terminological network has a scale-free and small-world nature, in common with other social networks. In addition, it is revealed that some network statistics are correlated with authorship and text quality (Antiqueira et al., 2007; Amancio et al., 2011).

The exploration of terminological networks based on the syntagmatic relations of terms, which reflect the mode of deployment of concepts in text, is a theoretical issue for the formal analysis of terminology. It is complementary to the terminological network based on the paradigmatic relation of terms, which reflect the conceptual structure in thesauri (Motter et al., 2002; Sigman and Cecci, 2002) or glossaries (Asaishi and Kageura, 2011; Kageura, 2012), as well as text-oriented studies such as text segmentation or the evaluation of readability (Hearst, 1997; Sato et al., 2008). In the terminological networks from text, two terms are connected if they are adjacent or co-occur in a sentence.

The dynamic modeling of the network has been also actively pursued (Barabasi and Albert, 1999). The growth of terminological networks that reflect the mode of deployment of concepts in texts, however, has not yet systematically been addressed although some exceptions exist (Masucci and Rodgers, 2006). Asaishi (2016) analyzed the growth of terminological networks of Japanese high school textbooks. This paper applies the same analysis to junior-high school textbooks to generalize the results and reveal differences according to the school level.

This paper is organized as follows. In Section 2, we describe the framework of the terminological network, which reflects the mode of deployment of concepts. In Section 3, we explain the data or terminology in textbooks and show the basic statistics of terminological networks as a whole. In Section 4, we observe the growth of terminological networks and compare the results between junior-high and high school textbooks in four domains. In Section 5, we summarize the results and outline future work.

2. Framework

In textbooks, concepts, which constitute the basic elements of knowledge, are deployed in accordance with the progression of knowledge described in them. Concepts are represented by terms, and these co-occur in discoursal units such as sentences or paragraphs. We assume that co-occurrence of terms in the discoursal unit reflects the arrangement of knowledge in textbooks. Therefore, we can analyze how the description of knowledge advances in textbooks by observing the co-occurrence of terms according to the progression of textbooks.

To analyze the mode of deployment of concepts, we construct terminological networks, because terms are not men-

tioned in isolation but are related to each other through discourse. While most of the previous studies regard all terms excluding functional items as vertices in the network (Amancio et al., 2011; Grabska-Gradzinska et al., 2012), we restrict vertices to index terms. Index terms are considered to represent the important concepts needed to read textbooks (Asaishi, 2011). We use "index term" and "term" interchangeably throughout the paper.

A terminological network grows as a reader reads through a textbook. Unlike many of the previous works mentioned above, we construct terminological networks with edges representing co-occurrence of terms in a paragraph, because the important and interesting discourse phenomena of text, e.g., subtopic shift, occur at the level of the paragraph (Hearst, 1997). Furthermore, it is suggested that the paragraph is an important unit to evaluate and improve the readability of text (Sakai, 2011).

Formally our terminological network is an undirected weighted graph with a vertex representing index terms and an edge representing co-occurrences in a paragraph. For ease of analysis, however, we will not apply weighting in this paper. We observe the growth of this terminological network from the following four viewpoints to analyze the general and fundamental characteristics.

1. size of the network, or how many terms are included in the network
2. scope of connection, or how many terms are connected to the components, and how many terms remain isolates in the network
3. strength of connection, or how strongly terms are interconnected in the network
4. centralization of network, or the way in which few terms are centered in the network

The growth of terminological networks is visualized using suitable measures which correspond to these viewpoints, and the results of junior-high and high school textbooks are compared in four domains.

3. Data

To describe the mode of deployment of concepts according to the school level, we compare the growth of terminological networks between two junior-high school textbooks, Science Field 1 and Science Field 2, and four high school textbooks, Physics, Chemistry, Biology and Earth Science. As Science Field 1 covers physics and chemistry while Science Field 2 covers biology and earth science, we compare the growth of the terminological networks of Science Field 1 with those of Physics and Chemistry, and the terminological networks of Science Field 2 with those of Biology and Earth Science. The market leader of the textbook publishing industry in Japan publishes all the textbooks. We construct the terminological networks in the following steps.

1. The body text is manually extracted from textbooks and prepare text file. Footnotes, ruby and mathematical and chemical formulas are disregarded.
2. Text is segmented into paragraph. We define the space at the beginning of line as the boundary of paragraph.
3. Index terms are extracted from each paragraph by checking the headings of back-of-the-book index.
4. Terminological networks with vertices representing index terms and edges representing the co-occurrence of terms in a paragraph are constructed.

Table 1 shows the basic statistics of terminologies and texts in each textbook. In Table 1, VN and N are the number of types and tokens of index terms, and P is the number of paragraphs. From Table 1, we can observe that high school textbooks have much more terminology and longer text than junior-high school textbooks. We can also observe that each term occurs 5〜14 times in the whole text, and that 5〜12 term tokens co-occur in a paragraph on average.

	VN	N	P	N/VN	N/P
Science Field 1	200	2647	345	13.24	7.67
Science Field 2	180	1408	289	7.82	4.87
Physics	289	3109	370	10.76	8.40
Chemistry	551	7887	674	14.31	11.70
Biology	517	4462	464	8.63	9.62
Earth Science	492	2647	440	5.38	6.02

Table 1: The basic statistics of terminology and textbooks

Table 2 shows the basic statistics of the terminological networks. In Table 2, $|G|$ and $||G||$ indicate the number of nodes and edges, respectively, $\#C$ is the number of connected components, and I is the number of isolates. From Table 2, we can observe that terminological networks consist of a few components and isolates as a whole, regardless of the domain and school level.

	$\|G\|$	$\|\|G\|\|$	$\#C$	I
Science Field 1	200	1896	2	3
Science Field 2	180	791	1	2
Physics	289	2287	1	1
Chemistry	551	8077	1	0
Biology	517	4616	2	2
Earth Science	492	2640	3	3

Table 2: The basic statistics of terminological networks

Furthermore, as is described in the next section, terminological networks consist of the largest component and a few smaller components and isolates, regardless of the domain and school level. Given this situation, we first examine the size and scope of networks by observing the overall network. Then we will focus on the largest component to examine the strength and centralization of the network because the characteristics of the overall network largely reflect the largest component.

4. Growth of terminological networks

In this section we observe the growth of terminological networks and compare the results of junior-high and high

school textbooks in four domains. In the following figures, the x-axis shows the number of paragraphs (%) and the y-axis shows the value of each measure.

4.1. Size of network

Firstly we use $|G|$ or the number of terms to observe the size of the terminological network. The leftmost column in Figure 1 shows the transitions of $|G|$ according to the growth of networks. From Figure 1, we can observe that $|G|$ increases linearly in all textbooks. This shows that new terms constantly appear when a reader reads through a textbook, regardless of the domain and school level.
The rate of increase differs in all domains. We can observe that $|G|$ increases more rapidly for high school textbooks than for junior-high school textbooks. This means that the rate of occurrence of new terms is faster in high school textbooks compared to junior-high school textbooks regardless of the domain. Among the four domains, the difference between Science Field 1 and Physics is smaller than in other domains.

4.2. Scope of connection

Secondly we use $|G_c|(\%)$ or the number of terms taking part in the connected components to observe the scope of connection in the terminological network. $|G_c|(\%)$ is calculated by adding the size of all connected components, or the size of the network excluding isolates. Formally $|G_c|(\%)$ is defined as follows:

$$\begin{aligned} |G_c|(\%) &= 100 \times \frac{\sum_{j=1}^{\#C} |G_j|}{|G|} \\ &= 100 \times \left(1 - \frac{I}{|G|}\right) \end{aligned}$$

where $|G_j|$ is the size of the j-th components ($j = 1, 2, ..., \#C$). The second and third columns from the left in Figure 1 show the transition of $|G_c|(\%)$ and I.
From Figure 1, we can observe that $|G_c|(\%)$ retains a very high value throughout the text in all textbooks, although it largely fluctuates at the beginning of the text. This means that almost all terms in the networks connect to at least one other term from the first appearance. As a matter of fact we can observe that I is less than five throughout the text in all textbooks.
We can observe that $|G_c|(\%)$ is lower for junior-high school textbooks than for high school textbooks in all domains, especially in the earlier part of the text. $|G_c|(\%)$ of junior-high school textbooks is approximately 85% to 90% in the earlier part of the text; this percentage is 95% to 100% for high school textbooks throughout the text. Among high school textbooks, $|G_c|(\%)$ of Chemistry remains at approximetely 100% throughout the text.
As we mentioned in the previous section, all terminological networks consist of the largest components, with few very small components or isolates by the end of the textbook. It is therefore useful to observe $|G_1|(\%)$ or the number of terms that constitute the largest component. $|G_1|(\%)$ is defined as follows:

$$|G_1|(\%) = 100 \times \frac{|G_1|}{|G|}$$

The rightmost column of Figure 1 shows the transitions of $|G_1|(\%)$. Unlike $|G_c|(\%)$, transitions of $|G_1|(\%)$ in junior-high school and high school textbooks differ vastly. We can observe that $|G_1|(\%)$ of junior-high school textbooks decreases to between 40% and 60% in the earlier or middle part of the text, while in high school textbooks, it remains at about 100% except for at the beginning of the text. This means that several connected components grow independently in junior-high school textbooks, while the largest component retains a dominant position in high school textbooks regardless of the domain.

4.3. Strength of connection

The third viewpoint or the strength of connection has three aspects: (1) strength of direct connection, or how many terms are directly connected to a term, (2) strength of indirect connection, or how many terms are required to mediate between two terms, and (3) strength of connection within a partial group. In this subsection we use mean degree, average path length, diameter and the cluster coefficients as suitable measures.
We use mean degree Z to observe the direct connection of terms. Z is defined as follows:

$$\begin{aligned} Z &= \frac{1}{|G|} \sum_{i=1}^{|G|} k_i \\ &= \frac{2||G||}{|G|} \end{aligned}$$

where k_i is the degree of vertex v_i ($i = 1, 2, ..., |G|$) in the network. The leftmost column in Figure 2 shows the transitions of Z. From Figure 2, we can observe that Z increases as a whole regardless of the domain and school level. This means that the average number of terms that are directly connected to a term increases. However, a detailed observation of the transitions of Z shows that the ratio of increase is not constant. Z increases rapidly in the former part of the text, especially in high school textbooks.
We can observe that Z has a higher value for high school textbooks than for junior-high school textbooks, with the exception of Physics. This means that many more terms are directly connected to a term in high school textbooks than in junior-high school textbooks in most domains.
We use average path length L and diameter d to observe the strength of indirect connections of terms. L is the mean of the shortest distance between all pairs of vertices, and d is the maximum distance. L and d are defined as follows;

$$L = \frac{2}{|G|(|G|-1)} \sum_{i<j<|G|} d(v_i, v_j)$$

$$d = max\{\, d(v_i, v_j) : \, v_i, v_j \in |G| \,\}$$

where $d(v_i, v_j)$ is the length of the shortest path between v_i and v_j. The second and third columns from the left in

Figure 2 show the transitions of L and d. From Figure 2, we can observe that L increases to between 3 and 4 at the beginning of a text regardless of the domain and school level. Then, L approaches a stable value around the middle of the text. The transitions of d are similar to those of L in all textbooks. These results mean that terms are indirectly connected mediated by a few terms in the largest component, even if the size of the terminological network becomes large.

The differences of L and d between junior-high and high school textbooks are small in physics and chemistry. In biology, the junior-high school textbook has a higher value of L and d in the latter part of the text. In earth science, by contrast, the high school textbook has a higher value of L and d in the earlier part of the text.

Lastly we use the cluster coefficient C to analyze the strength of the connection within the partial group in the network. C represents the probability that the adjacent vertices of a vertex are connected. C is defined as follows:

$$C = \frac{1}{|G|}\sum_{i=1}^{|G|} C_i$$

C_i is defined as follows.

$$C_i = \frac{t_i}{k_i(k_i-1)/2}$$

where t_i is the number of triangles that contain v_i. The rightmost column of Figure 2 shows the transitions of C. From Figure 2, we can observe that C retains a high value (0.6∼0.8) in all the textbooks, although it largely fluctuates in the beginning of the text. The differences of C between junior-high and high school textbooks are small in all domains. These results mean that terms are always strongly connected within partial groups of the network throughout the text, regardless of the domain and school level.

4.4. Centralization

The fourth viewpoint or centralization has two aspects: (1) degree centralization, or few terms are directly connected to many terms, and (2) betweenness centralization, or few terms facilitate indirect connections between terms. In this subsection we observe the transitions of four measures based on two ideas of centralization.

Before presenting centralization measures, we must confirm the centrality $C(v_i)$ of v_i in the network because centralization is calculated using the centrality of all terms in the network. Degree centralization is calculated using degree centrality $C_d(v_i)$, while betweenness centralization is calculated using betweenness centrality $C_b(v_i)$. $C_d(v_i)$ and $C_b(v_i)$ are defined as follows:

$$C_d(v_i) = \frac{k_i}{|G|-1}$$

$$C_b(v_i) = \frac{\sum_{k=1}^{|G|}\frac{g_{kl}(v_i)}{g_{kl}}}{(|G|-1)(|G|-2)/2}$$

where g_{kl} is the number of shortest paths between v_k and v_l, and $g_{kl}(v_i)$ is the number of shortest paths between v_k and v_l via v_i ($i \neq k \neq l$).

Wasserman and Faust (1994) proposed the idea of centralization as a variance of $C(v_i)$. We call the centralization measures based on this idea CC_{dW} and CC_{bW}, which are calculated by $C_d(v_i)$ and $C_b(v_i)$. They are defined as follows:

$$CC_{dW} = \frac{\sum_{i=1}^{|G|}(C_d(v_i) - \overline{C_d(v_i)})^2}{|G|}$$

$$CC_{bW} = \frac{\sum_{i=1}^{|G|}(C_b(v_i) - \overline{C_b(v_i)})^2}{|G|}$$

where $\overline{C_d(v_i)}$ and $\overline{C_b(v_i)}$ are the mean value of $C_d(v_i)$ and $C_b(v_i)$. The first and second columns on the left in Figure 3 show the transitions of CC_{dW} and CC_{bW}. We take the logarithm after adding 1 to each value, in order to avoid $\log 0$. From Figure 3 we can observe that both CC_{dW} and CC_{bW} decrease in all textbooks, although they largely fluctuate at the beginning of the text. This means that the degree centralization and betweenness centralization deteriorate. We can also observe that CC_{dW} and CC_{bW} of junior-high school textbooks tend to have higher values than those of high school textbooks except at the beginning of the text, in most domains.

Freeman (1979) also proposed the idea of centralization, which indicates the degree to which the largest value exceeds the centrality of other vertices. We call the centralization measures based on this idea CC_{dF} and CC_{bF}, which are calculated by $C_d(v_i)$ and $C_b(v_i)$. They are defined as follows.

$$CC_{dF} = \frac{\sum_{i=1}^{|G|}(C_d(n^*) - C_d(v_i))}{(|G|-1)(|G|-2)}$$

$$CC_{bF} = \frac{\sum_{i=1}^{|G|}(C_b(n^*) - C_b(v_i))}{(|G|-1)^2(|G|-2)/2}$$

where $C_d(n^*)$ and $C_b(n^*)$ are the largest value of $C_d(v_i)$ and $C_b(v_i)$, respectively. The third column from the left and rightmost column in Figure 3 show the transitions of CC_{dF} and CC_{bF}, respectively. From Figure 3, we can observe that the transitions of CC_{dF} differ greatly by textbook. A comparison of the junior-high and high school textbooks shows that the greater values change places in all domains but biology.

From Figure 3 we can also observe that CC_{bF} has the lowest value in the middle and latter part of the text in all textbooks. This means that the degree to which a specific term facilitates indirect connection between two terms is very weak after terminological networks grow to some extent. However, we can also observe that CC_{bF} of junior-high

school textbooks are higher than those of high school textbooks in earlier part of the text in all domains.

5. Conclusion

In this paper, we analyzed the mode of deployment of concepts when reading through textbooks. In order to do so, we constructed terminological networks guided by the discourse, with vertices representing index terms and edges representing co-occurrence of index terms in a paragraph. We compared the growth of terminological networks between junior-high and high school textbooks in four domains (physics, chemistry, biology and earth science). Our main results are summarized as follows.

1. The rate of occurrence of new terms in high school textbooks is faster than in junior-high school textbooks, regardless of the domain.
2. While several connected components grow independently in junior-high school textbooks, the largest component remains dominant in high school textbooks, regardless of the domain.
3. The average number of terms that are directly connected to a term is increasing, and many more terms are directly connected in high school textbooks than in junior-high school textbooks in all domains except physics. In addition, terms are indirectly connected through a couple of terms on average, and are strongly connected in partial groups throughout the text, regardless of domain and school level.
4. The degree of centralization (i.e., few terms are connected to many terms directly or facilitate indirect connections between terms) deteriorates, regardless of the domain and school level.

Following the results presented in this paper, a comparative analysis of the textbooks of other publishers, subjects (for example, social science and history), and school level (elementary school and university) is possible to generalize the results in this paper. Then we will analyze texts other than textbook, to reveal the terminology usage according to the various attributions of text. We will also explore the possible applications using the results, such as text segmentation or evaluating the coherence of text.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26880005.

7. Bibliographical References

Amancio, D., Altmann, E., Oliveira Jr., O., and Costa, L. d. F. (2011). Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, 13.

Amancio, D., Oliveira Jr., O., and Costa, L. d. F. (2012). Using complex networks to quantify consistency in the use of words. *Journal of Statistical Mechanics: Theory and Experiment*, pages 1–22.

Antiqueira, L., Nunesa, M., Oliveira Jr., O., and Costa, L. F. (2007). Strong correlations between text quality and complex networks features. *Physica A*, 373(1):811–820.

Asaishi, T. and Kageura, K. (2011). Comparative analysis of the motivatedness structure of Japanese and English terminologies. In *Proceedings of 9th International Conference on Terminology and Artificial Intelligence*, pages 38–44.

Asaishi, T. (2011). An analysis of the terminological structure of index terms in textbooks. In *Proceedings of 12th Conference of the Pacific Association for Computational Linguistics (PACLING 2011) (poster session)*.

Asaishi, T. (2016). Developmental process of knowledge in high school science textbooks: analysis of the growth process of a terminological network in text. *Journal of Japan Society of Library and Information Science*, 62(1):38–53.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286.

Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Grabska-Gradzinska, I., Kulig, A., Kwapien, J., and Drozdz, S. (2012). Complex network analysis of literary and scientific texts. *International Journal of Modern Physics C*, 23(7).

Hearst, M. (1997). Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Kageura, K. (2012). *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. John Benjamins.

Ke, X., Zeng, Y., Ma, Q., and Zhu, L. (2014). Complex dynamics of text analysis. *Physica A*, 415(1):307–314.

Masucci, A. and Rodgers, G. (2006). Network properties of written human language. *Physical Review E*, 74(2).

Masucci, A. and Rodgers, G. (2009). Differences between normal and shuffled texts: structural properties of weighted networks. *Advances in Complex Systems*, 12(1):113–129.

Motter, A., de Moura, A., Lai, Y.-C., and Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, 65. 065102.

Sakai, Y. (2011). Improvement and evaluation of readability of Japanese health information texts: an experiment on the ease of reading and understanding written texts on disease (in Japanese). *Library and Information Science*, (65):1–35.

Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008). Automatic assessment of Japanese text readability based on a textbook corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 654–660.

Sigman, M. and Cecci, G. (2002). Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences USA*, 99(3):1742–1747.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

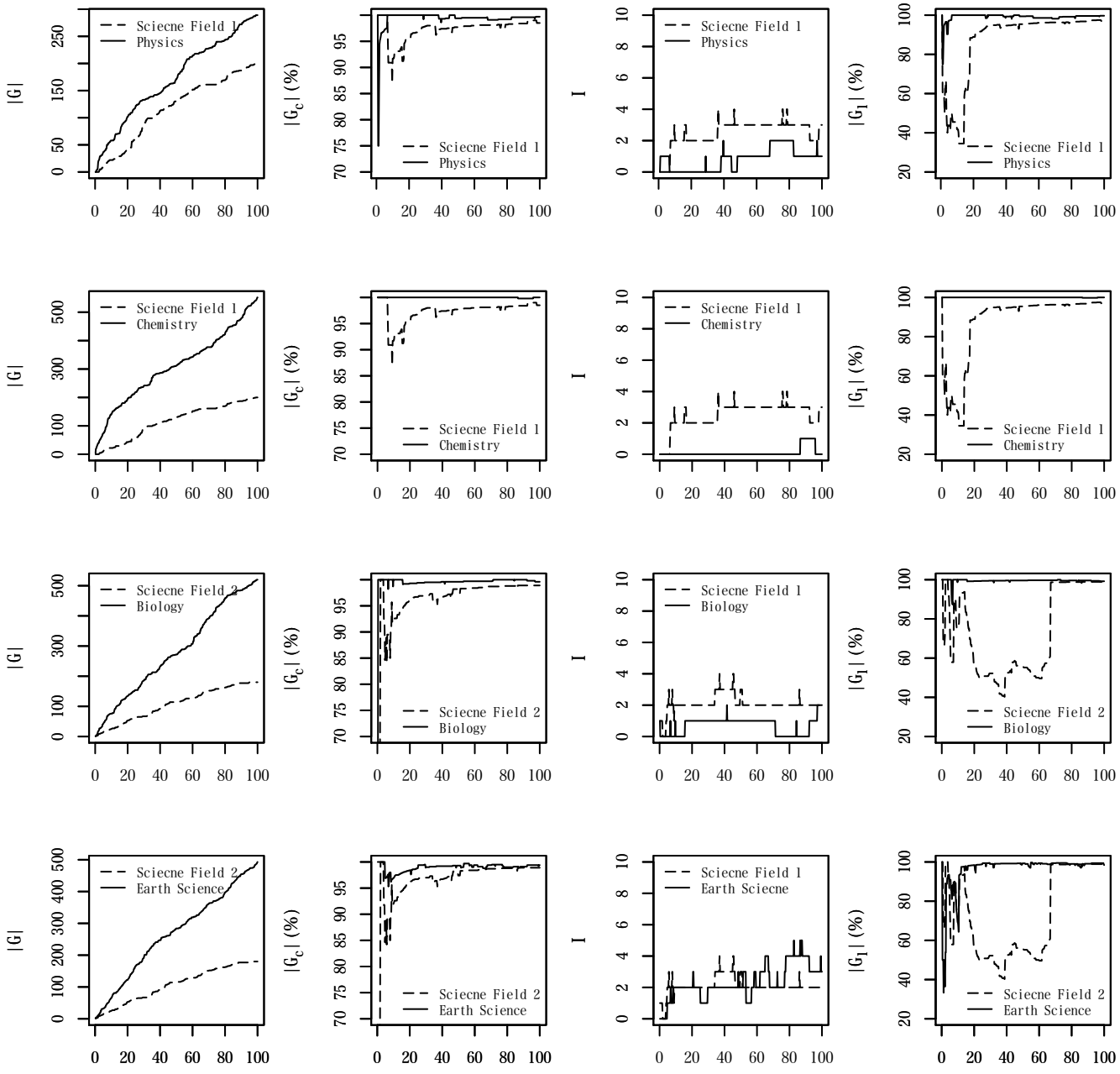

Figure 1: The network measures for size of network and scope of connection

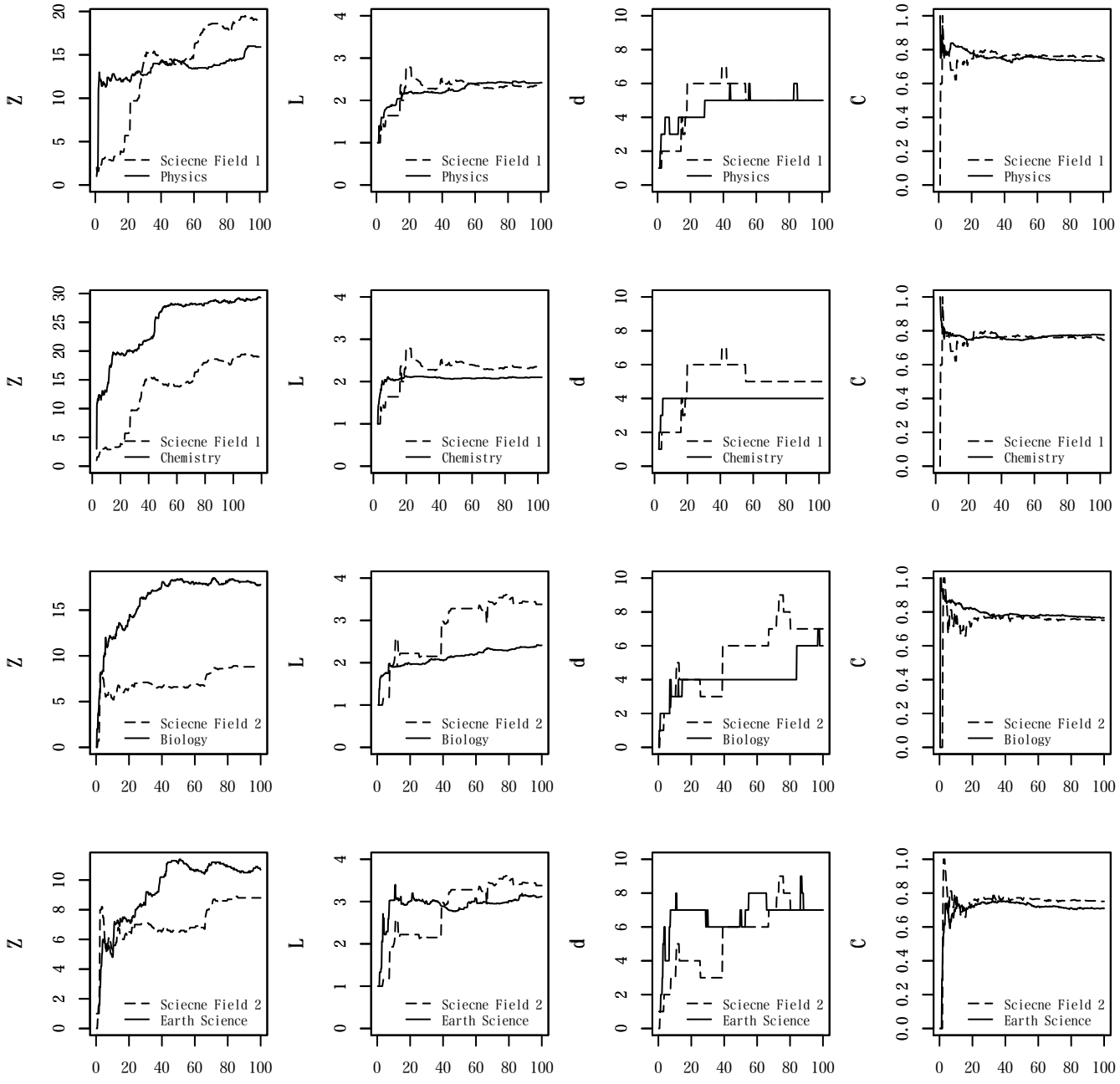

Figure 2: The network measures for strength of connection

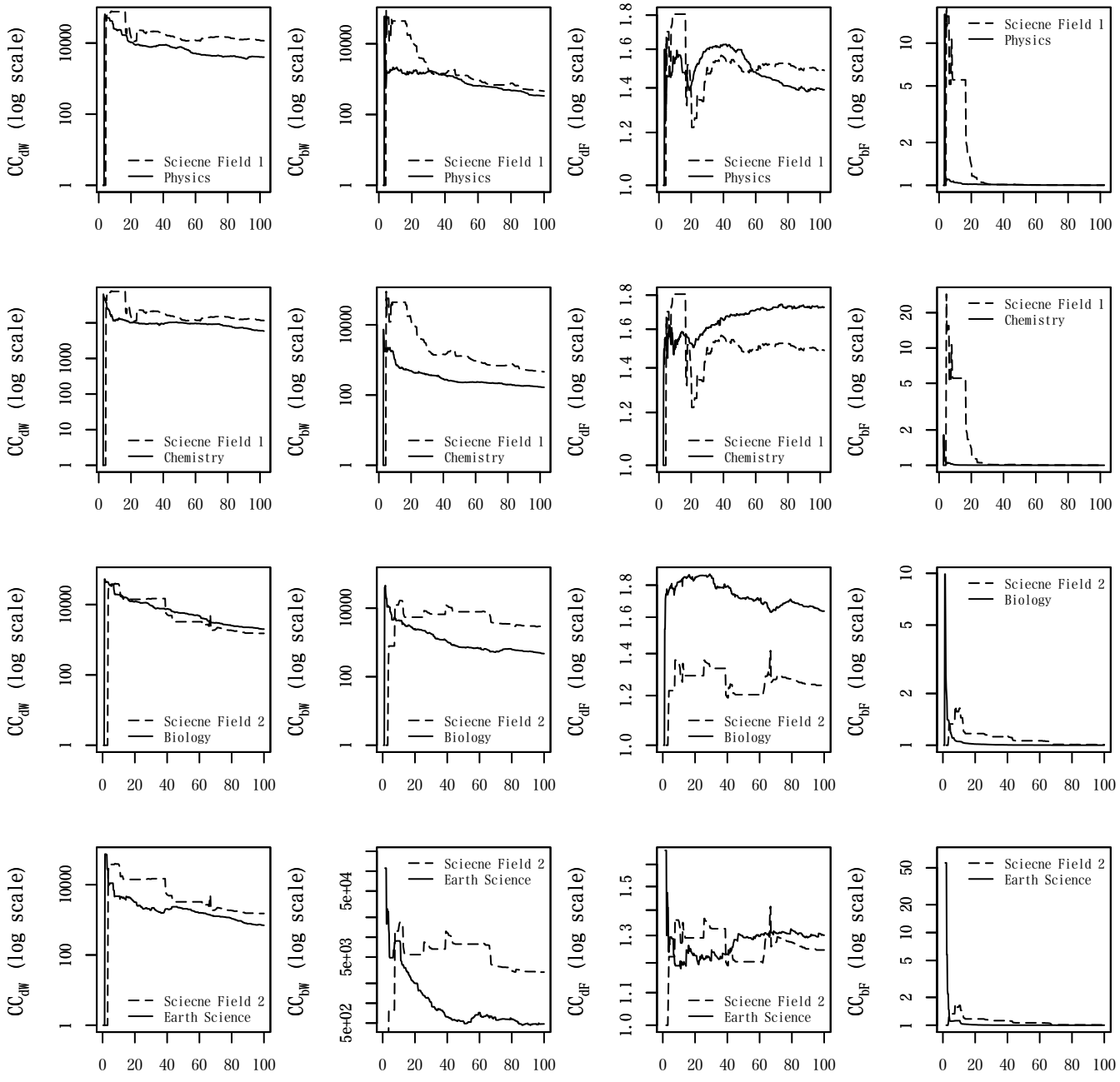

Figure 3: The network measures for centralization

Semantically Annotated Concepts in KDD's 2009-2015 Abstracts

Gabor Melli
OpenGov
Redwood City, CA, USA
e-mail: LangOnto2@gabormelli.com

Abstract

We introduce a linguistic resource composed of a semantically annotated corpus and a lexicalized ontology that are interlinked on mentions of concepts and entities. The corpus contains the paper abstracts from within the proceedings of ACM's SIGKDD conferences for the years 2009 through 2015. Each abstract was internally annotated to identify the concepts mentioned within the text. Then, where possible, each mention was linked to the appropriate concept node in the ontology focused on data science topics. Together they form one of the few semantic resources within a subfield of computing science. The joint dataset enables tasks such as temporal modeling of concepts over time, and the development of semantic annotation methods for documents with a large proportion of mid-level concept mentions. Finally, the resource also prepares for the transition into semantic navigation of computing science research publications. Both resources are publicly available at gabormelli.com/Projects/kdd/.

1. Introduction

We introduce a linguistic resource composed of a semantically annotated corpus, `kddcma2`, and a lexicalized ontology, `gmrkb2`, that are interlinked on mentions of concepts and entities. The corpus contains the 1,012 paper abstracts from the seventeen proceedings of ACM's SIGKDD's annual conferences for the years 2009 through to 2015. Each abstract was internally annotated to identify the concepts and entities mentioned within their text. Further, where feasible, each mention was linked to the appropriate record in the ontology, which is focused on the data science domain. To our knowledge the ontology and interlinked corpus is one of the few such resources for a computing science subdiscipline. Figure 1 illustrates the different items for the combined resource.

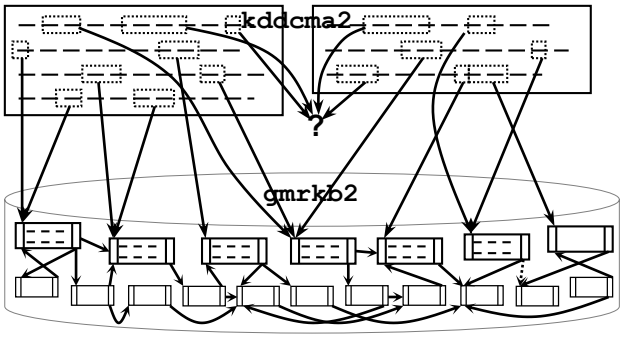

Figure 1 - Illustration of two abstracts with their concept mentions identified and linked to ontology nodes (that exist).

The resource has several possible applications. It directly enables the temporal modeling of topic trends that arise in the field. The dataset can also indirectly assist with the evaluation of terminology mining systems (Morin el al., 2007), and the semi-automated annotation of published research (Melli & Ester, 2010).

The remainder of the paper describes the joint resource in more detail. We begin with the corpus, then present the ontology and finally describe their interlinking.

2. The `kddcma2` corpus

The `kddcma2` corpus is composed of the 1,012 publicly accessible paper abstracts from the seventeen published proceedings for the years of 2009 to 2015 from the annual conferences hosted by ACM's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)[1]. The corpus significantly extends the `kdd09cam1` corpus with only 139 annotated abstracts from the KDD-2009 conference (Melli, 2010).

Each abstract was first annotated to identify the concepts and entities mentioned within their text. A *mention* is defined to be a substring of text characters that begins and ends a possible reference to an ontology node. Mentions can range from being references to specific entities such as named algorithms or publications, to abstract concepts such as algorithm and task types. Mentions can also range from being terminological units such as "*ER*" and "*mention linking*", to being entire terminological phrases, such as: "*entity resolution (ER)*" and "*problem of entity mention linking*". Figure 2 illustrates the result of the process.

Figure 2 - Annotation sample of two sentences

Unannotated Text: "*Collaborative filtering is the most popular approach to build recommender systems and has been successfully employed in many applications. However, it cannot make recommendations for so-called cold start users that have rated only a very small number of items.*"

Mention Annotation: [[*Collaborative filtering*]] *is the most popular* [[*approach*]] *to build* [[*recommender systems*]] *and has been successfully* [[*employed*]] *in many* [[*application*]]s. </s> *However, it cannot make* [[*recommendation*]]s *for so-called* [[*cold start user*]]s *that have* [[*rated*]] *only a* [[*very small*]] [[*number* of *items*]]. </s>

This corpus is akin to annotated corpi for biomedicine such as GENIA (Kim et al., 2003) and BioCreAtIvE (Hirschman, Yeh et al, 2005), and also the ERD-2014 corpus (Carmel et al., 2014) for common-place entity types. However, most of the annotated mentions in those corpi are of specific entity instances, and typically of "named" entities, such as the names of people, molecules, organisms, and locations. The `kddcm2` corpus on the other hand contains relatively few named entity mentions, and in cases where proper

[1] a premier peer-reviewed conference dl.acm.org/id=RE329

nouns are used, such as *Gibbs* and *Markov*, they are typically embedded within a terminological phrase, as in "*Gibbs sampling method*" and "*hidden Markov model*".

Each abstract was annotated using the following three-step process. First, they were automatically pre-annotated by the mention recognizer of the SDOI system (Melli, 2012) which applies a trained conditional random field (CRF-based) chunker. Next, we used Amazon Mechanical Turk workers to identify any obvious chunking errors. Often these errors were due to unexpected use of non-alphanumeric characters, such as abbreviations, quotes, parentheses, slashes and dashes. Finally, the author reviewed each abstract to remedy these errors and to add domain-specific repairs. This third step consumed approximately one minute per abstract.

Table 1 summarizes key statistics about the corpus's mentions. Of the 55,179 mentions approximately two thirds are single tokened, such as "*data*", "*C4.5*", and "*f-measure*", and the remaining third are multi-tokened, such as "*real-valued data set*" and "*minimal biclique set cover problem*".

Documents		1,012	PER DOC. min|med|max
Sentences		8,656	3 | 8 | 21
Tokens		212,678	103 |209| 363
Concept Mentions	(100%)	55,179	22 | 54| 102
Single Token	(~66%)	36,501	13 | 30| 65
Multi Token	(~33%)	18,823	6 | 21| 39

Table 1 – Summary statistics of the `kddcma2` corpus, including the minimum, median, and maximum per abstract.

3. The `gmrkb2` knowledge base

The second component of the offered resource is the `gmrkb2` lightweight lexicalized ontology that is focused on the domain of data-driven decision making ("data science"). The ontology also includes many concepts from related fields such as machine learning, optimization, Bayesian inference, hypothesis testing, and natural language processing. The ontology is derived from a snapshot of the gmrkb.com semantic wiki that is designed to be both human-readable (a wiki) and semi-structured enough to be convertible into machine processable resource. The semantic wiki was largely created by the author with assistance from formatting editors. The ontology significantly extends the quantity (and quality) `kddo1` ontology of (Melli, 2010) from 5,067 to 13,342 concepts.

When compared to other lexically-rich domain-specific semantic resources such as the Gene Ontology (geneontology.org) and the MeSH controlled vocabulary (nlm.nih.gov/mesh/), the concepts and relationships in `gmrkb2` tend to involve more mid-level concepts such as *textual data*[2] and *minimal biclique set cover problem*[3], though it also includes many leaf-level "named" entities such as: *EM*, *C4.5*, *CRF* and *word2vec* and high-level concepts such as: *algorithm* and *abstract entity*.

The semantic wiki uses a controlled English and structure proposed in (Melli & McQuinn, 2008) where each concept record contains: 1) A unique preferred name; 2) A definitional sentence of the form of "`X` *is a type of* `Y` *that* …"; 3) Words that are commonly synonymous with the concept; 4) A context section with statements relating the concept to other concepts in the ontology; 5) Examples and counter-examples of other related concepts; 6) A set of related concepts whose relationship has not been formally defined; and 7) Where possible, some helpful quoted text from external resources such as published research papers, Wikipedia and other web-accessible resources such as the concepts in The Encyclopedia of Machine Learning (Sammut & Webb, 2011). Table 2 summarizes key statistics of the ontology.

Concepts in ontology			13,342
Internal links			79,483
	Min	Median	Max
Links into a concept	0	3	364
Links out of a concept	3	5	535
Synonyms per concept	0	1	8

Table 2 – Summary statistics of the `gmrkb2` ontology

4. Interlinked Information

Finally, the annotated mentions within the corpus were interlinked with concepts within the ontology. We accomplished this using the popular wikification[4] format extensively used in wikis such as Wikipedia. If the raw mention is already naturally linked to the correct concept then no further annotation was performed. When the mention's text was an ambiguous referencer however, or when the phrasing was very idiosyncratic, then we used the vertical bar character (|) linking method[5]. As seen in Table 3, many of the mentions were unlinkable to an ontology record. For example, the ontology did not yet have a record for the concept of a *cold start user*.

Documents	2,163	PER DOC. min|med.|max
Linked mentions	51.7% 61,152	9 | 25 | 71
Unlinkable mentions	48.3% 57,096	3 | 24 | 52
Distinct concepts linked to by corpus	3,231	9 | 19 | 50
Concepts uniquely linked to by a single document		0 | 3 | 22

Table 3 – Summary statistics of the links from the `kddcma2` corpus to the `gmrkb2` ontology, including the minimum, median, and maximum per abstract.

The linking step also occasionally involved the splitting of mentions with very long compositional structure. For example "… *[[multi-label text classifier]]s*." was split into "… [[`Multi-Label Classifier`|*multi-label*]] [[*text classifier*]]s" in order to more realistically connect the document to a human-scale ontology. Overall, linking task required approximately three minutes per abstract. Figure 3 presents more examples of the linking annotation.

[2] gmrkb.com/textual_data
[3] gmrkb.com/minimal_biclique_set_cover_problem
[4] wiktionary.org/wiki/wikify#Verb
[5] wikipedia.org/wiki/Help:Wiki_markup#Free_links

Link Annotation: [[Collaborative Filtering Algorithm|*Collaborative filtering*]] *is the most popular* [[Item Recommendation Algorithm|*approach*]] *to build* [[*recommender system*]]*s and has been successfully* [[Computing System Application Act|*employed*]] *in many* [[Recommender-based Application|*application*]]*s*. </s> *However, it cannot make* [[Item Recommendation|*recommendations*]] *for so-called* [[*cold start user*]]*s that have* [[Item Rating Act|*rated*]] *only a very small* [[Quantity|*number*]] *of* [[Item Record|*item*]]s. </s>

Figure 3 - Annotation sample of interlinking two sentences

5. Conclusion and Future Work

In this work we described a new publicly available linguistic resource composed of a corpus and ontology from the field of data science: `kddcma2` and `gmrkb2`. The dataset represents one of the largest attempts to semantically interlink computing science literature at the level of concept mentions.

In the near future we plan to report on a temporal topic modeling analysis of concept cluster trends in the field. We further plan to release an updated version of the SDOI pre-annotation algorithm that is retrained using this new data and to release and RDF version of the resulting relations. Looking ahead, we would like to more formally align the ontology to existing resources such as the `SUMO`[6] top-level ontology and the OntoDM domain-specific ontology (Panov et al., 2014). In the longer term we aim to support digital libraries to help researchers navigate scientific literature at a semantic level as proposed in (Renear & Palmer, 2009) and (Clark et al., 2014).

6. Acknowledgements

We thank the authors of KDD papers who participated in the review of the annotated versions of their paper abstracts.

7. References

J. Baumeister, J. Reutelshoefer, and F. Puppe. (2011). "KnowWE: A Semantic Wiki for Knowledge Engineering." In: Applied Intelligence Journal, 35(3).

D. Carmel, M-W. Chang, E. Gabrilovich, et al. (2014). "ERD'14: Entity Recognition and Disambiguation Challenge." In: SIGIR Forum Journal, 48(2).

T. Clark, P. Ciccarese, and C. Goble. (2014). "Micropublications: A Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications." In: J. of Biomedical Semantics, 5(1).

A. Csomai, and R. Mihalcea. (2008). "Linking Documents to Encyclopedic Knowledge." In: IEEE Intelligent Systems 23(5)

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. (2005) "Overview of BioCreAtIvE: critical assessment of information extraction for biology." In: BMC Bioinformatics 2005, 6(Suppl 1):S1

E. Morin, B. Daille, K. Takeuchi, and K. Kageura. (2007). "Bilingual Terminology Mining-using Brain, Not Brawn Comparable Corpora." In: Proc. of ACL 2007.

J-D Kim, T. Ohta, Y. Teteisi, and J. Tsujii. (2003). "GENIA Corpus - a semantically annotated corpus for bio-textmining." In: Bioinformatics. 19(suppl. 1).

G. Melli. (2010). "Concept Mentions within KDD-2009 Abstracts (kdd09cma1) Linked to a KDD Ontology (kddo1)." In: Proceedings of LREC 2010.

G. Melli, and M. Ester. (2010). "Supervised Identification of Concept Mentions and their Linking to an Ontology" In: LREC-2010

G. Melli, and J. McQuinn. (2008). "Requirements Specifica. Using Fact-Oriented Modeling: A Case Study and Generalization." In: Object-Role Modeling 2008

G. Melli. (2012). "Identifying Untyped Relation Mentions in a Corpus Given An Ontology." In: TextGraphs-7.

P. Panov, L. Soldatova, and S. Džeroski. (2014). "Ontology of Core Data Mining Entities." In: Journal of Data Mining and Knowledge Discovery, July 2014.

A.H. Renear, and C.L. Palmer. (2009). "Strategic Reading, Ontologies, and the Future of Scientific Publishing." In: Science, 325(5942).

C. Sammut, and G.I. Webb. (2011). "Encyclopedia of Machine Learning." Springer.

S. Schaffert. (2006). "IkeWiki: A Semantic Wiki for Collaborative Knowledge Management." In: Procs. of STICA -2006.

[6] http://www.ontologyportal.org/

Neural Network Based Approach for Relational Classification for Ontology Development in Low Resourced Indian Language

Bhaskar Sinha, Somnath Chandra

Department of Electronics and Information Technology,

New Delhi, India

bhaskar_sindel@hotmail.com, somnath.chandra@gmail.com

Abstract

Processing natural language for text based information especially for low resource languages is challenging task. Feature extraction and classification for domain specific terms using Natural Language Processing (NLP) techniques such as pre-processing task, processing of semantic analysis etc., provides substantial support for better evaluation techniques for the accuracy of natural language based electronic database. In this paper we are exploring Neural Network based machine learning approach for Indian languages relation extraction. Convolution Neural Network (CNN) learning method is applied to semantic relational information extracted from domain specific terms, matching with multilingual IndoWordNet database. Results of machine leaning based techniques outperform with significant increase over other classification methods such as SVM, Normalized Web Distance (NWD) and statistical methods of evaluation. The objective of using this technique along with semantic web technology is to initiate a proof of concept for ontology generation by extraction and classification of relational information from IndoWordNet. This paper also highlights domain specific challenges in developing ontology in Indian languages.

Keywords: Unstructured data, IndoWordNet database, Machine learning method, Semantic relations, Ontology, Metadata.

1. Introduction

Natural Language Processing (NLP) for low resource Indian regional languages is quite lengthy and difficult task. Despite the fact that India is very rich with diverse multicultural and multilingual country, very few developments have been progressed in this regard to enrich low resource language. This necessitates the need for lexical and semantic development of electronic database to integrate cultural diversity and exchange of language resources. IndoWordNet (P. Bhattacharyya, 2010) promises to support lexical and semantic multilingual database interface, but it is not sufficient enough to process NLP related functional tasks on which various application interface depends. On the other hand, insufficient and unstructured electronic resource data, limits the reusability and utility of further resource maximization on which lots of language based applications depend, such as agriculture query, weather forecasting, soil and fertilizer testing etc.

Focusing on the above issues and necessities altogether, first, we have identified one of the India's major domains such as *Agriculture*, where many low resource Indian regional languages are spoken which needs to be linked. Secondly, we applied lexical and semantic level of filtering techniques of NLP using java program on *two hundred* domain specific *terms* and/or *concepts* matched and mapped to IndoWordNet to get filtered terms. During this semantic analysis process, we captured relational features of each term/concept for further processing. Next we applied various methods of statistical and machine learning technique such as SVM, which justifies the heuristic accuracy of IndoWordNet on the basis of classification of features. Further, we compared our natural language heuristic IndoWordNet database with global web engine database by applying NWD method (Sinha, Garg, and Chandra, 2016).

Finally, we applied Neural Network (NN) technique to accquire much improved result than earlier used classification based technique of SVM and traditional methods. Also, this reduces the cost of errors for pre-processing tasks and significantly improves learning accuracy. This paper focuses on Convolutional Neural Network (CNN) technique of NN, applied on pre-processed filtered features of each term used as input vector to neural nodes with added weight. These layered neural nodes go through learning process, mimicking similar to human natural brain function of learning. Finally, it delivers a very close matching result of predicted output. This classified relational features favour in generation of relational hierarchy and hence as a resource for generation of Ontology of interested domain as shown in figure 9.

Furthermore, this semantic relational hierarchy could be processed to create structured machine readable formatted metadata of RDF/OWL (Resource Description Framework/Ontology Web Language) (BERNERS-LEE, 2009) form. This eventually meets the reusability and utilization of resource through user application interface which claims the success of semantic web technology.

This paper is organized as follows: In section 2, we cover related previous works, section 3 briefs about other techniques that we have experimented earlier; section 4 covers architecture of data extraction, section 5 explores the results and analysis. Section 6 describes challenges and problems in extraction of relational information. Section 7 concludes with discussion of future scope of work.

2. Related Work

Relational information extraction for NLP, especially using machine learning techniques and generating ontology of specific domain was first suggested in OntoWeb Consortium in 2003. Prior to this (Hsu and Lin, 2002), worked for multi class classification using SVM learning method after that it was carried out by (Buitelaar, et al, 2005) and further by (Zhou, 2007). All are based on classification of textual as well as image and later relational information. Machine learning for deep architecture (Bengio, 2009), focused on textual information retrieval and extraction using feature based machine learned

techniques that was extended and added for relational extraction from text to generate ontology (Wilson, 2012). Deep learning using NN methods (Bengio et al., 2013) a greedy layered unsupervised pre-training, which tries to learn a hierarchy of features one level at a time and this helped to develop an ontological model from featured relation. This added a huge interest for generation and development of ontology based applications using semantic web technology including NLP, which basically rely on classification of machine learning techniques and became a new emerging research area for application development research community.

In this regard, Indian language based parallel work for relational extraction, using un-supervised learning techniques (Pawar, Bhattacharyya, et al., 2013) and pattern languages based ontology (Chaturvedi, and Prabhakar, 2013) has initiated. Some other works related to speech recognition, text classification etc. is on its way while using machine learning and classifying techniques. But to the best of my knowledge, no significant work for Indian regional language based domain specific ontology has successfully been developed using NN.

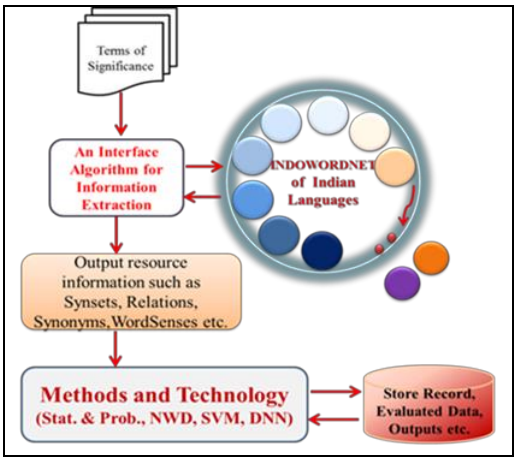

Figure 1: A generic model for our experiment. After pre-processing tasks, output is feed as input to learning module. This learning module applies various methods to get closer to the predicted avg. accuracy. (*Ref:* Sinha, Garg Chandra, et al., 2016).

3. Other Methods and SVM Learning Method

In a brief, we have experimented with other machine learning as well as statistical methods (Sinha, Garg Chandra, 2016) and compared the accuracy of IndoWordNet database based on extracted features of agricultural terms matched and mapped to IndoWordNet. We found average accuracy of IndoWordNet database is 40% with statistical and 65% using probabilistic model. Also, by applying and comparing with NWD method (Vitanyi, et al., 2008), it produces the avg. accuracy of 50.56%. Whereas applying machine learning technique such as using SVM method, the avg. accuracy is 71.87%. This added further interest to experiment with NN based approach. In our next section 4, we are focusing on discussions for experimental evaluation using Convolutional Neural Network (CNN).

3.1 Choosing Neural Network Technique

Above mentioned techniques and methods in section 3, requires lot of pre-processing tasks to handle manually with addition of extensive experimental error such as local minima or maxima, memory usage, outbound non-linearity issues etc., whereas NN based methods of machine learning techniques are easier to handle functional computations. Also, each neuron is capable of carrying out some tasks having biased weight which mimics human brain of learning approach. Each hidden input layer group divides its functional task automatically and job is done to produce as output with closer results of high accuracy.

4. Architecture and Feature Classification using Convolutional Neural Network

Convolutional neural network (CNN) is a type of deep learning classification algorithm that can learn useful features from raw data by its own. Learning is performed by tuning its weights. CNNs consist of several layers, that are usually convolution and subsampling layers interdependent on each other. Figure 2 below shows a novel model (Collobert, R., J. Weston, et al., 2011) for layered neural network architecture.

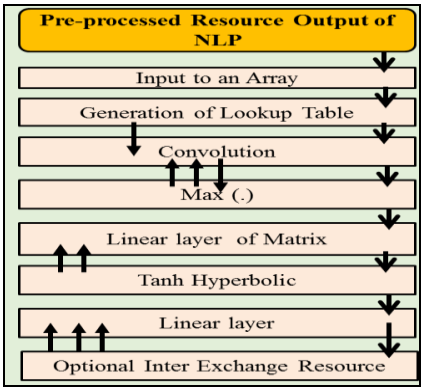

Figure 2: Layered architecture of Convolution Neural Network. (*Ref:* Collobert et al., 2011).

Convolutional layers have trainable filters that are applied across the entire input. For each filter, each neuron is connected only to a subset of the neurons in the previous layer. The weights are shared across neurons figure 4, which help filters to learn frequent patterns that happen in any part of the feature table. The feature learning process can be purely unsupervised, which can take advantage of huge unlabelled data. The feature learning neural node tries to learn a new makeover of the formerly learned features at each level, which is able to recreate the original data.

4.1 Workings with CNN

Input layer as feeder to vector array data, taken out from pre-processed extracted features of each term/concept as in figure 3. Each featured vector placed with index position $\{W^1_1 \ldots\ldots W^K_N\}$ in the matrix. Lookup table array extracts features from matrix as in figure 5 and places into a table for local and global layer use. These index vector locations take term/word embedding as input, and then apply interleaved convolution and pooling operations, followed by fully connected layers. Connected layer helps in further processing during back propagation training function of featured vector as per demand.

$$LT_W(w) = \langle W \rangle^1_w \quad (1)$$

This expression is used for each word with indices placed in the lookup table. Linear layer supports functional

transformation over input vector data as feeder for several or fixed size neural network layer.

Hindi-English Terms	Onto_ nodes	Null_P ointers	Distinct_ Relations	Hype rnym	Hypo nym	Holon yms	Grad_ State	Mero nyms	Modifie d_Noun	Attri butes	Ability_ Verb	Anto nyms	Causa bility	Entail ment	Function al_Verb
अनाज (Grain)	3	0	6	1	40	5	0	1	0	0	0	0	0	0	0
भोजन (Food)	5	1	3	4	27	0	0	0	0	0	0	0	0	0	0
फल (Fruit)	19	0	5	13	51	1	0	0	0	0	0	0	1	0	0
आम (Mango)	8	0	8	2	186	3	0	2	4	0	0	5	0	0	0
धान (Paddy)	4	1	5	1	16	1	0	2	0	0	0	0	0	0	0
गेहूँ (Wheat)	4	1	7	1	3	3	0	1	0	0	0	0	0	0	0
चना (ChickPea)	5	1	6	3	2	2	0	1	0	0	0	0	0	0	0
खाद (Fertilizer)	1	0	3	1	5	0	0	0	0	0	0	0	0	0	0
कपास (Cotton)	3	0	5	2	7	1	0	1	0	0	0	0	0	0	0
पेड़ (Tree)	1	2	5	1	288	0	0	6	0	0	0	0	0	0	0
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Figure 3: Pre-processed NLP task on Hindi *agricultural term* after matched and mapped to IndoWordNet. Also, each individual features have index positions in memory with start and end value for each synonym in a synset making a resource for vector table.

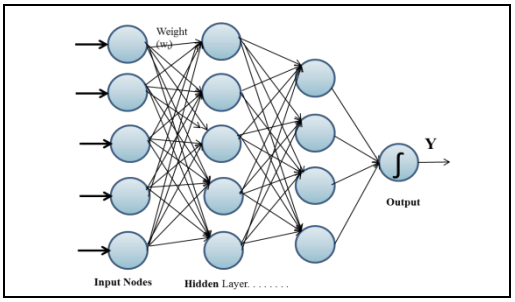

Figure 4: Featured input vector nodes undergoes through learning process including hidden layers with biased weight adjustments.

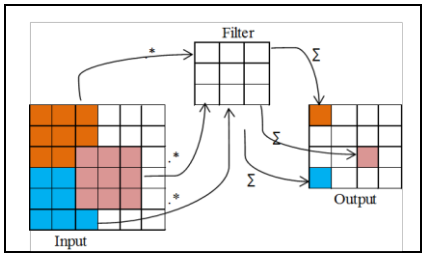

Figure 5: Training input data is looked up into lookup table to get support for an intermediate functional process of learning and classification which filters and places to output vector array during runtime state.

$$f_\theta^{\,l} = W^l f_\theta^{\,l-1} + b^l \qquad (2)$$

Here hyper parameter *l* is devoted to add number of hidden layer unit is used (figure 4). For convolutional approach, it looks for sequence of single term or window vector in a batch combining features to global vector. In that case we can use the same expression (2) for convolutional neural network repeatedly. Third layer uses Tanh hyperbolic tangent function to avoid network linearity so as to reduce the cost of computability.

$$[f_\theta^{\,l}]_i = \mathrm{Tanh}([f_\theta^{\,l-1}]_i) \qquad (3)$$

Training of neural network is done by maximizing the likelihood over training data using gradient ascent. It can be expressed as :

$$\theta \mapsto \sum \log p(y \mid x, \theta) \qquad (4)$$

where x represents the training word or a sequence of words associated features, and y corresponding tag. Learning process consists of forward and backward passes that repeat for all objects in a training set. On the forward pass each layer transforms the output from the previous layer according to its function. The output of the last layer is compared with the label values and the total error is computed. On the backward pass the corresponding change happens with the derivatives of error with respect to outputs and weights of this layer. Probability is calculated using gradient ascent of regression program module.

5. Results and Analysis

We fed two hundred agricultural domain specific Hindi terms as input data to our java program based on (Collobert et. al., 2011) neural network. It goes through the process of learning each intermediate featured parameters and classifying uniquely producing an output with significantly improved results of accuracy 77.3% for our IndoWordNet. Definitely more training dataset with added features in each synset for each term improves learning probability, hence results a better classified output as compared to other machine learning methods. The result is shown in figure 6.

Number of Iterations: 52
Learning rate: 1.0E+00
Training time: 2.76s
Training set accuracy: 70.9%
Cross Validation set accuracy: 77.3%

Figure 6: Results showing classified accuracy of features for IndoWordNet database while applying CNN technique.

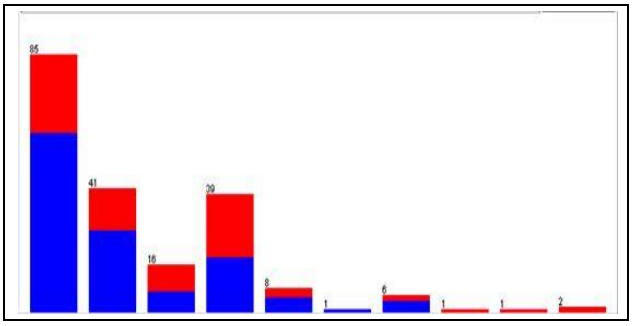

Figure 7: Stacked bar chart with label count such as 85, 43, 16, 39, 8, 1, 6, 1, 1, 2 respectively shows classified negative (blue) and positive (red) color for discretized filter used in Weka environment.

A summary of quantitative and qualitative measures of accuracy is also shown below which has been executed in Weka environment to compare with above Java based results. A parameterized function multilayer perceptron with 5-fold scaling factor has been set for execution to obtain classified featured results. Figure 7 shows stacked bar chart for classification. Figure 8 shows classified instance of relational features from training datasets.

```
=== Stratified cross-validation ===
      Total Number of Instances      :  200
=== Detailed Accuracy By Class ===
TP     FP    Precision  Recall   F-Measure  ROC   Class
0.942  0.43     0.77     0.942     0.848     0.793    tested-ve
0.57   0.058    0.865    0.57      0.687     0.793    tested+ve
Weighted Avg.:
0.795  0.283    0.808    0.795     0.774     0.793
```

Out of the classified features, we extracted semantic relations for the generation of *Ontology* of agricultural terms shown in figure 9, which are semantically related nodes of its synsets. This has further developed in Protégé environment and hence metadata is created which helps in

development of many applications. This machine readable metadata can be helpful in *Query processing, Tourism and Other knowledge extraction* based applications.

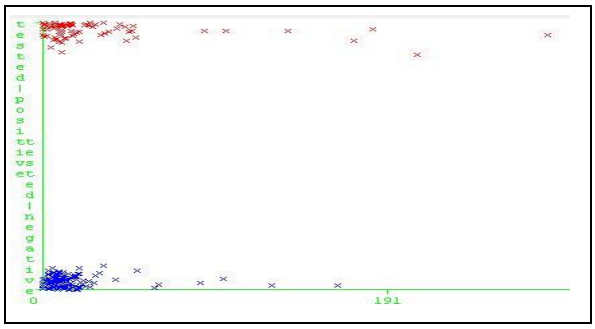

Figure 8: One of the classified instances of relational feature with positive in red and negative in blue from *Agricultural* terms.

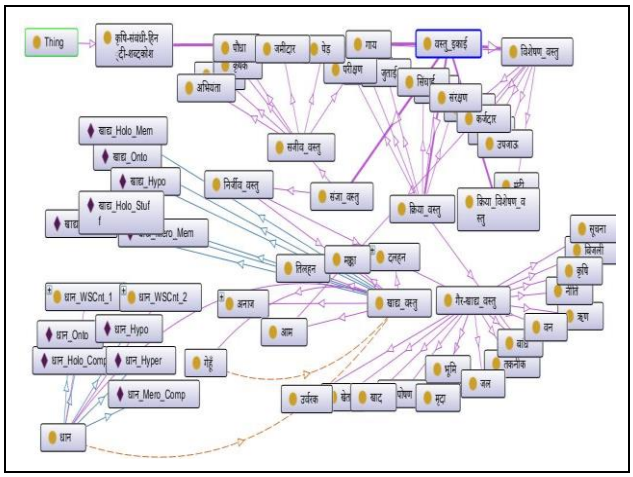

Figure 9: An *Ontology of Hindi Agricultural* term of semantically related features mapped to IndoWordNet and extracted from it.

6. Challenges and Issues

Multilingual IndoWordNet database is populated semantically and lexically sparse and it is semantically difficult to differentiate the meaning among other closer regional languages, leading to meronymy ambiguity.
Issues and restrictions which we have come across during our experiments with current available electronic source:

1. Unstructured and unformatted electronic database of existing specific domain (generally in flat file format).
2. Insufficiently populated existing electronic language resource database (*IndoWordNet* is sparse and smaller database).
3. Lack of support for pre-processing local language resource tools (limitations with language parsers etc.).
4. Synchronization of methodology and technology is not established properly(missing of W3C standard).

7. Conclusion and Future Work

In this paper we have presented a Neural Network based machine learning technique that has been applied to train and learn featured relations from IndoWordNet of matching terms of *agriculture domain* in Hindi language to get predicted result of interest showing the accuracy of IndoWordNet database. It could be any Indian regional language and/or domain of interest. Although, IndoWordNet is a small heuristic database in comparison to other web engine database, eventually this limits us to utilize the full potential of resource benefit of natural language database. In our future work, we would focus on our interest for automatization of ontology generation of W3C standard for linking with global source of FAO using Linked Open Data (LOD) of semantic web technology.

8. Acknowledgements

We are thankful to TDIL, DEITY, Govt. Of India for their extensive logistic as well as providing data and related information resource support that has helped us to experiment and writing this paper successfully.

Bibliographical References

BERNERS-LEE, T. (2009). *Putting Govt. Data online,* 2009 http://www.w3.org/DesignIssues/GovData.htm.

Bhattacharyya P. (2010). *IndoWordNe*t, Lexical Resources Engg. Conference, (LREC 2010), Malta.

Bengio, Y. (2009). *Learning deep architectures for AI*. Foundations & Trends in Mach. Learn. 2(1):1–127.

Bengio, Y., A. Courville, et al. (2013). *Representation learning: A review and new perspectives*.

Buitelaar, P., et al. (2005). *Ontology learning from text: An overview*. In Ontology Learning from Text: Methods, Evaluation and Applications", Eds. IOS Press, Amsterdam.

Chaturvedi, A., Prabhakar, T. V. (2013). *Ontological Visitor Pattern,* Indian Conference on Patterns and Pattern Languages of Programs, (GuruPLoP'13).

Collobert., R., J. Weston, et al. (2011). *Natural language processing (almost) from scratch*. The Journal of Machine Learning Research 12: 2493-2537.

Sinha, B., Garg, M., Chandra, S. (2016). *Identification and Classification of Relations for Indian Languages using Machine Learning Approaches for Developing a Domain Specific Ontology,* In Proc. of IEEE Intl. Conf. on Computational Tech. in Info. and Comm. Tech. (ICCTICT'March'16), New Delhi, pp. 508-513.

Hsu, C. W. and Lin, C. J., A. (2002). *Comparison of methods for multi-class support vector machines*, IEEE Transactions on Neural Networks, 13, pp. 415–425.

Pawar. S., Bhattacharyya P., et al. (2013). *Semi-supervised Relation Extraction using EM Algorithm*, 2013, International Conference on NLP(ICON'2013),India.

Vitanyi P. M. B., et al. (2008). *Normalized Information Distance*, In Information theory and statistical learning, eds. F. Emmert-Streib and M. Dehmer, pp. 39-71.

Wilson, W., et al. (2012). *Ontology learning from text: A look back and into the future*, ACM Comput. Surv. 44, 4, Article 20, 36 pages.

Zhou, L. (2007). *Ontology learning: State of the art and open issues*, Info. Technol. Manage. *8,* 3, pp. 241–252.

Language Resource References

Glossary of Agriculture (1996). Computerised Database. Commision for Scientific & Technical Terminology, Ministry of Human Resources Development (Dept. of Education), Govt of India.

Gentle with the Gentilics

Livy Real, Valeria de Paiva, Fabricio Chalub, Alexandre Rademaker
IBM Research Brazil, Nuance Communications USA, IBM Research Brazil, IBM Research and FGV/EMAp, Brazil
livym@br.ibm.com, valeria.depaiva@gmail.com, fchalub@br.ibm.com, alexrad@br.ibm.com

Abstract

To get from 'Brasília is the Brazilian capital' to 'Brasília is the capital of Brazil' is obvious for a human, but it requires effort from a Natural Language Processing system, which should be helped by a lexical resource to retrieve this information. Here, we investigate how to deal with this kind of lexical information related to location entities, focusing in how to encode data about demonyms and gentilics in the Portuguese OpenWordnet-PT.

Keywords: lexical resources, ontology, gentilics, Wordnet

1. Introduction

Inferring from 'Brasília is the Brazilian capital' that 'Brasília is the capital of Brazil' is an obvious task for a human, but doing it automatically in an Natural Language Processing (NLP) system requires some effort. Having this kind of information encoded in a lexical resource can help in several tasks, such as information retrieval, lexical disambiguation and textual entailment. However, deciding which kind of ontological information should be present in lexical resources, be they wordnet-like, or specific knowledge bases, such as DBpedia[1], Wikidata[2], OpenStreetMap[3], or Geonames[4] is a complex decision. We deal in this paper mostly with *gentilics*, a class of pertainym adjectives that sits in between lexical and ontological knowledge and whose proper linguistic treatment requires access to ontological resources such as linked geo-spatial data and formal ontologies. Thus we investigate how to deal with lexical information closely related to location entities, mainly focusing in how to encode these data in the Portuguese OpenWordnet-PT (de Paiva et al., 2012).

OpenWordNet-PT (OWN-PT) is a freely available wordnet for Portuguese.[5] OWN-PT was originally developed as a syntactic projection of the Universal WordNet (UNW) (de Melo and Weikum, 2009). Just like EuroWordNet (Vossen, 1998), OWN-PT was as much as possible built merging 'existing resources and databases with semantic information developed in various projects'. One reason to pay special attention to this wordnet for Portuguese is its connection to several other lexical resources based on Princeton WordNet (PWN) and on Linked Open Data (LOD) principles (Chiarcos, 2012). OWN-PT is available as an RDF/OWL download, following and expanding, when necessary, the original mappings proposed by (van Assem and Schreiber., 2006). Also OWN-PT is linked to the Suggested Upper Merged Ontology (SUMO)[6] (Niles and Pease, 2001) and to the Open Multilingual WordNet (OMW) project[7] (Bond and Foster, 2013). Since OMW merges dozens of wordnets, ways of improving each one of these wordnets might percolate to the other ones. Moreover, the issues we discuss in this work affect all of these other (merged or not) lexical resources, as we hope it is made clear in the sequel.

To start thinking about this kind of lexical geo-related information, we decided to investigate relational adjectives. Since these are traditional adjectives, they should appear in a lexical resource, but they are closely related to what we understand as ontological knowledge. Information about this kind of adjectives comes in PWN in a separated lexicographer file called ADJ.PERT (pertainym adjectives). Pertainyms are a class of adjectives that are associated with a base noun by the relation ADJECTIVEPERTAINSTO, such as the pairs *Brazilian/Brazil* and *fictional/fiction*. Thus a pertainym is an adjective, which can be defined as 'of pertaining to' another word.

The PWN lexicon has 3661 adjective pertainyms, of which 2617 had no translation to Portuguese in our OpenWordNet-PT lexical database in May 2015. We started working on pertainyms, but discovered that *gentilics*, a subclass containing adjectives pertaining only to *locational nouns*, offered enough challenges. Thus this note describes the work we did to produce the necessary translations to complete the Portuguese OpenWordNet-PT, as well as the work on improving the theoretical understanding of pertainyms and gentilics in this resource.

2. Pertainyms, Demonyms and Gentilics

Wikipedia tells us that 'demonym' is a word created to identify residents or natives of a particular place. A 'demonym' is also usually derived from the name of that particular place. Examples of demonyms include *Chinese* for the natives of China, *Swahili* for the natives of the Swahili coast, and *American* for the natives of the United States of America, or sometimes for the natives of the Americas as a whole. Just as *Americans* may refer to two different groups of natives, some particular groups of people may be referred to by multiple demonyms. For example, the natives of the United Kingdom are the *British* or the *Britons*.

The word gentilic comes from the Latin *gentilis* ('of a clan') and the English suffix *-ic*. The word demonym was derived from the Greek word meaning populace (*demos*) with the suffix for name (*-onym*). For English and Portuguese there is a generalized, but principled ambiguity: when we say *Brazilian/brasileiro*, without any context, we mean either

[1] http://wiki.dbpedia.org/
[2] https://www.wikidata.org/
[3] https://www.openstreetmap.org
[4] http://www.geonames.org/
[5] http://wnpt.brlcloud.com/wn/
[6] http://www.ontologyportal.org
[7] http://compling.hss.ntu.edu.sg/omw/

the noun or the adjective: {09694894-N BRAZILIAN — BRASILEIRO — A NATIVE OR INHABITANT OF BRAZIL} or {02966829-A BRAZILIAN — BRASILEIRO — OF OR RELATING TO OR CHARACTERISTIC OF BRAZIL OR THE PEOPLE OF BRAZIL}. To clearly distinguish pertainyms, which are adjectives, from the nouns (associated with a location), here we call adjectives *gentilics* and the associated location specific relational nouns *demonyms*. We are interested in discussing the adjectives, more than the nouns, but both bring to the fore one of the important issues that we grapple with: what is linguistic knowledge vs. world knowledge? How much of world knowledge needs to be present in a lexical-ontological resource such as a wordnet? GeoWordNet (Giunchiglia et al., 2010) is a resource that fully merges the GeoNames database, Princeton WordNet 1.6 and the Italian portion of MultiWordnet (Pianta et al., 2002), but perhaps a wordnet does not need to have much geographical information. Since there are many geographic databases, they could be used instead of growing the number of synsets referring to locations within the lexicon itself. This is the approach taken for instance in (Frontini et al., 2013), which transforms the GeoNames ontology into GeoDomainWN, a linguistic linked open data resource, linking both PWN and the Italian WordNet to GeoNames.

Language is tied up to culture and clearly when discussing the meanings of words in Portuguese we need to deal with meanings that do not exist in English (and, in general, are not present in general knowledge bases). The most obvious of these meanings are related to pertainyms, mostly to places (gentilics) but also to religions, styles of philosophy, music, etc. Examples in English include *Buddhist, Socratic, Wagnerian, Darwinian*, etc. Examples in Portuguese include *macumbeiro* (someone who practices *macumba*, a Brazilian religious practice, a mixture of African religions and Catholicism); *machadiano* (from Machado de Assis, one of the greatest Brazilian novelists); *tropicalista* (from *Tropicália*, a musical movement).

A few of the essentially Brazilian words have made their way into English. The word *samba* for instance appears in PWN within three synsets: {01896881-V SAMBAR — DANCE THE SAMBA}, {00537192-N SAMBA — A LIVELY BALLROOM DANCE FROM BRAZIL} and {07056895-N SAMBA — MUSIC COMPOSED FOR DANCING THE SAMBA}. Most Brazilians would agree that these three kinds of senses (the kind of music, the kind of dance, and the action of dancing) exist in Portuguese, however some might disagree with the glosses: samba is not necessarily for dancing. Also we need derived words like *sambista* (someone who dances or composes samba), and compounds like *samba-choro, samba canção, escola de samba*, etc.[8]

The issue of making the Portuguese wordnet culturally relevant to Brazilians is of paramount importance to us. Given that the development of OpenWordnet-PT was motivated by its use in information extraction from the Brazilian Dictionary of Historical Biographies (DHBB) (de Paiva et al., 2014), it needs several gentilics that are not present in PWN. For example, to process a very typical sentence from DHBB, as "[...] o deputado federal **pernambucano** Fernando Lira [...] votou a favor da emenda da reeleição [...]" (*The congressman **from Pernambuco** Fernando Lira voted in favor of the reelection amendment.*", gentilics information is required. For this kind of corpus the work of adding Portuguese gentilics seems a manageable task and an easy introduction to creating our own essentially Portuguese synsets, that we knew from the beginning we would need in the fullness of time.

3. Completing OWN-PT

Before starting creating new synsets for the gentilics of the states in Brazil (e.g. *paulistano, amazonense*) we needed to complete the gentilics present in PWN synsets, but with no Portuguese words in the corresponding OWN-PT synset.

Given our choice of encoding OpenWordnet-PT in RDF (Rademaker et al., 2014), SPARQL (Prud'hommeaux and Seaborne, 2008) queries can be created to find the pertainym synsets with no Portuguese words and relate them to gentilics and demonyms. That is, we can formulate a query that retrieves all pairs of synsets (s_1, s_2) that have senses related by the relation ADJECTIVEPERTAINSTO, where the first synset s_1 corresponds to the gentilic and the second synset s_2 is the place it is associated with (originally defined in the PWN lexicographer file NOUN.LOCATION).

Searching for Portuguese empty gentilic synsets and completing them was the first step of our methodology. Adding the missing Portuguese words to the OWN-PT synsets equivalent to the PWN synsets though is a manual labor. Some 400 gentilics had to be added, as the semi-automatic construction process had not found them. There is no general affix rule that captures in Portuguese all possible (and the right ones) gentilics, since this morphological process can occur via, at least, six main different suffixes — namely *-ês, português, 'Portuguese'*, *-ano, haitiano, 'Haitian'*, *-ino, argentino, 'Argentinian'*, *-eiro, brasileiro, 'Brazilian'*, *-ão, afegão, 'Afghan'* and *-ense, angolense, 'Angolan'* — with no standard syntactic-semantic pattern to be followed. There are also suffixes that can produce gentilics in a non regular way, e.g. *-ista, sul-africanista, 'South-African'* and *-enho, caribenho, 'Caribean'*. Moreover, in Portuguese, the zero-suffix (also called regressive morphological process) is highly productive and gives us gentilics, such as *bósnio, 'Bosnian'* and *búlgaro, 'Bulgarian'*. There are still some lexicalized forms, which are not morphologically related to the location nouns that they refer to, such as *barriga-verdes* ('green-bellies'), for the natives of the state of 'Santa Catarina' and *capixabas*, for the natives of the state of 'Espírito Santo'. All these issues turn the automatic processing of detecting or creating gentilics a challenge. The work here takes the alternative route of checking and completing the required synsets with the right gentilics, as suggested by PWN and Wikipedia. A preliminary list of verified entries was obtained from Portuguese DBpedia Sparql Endpoint [9].

[8] Both *samba-choro* and *samba canção* are not for dancing, mostly. *Escola de samba*, 'Samba School', is a group of people that practices samba and performs it once a year in *sambódromos*, huge spaces prepared to receive samba schools during Carnival.

[9] `http://pt.dbpedia.org/sparql`

4. New synsets

As expected PWN does not have most of the gentilics related to Brazilian culture and language. Actually PWN does have only one gentilic specific to Brazil, the word *carioca*, which is in the appropriate demonym synset {09695019-N CARIOCA — CARIOCA — A NATIVE OR INHABITANT OF RIO DE JANEIRO} but it does not have all the other 26 demonyms for the other Brazilian states, for example. The English PWN does not list Brazilian gentilics, since they are not part of the English language, but clearly they ought to be in the OWN-PT, a Portuguese wordnet, as they are an important part of our lexicon.

Despite a long list of gentilics to be found in the "Dicionário de Gentílicos e Topónimos" ('Dictionary of Gentilics and Toponyms'), kindly provided by the "Portal da Língua Portuguesa" ('Portal of the Portuguese Language'),[10] we do not want to have all of these gentilics in our knowledge base, as mostly, they are regular and would not be very useful for our main task of reasoning with language. We needed to come up with useful criteria to decide on the 'notoriety' of words that justify creating a synset for them, to borrow a concept from Wikipedia.

So we started investigating the Wikipedia list of gentilics for nations and the list of Brazilian gentilics which includes all adjectives related to states, capitals and most important cities in Brazil. Wikipedia actually offers a reasonable amount of Brazilian relevant terms that could be linked to OWN-PT. Table 1 presents some numbers.

Number of Gentilics	Locations
27	States of Brazil
455	World countries
532	Brazilian cities
288	cities in the state of Minas Gerais
93	cities in the state of Rio de Janeiro
274	cities in the state of São Paulo

Table 1: Table of Gentilics extracted from Wikipedia/DBpedia

So far our work in OpenWordnet-PT has been focusing on adding Portuguese words to OpenWordnet-PT synsets related to PWN synsets, postponing the creation of new synsets. Adding Brazilian gentilics to OpenWordnet-PT seems a good way to start adding synsets for Portuguese specific concepts since they have established and regular relations to their related nouns and are easily inserted in PWN's hierarchy. This information (lexical entries of gentilics, and also, of demonyms) is easily retrievable from DBpedia, as it links location articles, as for example *Brazil*, to its demonym (*Brazilian*), via a OWL:DEMONYM relation. DBpedia-PT offers all gentilics we think we need at the moment, thus we are now investigating how to link DBpedia-EN, PWN, DBpedia-PT and OWN-PT. We believe it is better to link all those resources than try to merge and disambiguate their actual state within OWN-PT. In the one hand, DBpedia has most of the Wikipedia content we are interested in and is often updated. On another hand, Wikipedia infoboxes still lack an uniform treatment for gentilics and demonyms — some of them actually bring plurals, *Brasileiros*, and feminine and masculine forms in different patterns, as *Australiano, Australiana vs Espanhol(a)* — which we do have in OWN-PT. We expect to improve the present state of both resources by linking them. A preliminary proposal of how to link those resources is found in Figure 1.

[10] http://www.portaldalinguaportuguesa.org/.

5. SUMO and World Knowledge

Most of our work in lexical resources is directed towards using these resources in Knowledge Representation. A question then poses itself, how many locations, countries, cities, etc should we have in the lexicon; how many of these should be in other ontologies or gazetteers or taxonomies? The promise of the Open Linked Data project is that we can outsource some of the work related to the ontologizing of these locations to others, for example GeoNames or DBpedia. But demonyms and gentilics still have to be in the lexicon. They can not always be derived from their related nouns and they are not named entities that one could keep track of in other instance-based ontological resources. At least the Academia Brasileira de Letras (Brazilian Academy of Letters), the official keeper of the Brazilian Portuguese language, lists gentilics and demonyms, but does not list names of places in its official vocabulary list.

Given our use of linked data and given the easy access to the mappings of PWN into SUMO (Niles and Pease, 2003), we have decided to investigate how the mapping of new possible synsets to SUMO would proceed. While it is desirable to link all languages via OMW, there some difficulties, when synsets exist in one language but not in another. One possible approach is to create new synsets in PWN, but creating synsets that are not used in the language is problematic, since a wordnet is supposed to be a representation of language as actually used. A more principled approach might be to create an Interlingua index (Pease and Fellbaum, 2010; Bond and Fellbaum, 2016) that can be the union of all the concepts that are lexicalized in different languages. While demonym noun synsets in PWN are mostly mapped to SUMO as an instance of the `EthnicGroup` concept, gentilic adjectives are not consistently mapped. Table 2 shows some numbers of the mappings from PWN of gentilics and associated noun synsets into SUMO concepts.

SUMO Concept	PWN Gentilic	PWN noun.location
Nation	172	20
'Specific Places'	7	199
GeographicArea	21	35
LandArea	27	64
GeopoliticalArea	33	10
City	30	37
Island	14	45
EthnicGroup+Human	13	0
Others	92	0

Table 2: Mappings from PWN synsets to SUMO concepts

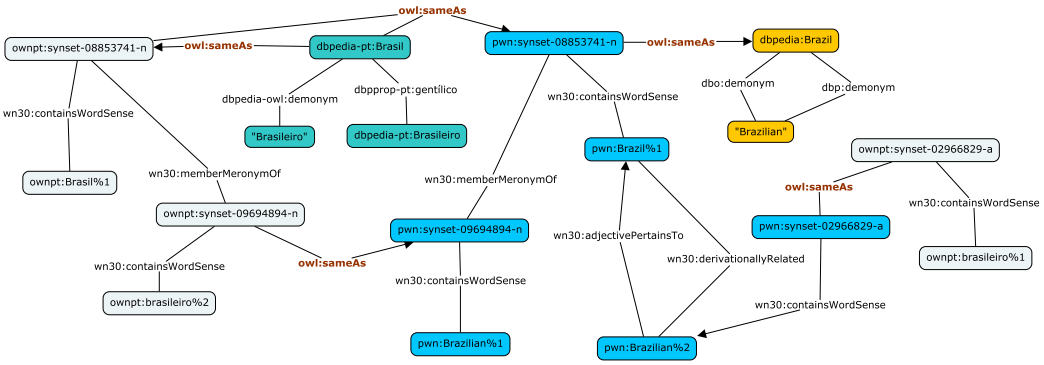

Figure 1: Connecting DBpedia with PWN and OWN-PT

In the column 'SUMO Concept', the label 'Specific Places' stands for specific places that are also specific concepts, such as `Paris`, `Brazil` and `SouthAfrica`. We can see that almost half of the nouns that we deal with are mapped, as expected, to their specific place concept: the synset {08853741-N BRAZIL — BRASIL — THE LARGEST LATIN AMERICAN COUNTRY AND THE LARGEST PORTUGUESE SPEAKING COUNTRY IN THE WORLD} is correctly mapped to concept of `Brazil`. However, while the synset for *Paris* is mapped to the concept `ParisFrance`, the synset for *Venice* is mapped into `PortCity`, a city which has a port. The PWN to SUMO mappings (as well as SUMO itself) have been constructed over a period of 15 years. Even when a precise SUMO concept is available, its corresponding WordNet mapping may not have been updated. Although SUMO has a proper treatment of many concepts, many are also missing and some are mapped to an overly general definition. Almost half of the mappings of the gentilics go to an instance of the concept `Nation`, as they are related to nouns that are instances of nations. One might expect that gentilic adjectives (e.g. 'Brazilian' in *Brazilian cuisine*) would be mapped to a relation, relating the type of the object it applies to (`Cuisine` is a class in SUMO) to the generic property of being associated with that place, in this case, Brazil. Instead, the gentilic adjectives are mapped at the moment to the geographical and abstract concepts they are associated with, such as `Nation`, `Island` and `LandArea`. These mappings are somewhat inconsistently done as well. Were they to be more consistent, one could perhaps argue that the ontology itself did not need to have relational concepts, that the location is meaningful enough. However the consistency of the mappings itself is complicated, for example, gentilics related to island places are not necessarily mapped into `Island`: the adjective *Seychellois* is mapped into `LandArea` (as the Seychelles are an archipelago), while *Tobagonian* is mapped into `Island` but *Mauritanian* into `Nation`, even if these three places are island-like.

The actual mapping implicitly tells us that gentilic is a relation between an entity and a location. While this seems generally correct, there are many cases where this seems wrong. Examples include nomadic people like 'gypsies' or 'Bedouins', not to mention all the Brazilian native tribes. We would prefer not to be too specific, as demonyms and gentilics do not carry only the meaning of the place where someone lives or was born, as a preliminary view suggests.

6. Conclusions

Gentilics are an interesting and useful phenomenon to investigate, when considering the frontiers of lexical resources and world ontologies. First they are clearly lexical, but related to locations, which are named entities and hence more akin to world knowledge than lexical knowledge. Then they are somewhat easier adjectives to deal with, as one does not have to worry too much about scales of being *paulista* 'of São Paulo', for example. Then they are slightly more amenable to Knowledge Representation methods and tools, as one can, as in the SUMO mapping available, use the location itself as a proxy for the adjective, relaying in some other language processing.

For our own driving application to the corpus of biographies in (de Paiva et al., 2014) they seem very useful, as historical data needs to be geographically located. Finally, as a way of starting creating new synsets, they seem a safe bet, as they are sandboxed, as they ought to be all in the class of pertainyms and all related to locational nouns.

We leave as future work the task of adding the most relevant Portuguese gentilics for other lusophone cultures different from the Brazilian one, that is the gentilics most relevant for places in Portugal or Mozambique, say. We would also like to discuss with the SUMO team the best way of improving the mapping of gentilics to SUMO. This includes fixing bugs in the SUMO-WN mappings but more importantly, adding definitions to SUMO itself. While we will save a detailed treatment for a latter paper, this might require using the full expressivity of higher order logic to use modal and temporally qualified expressions. Functions are also heavily employed so we would like to create PERSON-OF-REGION-FUNCTION with a geographical argument, without having to laboriously reify not only every country or region but also the notion of being from a region or typical of a region. As a result, these definitions may have to use SUMO's expressive logic rather than a simpler language like a description logic. Finally, we would like to evaluate how much the quality of the treatment of meanings on our historical corpus Brazilian Dictionary of Historical Biographies (DHBB) increases if we have relational information in the OWN-PT lexical base.

7. Bibliographical References

Bond, Francis, P. V. J. M. and Fellbaum, C. (2016). Cili: the collaborative interlingual index. In Christiane Fellbaum Piek Vossen Verginica Barbu Mititelu, Corina Forăscu, editor, *Proceedings of the Eighth Global WordNet Conference*, Romania.

Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, Sofia. ACL.

Chiarcos, C. (2012). *Linked Data in Linguistics*. Springer.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *Proc. of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

de Paiva, V., Oliveira, D., Higuchi, S., Rademaker, A., and de Melo, G. (2014). Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE, oct.

Frontini, F., del Gratta, R., and Monachini, M. (2013). Linking the geonames ontology to wordnet. In *6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Giunchiglia, F., Maltese, V., Farazi, F., and Dutta, B. (2010). Geowordnet: A resource for geo-spatial applications. In Lora Aroyo et all., editor, *The Semantic Web: Research and Applications*, volume 7th Extended Semantic Web Conference, ESWC 2010. Springer.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Chris Welty et al., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*. FOIS-2001.

Niles, I. and Pease, A. (2003). Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages pp 412–416.

Pease, A. and Fellbaum, C. (2010). Formal ontology as interlingua: The sumo and wordnet linking project and globalwordnet. In C. R. et al Huang, editor, *Ontologies and Lexical Resources*. Cambridge University Press, Cambridge.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: Developing and aligned multilingual database. *Proceedings of the First International Conference on Global WordNet*, pages pp. 293–302.

Prud'hommeaux, E. and Seaborne, A. (2008). Sparql query language for rdf. w3c recommendation, january 2008. Technical report, W3C.

Rademaker, A., de Paiva, V., de Melo, G., Coelho, L. M. R., and Gatti, M. (2014). OpenWordNet-PT: A project report. In Heili Orav, et al., editors, *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan.

van Assem, Mark, A. G. and Schreiber., G. (2006). Rdf/owl representation of wordnet. *W3c working draft, World Wide Web Consortium*.

P. Vossen, editor. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

The Importance of Being Referenced: Introducing Referential Semantic Spaces

Dante Degl'Innocenti, Dario De Nart and Carlo Tasso
Department of Mathematics and Computer Science
University of Udine
deglinnocenti.dante@spes.uniud.it, dario.denart@uniud.it,
carlo.tasso@uniud.it

Abstract

The Web is constantly growing and to cope with its ever-increasing expansion semantic technologies are an extremely valuable ally. The major drawback of such technologies, however, is that providing a formal model of a domain is time consuming task that requires expert knowledge and, on the other hand, extracting semantic data from text in an automatic way, although possible, is still extremely hard since it requires extensive human-annotated training corpora and non trivial document pre-processing. In this work we introduce a vector space representation of concept associations that can be built in an unsupervised way with minimal pre-processing effort and allows for associative reasoning supporting word sense disambiguation and related entity retrieval tasks.

Keywords: Semantic Networks, Vector Space, Text Processing, Theory

1. Introduction

Over the last years the interest for semantic technologies has grown at a steady pace due to their promise of effectively tackling the evergrowing amount of data of the Web. Modelling the actual semantics of documents is indeed an extremely valuable task that can provide significant benefits to information retrieval, filtering, and categorisation. Recently Semantic technologies have become increasingly popular also in commercial applications and have been adopted by several major players, like Facebook that introduced back in 2010 the *OpenGraph* format, shortly followed by Google's *Schema.org*. Popular search engines like Duckduckgo, Bing, and Google include in search results the so-called *infoboxes*, also known as zero-click searches, that leverage semantic datasets such as DBpedia and Freebase.

These technologies, however, have one major drawback: they require extensive annotation and the deeper the semantics to be expressed, the more complex is the annotation to be produced. Due to its complexity, semantic annotation requires either human intervention or complex artificial intelligence techniques that generally require extensive training data. While in the recent years it has proven to be true that *a little semantics goes a long way*[1] a Web-scale annotation is still impractical. On the other hand, it is possible to extract semantics from distributional features of text documents with no need for explicit annotation, leveraging on vector space representations of documents and terms. This approach has some downsides when compared to first class semantic annotations, in particular it does not provide explicit semantics, but rather hints of the actual semantics, moreover it requires a great amount of raw data to produce some interesting results. However, in our opinion, these are not severe limitations since (a) we assume that even shallow semantics can produce great benefits when applied to enough data, and (b) the Web produces gigabytes of raw, uncategorised and unannotated textual data on daily basis.

In this paper we introduce a novel technique to model into a vector space the semantic knowledge contained in a text corpus annotated with hyperlinks like any HTML page. Such technique leverages the idea that entities referenced by a similar set of documents may yield similar meaning. The resulting vector space bears significant resemblance with user/item spaces commonly used in recommender systems and can be exploited to implement semantics driven NLP tasks such as word sense disambiguation, evaluation of semantic distance between concepts, and related concepts retrieval.

[1] Also known as the Hendler hypothesis (Hendler, 2007)

2. Related Work

Several works in the literature address the problem of representing the semantic knowledge that can be found in text documents. Traditional approaches leverage keywords or keyphrases i.e. short phrases, typically made of one to four words which identify a concept. Over the years many automatic keyphrase extraction tools, such as Distiller (Degl'Innocenti et al., 2015; Degl'Innocenti et al., 2014) and KEA (Witten et al., 1999) were developed. Keyphrases, however are a flat representation of the text from which they are extracted, providing very little semantics, nonetheless they are still the most practical and widespread technique to represent the textual content of Web pages. On the other end of the semantic spectrum, the most formal representation of semantic knowledge is the one provided by the Semantic Web stack through technologies such as RDF and OWL that allow for formal knowledge modelling based on description logics. Manual generation of RDF triples, however, is expensive and needs expert knowledge, therefore automatic approaches are needed to cope with the current scale of the Web. The most well known RDF knowledge base is DBpedia (Bizer et al., 2009), a Wikipedia semantic mirror built automatically by exploiting information included in structured sections of Wikipedia articles. DBpedia is currently the reference knowledge base for unstructured text annotation, and over the last years several tools have been proposed to link texts with DBpedia entities, like DBpedia Spotlight (Mendes et al., 2011). Knowledge based agnostic tools for

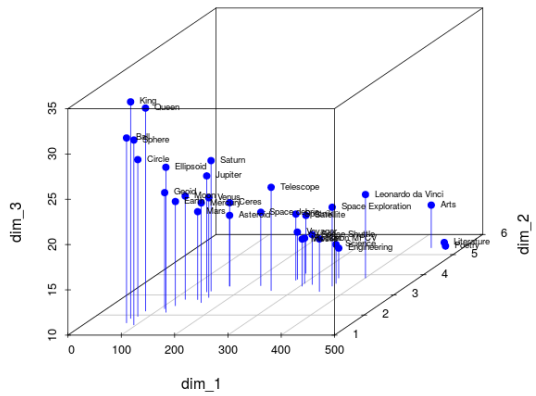

Figure 1: An instance of a 3 dimensional semantic space.

text annotation have been proposed as well, like Agdistis (Usbeck et al., 2014) a graph based method for automatic word sense disambiguation. Large knowledge bases can also be used to allow exploratory search, like in (Musetti et al., 2012) where a tool for finding DBpedia entities related to a query element is presented.

Vector Space Model (herein VSM) approaches are an alternative to explicit and formal representations such as the one provided by Semantic Web technologies, but capable of carrying more semantics than mere keyphrases. The basic idea is that entities, instead of being described by a set of predicates, are represented as a vector in a space with a finite number of dimensions. VSM are highly connected to the *distributional hypothesis* of linguistics, which claims that words that occur in similar contexts tend to have similar meanings (Harris, 1954). Some author (Novak, 2010) also defines the meaning of a concept as the set of all propositions that contain that concept. Figure 1 shows an instance of a vector semantic space with 3 dimensions, where similar concept are located near each other. The VSM was first developed for the SMART information retrieval system (Salton, 1971) and it is now widely used in many different fields. VSMs are used to support several different tasks, such as document retrieval, document clustering, document classification, word similarity, word clustering, word sense disambiguation, etc. The most notable advantage of these techniques over formal representations is that vector spaces can be built in a totally automated and unsupervised way. For a deeper and more exhaustive survey of vector spaces and their usage in state of the art systems, we address the interested reader to (Turney and Pantel, 2010), (Manning et al., 2008), and (Levy et al., 2015).

3. Enter the Reference Hypothesis

As shown in the previous section, most literature work on word spaces leverages the distributional hypothesis, that is that words occurring in similar contexts may yield similar meaning. However to overcome their limitations and to exploit the potential of hypertextual connections we introduce a new hypothesis: the *Reference Hypothesis*. We assume

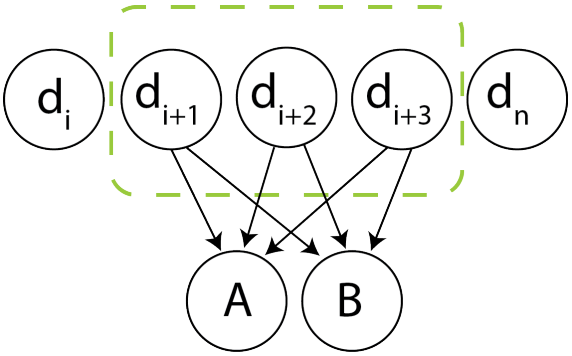

Figure 2: Two entities referenced by the same set of documents.

that entities that are referenced in a similar set of documents might yield strong semantic affinity. For instance, in Figure 2 two entities (A and B) are referenced in the same documents: this implies a semantic affinity between A and B.

This assumption is motivated by the fact that intuitively referencing something in a document implies the referenced item to be relevant in the context of the document, therefore entities that get constantly referenced together are relevant in the same contexts, hence they might be semantically related.

This hypothesis can be seen as a generalised version of the aforementioned distributional hypothesis, however we would like to stress how even though words can be seen as entities, entities can be intended as way more abstract items, for instance other documents or ontology entries. For instance, the reference hypothesis applies to the scientific literature since articles citing similar sources are very likely to deal with similar topics. Other works in literature embrace this assumption though not formalising it, such as (Huynh et al., 2012) wherein a scientific paper recommender system exploiting co-citation networks is presented.

3.1. Building a referential space

Building a vector space exploiting the Reference Hypothesis is straightforward once a large enough corpus of documents annotated with hyperlinks is provided. Within the corpus two sets must be identified: the *entity set* E and the *document set* D; the first includes all the referenced entities, while the latter the considered annotated documents. The vector space is represented with an $E \times D$ matrix that initially is a zero matrix. Iteratively, for each $d \in D$ all the references to elements in E are considered, and for each $e \in E$ referenced in d, the (e, d) cell of the matrix is set to 1. Since referencing a given entity only once in a document is a typical best practice in several domains[2] we are not considering how many times e is referenced in d. Once all documents are processed we obtain a matrix where each row represents all the references to a given entity: we call such matrix *Reference Matrix* and the vector space it gen-

[2]For instance in Wikipedia only the first time an entity is referenced it is annotated with an hyperlink, and in literature bibliographies have no duplicate entries.

erates *referential space*.

It is important to point out that, as long as the considered corpus is made of HTML pages, there is no need of annotating the texts. Hyperlinks can be conveniently parsed without performing NLP tasks such as tokenization, stemming, linguistic analysis, and so on. Furthermore, there exist on the Web corpora of cross-referenced hypertext documents, where documents form a network of connections. Such corpora are particularly interesting to analyse under the Reference Hypothesis, because $E \equiv D$ since any entity is also a document, resulting in a square, although not symmetrical, matrix. On the other hand, if the considered corpus is not annotated with hyperlinks, there exist technologies such as *TagMe* or *Babelfy* that allow automatic annotation with links to ontology entries. In this scenario, however, heavy NLP is involved and for a very large corpus this solution might be impractical.

Another relevant feature of hypertextual connections is that they are provided with a *surface* label, that is a word or a string of words to be clicked to open the linked page. The surface label represents the natural language label associated to the linked entity and typically this is not a symmetrical relationship: an entity can be referred with different surface labels as well as a surface label can link to different entities in different contexts. Entities represent the meaning of surface labels, while surface labels represent the signifier of entities. We call the multiplicity of meanings of a surface label its *ambiguity*.

3.2. Evaluating similarity

Evaluating the similarity of two entities in the vector space ultimately reduces to computing some distance between their vectors. Countless different metrics exist in the literature, which are useful to assess the similarity between two vectors, such as norms, cosine similarity that estimates the angle between the vectors, hamming distance that estimates the lexical distance between the vectors, and many others conveniently surveyed in (Wang et al., 2014). All these metrics can be used in the Reference Matrix, however we prefer the Jaccard similarity coefficient (also known as Jaccard index (Jaccard, 1902)), that is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

where A and B are sets of items. Since each entity $e_i \in E$ can be considered a binary vector, it can also be expressed as the set that contains all the document $d_j \in D$ such that $(e_i, d_j) = 1$ in the Reference Matrix. This choice is motivated by the intimate simplicity of such metric that scales well to large sparse matrices since it ignores zero values, and easily translates into computational efficiency. Furthermore, the computation of the Jaccard index can be reduced to constant time using optimisations such as MinHash (Broder, 1997), and has the desirable property to be always comprised between zero and one. The similarity of two equal sets is one, whereas the similarity between two sets that have no elements in common is zero. Finally, it is known in the literature that Jaccard index performs better than other methods for finding word similarities in VSM approaches (Lee, 1999; Manning et al., 2008).

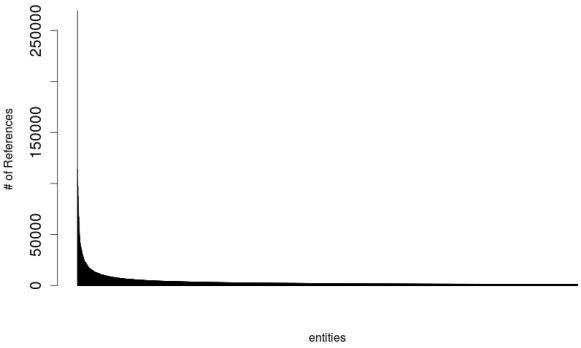

Figure 3: Distribution of page references in the 5000 most referenced Wikipedia pages.

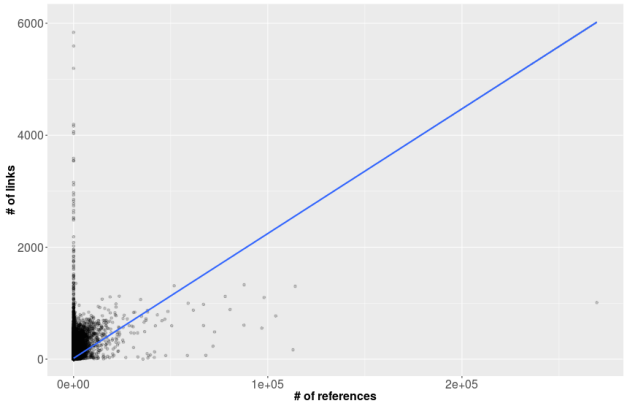

Figure 4: Distribution of page references and links in the page.

4. A real world example: the Wikipedia reference space

Wikipedia is actually the largest human annotated cross-referenced[3] text corpus that can be practically downloaded and which is freely available on the Internet. This allows an extremely relevant case study due to both the good properties of the corpus and its size. For the purposes of this paper we are considering just the English Wikipedia that consists of over 8 millions articles. All Wikipedia articles are considered to identify the document set that, being Wikipedia cross-referenced, equals to the entity set as well; revision pages and other documents that have no encyclopedic value are not considered. The vector space is then constructed as illustrated in the previous section by parsing all articles and the final result is a square matrix, wherein each article is associated to a set of other articles referencing it.

The dimensionality of such a matrix is over 8 millions, which is the count of English Wikipedia's encyclopedic articles[4]. This Reference Matrix is also highly sparse, with few entities being frequently referenced (with a peak of

[3]Here we intend as reference any hyperlink present on the page, not limiting to the homonymous section commonly included in Wikipedia articles.

[4]All the statistics provided in this section refer to a Wikipedia snapshot taken in September 2015.

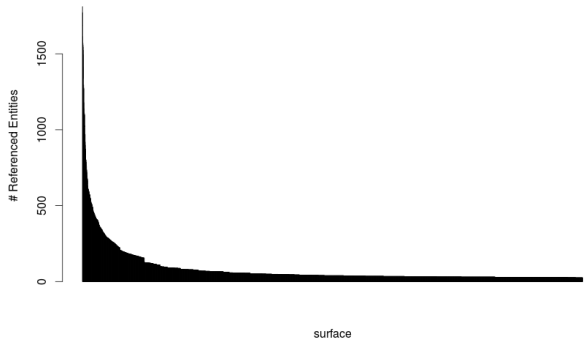

Figure 5: Number of interpretations of the 5000 most ambiguous surface labels in Wikipedia.

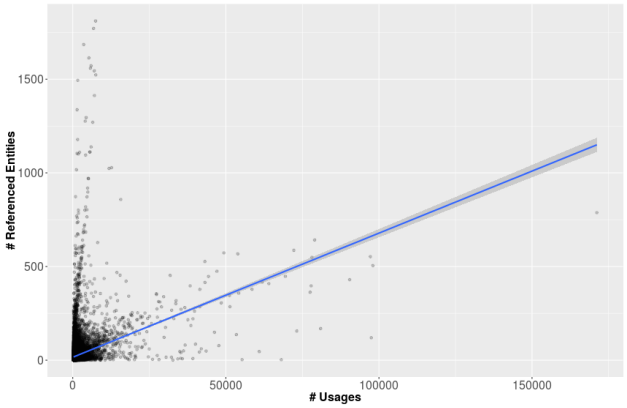

Figure 6: Distribution of surface labels usage and their ambiguity.

over 269000 references for the article "*United States*") and the vast majority getting only a handful of references: while the average number of references to an entity is 9.77, the median is only two, and the 75% of the considered entities have at most four references. This fact is illustrated in Figure 3 where the distribution of references to the 5000 most referenced Wikipedia pages is shown and the power law like trend is evident. There is also a loose correlation between how many links are included in a page and how many times that page is referenced, indicating that frequently referred entities often correspond to articles that point to many other entities. This situation is pictured in Figure 4 where it is evident that most pages have few links and are seldom referenced as well, while only a small set of articles holds the majority of connections.

During the construction of the reference matrix we also build a map of associations between entities and surface labels. This data structure (called *SE-table"*) will be essential for the purpose of word sense disambiguation, since a situation where the same surface label references multiple entities implies polisemy, vice-versa having multiple surface labels for the same entity implies synonymy. The parsing of the whole English Wikipedia generates 14 millions of surface label-entity associations, providing a referenced entity for 12 millions of distinct terms. From this structure we can obtain interesting distributional information over Wikipedia, like the fact that surface labels have an average of 1.24 meanings, while entities are referenced by an average of 1.88 distinct surface labels. This distribution, however has a huge variability with peaks of over 1800 interpretations for a single term and a vast majority of surface labels yielding a single meaning. Figure 5 provides a clear picture of this situation, showing the distribution of the number of meanings among the 5000 most ambiguous surface labels, which means approximately only the 0.04% of the surface labels found in the parsed articles.

It could be tempting to assume that given a random document it is highly unlikely to find within its text highly polysemous surface labels, however these labels are also the most frequently used in the considered document corpus. In fact, the ambiguity of a surface label and its number of occurrences in the corpus have a correlation of 0.38, that, together with the power law like distribution of both ambiguity and occurrences implies that, out of the total 12 millions of surface labels, the vast majority yields one meaning and it is seldom used, as clearly visible in Figure 6.

5. Exploiting local properties in the referential space

In this section we present how the referential space can be exploited to perform complex information retrieval and natural language tasks that imply an underlying semantic interpretation layer. The considered tasks are related entity search, that consists in finding similar entities to the one given as input; and word sense disambiguation, that involves associating to a string its precise meaning in a specific context. Referential spaces provide a spatial representation of entities we can exploit to pinpoint sets of related entities under the assumption that semantically related entities will be close to each other.

5.1. Related entity retrieval

Related entity retrieval consists in finding a set of entities semantically related to a specific entity used as query and it is what several search engines, like Bing and Google, nowadays do in their zero-click information boxes exploiting semantic knowledge extracted from DBpedia or Freebase. For instance, when searching for "Tim Berners-Lee", the inventor of the HTTP protocol, Google suggests, among others "Robert E. Kahn" and "Vint Cerf" who invented the TCP-IP protocol. After a few interactions, it can be easily noticed by any user of such systems that these related entities generally belong to a single class, for instance searching for a person results in retrieving related people and searching for a movie will result in the retrieval of more movies. On the other hand explorative search tools allow the retrieval of different classes of entities, however the user has to choose a "direction" (that under the hood is an RDF property) in which he intends to expand his search. Referential spaces, instead, allow a transparent multidimensional search leveraging the assumption that semantically related entities will appear close to each other in the space. Once a Reference Matrix is built, it can be used to retrieve for a given element e in the entity set E that represents the

query item, a neighbourhood of related entities. To perform this task the element e is compared to all the other elements of the matrix using a similarity metric (see Section 3.2.), this allows a ranking among the elements of E according to their relatedness with e. The n most related elements to e can then be retrieved. This procedure bears a strong resemblance with item-item collaborative filtering algorithms, in fact both tackle the problem of retrieving related or similar items. The complexity of the procedure is linear on the number of entities in the Reference Matrix, moreover the task can also be easily parallelised and the result pre-computed and stored as it happens in collaborative item-item recommender systems.

Back to the "Tim Berners-Lee" example, our method retrieves in the Wikipedia referential space, among others, "Robert Cailliau" who assisted Tim Berners-Lee, "World Wide Web", and "CERN", offering a variety of related concepts that are not limited to a single class, but include in this case people, organisations, and technologies. Generally speaking, our technique does not impose any boundary upon the kind of entities retrieved, allowing the discovery of unknown and unexpected connections. Another major advantage of this technique is that it allows the creation of a neighbourhood of related entities for any point in the referential space. This means that the query vector is not bound to represent a entity.

5.2. Word sense disambiguation

Word sense disambiguation consists in associating to a string a precise meaning, i.e. the actual entity it refers to. In the context of the Web entities are represented by URIs, so word sense disambiguation on the Web boils down to associating a URI to a string, that is what hypertextual links do, thus annotating text documents with hyperlinks. In state-of-art systems this task is mostly performed by leveraging the network structure of ontologies or the distribution of words surrounding the candidate surface label. Referential spaces allow for a different approach that leverages Occam's razor: given a set of surface labels, each with a set of possible meanings, the interpretation that is more likely to be correct is the one that involves the more related entities. This assumption is motivated by the fact that in the context of a document there generally is a logical thread uniting its entities, while totally unrelated concepts are unlikely to be mentioned in the same text; given a large enough document corpus, like Wikipedia, such logic connections might be embedded in the referential space making the related entities occupy nearby regions. In the referential space the similarity between entities is assessable by measuring their distance, therefore we look for interpretations whose involved entities are close to each other.

To realise this task both the Reference Matrix and the association map between surface labels and entities are exploited. Once the candidate entities in the text are spotted the latter data structure is used to retrieve the possible meanings of the spotted surface labels.

Each possible interpretation is computed, resulting in a set of $\prod_{i=1}^{n} |M_i|$ interpretation where n is the number of surface labels to be considered and M_i is the set of meanings of the surface label i. Since each entity can be seen as a point

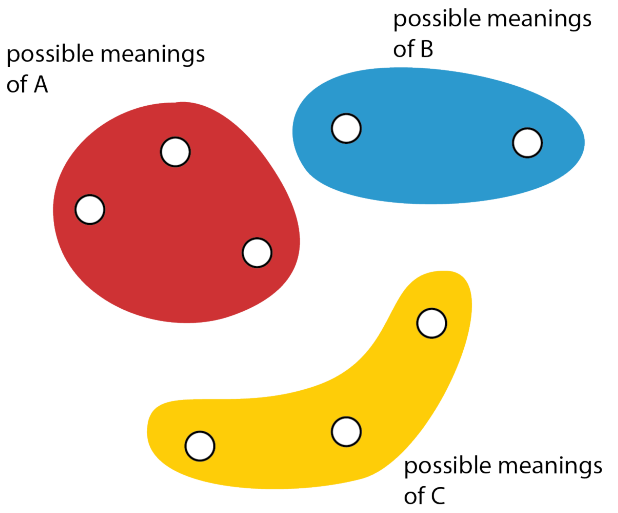

Figure 7: The possible entities referenced by three surface labels.

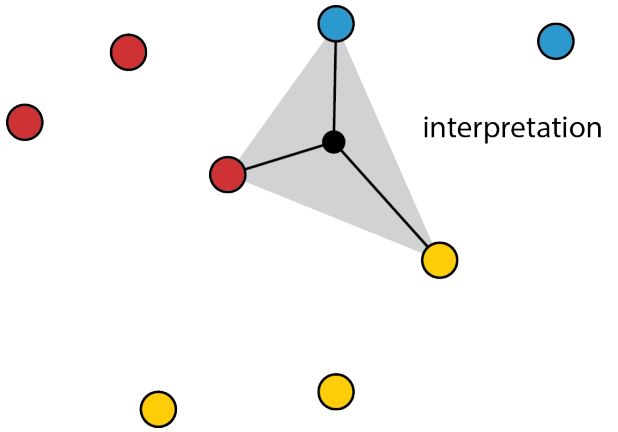

Figure 8: An interpretation of the three surface labels A, B, and C.

in the referential space, as shown in Figure 7, each interpretation can be seen as an hypersolid, and the most plausible interpretation is, according to our assumptions, the one which has the smallest volume. Determining the volume of such figures is, however problematic, so an heuristic is used: for each interpretation a *centroid point* is evaluated by averaging the vectors that represent its vector entities, as shown in Figure 8, then, is evaluated the similarity between such point and all vertexes. The interpretation that yields the minimum sum of such distances is considered the most plausible.

This solution may not be efficient from a computational point of view, but addresses one major drawback of state of the art systems: interpreting lists of surface labels with little or no word context. In such a situation most word sense disambiguation algorithms prove ineffective since they need a context of function words around the surface label to interpret it. With the proposed approach, instead even flat lists of keyphrases can be interpreted with no need for textual context.

6. Conclusions

In this work we introduced a novel data structure to represent the semantic knowledge embedded in references inside hypertextual documents, showing how it can support tasks like similar entity search and word sense disambiguation. Wrapping up, Referential spaces are built in an unsupervised way on the top of HTML pages, an extremely available raw material, need minimal processing effort to be built, and are able to carry non trivial distributional semantics, including tacit knowledge embedded in the link structure of Web documents. Such representation of entities and documents, being a vector space, allows the assessment of distance between its points, which allows the identification of neighbourhoods of related entities. Such neighbourhoods can be used to perform tasks such as related entity retrieval and word sense disambiguation.

The presented data structures and algorithms are still under experimentation and a proper evaluation of how similar entity retrieval and word sense disambiguation perform with respect to state of the art systems has not been completed yet. However, due to the simplicity of building a referential space and the tacit nature of the knowledge contained in hypertextual references, we believe that these techniques can be employed in different domains and different applications from the ones current state of the art systems are designed for. In fact a referential knowledge base for a new domain can be created by simply crawling a large enough number of Web pages dealing with the desired topics. Such an economical knowledge base can provide support for tasks that usually require expansive ontological knowledge, extensive linked data sets or a great amount of human-annotated training data. Moreover, while ontologies and linked data, though carrying much more semantics and allowing for deeper automated reasoning, do not cope well with change, requiring expert work to address changes in the domain they describe, referential spaces can be easily updated by adding new documents to the corpus they are built upon. The capability of referential spaces to model easily new domains, adapt to changes, and allow the integration of heterogeneous documents as long as they include hypertextual connections makes them an extremely interesting abstraction that, aside from easing a number of complex artificial intelligence tasks, might also foster the creation of new ontologies and linked data in a complete automated way. After all, a little semantics goes a long way.

7. Bibliographical References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

Broder, A. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29, Jun.

Degl'Innocenti, D., Nart, D. D., and Tasso, C. (2014). A new multi-lingual knowledge-base approach to keyphrase extraction for the italian language. In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 78–85.

Degl'Innocenti, D., De Nart, D., and Tasso, C. (2015). A novel knowledge-based architecture for concept mining on italian and english texts. In Ana Fred, et al., editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 553 of *Communications in Computer and Information Science*, pages 132–142. Springer International Publishing.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hendler, J. (2007). The dark side of the semantic web. *IEEE Intelligent Systems*, (1):2–4.

Huynh, T., Hoang, K., Do, L., Tran, H., Luong, H., and Gauch, S. (2012). Scientific publication recommendations based on collaborative citation networks. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 316–321. IEEE.

Jaccard, P. (1902). Lois de distribution florale. *Bulletin de la Socíeté Vaudoise des Sciences Naturelles*, (38):67–130.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM.

Musetti, A., Nuzzolese, A. G., Draicchio, F., Presutti, V., Blomqvist, E., Gangemi, A., and Ciancarini, P. (2012). Aemoo: Exploratory search based on knowledge patterns over the semantic web. *Semantic Web Challenge*, 136.

Novak, J. D. (2010). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Routledge.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S., and Both, A. (2014). Agdistis-graph-based disambiguation of named entities using linked data. In *The Semantic Web–ISWC 2014*, pages 457–471. Springer.

Wang, J., Shen, H. T., Song, J., and Ji, J. (2014). Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927*.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255. ACM.

Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events

Haizhou Qu, Marcelo Sardelich, Nunung Nurul Qomariyah and Dimitar Kazakov
Artificial Intelligence Group, Department of Computer Science, University of York, UK
hq524, msn511, nq516, dimitar.kazakov@york.ac.uk

Abstract

This article describes a novel dataset aiming to provide insight on the relationship between stock market prices and news on social media, such as Twitter. While several financial companies advertise that they use Twitter data in their decision process, it has been hard to demonstrate whether online postings can genuinely affect market prices. By focussing on an extreme financial event that unfolded over several days and had dramatic and lasting consequences we have aimed to provide data for a case study that could address this question. The dataset contains the stock market price of Volkswagen, Ford and the S&P500 index for the period immediately preceding and following the discovery that Volkswagen had found a way to manipulate in its favour the results of pollution tests for their diesel engines. We also include a large number of relevant tweets from this period alongside key phrases extracted from each message with the intention of providing material for subsequent sentiment analysis. All data is represented as a ontology in order to facilitate its handling, and to allow the integration of other relevant information, such as the link between a subsidiary company and its holding or the names of senior management and their links to other companies.

Keywords: Financial forecasting, stock prices, Twitter, ontology

1. Introduction

On 18 Sep 2015, Volkswagen, one of the world's largest and best known automakers, was named by the US Environment Protection Agency (EPA) as being in breach of its regulations concerning the amount of pollution from diesel engines. Volkswagen had manipulated the outcomes of a vehicle emission test by detecting the specific conditions under which the test took place, and adjusting the performance of its diesel engines in other to meet the required pollution targets, while the same vehicle might fail those targets by a vast margin in actual driving conditions.

Several recent models, including Golf, Polo and the Passat equipped with certain diesel engines were confirmed to contain cheating software that would reduce harmful emissions. The revelation led to a fall of more than 30% of the VW stock price in a single day, which continued to fall in the following weeks, as news was gradually released about the number and seniority levels of people who had knowledge of the deception, until the company CEO himself decided to resign and apologise. There was a prolongued period of uncertainty regarding the spread of this deception across the different continents, and the prices continued to tumble as it became clear that it was not limited to the US market. In addition, subsidiary brands, such as Audi, Seat and Škoda soon revealed the existence of similar practices, with a corresponding effect on their own sales figures and share prices.

We observed these events and collected relevant tweets for the period 15–30 Sep 2015, as well as the minute by minute intra-day stock market prices for Volkwagen (stock symbol $VW, as traded on the Frankfurt Stock Exchange), Ford ($F, NYSE) as an example of an automotive company with no links to the scandal, and the same type of data for the S&P500 stock market index (SPY), providing a baseline for comparison with the US economy as a whole. The Twitter data was then enhanced with the addition of extracted key phrases suitable for sentiment analysis, and the entire dataset was stored as an ontology[1].

2. Financial Forecasting

Since the advent of the stock markets, studying and predicting the future of companies and their share price have been the main tasks facing all market participants. It is extremely difficult to achieve an accurate model that remains reliable over time. There is a very famous yet controversial Efficient Market Hypothesis (EMH) (Fama, 1965), which comes in three forms: weak, semi-strong and strong. If the weak form holds true, stock price cannot be predicted using history prices. The semi-strong form of EMH suggests that stock price reveals all publicly available information. The strong form implies that stock prices will always reflect all information including any hidden information, including even insider's information, if the hypothesis holds. Numerous studies show that EMH does not always hold true (Grossman and Stiglitz, 1980; Haugen, 1995; Shleifer, 2000; Shiller, 2003; Butler and Kazakov, 2012). In all cases, attempts to model and forecast the market are based on time series containing the prices of relevant stock along with other relevant information, which often includes indicators of the general state of the market to allow the evaluation of the relative performance of a given company with respect to the general market trends.

3. Mining Twitter

Along with the development of Social Networking, Twitter has became one of the most popular ways for people to publish, share and acquire information. The two characteristics of this service, instantaneity and publicity, make it a good resource for studying the behaviour of large groups of people. Makinf predictions using tweets has proved a popular research topic. Asur and Huberman (2010) used tweet rate

[1] See the data available at `http://j.mp/FinancialEventsOntology`.

time series to forecast movie sales, with the result outperforming the baseline market-based predictor using HSX,[2] the gold standard of this industry. O'Connor et al. (2010) presented a way to use tweets to predict the US presidential polls. The authors concluded that evolution of tweet sentiment is correlated with the results of presidental elections and also with presidential job approval. Tumasjan et al. (2010) used a much smaller dataset of tweets to forecast the result of the 2009 German elections. Eichstaedt et al. (2015) studied the use of sentiment keywords to predict country level heart disease mortality.

Information extraction from social media can be rather challenging, due to the fact that the texts are very short (up to 140 characters only), noisy and written in an informal style, which often contains bad spelling and non-standard abbreviations (Piskorski and Yangarber, 2013).

4. Ontologies For Financial Data

Ontologies are powerful Artificial Intelligence approach to representing structured knowledge. Their use can also facilitate knowledge sharing between software agents or human users (Gruber, 1993). They are often used in text mining to represent domain knowledge, but their use to describe dynamic processes like time series has been much more limited. The use of ontologies has already been considered in the context of Twitter, as well as in the domain of financial news. For instance, Kontopoulos et al. (2013) discuss the benefits of their use when calculating a sentiment score for Twitter data. Mellouli et al. (2010) describe a proposal for an ontology with 31 concepts and 201 attributes for financial headline news. Lupiani-Ruiz et al. (2011) present an ontology based search engine for financial news. Cotfas et al. (2015) have used ontologies to model Twitter sentiments, such as happiness, sadness or affection. Lee and Wu (2015) developed a framework to extract key words from online social messages and update related event ontologies for fast response to unfolding events.

5. The VW Pollution Scandal Dataset

Despite the substantial amout of research on Twitter data in recent years (Bollen et al., 2011; Wolfram, 2010; Zhang et al., 2011; Si et al., 2013), there are very few publicly available datasets for academic research, with some of the previously published datasets becoming unavailable for various reasons. Yang and Leskovec (2011) provide a large Twitter dataset which has 467 million tweets from 20 million users from 1 June to 31 Dec 2009, or 7 months in total, representing an estimated 20–30% of all tweets published during this period. Go et al. (2009) provide a Twitter dataset labelled with sentiment polarity (positive, neutral or negative), and also split into a training set of 1.6 million tweets (0.8 million positive and 0.8 million negative), and a manually selected test set with 182 positive tweets, and 177 negative tweets.

So far, there has not been a publicly available Twitter dataset, which is aligned with company stock prices. We aim to address this gap, with a focus on an extreme financial event, which could prove helpful in revealing the interplay between financial data and news on social media.

[2]Hollywood Stock Exchange

We collected tweets and retweets from 00:00h EDT on 15 Sep 2015 until 23:59h EDT on 30 Sep 2015.[3] In order to retrieve only relevant tweets, we queried the Twitter API using the tags and keywords listed in Table 1.

Table 1: Tags and keywords for the selection of tweets

Tag/keywords	
@vw	#volkswagen
$vow	#volkswagengate
$vlkay	#volkswagencheat
#vw	#volkswagendiesel
#vwgate	#volkswagenscandal
#vwcheat	#dieselgate
#vwdiesel	emission fraud
#vwscandal	emission crisis

One encouraging observation about this dataset is that it contained tweets with relevant information that predated the official EPA announcement that started the VW diesel engine pollution scandal, as shown below.

Published at 2015, September 18, 10:56:35 EDT

> EPA[4] set to make announcement on major automaker $GM $F $TM $FCAU $HMC $NSANY $TSLA $VLKAY $DDAIF $HYMLF http://t.co/02hNHKq9cx

Published at 2015, September 18, 11:47:58 EDT

> .@EPA to make announcement regarding a "major automaker" at 12 noon today. Source says it will involve @VW. No details yet. Stay tuned.

Published at 2015, September 18, 11:51:42 EDT

> Inbox: EPA, California Notify Volkswagen of Clean Air Act Violations

The first and second tweet did not clearly state that Volkswagen was exactly the automaker, the third tweet is the first one with a clear statement which is ahead of EPA official announcement.[5]

A total of 536,705 tweets were extracted. We have chosen the third tweet as a point in time to split the data into the period 'before the news was out', and the one that followed, resulting in 51,921 tweets before11:51:42 on 18 Sep 2015, and 484,784 after that time. Figure 2 shows a histogram of the number of tweets over each 12h period. A brief timeline of relevant events of the Volkswagen scandal according to Kollewe (2015) is listed below:

18 Sep EPA announces that Volkswagen cheated on the vehicle polluiton test. 482,000 VW diesel cars are required to be recalled in the US.

[3]Earlier tweets were also included if they were retweeted during the indicated time interval.

[4]US Environmental Protection Agency

[5]The attentive reader will find it interesting to compare the timing of the EPA announcement with the closing for the weekend of the Frankfurt stock exchange on that Friday.

20 Sep VW orders an external investigation and CEO apologizes to public.

21 Sep Share price drops by 15 billion Euros in minutes after the Frankfurt stock exchange opens.

22 Sep VW admits 11 million cars worldwide fitted with cheating devices. The CEO says he is "endlessly sorry" but will not resign. The US chief, Michael Horn, says the company "totally screwed up".

23 Sep The CEO quits but insists he is "not aware of any wrongdoing on his part". Class-action lawsuits are filed in the US and Canada and criminal investigations are launched by the US Justice Department.

24 Sep Officical confirms that VW vehicles with cheating software were sold across Europe as well. The UK Department for Transport says it will start its own inquiry into car emissions, as VW faces a barrage of legal claims from British car owners.

26 Sep Switzerland bans sales of VW diesel cars.

28 Sep German prosecutors launch an investigation of VW ex-CEO Winterkorn.

30 Sep Almost 1.2 million VW diesel vehicles in the UK are affected by the scandal, more than one in ten diesel cars on Britain's roads.

We have extended the Twitter dataset with a set of key phrases of length 2 that are potentially relevant to sentiment analysis. In this, we followed the approach discussed by Turney (2002). The main idea is to identify syntactic patterns that are considered suitable to matching subjective opinions (as opposed to objective facts). The resulting candidates for such *polarity_keywords* are linked in the database to the tweet from which they were extracted. This approach can be compared to another related approach to opinion extraction from financial news (Ruiz et al., 2012), in which sentiment gazetteers were also used to indicate the news polarity. Here the decision about polarity has not been made, but is left to future users of the data.

To extract the keywords in question, we employed the Stanford Part-Of-Speech (POS) tagger and Tgrep2 tool to extract the tag patterns proposed by Turney (2002), as listed in Table 2. About a third of all messages were annotated with pairs of key words as a result of the above mentioned procedure. In Table 3 we list the 20 most common pairs: on the whole, they appear quite specific and well correlated with the corpus topic.

In addition to the Twitter data, our dataset includes price information on the per-minute basis for Volkswagen (symbol: VOW.DE) shares and those of Ford (symbol: F) as an example of an automaker unaffected by the scandal. In addition, we have included S&P500 data (American Stock Market Index, symbol: SPY) as an indication of the state of the markets as a whole during the period in question. The data, as available from a number of public websites, includes time stamps, the 'open' and 'close' price, as well as the 'high' and 'low' price for the given one minute interval.

Figure 1 shows a comparison of Buy-and-hold[6] cumulative returns of those three securities during 15-30 Sep. 2015.

6. Ontology Representation and Sample Queries

The hierarchy of classes representing the dataset is shown in Figures 3. The **Event** class has three properties: *datetime*, *epoch* and *duration*. The *epoch* property is the number of seconds elapsed from 1st January 1970 00:00 UTC, which provides a common timeline between individuals. The *duration* property describes how long an event lasts and in our dataset, we use second as the timing unit. The **Event** class has two subclasses: **Tweet** and **OHLC**[7]. **Tweet** contains all the individuals storing tweets with their properties: *id*, *username*, *url*, *sourceUrl*, *numberOfRetweet* and *polarityKeyword*. **OHLC** contains all the individuals of stock price of specific company or market index. Each of them has the following properties: *high*, *low*, *open*, *close*, *symbol* and *isin*[8] (See Listing 1).

Listing 1: Individuals of **OHLC** and **Tweet**, shown in turtle format.

```
@prefix nsp: <http://example.org/vwevent2015/property/> .
@prefix nss: <http://example.org/vwevent2015/ontology/OHLC/> .
@prefix nst: <http://example.org/vwevent2015/ontology/Tweet/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# An individual of OHLC
nss:f-1442323800 nsp:close "13.84"^^xsd:float ;
    nsp:datetime "2015-09-15T09:30:00-04:00" ;
    nsp:duration "60"^^xsd:xsd:unsignedLong ;
    nsp:epoch "1442323800"^^xsd:unsignedLong ;
    nsp:high "13.86"^^xsd:float ;
    nsp:low "13.79"^^xsd:float ;
    nsp:symbol "F" ;
    nsp:isin "US3453708600" ;
    nsp:open "13.8"^^xsd:float ;
    nsp:return "0.00289855072464"^^xsd:float .

# An individual of Tweet
<http://example.org/vwevent2015/ontology/Tweet/646575192907644928> nsp:
      datetime "2015-09-23T02:41:53-04:00" ;
    nsp:epoch "1442990513"^^xsd:unsignedLong ;
    nsp:id 646575192907644928 ;
    nsp:numberOfRetweet "0"^^xsd:unsignedLong ;
    nsp:polarityKeyword "criminal charges" ;
    nsp:sourceUrl <http://twitter.com/brian_poncelet/status
          /646575192907644928> ;
    nsp:url <http://twitter.com/Brian_Poncelet/status/646575192907644928> ;
    nsp:username "brian_poncelet" .
```

Representing our data as an ontology makes it possible to be queried in a flexible and powerful fashion, allowing its users to link the textual and time series data in a seamless way. Here are some examples of SPARQL queries seeking to extract useful features through the use of both polarity keywords and stock price movements.

Query 1 This SPARQL query will extract the tweets whose time stamp coincides with a drop in the Volkswagen stock price by more than 1%, ranked by *numberOfRetweets*.

The results of this query 1 are shown in listing 3. In order to improve readability, returns only show three decimal places, and *datetime*s are reformatted not to show the year.

[6]Buy-and-hold is a trading strategy, typically for benchmarking purposes, that considers the performance of buying the security and holding it for the whole period of analysis. Cumulative return on day i: $r_i = (price_i - price_{buy})/price_{buy}$.

[7]OHLC stands for open, high, low and close price of stock price during a period of time.

[8]ISIN refers to International Securities Identification Numbers, which provides a unique identification for each security.

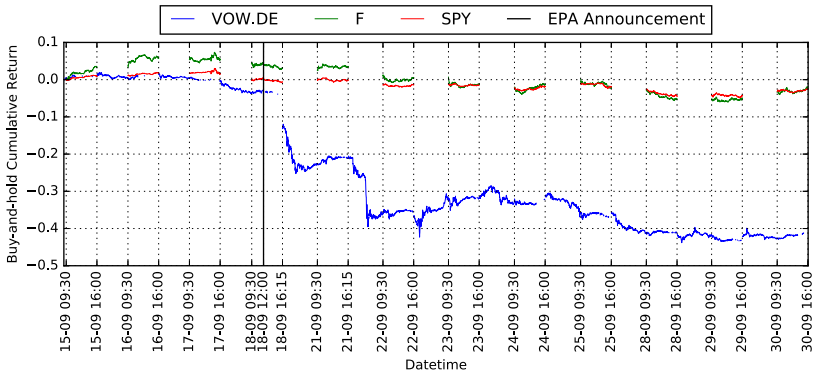

Figure 1: Buy-and-hold cumulative returns of Volkswagen stock, Ford stock and S&P500 during 15-30 September 2015.

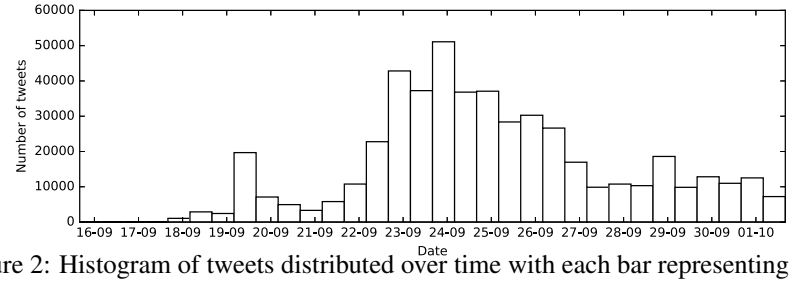

Figure 2: Histogram of tweets distributed over time with each bar representing 12 hours.

Listing 2: Query 1

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT
  ?username
  (?id AS ?tweet_id)
  ?return
  (?numberOfRetweet AS ?nbRt)
  ?datetime
  (group_concat(distinct ?pk;separator=", ") as ?polarityKeywords)
WHERE{
  ?ohlc nsp:epoch ?ohlc_epoch .
  ?ohlc nsp:return ?return .
  ?ohlc nsp:symbol "VOW.DE" .
  FILTER(?return < -0.01)
  ?tweet nsp:epoch ?tweet_epoch .
  ?tweet nsp:datetime ?datetime .
  ?tweet nsp:numberOfRetweet ?numberOfRetweet.
  ?tweet nsp:url ?url .
  ?tweet nsp:sourceUrl ?sourceUrl .
  ?tweet nsp:username ?username .
  ?tweet nsp:id ?id .
  ?tweet nsp:polarityKeyword ?pk .
  FILTER EXISTS{?tweet nsp:polarityKeyword ?pk}
  FILTER(
    ?url = ?sourceUrl
    && xsd:integer(?numberOfRetweet) >= 5
    && xsd:integer(?tweet_epoch) <= xsd:integer(?ohlc_epoch) + 60
    && xsd:integer(?tweet_epoch) >= xsd:integer(?ohlc_epoch)
  )
}
GROUP BY ?username ?id ?return ?numberOfRetweet ?datetime
ORDER BY DESC(xsd:integer(?numberOfRetweet)) ?return
LIMIT 10
```

Listing 3: Result of Query 1

	username	tweet_id	return	nbRt	datetime	polarityKeywords
1	business	646586797636644864	-0.023	113	09-23 03:28:00	as much
2	newswaala	646580334616645633	-0.011	30	09-23 03:02:19	high emissions, first detected
3	twistools_en	646586860173688832	-0.023	8	09-23 03:28:15	national embarrassment
4	nytimesbusiness	646260916129005568	-0.022	6	09-22 05:53:04	diesel cars, little effect
5	speedmonkeycouk	648435351476957184	-0.011	6	09-28 05:53:29	now being

Listing 4: Query 2

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT
  ?pk
  (COUNT(?pk) AS ?count)
WHERE{
  {
    SELECT
      (xsd:unsignedLong(xsd:float(?ohlc_epoch)/60.0) AS ?ohlc_minute)
      (xsd:unsignedLong(xsd:float(?tweet_epoch)/60.0+1.0) AS ?tweet_minute)
      ?pk
      ?return
    WHERE{
      ?ohlc nsp:epoch ?ohlc_epoch ;
        nsp:return ?return ;
        nsp:symbol "VOW.DE" .
      FILTER(?return <= -0.02)
      ?tweet nsp:epoch ?tweet_epoch ;
        nsp:polarityKeyword ?pk ;
    }
    HAVING(?ohlc_minute=?tweet_minute)
  }
}
GROUP BY ?pk
ORDER BY DESC(?count)
LIMIT 20
```

Listing 5: Result of Query 2

	pk	count
1	worldwide fitted	41
2	as much	23
3	entire auto	14
4	sure people	11
5	first detected	10
6	high emissions	10
7	multiple probes	7
8	totally screwed	7
9	chief executive	5
10	diesel cars	5
11	early trading	5
12	here come	5
13	national embarrassment	5
14	little effect	4
15	not sure	4
16	also installed	3
17	false emission	3
18	internal investigations	3
19	just lost	3
20	absolutely foolish	2

Expression	Word1	Word2	followed by
(JJ.(NN|NNS))	JJ	NN or NS	no restrictions
(RB.(JJ!.(NN|NNS)))	RB	JJ	not NN nor NNS
(RBR.(JJ!.(NN|NNS)))	RBR	JJ	not NN nor NNS
(RBS.(JJ!.(NN|NNS)))	RBS	JJ	not NN nor NNS
(JJ.(JJ!.(NN|NNS)))	JJ	JJ	not NN nor NNS
(NN.(JJ!.(NN|NNS)))	NN	JJ	not NN nor NNS
(NS.(JJ!.(NN|NNS)))	NS	JJ	not NN nor NNS
(RB.(VB|VBD|VBN|VBG))	RB	VB, VBD, VBN or VBG	no restrictions
(RBR.(VB|VBD|VBN|VBG))	RBR	VB, VBD, VBN or VBG	no restrictions
(RBS.(VB|VBD|VBN|VBG))	RBS	VB, VBD, VBN or VBG	no restrictions

Table 2: Extracted Word1+Word2 keyphrases using *Tgrep2* expressions

keywords	count		
diesel scandal	3993	diesel deception	1294
chief executive	3835	multiple probes	1189
diesel emissions	3280	electric car	1166
diesel cars	2980	new tech	1110
sure people	2801	clean diesel	1059
new boss	2407	criminal probe	1037
totally screwed	2208	finally be	953
clean air	1919	fresh start	908
as many	1449	refit cars	898
criminal charges	1323	diesel vehicles	890

Table 3: 20 most common pairs of keywords extracted from the Twitter data.

Figure 3: VW Event Ontology Classes

The first tweet was published by Bloomberg (@business):

> CEO Martin Winterkorn faces a showdown with #Volkswagen's board later http://bloom.bg/1FdA4sA

Tweet No. 4 came from Business news of NY Times (@nytimesbusiness):

> Volkswagen's recall troubles may have little effect on China: It sells almost no diesel cars in the country. http://nyti.ms/1Jmipd8

Apart from main public media accounts, we found that among the authors of those tweets are also an indian media (No. 2), a marketing account (No. 3), a motor amateur (No. 5). This indicates our dataset contains information from a range of sources that provide potentially useful information on this event.

Query 2 We have also been able to check whether some of the keywords are associated with specific stock price movements by using the following SPARQL query, which aims to retrieve the keywords associated on drops in Volkswagen price greater than 2% within any one-minute-period.

The result of Query 2 shows that in most cases, the worst drops in VW price coincide with keywords expressing negative sentiment or referring to some of the specific facts of the scandal (e.g. "worldwide fitted", "diesel cars").

Query 3 For users with access to twitter contents (mapped to nsp:content), listing 6 shows the potential usage of connecting with other existing ontologies to combine domain knowledge with stock price time series: *get the average one minute return of stock the surname of a key person* (CEO for example) *appears in the tweets.*

Listing 6: Query 3

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX db: <http://dbpedia.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT
  ?sn (AVG(?return) AS ?avgReturn)
WHERE{
  SERVICE <http://dbpedia.org/sparql/>{
    ?company dbo:keyPerson ?person .
    ?person foaf:surname ?surname .
    BIND(LCASE(STR(?surname)) AS ?sn)
    FILTER(?company=<http://dbpedia.org/resource/Volkswagen>)
  }
  ?ohlc nsp:epoch ?ohlc_epoch .
  ?ohlc nsp:return ?return .
  ?ohlc nsp:symbol "VOW.DE" .
  ?tweet nsp:epoch ?tweet_epoch .
  ?tweet nsp:content ?content .
  ?tweet nsp:url ?url .
  ?tweet nsp:sourceUrl ?sourceUrl .
  ?tweet nsp:id ?id .
  ?tweet nsp:numberOfRetweet ?numberOfRetweet
  FILTER(?url=?sourceUrl && ?numberOfRetweet > 100)
  FILTER(CONTAINS(LCASE(?content), ?sn))
  FILTER(
    xsd:integer(?ohlc_epoch) >= xsd:integer(?tweet_epoch) &&
    xsd:integer(?ohlc_epoch) <= xsd:integer(?tweet_epoch) + 60
  )
}
GROUP BY ?sn
```

7. Conclusion and Future Works

With the advantages of ontology representation, discovering useful information in time-labelled text data (tweets) and numerical time series (stock prices) becomes an easier task. Both queries and dataset can be easily modified or extended. On the other hand, copyright issues with Twitter data put limits to displaying and sharing information in a more straightforward way, and restrict us to only displaying tweet IDs in our dataset.

The polarity keywords are a useful feature, despite the unsupervised way in which they were extracted. Our future work will focus on adding to the range of features available in the dataset.

We also want to assess our work in connection with other related ontologies for stock markets[9] (Alonso et al., 2005) and companies[10] as described in DBpedia. Such integration for example should allow one to recognise Volkswagen Group as an entity of Public Company in DBpedia[11], where we can find information about their assets, revenue, owner, holding company, products and many more. This type of information would potentially allow one to automatically link one company affected by adverse events to, say, its subsidiary companies, which one may expect also to feel the repercussions of such events. Indeed, Audi, Seat and Škoda, all subsidiary companies of VW Group, were all eventually linked to the diesel engine cheating software scandal. More recent news from France has shown that any results from our data could also find use to handle other related news from the automotive industry. We hope that our work will encourage more interesting research in the financial domain as a whole.

8. References

Alonso, L., Bas, L., Bellido, S., Contreras, J., Benjamins, R., and Gomez, M. (2005). WP10: Case Study eBanking D10. 7 Financial Ontology. *Data, Information and Process Integration with Semantic Web Services, FP6-507483*.

Asur, S. and Huberman, B. A. (2010). Predicting the Future with Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, volume 1, pages 492–499. IEEE.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1 – 8.

Butler, M. and Kazakov, D. (2012). Testing Implications of the Adaptive Market Hypothesis via Computational Intelligence. In *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2012 IEEE Conference on*, pages 1–8. IEEE.

Cotfas, L.-A., Delcea, C., Roxin, I., and Paun, R., (2015). *New Trends in Intelligent Information and Database Systems*, chapter Twitter Ontology-Driven Sentiment Analysis, pages 131–139. Springer International Publishing, Cham.

[9] http://dbpedia.org/page/Stock_market

[10] http://dbpedia.org/ontology/company

[11] http://dbpedia.org/resource/Volkswagen

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., et al. (2015). Psychological Language on Twitter Predicts County-level Heart Disease Mortality. *Psychological Science*, 26(2):159–169.

Fama, E. F. (1965). The Behavior of Stock-market Prices. *Journal of Business*, 38(1):34–105.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, 1:12.

Grossman, S. J. and Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, pages 393–408.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220.

Haugen, R. A. (1995). *The New Finance: the Case Against efficient markets*. Prentice Hall Englewood Cliffs, NJ.

Kollewe, J. (2015). Volkswagen Emissions Scandal Timeline. http://www.theguardian.com/business/2015/dec/10/volkswagen-emissions-scandal-timeline-events Accessed: Jan. 08, 2016.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based Sentiment Analysis of Twitter Posts. *Expert Systems with Applications*, 40(10):4065–4074.

Lee, C.-H. and Wu, C.-H. (2015). Extracting Entities of Emergent Events from Social Streams Based on a Data-Cluster Slicing Approach for Ontology Engineering. *International Journal of Information Retrieval Research*, 5(3):1–18, July.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., and Camón-Herrero, J. B. (2011). Financial News Semantic Search Engine. *Expert Systems with Applications*, 38(12):15565–15572.

Mellouli, S., Bouslama, F., and Akande, A. (2010). An Ontology for Representing Financial Headline News. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2–3):203–208.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, 11(122-129):1–2.

Piskorski, J. and Yangarber, R. (2013). Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 23–49. Springer.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. (2012). Correlating Financial Time Series with Micro-blogging Activity. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12*, page 513.

Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, pages 83–104.

Shleifer, A. (2000). *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford University Press.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. In *ACL (2)*, pages 24–29.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *ICWSM*, 10:178–185.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wolfram, M. S. A. (2010). *Modelling the Stock Market using Twitter*. Master thesis, The University of Edinburgh.

Yang, J. and Leskovec, J. (2011). Patterns of Temporal Variation in Online Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186. ACM.

Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, 26:55–62.

×-*Protégé*
An Ontology Editor for Defining Cartesian Types to Represent n-ary Relations

Christian Willms, Hans-Ulrich Krieger, Bernd Kiefer
German Research Center for Artificial Intelligence (DFKI)
Campus D3 2, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
christian.willms@dfki.de, krieger@dfki.de, kiefer@dfki.de

Abstract

Arbitrary n-ary relations ($n \geq 1$) can, in principle, be realized through binary relations obtained by a reification process which introduces new individuals to which the additional arguments are linked via "accessor" properties. Modern ontologies which employ standards such as RDF and OWL have mostly obeyed this restriction, but have struggled with it nevertheless. In (Krieger and Willms, 2015), we have laid the foundations for a *theory-agnostic* extension of RDFS and OWL and have implemented in the last year an extension of *Protégé*, called ×-*Protégé*, which supports the definition of Cartesian types to represent n-ary relations and relation instances. Not only do we keep the distinction between the domain and the range of an n-ary relation, but also introduce so-called *extra* arguments which can be seen as position-oriented unnamed *annotation* properties and which are accessible to entailment rules. As the direct representation of n-ary relations abolishes RDF triples, we have backed up ×-*Protégé* by the semantic repository and entailment engine *HFC* which supports tuples of arbitrary length. ×-*Protégé* is programmed in Java and is made available under the Mozilla Public License.

Keywords: ontology editor, ×-*Protégé*, Cartesian types, n-ary relations, RDF, RDFS, OWL, n-ary Description Logics.

1. Decription Logics, OWL, and RDF

Relations in description logics (DLs) are either unary (so-called *concepts* or *classes*) or binary (*roles* or *properties*) predicates (Baader et al., 2003). As the designers of OWL (Smith et al., 2004; Hitzler et al., 2012) decided to be compatible with already existing standards, such as RDF (Cyganiak et al., 2014) and RDFS (Brickley and Guha, 2014), as well as with the universal RDF data object, the *triple*,

subject predicate object

a unary relation such as $C(a)$ (class membership) becomes a binary relation via the RDF type predicate:

a rdf:type C

For very good reasons (mostly for decidability), DLs usually restrict themselves to decidable function-free two-variable subsets of first-order predicate logic. Nevertheless, people have argued ver early for relations of more than two arguments (Schmolze, 1989), some of them still retaining decidability and coming up with a better memory footprint and a better complexity for the various inference tasks (including querying) than their triple-based relatives (Krieger, 2012; Krieger, 2014). This idea conservatively extends the standard *triple-based* model towards a more general *tuple-based* approach ($n + 1$ being the arity of the *predicate*):

subject predicate $object_1 \ldots object_n$

Using a standard relation-oriented notation, we often interchangeably write

$p(s, o_1, \ldots, o_n)$

Here is an example, dealing with *diachronic* relations (Sider, 2001), relation instances whose object values might change over time, but whose subject values coincide with each other. For example (quintuple representation),

peter marriedTo liz 1997 1999
peter marriedTo lisa 2000 2010

or (relation notation)

marriedTo(*peter, liz, 1997, 1999*)
marriedTo(*peter, lisa, 2000, 2010*)

which we interpret as the (time-dependent) statement that *Peter* was married to *Liz* from 1997 until 1999 and to *Lisa* from 2000–2010.

In a triple-based setting, semantically representing the same information requires a lot more effort. There already exist several approaches to achieve this (Welty and Fikes, 2006; Gangemi and Presutti, 2013; Krieger and Declerck, 2015), all coming up with at least one brand-new individual (introduced by a hidden existential quantification), acting as an *anchor* to which the object information (the range information of the relation) is bound through additional properties (a kind of *reification*). For instance, the so-called *N-ary relation encoding* (Hayes and Welty, 2006), a W3C best-practice recommendation, sticks to binary relations/triples and uses *container* objects to encode the range information (ppt1 and ppt2 being the *new* individuals):

peter marriedTo ppt1
ppt1 rdf:type nary:PersonPlusTime
ppt1 nary:value liz
ppt1 nary:starts "1997"^^xsd:gYear
ppt1 nary:ends "1999"^^xsd:gYear
peter marriedTo ppt2
ppt2 rdf:type nary:PersonPlusTime
ppt2 nary:value lisa
ppt2 nary:starts "2000"^^xsd:gYear
ppt2 nary:ends "2010"^^xsd:gYear

As we see from this small example, a quintuple is represented by five triples. The relation name is retained, however, the range of the relation changes from, say, Person to the type of the container object which we call here PersonPlusTime.

Rewriting ontologies to the *latter* representation is an unpleasant enterprise, as it requires further classes, redefines property signatures, and rewrites relation instances,

as shown by the marriedTo example above. In addition, reasoning and querying with such representations is extremely complex, expensive, and error-prone.

Unfortunately, the *former* tuple-based representation which argues for additional (temporal) arguments is **not** supported by *ontology editors* today, as it would require to deal with general n-ary relations ($n \geq 2$). *×-Protégé* fills exactly this gap.

2. Further Motivation

×-Protégé supports the definition of Cartesian types, composed from standard OWL classes and XSD datatypes. Given Cartesian types and by keeping the distinction between the *domain* $\mathbb{D}$ and the *range* $\mathbb{R}$ of a binary property p, it is now possible to define $m + n$-ary relations $p \subseteq \mathbb{D}_1 \times \cdots \times \mathbb{D}_m \times \mathbb{R}_1 \times \cdots \times \mathbb{R}_n$.

The deeper reason why it is still useful to separate domain and range arguments from one another is related to the so-called *property characteristics* built into OWL, e.g., symmetry or transitivity. This ultimately allows us to generalize the corresponding entailment rules, by replacing atomic classes with Cartesian types. For instance, entailment rule rdfp4 for *transitive* properties p from (ter Horst, 2005)

$$p(x, y) \wedge p(y, z) \rightarrow p(x, z)$$

can be generalized as ($m = n = o$)

$$p(\times_{i=1}^{m} x_i, \times_{j=1}^{n} y_j) \wedge p(\times_{j=1}^{n} y_j, \times_{k=1}^{o} z_k) \rightarrow p(\times_{i=1}^{m} x_i, \times_{k=1}^{o} z_k)$$

×-Protégé not only keeps the distinction between the *domain* and *range* arguments of a relation, but also provides further distinct *annotation*-like arguments, called *extra* arguments which have been shown useful in various situations and which are accessible to entailment rules of the above kind. Consider a binary *symmetric* property q which we would like to generalize by the concept of *valid time* (the time in which an atemporal statement is true), thus the corresponding entailment rule needs to be extended by two further temporal arguments b and e:

$$q(x, y, b, e) \rightarrow q(y, x, b, e)$$

By assuming that the temporal arguments are part of the domain and/or range of q, we are running into trouble as symmetric properties require the same number of arguments in domain and range position. Thus, we *either* need to adjust this rule, i.e.,

$$q(\underline{x, b, e}, \underline{y, b, e}) \rightarrow q(\underline{y, b, e}, \underline{x, b, e})$$

or assume that b and e have a special "status". We decided for the latter and call such information *extra arguments*. As an example, the former marriedTo relation (a symmetric relation) is of that kind, thus having the following relation signature (assuming a biography ontology with class Person):

$$\underset{\textit{domain}}{\underline{\text{Person}}} \times \underset{\textit{range}}{\underline{\text{Person}}} \times \underset{\textit{2 extra arguments}}{\underline{\text{xsd:gYear} \times \text{xsd:gYear}}}$$

Other *non-temporal* examples of *extra arguments* might involve *space* (or *spacetime* in general), using further XSD custom types, such as point2D or point3D, in order to encode the position of a moving object over time (Keshavdas and Kruijff, 2014).

More linguistically-motivated examples include the *direct* representation of ditransitive and ergative verb frames, including adjuncts (Krieger, 2014). We will present an example of this at the end of Section 7. when defining the quaternary relation obtains. Such kinds of properties are often wrongly addressed in triple-based settings through *relation composition*, applied to the second argument of the corresponding binary relation. This does *not* work in general, but only if the original relation is *inverse functional*.

As a last example, we would like to mention the *direct* representation of *uncertain* statements in medicine or technical diagnosis in an extension of OWL (Krieger, 2016) which is far superior to various encodings described in (Schulz et al., 2014) which have accepted the boundaries of RDF triples in order to be compatible with an existing standard.

3. *Protégé*, *×-Protégé*, and *HFC*

Protégé is a free, open source ontology editor, providing a graphical user interface to define and inspect ontologies (http://protege.stanford.edu). *Protégé* version 4 has been designed as a modular framework through the use of the OSGi framework as a plugin infrastructure (https://www.osgi.org/developer/). For this reason, *×-Protégé* has been implemented as an EditorKitFactory plugin for *Protégé*, replacing the built-in OWL EditorKitFactory. The EditorKit is the access point for a particular type of model (in our case, a model based on n-tuples) to which a GUI has access to.

×-Protégé is divided into three separate components (Figure 1, large right box). The "bottom" layer is realized by *HFC* (Krieger, 2013), a bottom-up forward chainer and semantic repository implemented in Java which is comparable to popular systems such as Jena and OWLIM (http://www.dfki.de/lt/onto/hfc/). *HFC* supports RDFS and OWL reasoning à la (Hayes, 2004) and (ter Horst, 2005), but at the same time provides an expressive language for defining custom rules, involving functional and relational variables, complex tests and actions, and the replacement of triples in favour of tuples of arbitrary length. The query language of *HFC* implements a subset of SPARQL, but at the same time provides powerful custom M:N aggregates ($M, N \geq 1$), not available in SPARQL.

The data read in by *HFC* is preprocessed and transformed into an *×-Protégé* model. Among other things, it contains inheritance hierarchies for classes and properties which are directly used to visualize the ontology in the graphical user interface of *×-Protégé*.

This GUI consists of several workspaces (similar to *Protégé*, version 4.3), presenting the ontology itself, the classes, the properties, and the instances. User actions result in an update of the model and *HFC*'s n-tuple database.

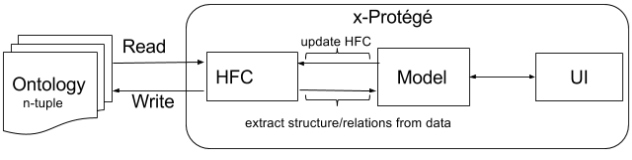

Figure 1: The three-layered structure of *×-Protégé*.

In the next section, we will look into some of these workspaces (or tabs), assuming the marriedTo example from Sections 1. and 2.

4. Class Tab

When starting ×-*Protégé* the class hierarchy consists of a unique, most general type, called Thing+ in the GUI which subsumes every other Cartesian type and which can be formally defined as

$$\text{Thing+} := \bigsqcup_{i=1}^{k} (\text{owl:Thing} \sqcup \text{xsd:AnyType})^i$$

For a given ontology, k is fixed (finite, of course). Initially, Thing+ has two direct subtypes, viz., owl:Thing and xsd:AnyType. *HFC* already provides a set of built-in XSD subtypes, such as xsd:gYear (Gregorian Year) or xsd:int (4 Byte integers), but also defines non-standard datatypes, such as xsd:monetary. As in a pure OWL setting, owl:Thing and xsd:AnyType are incompatible, but xsd:AnyType is made available under Thing+ in order to define Cartesian types, such as xsd:gYear × xsd:gYear for the two extra arguments of the marriedTo relation (or even Person × xsd:gYear × xsd:gYear for the sexternary relation q in Section 2.). This small type hierarchy is depicted in Figure 2.

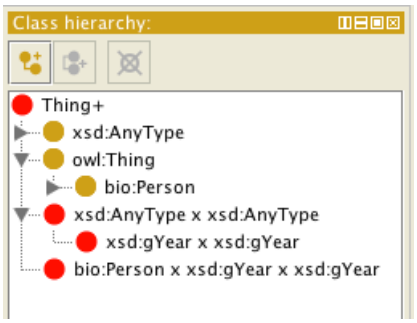

Figure 2: The class hierarchy for the marriedTo example.

Note that the non-singleton Cartesian types are highlighted using red colour and that xsd:gYear × xsd:gYear is correctly classified as a subclass of the Cartesian type xsd:AnyType × xsd:AnyType.

5. Property Tab

As in OWL, we distinguish between the property characteristics owl:DatatypeProperty and owl:ObjectProperty. We group these two classes under the super-property MixedProperty, as we do allow for further "mixed" property characteristics; e.g., properties which are instantiated with an XSD atom in first place or properties with Cartesian domain and range types which are a mixture of OWL classes and XSD types (and thus are neither datatype nor object properties). Since the quaternary relation marriedTo (binary relation plus two extra args) maps URIs onto URIs, it is classified as an object property (remember, the extra args neither belong to the domain nor range of a property). However, the ternary relation hasAge (binary relation plus one extra args) is a datatype property as it maps URIs onto XSD ints (the extra arg is the *transaction time*, the time when the birthdate was entered to *HFC*); cf. Figure 3.

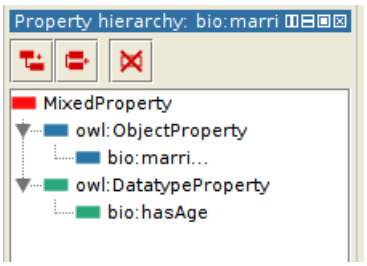

Figure 3: The property hierarchy for the marriedTo and hasAge relations.

Figure 4: The property signature for the marriedTo relation.

When defining a new property, a user is required to choose the right Cartesian types to complete the property signature. This is displayed in Figure 4 for the marriedTo relation. Depending on the kind of property, an ontology engineer is even allowed to associate further property characteristics with a property under definition; see Figure 5.

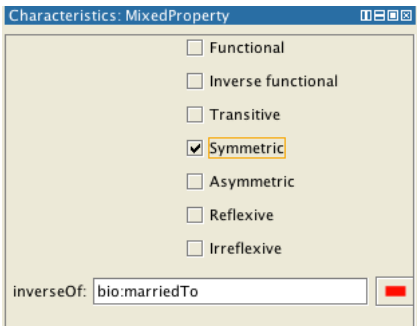

Figure 5: Further potential property characteristics for the marriedTo relation.

6. Instance Tab

We complete the overview of the workspace tabs by coming back to *Peter* and his relation to *Liz* and *Lisa* (cf. Section 1.). From the instance tab, we learn about his two marriages and that he is currently 53 years old (see Figure 6).
The symmetry of the marriedTo relation (see Figure 5) further guarantees that *Peter* is listed in the instance tabs of *Liz* and *Liza* as well.

7. N-Tuples & I/O Formats

As ×-*Protégé* allows us to deviate from pure binary relations, certain adjustments to the *N-triples* format (Carothers and Seaborne, 2014) are necessary, especially as extra arguments need to be represented. Assume a quaternary relation obtains between a person and a degree obtained from an educational organization at a specific time:

$$\text{obtains} \subseteq \underbrace{\text{Person}}_{\mathbb{D}} \times \underbrace{\text{Degree} \times \text{School}}_{\mathbb{R}_1 \times \mathbb{R}_2} \times \underbrace{\text{xsd:date}}_{\mathbb{A}}$$

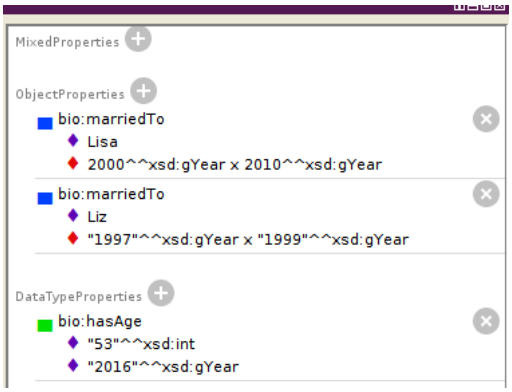

Figure 6: Facts about *Peter*.

In order to let the system know of how many arguments the domain, the range, and the extra part of a relation is composed of, we add further length-related information (infix notation):

```
obtains rdfs:domain Person
obtains rdfs:range Degree School
obtains nary:extra xsd:date
obtains nary:domainArity "1"^^xsd:int
obtains nary:rangeArity "2"^^xsd:int
obtains nary:extraArity "1"^^xsd:int
```

Notice that the rdfs:range keyword directly above is followed by *two* classes: Degree and School ($= \mathbb{R}_1 \times \mathbb{R}_2$). Not only is this kind of representation used in the RBox of an ontology, but also in the TBox, e.g.

```
Degree School rdfs:subClassOf owl:Thing owl:Thing
```

as

$$\text{Degree} \times \text{School} \sqsubseteq \top \times \top$$

is the case. ABox information is also affected by this style of representation, as, for instance

```
peter obtains phd stanford "1985"^^xsd:date
```

Besides providing such an (asymmetric) *infix* representation, ×-*Protégé* let the user decide whether a *prefix* representation is more appropriate for him/her. So, for instance, the last ABox statement above would then become

```
obtains peter phd stanford "1985"^^xsd:date
```

We finally like to stress the fact that once one decided to go for a direct representation of additional arguments and reason upon them, queries and rules will usually intermix tuples of different length. For example, in a *valid time* approach *universal* information from the TBox and RBox of an ontology is encoded as triples, whereas *assertional* knowledge will be represented as quintuples (Krieger, 2012); see *HFC* rule at the end of Section 8.

8. Future Work

Since ×-*Protégé* already uses functionality from *HFC* (see Section 3.), we would like to add further *query* and *rule definition* tabs to the next major version of ×-*Protégé* to support the construction of *HFC* queries and rules (see the two examples below).

The query support in ×-*Protégé* will ease the definition of SPARQL-like queries in *HFC* over n-tuples, using keywords such as SELECT, SELECTALL (for the *multiply-out* mode in *HFC* in case equivalence class reduction is enabled), DISTINCT, WHERE, FILTER, and AGGREGATE. Depending on the property signatures, ×-*Protégé* will then alarm a user if too less, too many, or wrong arguments have been specified in WHERE clauses, FILTER tests, or AGGREGATE functions. This helps to simply the construction of a query such as

```
SELECT DISTINCT ?partner
WHERE peter marriedTo ?partner ?start ?end
FILTER GreaterEqual ?start "1998"^^xsd:gYear &
       LessEqual ?end "2005"^^xsd:gYear
AGGREGATE ?noOfPartners = Count ?partner
```

which computes how many times *Peter* was married to distinct women between 1998 and 2005. The results of such queries (viz., tables) will also be displayed in this tab.

The rule support will provide means to define, maintain, and extend RDFS, OWL, and custom rule sets. Again, as is the case for queries, clauses, @test, and @action sections of rules in *HFC* will benefit from checking for the right number of arguments. For instance, the valid time extension of the entailment rule for *transitive* properties (ter Horst, 2005) in *HFC* looks as follows (Krieger, 2012):

```
?p rdf:type owl:TransitiveProperty     // triple
?x ?p ?y ?start1 ?end1                 // quintuple
?y ?p ?z ?start2 ?end2
→
?x ?p ?z ?start ?end
@test                                  // 3 LHS tests
?x != ?y
?y != ?z
IntersectionNotEmpty ?start1 ?end1 ?start2 ?end2
@action                                // 2 RHS actions
?start = Max2 ?start1 ?start2          // new RHS variable
?end = Min2 ?end1 ?end2                // new RHS variable
```

In both cases, we would also like to provide a *completion* mechanism for properties and URIs, as well as for external tests (see @test above) and value-returning functions (see @action above), an extremely useful functionality known from programming environments.

Our ultimate goal is thus to offer ×-*Protégé* as a front-end GUI for ontology-based systems, based on *HFC*.

9. Download

×-*Protégé* version 1.0 as of Monday Feb 15, 2016 can be downloaded from https://bitbucket.org/cwillms/x-protege/downloads/ and is made available under the Mozilla Public License. Here, you will also find a preliminary version of the user guide.

10. Acknowledgements

The research described in this paper has been partially financed by the European project PAL (Personal Assistant for healthy Lifestyle) under Grant agreement no. 643783-RIA Horizon 2020. This work was conducted using the *Protégé* resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. We would like to thank the three reviewers for their detailed and useful suggestions.

11. Bibliographical References

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (2003). *The Description Logic Handbook*. Cambridge University Press, Cambridge.

Brickley, D. and Guha, R. (2014). RDF Schema 1.1. Technical report, W3C.

Carothers, G. and Seaborne, A. (2014). RDF 1.1 N-Triples. a line-based syntax for an RDF graph. Technical report, W3C.

Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax. Technical report, W3C.

Gangemi, A. and Presutti, V. (2013). A multi-dimensional comparison of ontology design patterns for representing n-ary relations. In *39th International Conference on Current Trends in Theory and Practice of Computer Science*, pages 86–105.

Hayes, P. and Welty, C. (2006). Defining N-ary relations on the Semantic Web. Technical report, W3C.

Hayes, P. (2004). RDF semantics. Technical report, W3C.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2012). OWL 2 web ontology language primer (second edition). Technical report, W3C.

Keshavdas, S. and Kruijff, G.-J. (2014). Functional mapping for human-robot cooperative exploration. *International Journal of Computer and Applications*, 36(1).

Krieger, H.-U. and Declerck, T. (2015). An OWL ontology for biographical knowledge. representing time-dependent factual knowledge. In *Proceedings of the Workshop on Biographical Data in a Digital World*.

Krieger, H.-U. and Willms, C. (2015). Extending OWL ontologies by Cartesian types to represent N-ary relations in natural language. In *Proceedings of the IWCS Workshop on Language and Ontologies*.

Krieger, H.-U. (2012). A temporal extension of the Hayes/ter Horst entailment rules and an alternative to W3C's n-ary relations. In *Proceedings of the 7th International Conference on Formal Ontology in Information Systems (FOIS)*, pages 323–336.

Krieger, H.-U. (2013). An efficient implementation of equivalence relations in OWL via rule and query rewriting. In *Proceedings of the 7th IEEE International Conference on Semantic Computing (ICSC)*, pages 260–263.

Krieger, H.-U. (2014). A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in RDF and OWL. In *Proceedings of the 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA)*.

Krieger, H.-U. (2016). Capturing graded knowledge and uncertainty in a modalized fragment of OWL. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 19–30.

Schmolze, J. G. (1989). Terminological knowledge representation systems supporting n-ary terms. In *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 432–443.

Schulz, S., Martínez-Costa, C., Karlsson, D., Cornet, R., Brochhausen, M., and Rector, A. (2014). An ontological analysis of reference in health record statements. In *Proceedings of the 8th International Conference on Formal Ontology in Information Systems (FOIS 2014)*.

Sider, T. (2001). *Four Dimensionalism. An Ontology of Persistence and Time*. Oxford University Press.

Smith, M. K., Welty, C., and McGuinness, D. L. (2004). OWL Web Ontology Language Guide. Technical report, W3C.

ter Horst, H. J. (2005). Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary. *Journal of Web Semantics*, 3:79–115.

Welty, C. and Fikes, R. (2006). A reusable ontology for fluents in OWL. In *Proceedings of 4th FOIS*, pages 226–236.

A Lexical Ontology to Represent Lexical Functions

Alexsandro Fonseca[1], Fatiha Sadat[1], François Lareau[2]
[1]Université du Québec à Montréal, [2]Université de Montréal
[1]201 Président Kennedy, Montreal, Canada, [2]C.P. 6128 succ. Centre-Ville, Montreal, Canada
affonseca@gmail.com, sadat.fatiha@uqam.ca, francois.lareau@umontreal.ca

Abstract

Lexical functions are a formalism that describes the combinatorial, syntactic and semantic relations among individual lexical units in different languages. Those relations include both paradigmatic relations, i.e. vertical or "in absence", such as synonymy, antonymy and meronymy, and syntagmatic relations, i.e. horizontal or "in presence", such as intensification (*deeply committed*), confirmative (*valid argument*) and support verbs (*give an order*, *subject to an interrogation*). We present in this paper a new lexical ontology, called Lexical Function Ontology (LFO), as a model to represent lexical functions. The aim is for our ontology to be combined with other lexical ontologies, such as the Lexical Model for Ontologies (lemon) and the Lexical Markup Framework (LMF), and to be used for the transformation of lexical networks into the semantic web formats, enriched with the semantic information given by the lexical functions, such as the representation of syntagmatic relations (e.g. collocations) usually absent from lexical networks.

Keywords: Lexical Functions, Ontology, Lexical Relations, Lexical Network, Lexical Semantics, Collocations, Multiword Expressions

1. Introduction

We present in this paper an ongoing project that aims to represent the lexical functions (Mel'čuk, 1996) of the Meaning-Text Theory (MTT) (Mel'čuk, 1997) as a lexical ontology, called Lexical Functions Ontology (LFO).

A lexical ontology is a representation of the different aspects of the lexicon, such as meaning, morphology, part of speech, as well as the relation among lexical units, such as syntactic, semantic and pragmatic relations, using the semantic web formalisms (RDF/OWL languages).

Our objective in this project is to use this ontology in order to represent the relations among lexical units in lexical networks, especially in those networks based on lexical functions (LFs), such as the *Réseau lexical du français* (RLF) (Lux-Pogodalla and Polguère, 2011). However, it can also be used to represent different lexical relations in other lexical networks, such as WordNet in RDF/OWL format. This is an important aspect, since most of the existing lexical networks do not implement the syntagmatic information (Schwab et al., 2007) provided by some of the LFs. Moreover, we show how this model can be used to represent collocations in a lexical network, since the relation among lexical units in a collocation is a syntagmatic relation (Mel'čuk 1998).

We do not intend to recreate lexical representations already realized by previous works, such as lemon (McCrae et al., 2012), LexInfo (Buitelaar, 2009) or LMF (Francopoulo, 2007). Our proposal is to use, whenever possible, the lexical information already implemented by those models, such as the classes "LexicalEntry" and "LexicalSense" in the lemon model, and create the necessary classes for the implementation of lexical functions information.

2. Foundations and related work

We present in this section the theoretical information about lexical functions and related work.

2.1. Lexical functions

Bolshakov and Gelbukh (1998) defined a lexical function (LF) as a formalism for the description and use of combinatorial properties of individual lexemes. A more technical definition, given by Mel'čuk (1998), says that a "*Lexical Function f* is a function that associates with a given lexical unit *L*, which is the argument, or keyword, of *f*, a set $\{L_i\}$ of (more or less) synonymous lexical expressions – the *value* of *f* – that are selected contingent on *L* to manifest the meaning corresponding to *f*:

$$f(L) = \{L_i\}$$

The LFs considered in this paper are the standard ones, differentiated from the non-standard by the fact that the former can be coupled with a higher number of possible keywords and value elements (Mel'čuk 1998). For example, the LF *Magn*, which represents the sense 'intensification', can be coupled with many keywords (e.g. $shave_N$, *easy*, *to condemn*, *naked*, *thin*, *to rely*, and many others) to give different values : Magn(*shave*) = {*close*, *clear*}; Magn(*easy*) = {*as pie*, *as 1-2-3*}; *Magn*(*to condemn*) = *strongly*; Magn(*naked*) = *stark*; Magn(*thin*) = *as a rake*; *Magn*(*to rely*) = *heavily*; (Mel'čuk 1998). On the other hand, the sense "*additionné de…"* (with the addition of…) is a non-standard LF in French, because it can only be coupled with a few number of keywords (*café*; *fraises*; *thé*), to create the expressions: *café crème*, *fraises à la crème* (and not **café à la crème*, **fraises crème*); *café au lait*; *café arrosé*; *café noir*; *thé nature*; etc (Mel'čuk 1992).

About 70 simple standard LFs have been identified (Kolesnikova, 2011). Complex LFs are formed by the combination of simple standard ones.

LFs can be classified as paradigmatic or syntagmatic, according to the kind of lexical relation they model. Figure 1 illustrates the difference between the two kinds of relations. The paradigmatic LFs model the vertical, "in absence" or "in substitution" relation among lexical units (Saussure, 1983). For example, antonymy, Anti(*big*) = *small*; synonymy, Syn(*car*) = *automobile*; hyponymy, Hypo(*feline*) = {*cat*, *tiger*, *lion*, etc.}. Syntagmatic LFs model the horizontal, "in presence" or "in composition" relations among lexical units (Saussure, 1983). For example: magnification, Magn(*committed*) = *deeply*; confirmation, Ver(*argument*) = *valid*; laudatory, Bon(*advice*) = {*helpful*, *valuable*}. And simple standard LFs can be combined to form complex ones. For instance, AntiBon(*criticism*) = *harsh*; AntiMagn(*similarity*) = *vague*.

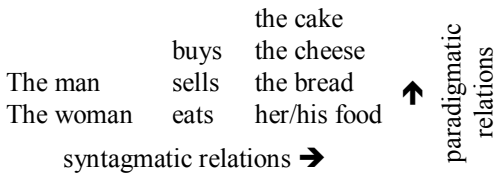

Figure 1: The difference between syntagmatic and paradigmatic relations. There are syntagmatic relations between "*The man*" and "*buys*", "*buys*" and "*the cheese*", "*The man*" and "*sells*", "*The woman*" and "*buys*", etc. There are paradigmatic relations between "*the cake*" and "*the cheese*", "*the cheese*" and "*the bread*", etc., which belong to the same syntactic/semantic fields and can substitute each other in a phrase.

Another important concept is that of semantic actant (Sem-actant) (Mel'čuk, 2004). In logic, a predicate is a falsifiable assertion. Each predicate has one or more arguments. For example, in the assertion "Rome is the capital of Italy", we can define the predicate 'capital' having two arguments, 'Rome' and 'Italy': *capital*(*Italy*, *Rome*).

In linguistics, the predicate is called "predicative sense" and the arguments are its "semantic actants". Each LF represents a different predicative sense and the semantic actants are represented by subscripts. For example, the LF *S* gives the equivalent noun of the value to which it is applied. S_1 gives the first actant (the one who executes the action), S_2 gives the second actant (the object of the action) and S_3 gives the third actant (the recipient of the action): S_1(*to teach*) = *teacher*; S_2(*to teach*) = {*subject*; *matter*}; S_3(*to teach*) = {*pupil*; *student*}. Other subscripts give circumstantial information. For example: S_{loc} – local of the action/event; S_{instr} – instrument used; etc.

LFs can be classified according to their semantic or syntactic behaviour. For example, in (Mel'čuk, 1998) we find the following classification:

- Semantic derivatives: S_1(*to teach*) = *teacher*; S_3(*to teach*) = *pupil*; S_{loc}(*to fight*) = *battlefield*; $S_{instr}(murder_{V,N})$ = *weapon;* $A_1(anger_N)$ = *angry;* Adv_1(*anger*) = *angrily*;
- Semi-auxiliary verbs: $Oper_1$(*support*) = [*to*] *lend* [~ *to N*]; $Oper_1(promise_N)$ = [*to*] *make* [ART ~]; $Func_2$(*proposal*) = *concerns* [*N*];
- Realization verbs: $Real_1$(*bus*) = [*to*] *drive* [ART ~]; $Real_2$(*bus*) = [*to*] *ride* [*on* ART ~]; $Real_1(promise_N)$ = [*to*] *keep* [ART ~];
- Modifiers: Magn(*injury*) = *serious*; Ver(*citizen*) = *loyal*; Ver(*argue*) = *convincingly*; Bon(*analysis*) = *fruitful*.

In this work, we combine different classifications of the LFs, especially those presented by Mel'čuk (1998) and Mel'čuk et al., (1995) to create the classes of our ontology.

2.2. lemon model

lemon (McCrae et al., 2012) is a model for sharing lexical information on the semantic web. It is based on earlier models, such as LexInfo (Buitelaar, 2009) and LMF (Francopoulo, 2007). As its main advantages over these previous models, we cite:

- separation between the linguistic and the ontological information;
- linguistic information, such as "partOfSpeech" and "writenForm" are represented as RDF properties, differently of LMF, which represent them as attributes of a property, which makes easier the use of other resources, like the SPARQL query language;
- lemon uses ISOCat, data categories homologated by ISO (for example, "partOfSpeech", "gender" and "tense");
- lemon is an easily extensible model;
- there are already many linguistic resources in lemon format, like WordNet and DBPedia Wiktionary.

Lexical units are represented in the lemon model using the classes "LexicalEntry" and "LexicalForm". The "LexicalEntry" class is connected to the lexical unit sense, which is represented by the "LexicalSense" class. The connection between the lemon model and external ontologies are made through this last class.

In our model, the keyword and the value of a LF will be represented as a lemon "LexicalSense" class. In MTT, the different senses of a word are represented by subscripts, using Roman and Arabic numbers and Latin letters (Mel'čuk 1995), which we illustrate here with an example. Consider the word "*ocean*". It has concrete senses, like "*a body of water that covers the planet*" and abstract senses, like in "*ocean of people*". In MTT, the concrete senses of "*ocean*" would be represented as "$Ocean_I$" and the abstract senses as "$Ocean_{II}$". Inside "$Ocean_I$" we could have subdivisions:

- $Ocean_{I.1a}$: "*extension of water that covers the planet*" (always in singular, referring to the entire body of water);
- $Ocean_{I.1b}$: the set of oceans in general (always in plural) – "*the oceans are becoming more polluted*.";
- $Ocean_{I.2}$: a part of $Ocean_{I.1a}$ in a specific region – *Atlantic Ocean*, *Pacific Ocean*, *Arctic Ocean*, etc.

In our model, the word "*ocean*" is represented by a lemon object "LexicalEntry" and $Ocean_{I.}$, $Ocean_{I.1a}$, $Ocean_{I.1b}$, $Ocean_{I.2}$ and $Ocean_{II}$ are each represented by a "LexicalSense" lemon object. The reason for this is explained as follows: the semantic connection represented by an individual LF is between senses, and not between lexical forms or lexical entries. Doing so, we can have an already disambiguated lexical network when connecting lexical units with a LF.

2.3. ILexicOn lexical ontology

Lefrançois and Gandon (2011) present a lexical ontology based on MTT for the construction of a dictionary.
Their approach is based on a three layers architecture:
- the meta-ontology layer;
- the ontology layer;
- the data layer.

The meta-ontology layer is formed by what the authors call meta-classes, which are super classes for the classes in the ontology layer. For example, the meta-class "ILexicalUnit" is a super class of all types of lexical units and the meta-class "ISemanticRelation" is a super-class for all the semantic relations appearing in the ontology layer. The ontology layer is formed by classes that represent concepts, such as "Entity", "Person" and "State". They are connected to the meta-class "ILexicalUnit" by a "is-a" relation and to each other by semantic relations that are instances of the ISemanticRelation meta-class. The data layer contains instances of the classes in the ontology layer. For example, "Mary01" can be an instance of the class "Person" and "Alive01" an instance of the class "Alive".
The authors justify those layers saying that this ensures three of the four redaction principles of an explanatory and combinatory dictionary (Mel'cuk et al, 1995), in MTT: the principles of formality, internal coherence and uniform treatment. The principle of exhaustivity is not ensured. In their model, the collocations and locutions are represented as dictionary entries of the keyword's collocation or locution.
The difference between their work and our model is that ILexicOn is intended to represent an entire dictionary following the Meaning-Text Theory precepts, while our model is intended to represent lexical relations in a lexical network. Moreover, using ILexicOn, collocations and locutions are represented as dictionary entries, while with our model, they will be represented as a graph, representing connections between lexical units.

3. The LFO Model

Figure 2 illustrates the LFO core model. The central class in our model is the "LexicalRelation". It connects to the LexicalFunction class, to the lexical relation type (which can be paradigmatic or syntagmatic), and to the value and to the key of the lexical relation (LR).
We decided to connect the LF keyword and the LF value using an intermediate class (LexicalRelation), instead of connecting them directly with the LexicalFunction class because in this way we can connect to the LexicalRelation information that is specific to the relation between two lexical units, independently of the LF connection them, and we can connect to a LF information that is independent of the lexical units that it connects. Also, the paradigmatic/syntagmatic information (LRType) is connected to the LexicalRelation class instead of being connected to the LexicalFunction class. Although the LFs usually have a definite type (paradigmatic or syntagmatic), some of them do not have, which will depend on the lexical units they model.
Figure 3 illustrates how the collocation "*close friend*" would be represented. It is modelled by the LF Magn (predicative sense = intensification): Magn($friend_{I.1}$) = $close_{III.1a}$; Since also Magn($friend_{I.1}$) = $good_{II}$, we could have another LexicalRelation (*Magn_02*) connecting the LexicalSense $good_{II}$ and the LexicalSense $friend_{I.1}$.

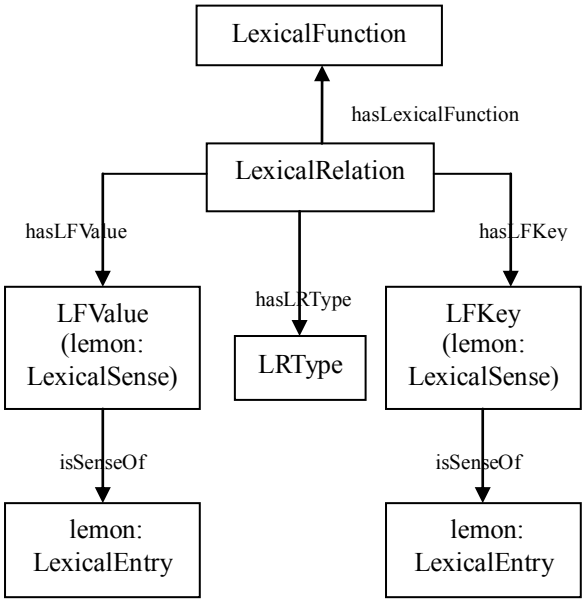

Figure 2: The LFO core model combined with the LexicalSense and LexicalEntry lemon objects

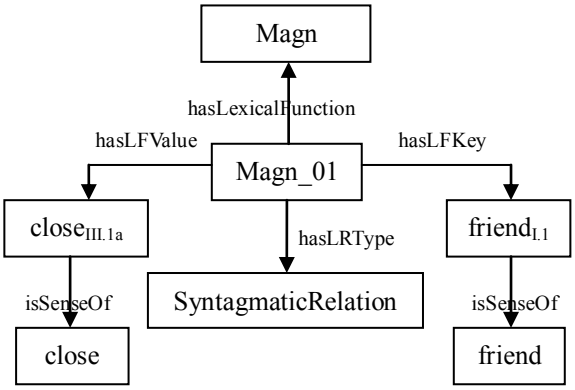

Figure 3: The representation of the collocation "*close friend*"

The lexical relation is connected to the value of the collocation using the property "hasLFValue" and to the keyword using the property "hasLFKey". The property "hasLRType" informs that the relation between

"*close*$_{III.1a}$" and "*friend*$_{I.1}$", modelled by the LF "Magn", is a syntagmatic relation.

Example 1 illustrates a possible RDF/OWL code of our previous example. As explained in Section 2.2, it is important to note that the lexical units that appear in our example, "*friend*$_{I.1}$", and "*close*$_{III.1a}$" will be modeled as "LexicalSense" and not as a "LexicalEntry" lemon object. This means that our model will connect to the lemon model via the sense of the lexical units. This allows the construction of already disambiguated lexical networks. Finally, the lexical variations (e.g. plural) can be treated at the level of the LexicalEntry lemon object, already implemented by the lemon model.

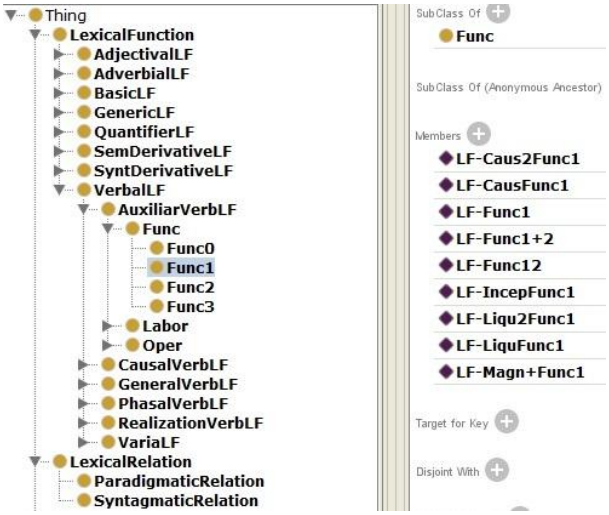

Figure 4: An overview of the LFs and some complex LFs implemented as members of the class *Func$_1$*

```
:close a lemon:LexicalEntry.
:closeIII.1a
    a lemon:LexicalSense;
    lemon:isSenseOf :close.
:friend a lemon:LexicalEntry.
:friendI.1
    a lemon:LexicalSense;
    lemon:isSenseOf :friend.
:LexicalFunction rdf:type owl:Class.
:AdjectivalLF rdf:type owl:Class;
    rdfs:subClassOf :LexicalFunction.
:Magn rdf:type owl:Class;
    rdfs:subClassOf :AdjectivalLF.
:LF-Magn rdf:type :Magn, owl:NamedIndividual.
:LexicalRelation rdf:type owl:Class.
:SyntagmaticRelation rdf:type owl:Class;
    :rdfs:subClassOf :LexicalRelation.
:Magn_01 a lfo:LexicalRelation;
    lfo:hasLRType  lfo:SyntagmaticRelation;
    lfo:hasLFValue :closeIII.1a;
    lfo:hasLFKey :friendI.1;
    lfo:hasLexicalFunction lfo:LF-Magn;
```

Example 1: The RDF/OWL representation, using the Turtle notation, of the collocation "*close friend*"

Figure 4 is an overview of how the simple LFs are organized by OWL classes and how the complex LFs are modelled as members of a class. For example, the LF *Func$_1$* has the following members: CausFunc$_1$, Func$_1$, Func$_{1+2}$, Func$_{12}$, etc. In total, about 600 simple and complex LFs are already represented in our model, which were extracted from a relational database representation of the RLF (Lux-Pogodalla and Polguère, 2011).

4. Conclusion and Future Work

We presented in this paper an ongoing project, called Lexical Function Ontology (LFO), aimed at the representation of the lexical functions of Meaning-Text Theory as a lexical ontology.

Most of the existing lexical networks lack important semantic information, especially the syntagmatic relations among lexical units. Lexical functions are a powerful tool for the representation of linguistic relations. In particular, syntagmatic lexical functions can fill the present gap in the representation of syntagmatic relations in lexical networks.

Moreover, the combination of the descriptive logic embedded in the OWL language with the semantic and syntactic information provided by lexical functions creates a strong tool for studying human reasoning and for interesting psycholinguistic studies.

Finally, this work can be seen as a new form for the representation of multiword expressions.

As future work, we intend to complete the representation of lexical functions in our model with the combinatorial, semantic, syntactic, and communicational perspectives presented by Jousse (2010). The implementation of such classification perspectives will allow invaluable semantic, syntactic and pragmatic information to be coded directly in a lexical network. For example, the LF *CausFunc1* has as perspectives:

- combinatorial: ThreeActants;
- part of speech of the value of the function: verb;
- semantics: represents a cause;
- target: the target of the function is its first actant;

This information can be connected to each LexicalFunction using owl:objectProperties. For example:

```
:CausFunc1 hasCombinatorialPersp :ThreeActants.
:CausFunc1 hasSemanticPersp :CauseSemPersp.
:CausFunc1 hasValuePOS isoCat:Verb.
:CausFunc1 hasTarget :FirstActant.
```

Also as a future work, we intend to use our model to transform the *Réseau Lexical du Français* (Lux-Pogodalla and Polguère, 2011), from its present relational database format to an ontology format (already in progress).

5. Acknowledgments

We want to thank Professor Alain Polguère for providing us with the datasets from RLF and the reviewers for helpful comments and suggestions.

6. Bibliographical References

Bolshakov, I. and Gelbukh, A. (1998). Lexical functions in Spanish. *Proceedings CIC-98, Simposium Internacional de Computación*, pp. 383-395. November 11-13, 1998, Mexico D.F.

Buitelaar, P., Cimiano, P., Haase, P. et Sintek, M. (2009). Towards linguistically grounded ontologies. In *L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554*, pp. 111-125, Spring-Verlag Berlin. Heidelberg 2009.

Francopoulo, G., Bel, N., Georg, M., Calzolari, N., Monachini, M., Pet, M. and Soria, C. (2007). Lexical markup framework: ISO standard for semantic information in NLP lexicons. In: *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*.

Jousse, A. (2010). Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales. Thèse de doctorat. Université de Montréal et Université Paris Diderot (Paris 7), 340 p.

Kolesnikova, O. Automatic extraction of lexical functions, PhD Thesis (2011). Instituto Politecnico Nacional – Centro de Investigacion en Computacion, Mexico, D.F., Mexico, 116 p.

Lefrançois, M. et Gandon, F. (2011). ILexicOn: Toward an ECD-compliant interlingual lexical ontology dscribed with semantic web formalisms. In *5th International Conference on Meaning-Text Theory (MTT '11)*, September 2011, pp. 155-164, Barcelona, Spain.

Lux-Pogodalla, V. and Polguère, A. (2011). Construction of a French lexical network: Methodological issues. In *Proceedings of the First International Workshop on Lexical Resources, WoLeR'11*. An ESSLLI 2011 Workshop, pp. 54-61, Ljubljana, Slovenia.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Lang Resources & Evaluation (2012)* 46:701–719.

Mel'čuk, I. (2004). Actants in semantics and syntax, I: Actants in semantics. *Linguistics* 42(1):1-66.

Mel'čuk, I. (1998). Collocations and lexical functions - *A.P. Cowie (Ed.), Phraseology. Theory, Analysis and Applications*, Oxford: Clarendon Press, pp. 23-53.

Mel'čuk, I. (1997). Vers une linguistique Sens-Texte. Leçon inaugurale. Paris: Collège de France, 78 p.

Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in the lexicon. In *L. Wanner (ed.): Lexical Functions in Lexicography and Natural Language Processing*, pp. 37-102, Amsterdam/Philadelphia: Benjamins.

Mel'čuk, I., Clas, A. et Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Coll. Champs linguistiques/Universités francophones, Louvain-la-Neuve/Paris, Éditions Duculot/AUPELF-UREF, 256 p.

Mel'čuk, I. (1992). Paraphrase et lexique: La théorie Sens-Texte et le *Dictionnaire explicatif et combinatoire* in Mel'cuk *et al*. 1992: 9-58.

Saussure, F. de (1983). Course in general linguistics. Eds. Charles Bally and Albert Sechehaye. Trans. Roy Harris. La Salle, Illinois: Open Court. 1983, 236p.

Schwab, D., Tze, L. L. et Lafourcade, M. (2007). Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. TALN'07: Traitement Automatique des Langues Naturelles, pp. 293-302, Jun 2007, Toulouse, France, ATALA.

Dutch Hypernym Detection: Does Decompounding Help?

Ayla Rigouts Terryn, Lieve Macken and Els Lefever
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
ayla.rigoutsterryn,lieve.macken,els.lefever@ugent.be

Abstract

This research presents experiments carried out to improve the precision and recall of Dutch hypernym detection. To do so, we applied a data-driven semantic relation finder that starts from a list of automatically extracted domain-specific terms from technical corpora, and generates a list of hypernym relations between these terms. As Dutch technical terms often consist of compounds written in one orthographic unit, we investigated the impact of a decompounding module on the performance of the hypernym detection system. In addition, we also improved the precision of the system by designing filters taking into account statistical and linguistic information. The experimental results show that both the precision and recall of the hypernym detection system improved, and that the decompounding module is especially effective for hypernym detection in Dutch.

Keywords: terminology, hypernym detection, decompounding

1. Introduction

Structured lexical resources have been proven essential for different language technology applications such as efficient information retrieval, word sense disambiguation or coreference resolution. Also from a business perspective, ontologies and user-specific taxonomies appear to be very useful (Azevedo et al., 2015). Companies like to have their own mono- or multilingual enterprise semantic resources, but manual creation of structured lexical resources appears to be a very cumbersome and expensive task. Therefore, researchers have started to investigate how terminological and semantically structured resources such as ontologies or taxonomies can be automatically constructed from text (Biemann, 2005).

This paper focuses on the task of automatic hypernym detection from text, which can be considered the central task of automatic taxonomy construction. Detecting hyponym–hypernym tuples consists in finding subtype-supertype relations between terms of a given domain of interest. Different approaches have been proposed to automatically detect these hierarchical relations between terms.

Pattern-based approaches, which are inspired by the work of Hearst (1992), deploy a list of lexico-syntactic patterns able to identify hypernym pairs in text. An example of these manually defined Hearst patterns is "*NP {, NP}* {,} or other NP*", as in "Bruises, wounds, broken bones or other injuries", which results in three hypernym pairs: (*bruise, injury*), (*wound, injury*) and (*broken bone, injury*). The lexico-syntactic approach of Hearst has been applied and further extended for English (Pantel and Ravichandran, 2004) and various other languages such as Romanian (Mititelu, 2008), French (Malaisé et al., 2004) and Dutch (Lefever et al., 2014a). Researchers have defined these lexico-syntactic patterns manually (Kozareva et al., 2008), but also statistical and machine learning techniques have been deployed to automatically extract these patterns and to train hypernym classifiers (Ritter et al., 2009).

Other researchers have applied **distributional approaches** to automatically find hypernym pairs in text (Caraballo, 1999; Van der Plas and Bouma, 2005). Distributional approaches start from the assumption that semantically related words tend to occur in similar contexts. The hypernym detection task is then approached as a clustering task, where semantically similar words are clustered together and the hierarchical structure of the clustering is used to express the direction of the hypernym-hyponym relation. An extension of this approach is the distributional inclusion hypothesis, which has been the inspiration to use directional similarity measures to detect hypernym relations between terms (Lenci and Benotto, 2012). More recently, the potential of word embeddings, which are word representations computed using neural networks, has been investigated to predict hypernyms (Fu et al., 2014; Rei and Briscoe, 2014).

The morphological structure of terms has also been used as an information source to extract hypernym–hyponym relations from compound terms (Tjong Kim Sang et al., 2011). These **morpho-syntactic approaches** have shown to be successful for technical texts, where a large number of the domain specific terms appear to be compounds (Lefever et al., 2014b). The latter approaches start from the assumption that the full compound can be considered as the hyponym, whereas the head term is then to be considered as the hypernym term of the compound. Other approaches use heuristics to extract hypernym relations from **structured (collaborative) resources** such as Wikipedia. Ponzetto and Strube (2011) use methods based on connectivity in the Wikipedia network and lexico-syntactic patterns to automatically assign hypernym labels to the relations between Wikipedia categories. Navigli and Velardi (2010) use word class lattices, or directed acyclic graphs, to develop a pattern generalization algorithm trained on a manually annotated training set, which is able to extract definitions and hypernyms from web documents.

In this paper we propose a domain-independent approach to automatically detect hyponym–hypernym relations between Dutch domain specific terms. To find hierarchical relations between terms, we propose a data-driven approach combining a lexico-syntactic pattern-based module, a morpho-syntactic analyzer and a decompounding module. Given the very productive compounding system in Dutch, we expect to improve the recall of the hypernym detection system by decompounding all Dutch terms. In addition, we

also investigate the impact of filtering the relations, based on the results of automatic terminology extraction.
The remainder of this paper is structured as follows. In Section 2., we describe the annotated data sets we constructed. Section 3. presents our system, which combines fully automatic term extraction with a data-driven hypernym detection approach. In Section 4., the experimental results are discussed, while Section 5. formulates some conclusions and ideas for future research.

2. Data Set Construction

In order to evaluate the impact of filtering and decompounding on the hypernym detection performance, we constructed a development and a test corpus. The development corpus consists of manually annotated, highly specialized texts for the dredging and financial domains in Dutch. Dredging texts are annual reports from a Belgian dredging company, whereas the financial texts are news articles from the business newspaper De Tijd. For the test corpus, we used a technical manual for mobile screens. The development corpus was used to tune the linguistic and statistical filtering of the terminology extraction output (See Section 4.2.). Both the development and test corpus were manually annotated with BRAT (Stenetorp et al., 2012). This web-based tool was used to annotate all terms and named entities and the hypernym relations between them. Figure 1 shows an example sentence in which 5 terms: *persoonlijke beschermingsmiddelen* (English: personal protective equipment), *beschermingsmiddelen* (English: protective equipment), *veiligheidsbril* (English: safety goggles), *bril* (English: glasses) and *handschoenen* (English: gloves) and the hypernym relations between those terms were annotated. The annotation results were then exported and processed into a gold standard. The manual annotation for the development and test corpus allowed us to measure both precision and recall. Another advantage was that the evaluation did not have to rely on general-purpose inventories, such as WordNet, and could therefore also accurately evaluate specialized terms which do not occur in lexical inventories. Table 1 gives an overview of the number of hypernym relations in the development en test corpora, which contain each around 10,000 tokens.

Gold Standard	# Relations in gold standard
Dredging (Development)	822
Financial (Development)	364
Technical manual (Test)	480

Table 1: Gold Standard relations per data set.

3. System Description

The hypernym detection system starts from a raw domain-specific corpus that is first linguistically preprocessed by means of the LeTs Preprocess toolkit (Van de Kauter et al., 2013), which performs tokenization, Part-of-Speech tagging, lemmatization and chunking. The preprocessed corpus is then the input for both the terminology extractor and the different modules of the hypernym detection system.

3.1. Terminology Extraction

In order to automatically extract domain specific terms from our corpus, we applied the TExSIS terminology extraction system (Macken et al., 2013). TExSIS is a hybrid system, which first generates syntactically valid candidate terms and then applies statistical filtering (termhood as implemented by Vintar (2010), log-likelihood, etc.), resulting in a list of domain specific single and multiword terms. Examples of terms extracted by TExSIS are 'rupsbanden' (caterpillar tracks) and 'brandstofinjectiepomp' (fuel injection pump). The resulting term lists are then used to filter the results of the hypernym detection. For example, the hypernym detection system might discover that 'language' is a hypernym of 'English'. Although this is correct, it is likely that for specialized technical texts, the user is not interested in this particular relation and wants to focus on terms that are relevant to the field, such as 'iron ore' as a hyponym of 'primary raw materials'.
Apart from this original filtering, analysing the results of the development corpus revealed some additional patterns that could be used to further tune the terminology extraction to the hypernym detection and improve the precision without hurting the recall. The first correlation we discovered between the extracted terms and the terms in the development gold standard, was the termhood score. The higher the termhood score of the extracted terms, the more likely the term would appear in our gold standard. Figure 2 shows the percentage of terms that were in the gold standard of the development corpus out of all the terms with a termhood score within a certain range.
Another correlation was discovered in the Part-of-Speech tags of the terms. Terms in the gold standard had less diverse PoS tags than terms that were not in the gold standard and were mostly restricted to less 'complicated' PoS categories, such as a single noun or an adjective-noun combination (see figure 3 and 4). More complex PoS categories such as 'Noun Preposition Determiner Adjective Noun' were never found in the gold standard. Based on this information, we experimented with different filters on the development corpus to find the ones that discarded the most irrelevant terms, without rejecting terms present in the gold standard.

3.2. Hypernym Detection

For the automatic extraction of Dutch hypernym relations, we combined the lexico-syntactic pattern-based approach and morpho-syntactic analyzer of Lefever et al. (2014a) with a newly developed hypernym detection module integrating decompounding information. The current system takes as input a list of automatically extracted terms and a linguistically preprocessed corpus, and generates a list of hyponym–hypernym tuples.

3.2.1. Pattern-based Module

The first hypernym detection module is a pattern-based module. The patterns are implemented as a list of regular expressions containing lexicalized strings (e.g. *like*), isolated Part-of-Speech tags (e.g. *Noun*) and chunk tags, which consist of sequences of Part-of-Speech tags (e.g. *Noun Phrase*). For a complete list of Dutch lexico-syntactic

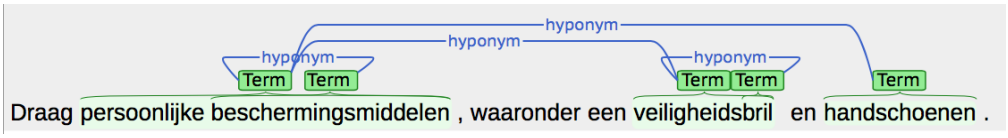

Figure 1: Annotation of terms and hypernym relations in BRAT.

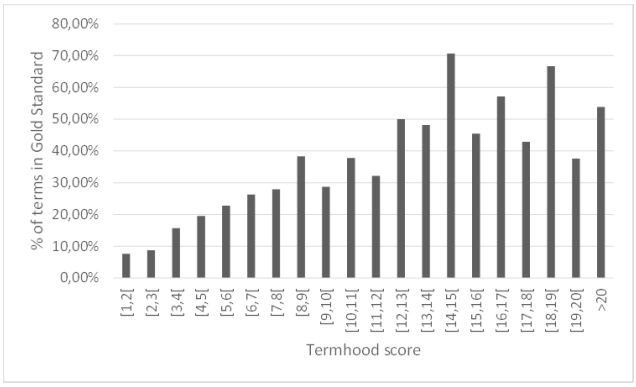

Figure 2: Likelihood that a term is in the gold standard of the development corpus.

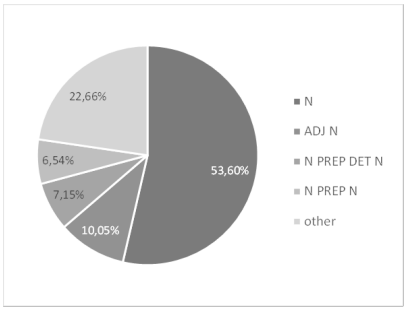

Figure 3: PoS sequences of terms not appearing in the hypernym gold standard of the development corpus (N: noun, ADJ: adjective, PREP: preposition, DET: determiner).

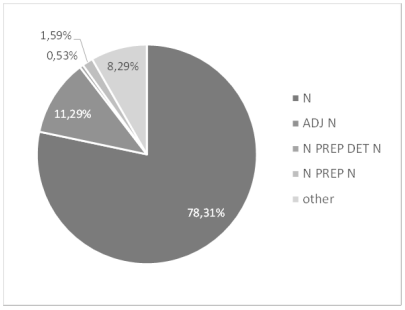

Figure 4: Part-of-Speech sequences of terms appearing in the hypernym gold standard of the development corpus.

patterns, we refer to Lefever et al. (2014a). An example of a Dutch pattern is *NP, zoals NP {,NP}* {(en/of) NP}*, matching the test sentence *veiligheidsfuncties, zoals noodstopknoppen en beschermknoppen* (English: *safety features such as emergency stop buttons and safety guards*) and resulting in the hypernym-hyponym tuples *(veiligheidsfuncties, noodstopknoppen)* and *(veiligheidsfuncties, beschermkappen)*.

3.2.2. Morpho-syntactic Module

The second module starts from the automatically generated term list to generate hypernym–hyponym tuples based on the morpho-syntactic structure of the terms. This approach is inspired by the head-modifier principle of Sparck Jones (1979), which states that the head of the compound refers to a more general semantic category, while the modifying part narrows down the meaning of the compound term. Three different morpho-syntactic rules were implemented:

single-word noun phrases. If Term1 is a suffix string of Term2, Term1 is considered as a hypernym of Term2. Example: *(mat, deurmat)* (English: *(mat, doormat)*).

multi-word noun phrases. If Term1 is the head term of Term2, Term1 is considered to be a hypernym of Term2. As the head of a noun phrase (NP) appears at the right edge of a multiword in Dutch and English, the last constituent of the NP is regarded as the head term. Example: *(afstandsbediening, draadloze afstandsbediening)* (English: *(remote control, wireless remote control)*).

noun phrase + prepositional phrase. If Term1 is the first part of Term2 containing a noun phrase + preposition + noun phrase, Term1 is considered as the hypernym of Term2. In Dutch, the head of a prepositional compound phrase is situated at the left edge of the compound term. Example: *(Raad, Raad van State)* (English: *(Council, Council of State)*).

A qualitative analysis revealed that the morpho-syntactic approach overgenerates because it has no information on the validity of the remaining part of the compound. As an example, we can refer to *soil*, which contains the term *oil*, but the remaining part *-s* is not a valid lexeme. In addition, the morpho-syntactic approach is also constrained by the occurrence of both the hypernym and hyponym term in the automatically extracted term list. To overcome both issues, we implemented a third module that takes into account decompounding information for all domain specific terms.

3.2.3. Decompounding Module

As already mentioned, the right-hand part of a compound in Dutch is the head and determines the meaning of the compound, e.g. *bediening+s+knoppen* (English: control buttons), are a special type of *knoppen* (English: buttons). This information can be used to find hypernyms, because the head of the compound is mostly also the hypernym of the compound, e.g. *knoppen* is a hypernym of *bedieningsknoppen*. Compounds can be nested and especially in technical texts, compounds of more than two components frequently occur, such as *[nood+stop]+knoppen* (English: [emergency stop] buttons). To add decompounding

# relations in gold standard = 480	Patterns	Morphosynt.	Patterns and Morphosynt.	Decompounder	All modules combined	Filter 1	Filter 2
Relations found	29	895	923	317	1091	970	961
Correct relations found	8	259	266	230	410	409	409
Precision	0.275862	0.289385	0.288191	0.725552	0.375802	0.421649	**0.425598**
Recall	0.016667	0.539583	0.554167	0.479167	**0.854167**	0.852083	0.852083

Table 2: Results for the test corpus.

information to our term list, we used the compound splitter of Macken and Teczan (in press), which is a hybrid compound splitter for Dutch that makes use of corpus frequency information and linguistic knowledge. The compound splitting tool determines a list of eligible compound constituents on the basis of word frequency information derived from a PoS-tagged Wikipedia dump of 150 million words extended with a smaller dynamically compiled frequency list derived from the extraction corpus. Part-of-speech information is used to restrict the list of possible constituents. The compound splitter selects the split point with the highest geometric mean of word frequencies of its parts (Koehn and Knight, 2003): $(\prod_{i=1}^{n} freq_p)^{1/n}$ in which *n* is the number of split points in the compound and $freq_p$ is the frequency of the component parts. The compound splitter allows a linking-s between two component parts and can be called recursively to deal with with nested compounds.
A qualitative analysis of the results revealed that this module may overgenerate as well, for example with words such as *hand+schoenen* (English: gloves, literal translation: *hand+shoes*). In this case, the decompounding module might wrongly say that *shoes* is a hypernym of *gloves*. Despite these exceptions, experiments with the development corpus showed that the precision of the decompounding module was still higher than that of the pattern-based and morphosyntactic modules for Dutch. In addition, it also deals with some of the shortcomings of the morphosyntactic module as it limits the number of split points, can correctly process compounds with a linking-s and can be called recursively for nested compounds.

4. Experimental Results

4.1. Improving Recall

To find more correct hypernym relations, we expanded the pattern-based and morphosyntactic modules with the decompounding module. Since compound terms are very common in Dutch, especially in technical texts such as the ones used for this experiment, this led to a significant increase in recall. The pattern-based and morphosyntactic modules combined were able to find 266 out of the 480 relations in the gold standard. The decompounding module on its own already found 230 correct relations and all three modules combined achieved a score of 410 correctly identified relations out of the 480 relations in the gold standard. This means a recall of 85%, an increase of 30% over the original system.

4.2. Improving Precision

The decompounding module did not overgenerate much and got a high recall, which already increased the precision of the combined system with 9%. However, by using additional filters based on the terminology extraction, the precision could be further improved. We implemented 2 different filters. The first filter was based purely on the PoS tags of the extracted terms: all terms which were assigned a PoS pattern that was not in the list created on the basis of the experiments carried out for the development corpus, were automatically filtered out. The list we used was: N, ADJ N, N N, N N N, N N N N, N PREP N, N CONJ-coord N, N PREP DET N, N PREP ADJ N. This filter works well, but may in some cases be too strict and discard some terms which are still relevant. That is why we developed an alternative filter, which filtered out all the same terms as the first filter, except if they had a termhood score of more than 10. Even though, in the case of our test corpus, the second filter made little difference, it may be a good safety precaution when recall is more important than precision. These filters can easily be adapted to focus more on precision or recall by adding or deleting certain PoS categories and by choosing a higher or lower minimum termhood score for the second filter.

5. Conclusions and Future Work

We presented proof-of-concept experiments for a data-driven hypernym detection system for Dutch, which starts from an automatically generated term list and combines a pattern-based module, a morpho-syntactic analyzer and a decompounding module. Both the precision and recall of the extracted hypernym relations clearly improved by adding the decompounding module. Precision was further improved by filtering the results of the terminology extraction based on Part-of-Speech and termhood score.
In future work, we will investigate whether our methodology, especially the decompounding module, works equally well on other languages with many compounds, such as German. In addition, we will work towards hypernym detection in multiple levels. For example, if terrier is a hyponym of dog, and dog is a hyponym of animal, then terrier is also a hyponym of animal. Ultimately, this can result in hypernym trees, which can be used to structure terminology databases. Finally, during the annotation, we came upon the problem of split compound terms, such as "gezondheid- en veiligheidsbescherming" (health and safety protection) or split multiword terms such as "elektrische en statische vonken" (electric and static sparks). It was difficult or even impossible to correctly annotate these terms with BRAT and our hypernym detection system cannot process these terms yet either. Nevertheless, this syntax is not uncommon and systems for annotation, automatic terminology extraction and hypernym detection could benefit from being able to process these complex terms and relations.

6. Bibliographical References

Azevedo, C., Iacob, M., Almeida, J., van Sinderen, M., Pires, F., L., and Guizzardi, G. (2015). Modeling Resources and Capabilities in Enterprise Architecture: A Well-founded Ontology-based Proposal for ArchiMate. *Information Systems*, 54:235–262.

Biemann, C. (2005). Ontology Learning from Text: A Survey of Methods. *LDV Forum*, 20(2):75–93.

Caraballo, S. (1999). Automatic Acquisition of a Hypernym-labeled Noun Hierarchy from Text. In *Proceedings of ACL-99*, pages 120–126, Baltimore, MD.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1199–1209, Baltimore, Maryland, USA.

Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the International Conference on Computational Linguistics*, pages 539–545.

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 187–193, Budapest, Hungary.

Kozareva, Z., Riloff, E., and Hovy, E. (2008). Semantic Class Learning from the Web with Hyponym Pattern Linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1048–1056, Columbus, Ohio, USA.

Lefever, E., Van de Kauter, M., and Hoste, V. (2014a). HypoTerm: Detection of Hypernym Relations between Domain-specific Terms in Dutch and English. *Terminology*, 20(2):250–278.

Lefever, E., Van de Kauter, M., and Hoste, V. (2014b). Evaluation of Automatic Hypernym Extraction from Technical Corpora in English and Dutch. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 490–497, Reykjavik, Iceland.

Lenci, A. and Benotto, G. (2012). Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of the first Joint conference on Lexical and Computational Semantics (*SEM)*, pages 75–79, Montréal, Canada.

Macken, L. and Tezcan, A. (in press). Dutch Compound Splitting for Bilingual Terminology Extraction. In Ruslan Mitkov et al., editor, *Multi-word Units in Machine Translation and Translation Technology*. John Benjamins.

Macken, L., Lefever, E., and Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora using Chunk-based Alignment. *Terminology*, 19(1):1–30.

Malaisé, V., Zweigenbaum, P., and Bachimont, B. (2004). Detecting Semantic Relations between Terms in Definitions. In *the CompuTerm workshop 2004: 3rd International Workshop on Computational Terminology*, pages 55–62.

Mititelu, V. (2008). Hyponymy Patterns. Semi-automatic Extraction, Evaluation and Inter-lingual Comparison. *Text, Speech and Dialogue: Lecture Notes in Computer Science*, 5246:37–44.

Navigli, R. and Ponzetto, S. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Pantel, P. and Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of HLT/NAACL-04*, pages 321–328, Boston, MA.

Ponzetto, S. and Strube, M. (2011). Taxonomy Induction based on a Collaborative Built Knowledge Repository. *Artificial Intelligence*, 175:1737–1756.

Rei, M. and Briscoe, T. (2014). Looking for Hyponyms in Vector Space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 68–77, Baltimore, Maryland, USA.

Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In *Proceedings of Association for Advancement of Artificial Intelligence Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93.

Sparck Jones, K. (1979). Experiments in Relevance Weighting of Search Terms. *Information Processing and Management*, 15:133–144.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 102–107, Avignon, France.

Tjong Kim Sang, E., Hofmann, K., and de Rijke, M. (2011). Extraction of Hypernymy Information from Text. In *Interactive Multi-modal Question-Answering. Series: Theory and Applications of Natural Language Processing*, pages 223–245. Springer-Verlag Berlin Heidelberg.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., and Hoste, V. (2013). LeTs Preprocess: The Multilingual LT3 Linguistic Preprocessing Toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Van der Plas, L. and Bouma, G. (2005). Automatic Acquisition of Lexico-semantic Knowledge for Question Answering. In *Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources*, Jeju Island, Korea.

Vintar, S. (2010). Bilingual Term Recognition Revisited. The Bag-of-equivalents Term Alignment Approach. *Terminology*, 16(2):141–158.