Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 115-118 recibido 01-12-12 revisado 21-01-13 aceptado 19-02-13

Multilingual Acquisition of Large Scale Knowledge Resources

Adquisición multilingüe de bases de conocimiento de gran escala

Montse Cuadros
Vicomtech-IK4
Mikeltegi Paselekua, 57
2009 Donostia-San Sebastián
mcuadros@vicomtech.org

Resumen: Tesis doctoral en Informática realizada por Montse Cuadros y dirigida por Dr. Lluís Padró y Dr. German Rigau. La defensa de la tesis fue en la facultad de Informática de la Universitat Politècnica de Catalunya el día 22 de noviembre de 2011. El tribunal estuvo formado por Dr. Horacio Rodríguez (Universitat Politècnica de Catalunya), Prof. Dr. Piek Vossen (Vrije Universiteit Amsterdam), Dra. Arantza Díaz de Ilarraza (Universidad del Pais Vasco), Dra. Irene Castellón (Universitat de Barcelona) y Dr. Roberto Navigli (Sapienza University of Rome), que le otorgaron la nota de Sobresaliente Cum Laude.
Palabras clave: Adquisición de conocimiento, adquisición de léxico, evaluación de recursos, WordNet, desambiguación de acepciones

Abstract: Ph. D. thesis in Computer Science written by Montse Cuadros under the supervision of Dr. Lluís Padró and Dr. German Rigau. The thesis defense was done on 22th November 2011 at the Computer Science Faculty of the Universitat Politècnica de Catalunya. The Doctoral Examination Committee was composed by Dr. Horacio Rodríguez (Universitat Politecnica de Catalunya), Prof. Dr. Piek Vossen (Vrije Universiteit Amsterdam), Dra. Arantza Diaz de Ilarraza (Universidad del Pais Vasco), Dra. Irene Castellón (Universitat de Barcelona) and Dr. Roberto Navigli (Sapienza University of Rome). The thesis was graded Cum Laude.
Keywords: Knowledge acquisition, lexical acquisition, resource evaluation, WordNet, word sense disambiguation

1 Introduction

The use of large-scale semantic resources, such as WordNet, has become a usual, often necessary, practice for most current NLP systems. Princeton WordNet (WN) is by far the most widely-used semantic resource in NLP.

However, even manually, the construction of large-scale semantic repositories for broad-coverage NLP is not a trivial task. It is quite difficult to acquire and consistently integrate large amounts of knowledge into an existing resource. The construction of large and rich knowledge bases for broad-coverage semantic processing takes a great deal of expensive manual effort and incurs large development costs. It involves large research groups for long periods of development. For instance, hundreds of person-years have been invested in the development of wordnets for various languages. In the case of the English WordNet, in more than ten years of manual construction (from 1995 to 2006, that is, from version 1.5 to 3.0), WordNet grew from 103,445 to 235,402 semantic relations[1], which represents a growth of around one thousand new relations per month. However, in 2008, the Princeton group released a new resource with 458,825 word forms from the WordNet definitions manually linked to its appropriate WordNet sense[2].

Furthermore, the relevant knowledge changes across domains and cultures and it has to be steadily kept up to date. New knowledge emerges day by day everywhere and has to be combined with the existing knowledge. For these reasons, knowledge acquisition is still a highly active area of research since the existing knowledge repositories do not

[1] Symmetric relations are counted only once.
[2] http://wordnet.princeton.edu/glosstag.shtml

ISSN 1135-5948

seem to be rich enough to support advanced concept-based NLP applications directly. It seems that such applications require more detailed general-purpose (and also domain-specific) semantic knowledge, which have to be built by automatic means to keep development cost and time inside affordable limits. Obviously, this fact has severely hampered the state-of-the-art of advanced NLP applications.

Thus, the automatic acquisition of the necessary knowledge from available resources, such as naturally occurring text, is one of the most challenging tasks in NLP since it requires some *knowledge understanding* capabilities, which is our final goal. This vicious circle is known as the *acquisition bottleneck*. The intrinsic cycling nature of the problem also suggests a cycling approach for solving it, with incremental iterations of *acquisition–identification* stages. Ideally, the process would start with a *minimal* knowledge base and the relevant resources containing the *implicit* knowledge to be acquired. Then, the automatic acquisition process might produce new content that should be *identified* with respect to the existing knowledge base. This identification process is necessary in order to facilitate the integration of the new knowledge into the existing one, to form a comprehensive and computationally useful knowledge base. Arguably, although these sub-tasks are undeniably difficult, combining them might simplify both.

Figure 1 shows the senses of *party* in WordNet 3.0[3]. From left to right the figure shows the senses, the total number of explicit semantic relations encoded for each synset, the new semantic relations gathered from the semantically tagged WordNet definitions[4] and the gloss. Consider the subtle distinctions among some of them. The first three senses are groups of people and the fourth refers to an entertaining event. Obviously, these senses are defining different aspects of related concepts. This is a major drawback when trying to acquire specific knowledge for each sense.

Hopefully, the semantic relations encoded for each sense can help its proper characterization. For instance, Figure 2 shows some of

Sense	#rel.	#gloss	Gloss
party^1_n	36	114	an organization to gain political power
party^2_n	18	27	a group of people gathered together for pleasure
party^3_n	9	41	a band of people associated temporarily in some activity
party^4_n	13	38	an occasion on which people can assemble for social interaction and entertainment
party^5_n	3	87	a person involved in legal proceedings:

Figura 1: Number of relations for party$_n$ in WordNet 3.0

the related concepts encoded in WordNet[5]. Additionally, this table also presents some of the relations captured by KnowNet (KN), a very large lexical knowledge base which has been derived during our research.

Sense	relation	Sense
party^1_n	hypernym	organization^1_n, organisation^1_n
	member-holonym	political_system^1_n, form_of_government^1_n
	hyponym	American_Labour_Party^1_n
	rgloss	machine^5_n, political_machine^1_n
	related-to	**election**^1_n, **political**^1_a,**vote**^1_v, **elect**^1_v
party^2_n	hypernym	social_gathering^1_n, social_affair^1_n
	hyponym	shindig^1_n, shindy^1_n
	hyponym	dinner^1_n, dinner_party^1_n
	hyponym	wedding^3_n, wedding_party^1_n
	related-form	party^1_v
	rgloss	carouse^1_n carousal^1_n bender^2_n toot^2_n, booze-up^1_n
	related-to	**invitation**^1_n,**ceremonial**^1_a,**cocktail**^1_n, **farewell**^2_n
party^3_n	hypernym	set^5_n, circle^2_n, band^1_n, lot^1_n
	hyponym	rescue_party^1_n
	rgloss	fairly^2_r, fair^2_r evenhandedly^1_r
	related-to	**camp**^4_n,**landing**^2_n, **stretcher**^3_n, **Olympiad**^2_n
party^4_n	hypernym	affair^3_n, occasion^2_n, social_occasion^1_n, function^6_n, social_function^1_n
	hyponym	birthday_party1^1_n
	hyponym	cocktail_party1^1_n
	related-form	party^1_v
	rgloss	party_game^1_n
	related-to	**nuptials**^1_n,**prom**^1_n,**reception**^2_n, **gift**^1_n
party^5_n	hypernym	person^1_n, individual^1_n, someone^1_n, somebody^2_n, mortal^1_n, soul^2_n
	hyponym	assignee^1_n
	domain	law^1_n, jurisprudence^2_n
	rgloss	submission^5_n
	related-to	**accountancy**^1_n, **appearance**^3_n, **attendance**^1_n, **court**^1_n

Figura 2: Some relations for party$_n$ in WordNet 3.0 and KnowNet(in bold)

1.1 Research goals

The main goal of the research presented in this thesis is to devise new methods and tools

[3] word$^{num}_{pos}$, where pos is the part-of-speech (n for nouns, v for verbs, a for adjectives and r for adverbs)

[4] That is, the number of glosses that include that particular sense annotated in its definition

[5] *rgloss* stands for reverse gloss. That is, the corresponding sense of party appears in its gloss. These relations are gathered from the manually sense-disambiguated glosses

for creating automatically new semantic relations between WordNet senses. That is, to accurately increase by automatic means the knowledge represented in WordNet.

In particular, our research requires the construction of new methods and tools for:

1. Acquiring relevant words from general or domain corpora for an specific WordNet word sense.
2. Identifying the *implicit* word senses of the acquired relevant words with respect to an *existing* knowledge base (in particular, WordNet).
3. Evaluating empirically the quality of the resulting *new* semantic relations in a controlled multilingual evaluation framework.

2 Thesis overview

The thesis is organised in seven chapters:

- **Chapter 1: Introduction**
  This chapter presents an overview of the thesis. It revises the motivation and presents the main contributions of the thesis to the state-of-the-art.
- **Chapter 2: State of the Art**
  This chapter reviews the state of the art. It revises the use of wide-coverage *semantic resources* in different NLP tasks. Furthermore, it presents the main methodologies, approaches and techniques used for *building large-scale knowledge resources* in general, manually and automatically. Finally, it overviews the existing *evaluation frameworks* used in the research field to assess the quality of the acquired knowledge.
- **Chapter 3: Knowledge Acquisition Method**
  This chapter describes the knowledge acquisition architecture developed in this research.
- **Chapter 4: Acquisition of topic signatures**
  This chapter reviews the different methods applied to acquire automatically topic signatures as well as the methodology for evaluating their quality.
- **Chapter 5: KnowNet**
  This chapter depicts the KnowNet building process and its grounding Word Sense Disambiguation algorithm, used to obtain word-sense relations from topic signatures acquired from general corpora.
- **Chapter 6: deepKnownet**
  This chapter explores a new method for building KnowNets, named deepKnowNets. Basically, instead of a Word Sense Disambiguation algorithm, the method exploits a graph-based similarity measure to rerank the topic signatures.
- **Chapter 7: Concluding remarks and future directions**
  This chapter draws the main conclusions of this thesis and outlines some further steps to follow.

3 Main Contributions

The knowledge acquisition bottleneck problem is particularly acute for open domain (and also domain specific) semantic processing. However, we acquired by fully automatic means highly connected knowledge bases, increasing the total number of semantic relations from less than one million (the current number of available relations in WordNet) to millions of new and accurate semantic relations between WordNet senses. The different versions of KnowNet seem to be a major step towards the autonomous acquisition of knowledge from text, since they are several times larger than the available knowledge resources which encode relations between WordNet senses, and the knowledge they contain outperforms any other resources when they are empirically evaluated in a common framework.

Firstly, in order to acquire relevant semantic relations from large text collections corresponding to general or particular domains, we apply several methodologies and settings to automatically acquire **topic signatures** (TS) (Cuadros, Padró, and Rigau, 2005; Cuadros, Padró, and Rigau, 2006). Originally, topic signatures were used to describe a set of words related to the same topic or domain, but in our case, the topic is a WordNet sense[6]. Thus, topic signatures are sets of words related to that particular WordNet sense. We

[6]The name of Topic Signature, instead of the more appropriate *concept signature*, *word sense signature* or *synset signature*, is maintained for consistency with the literature

use in this research the original topic signatures acquired from the web [7] together with new sets of automatically acquired topic signatures which result in new acquisition methods, new tools and different resources, including different types of corpora. All these topic signatures are compared in a common framework together with existing knowledge bases. **ExRetriever**[8] (Cuadros et al., 2004; Cuadros et al., 2005) is used for the automatic acquisition of examples of particular WordNet word senses. It is a tool to automatically extract a subcorpus of text examples from a large corpus (for instance, BNC, SemCor or the Web).

Secondly, in order to identify the *implicit* semantic relations encoded by a Topic Signature (the sets of words related to a particular WordNet sense) with respect an *existing* knowledge base (in this case, also WordNet), we apply a graph-based Word Sense Disambiguation (WSD) algorithm, SSI-Dijkstra (Cuadros and Rigau, 2008b), also developed in the framework of this thesis. SSI-Dijkstra is based on the Structural Semantic Interconnections (SSI) algorithm. The method uses a robust and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to semantically related words associated to a particular WordNet sense. In that way, the method identifies a particular WordNet sense for each word in the Topic Signature, converting the original list of concept-to-word relations into a list of concept-to-concept relations.

Thirdly, a variant of SSI-Dijkstra has been applied in a task to integrate a very large domain thesaurus with millions of Species into WordNet (Toral et al., 2010; Cuadros et al., 2010). The process disambiguate every taxonomy of species in several languages.

Finally, the full list of new concept-to-concept relations between WordNet senses forms new knowledge bases, which we call **KnowNet**[9] (Cuadros and Rigau, 2008b) and **deepKnowNet**. Different sets of new KnowNets are empirically evaluated in different evaluation frameworks (Cuadros and Rigau, 2008b; Cuadros and Rigau, 2008c; Cuadros and Rigau, 2008a; Cuadros and Rigau, 2008d; Agirre et al., 2010).

[7] http://ixa.si.ehu.es/Ixa/resources/sensecorpus

[8] http://www.lsi.upc.edu/ nlp/meaning/downloads.html

[9] http://adimen.si.ehu.es/web/KnowNet

References

Agirre, Eneko, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring Knowledge Bases for Similarity. In *Proceedings of LREC 2010. ISBN: 2-9517408-6-7. Pages 373–377.*

Cuadros, Montse, Jordi Atserias, Mauro Castillo, and German Rigau. 2004. Automatic Acquisition of Sense Examples Using Exretriever. In *Proceedings of IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*, pages 97–104, November.

Cuadros, Montse, Jordi Atserias, Mauro Castillo, and German Rigau. 2005. The MEANING approach for automatic acquisition of sense examples. *MEANING Workshop*, February.

Cuadros, Montse, Egoitz Laparra, German Rigau, Piek Vossen, and Wauter Bosma. 2010. Integrating a large domain ontology of species into WordNet. In *Proceedings of LREC 2010*, La Valleta, Malta.

Cuadros, Montse, Lluís Padró, and German Rigau. 2005. Comparing Methods for Automatic Acquisition of Topic Signatures. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'05)*, September.

Cuadros, Montse, Lluís Padró, and German Rigau. 2006. An Empirical Study for Automatic Acquisition of Topic Signatures. In *Proceedings of Third International WordNet Conference (GWC 06)*, pages 51–59, Jeju Island (Korea), January. ISBN 80-210-3915-9.

Cuadros, Montse and German Rigau. 2008a. *Bases de Conocimiento Multilíng ües para el Procesamiento Semántico a Gran Escala.* Procesamiento del Lenguaje Natural, Vol. 40, 35–42.

Cuadros, Montse and German Rigau. 2008b. KnowNet: A proposal for building highly connected and dense knowledge bases from the web. In *First Symposium on Semantics in Systems for Text Processing, STEP'08.*, Venice, Italy, September.

Cuadros, Montse and German Rigau. 2008c. KnowNet: using Topic Signatures acquired from the web for building automatically highly dense knowledge bases. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, August.

Cuadros, Montse and German Rigau. 2008d. Multilingual Evaluation of KnowNet. *Procesamiento del Lenguaje Natural*, 41.

Toral, Antonio, Monica Monachini, Claudia Soria, Montse Cuadros, German Rigau, Wauter Bosma, and Piek Vossen. 2010. Linking a domain thesaurus to WordNet and conversion to WordNet-LMF. In *Proceedings of ICGL 2010*, Hong Kong.