Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources

Spandana Gella°, Carlo Strapparava⋆, Vivi Nastase⋆
° University of Melbourne, Australia
⋆FBK-irst, Povo, Italy
sgella@student.unimelb.edu.au, strappa@fbk.eu, nastase@fbk.eu

Abstract

In this paper we present the mapping between WordNet domains and WordNet topics, and the emergent Wikipedia categories. This mapping leads to a coarse alignment between WordNet and Wikipedia, useful for producing domain-specific and multilingual corpora. Multilinguality is achieved through the cross-language links between Wikipedia categories. Research in word-sense disambiguation has shown that within a specific domain, relevant words have restricted senses. The multilingual, and comparable, domain-specific corpora we produce have the potential to enhance research in word-sense disambiguation and terminology extraction in different languages, which could enhance the performance of various NLP tasks.

Keywords: Semantic Domains, WordNet, Wikipedia

1. Introduction

Relying on a lexical resource, where words are associated with domain information in addition to senses, has been proven beneficial in many tasks, e.g. text categorization, word sense disambiguation. We use the term *domain* to denote a set of words between which there are strong semantic relations. The rationale is that domain labels, such as MEDICINE, ARCHITECTURE and SPORT, provide a useful way to establish semantic relations among word senses, which can be profitably used in many NLP tasks (Gliozzo and Strapparava, 2009). An approximation to domains are Subject Field Codes, used in Lexicography (e.g. in (Procter, 1978)) to mark technical usages of words. Although this information is useful for sense discrimination, in dictionaries it is typically used only for a small portion of the lexicon.

As they provide useful high level information, research has been carried out into adding domain labels to existing lexical resources. One successful outcome of this work is WORDNET DOMAINS (Magnini and Cavaglià, 2000), which provides a domain annotation for every synset in WORDNET. In the recent versions of Princeton WORDNET (Fellbaum, 1998), the *topic* relation has been partially introduced. It connects synsets with synsets that denote "domains", such as *medicine, architecture*.

The online encyclopedia Wikipedia contains a large amount of articles and information and it has become a valuable resource in major NLP tasks. Thus there was a growing interest for the automatic acquisition of knowledge from that resource (Suchanek et al., 2008) and because WordNet and Wikipedia complement each other in several aspects, many research efforts have been done in mapping WordNet synsets with a large number of instances from Wikipedia, e.g. (Ponzetto and Navigli, 2009; Fernando and Stevenson, 2012) and many others. While not intended as such, Wikipedia categories lend themselves naturally as domains. They have been used to generate domain-specific thesauri (Milne et al., 2006) and conceptual taxonomies (Ponzetto and Strube, 2007).

In this paper, we describe an integrated resource that maps these three domain labelings onto each other, and the comparable multilingual domain-specific corpora, built automatically by extracting the Wikipedia articles under each domain. While such a domain-labeled corpus can be built for any language, the resource we describe contains European languages that are not only present in Wikipedia, but also have a wordnet version, synset-aligned to the Princeton WORDNET, e.g. MultiWordNet (Pianta et al., 2002). First we map WordNet domain labels from WORDNET DOMAINS onto WordNet topic synsets. This mapping plus the fact that we work with synset-aligned wordnet versions, gives us multilingual domain labels. Thus we mapped the domain labels onto Wikipedia categories. The direct mapping combined with cross-language links in Wikipedia category pages provide high confidence *domain labels* $\rightarrow$ *Wikipedia category* mapping. In addition, redundant information can be used to bootstrap mappings for the (few) cases where a direct match could not be found. The process is explained in detail in Section 3.

The resource will be freely available as an extension of WORDNET DOMAINS.

2. Semantic Domains Resources

2.1. WordNet Domains

In WORDNET DOMAINS the annotation methodology was mainly manual and based on lexico-semantic criteria which take advantage of the already existing conceptual relations in WORDNET. About 200 domain labels were selected from a number of dictionaries and then structured in a taxonomy according to the Dewey Decimal Classification (DDC (Dewey, 1876)). The DDC classification was chosen as it guarantees good coverage, has historically proven its usefulness, and it is easily available. The fact that DDC classification is used to classify textual material can turn out to be an advantage: the current effort in digitizing printed textual data, already assigned DDC labels, can produce large, already categorized data for computational linguistics tasks. The annotation task in WORDNET DOMAINS consisted of inter-

Domain	#Syn	Domain	#Syn	Domain	#Syn
Factotum	36820	Biology	21281	Earth	4637
Psychology	3405	Architecture	3394	Medicine	3271
Economy	3039	Alimentation	2998	Administration	2975
Chemistry	2472	Transport	2443	Art	2365
Physics	2225	Sport	2105	Religion	2055
Linguistics	1771	Military	1491	Law	1340
History	1264	Industry	1103	Politics	1033
Play	1009	Anthropology	963	Fashion	937
Mathematics	861	Literature	822	Engineering	746
Sociology	679	Commerce	637	Pedagogy	612
Publishing	532	Tourism	511	Computer_Science	509
Telecommunication	493	Astronomy	477	Philosophy	381
Agriculture	334	Sexuality	272	Body_Care	185
Artisanship	149	Archaeology	141	Veterinary	92
Astrology	90				

Table 1: Distribution of some domains over WORDNET synsets.

preting a WordNet synset with respect to the DDC classification. The annotation methodology (Magnini and Cavaglià, 2000) was primarily manual and was based on lexicon-semantic criteria that take advantage of existing conceptual relations in WORDNET. In Table 1 we report the distribution of a sample of domains over WORDNET synsets[1]. WORDNET DOMAINS is freely available at `http://wndomains.fbk.eu/index.html`.

2.2. Princeton WordNet Topics

Starting from version 3.0, Princeton WordNet has associated *topic* information with a subset of its synsets. This topic labeling is achieved through pointers from a source synset to a target synset representing the topic, and it was developed independently from WORDNET DOMAINS. Topic information is thus seen as a new relation among synsets (**;c** Domain_of_synset or *topic*, and the inverse **-c** Member_of_this_domain). Compared to WORDNET DOMAINS where the domains are labels on the synsets, in this case the topics/domains are specifically marked synsets in the WORDNET noun hierarchy. In WORDNET 3.0 there are 440 topics/domains. We have automatically mapped these topics to corresponding WordNet domains using exact match, pattern match on synset gloss and the similarity of latent semantic vectors (built on the Brown corpus) of WordNet topic synsets and the WORDNET DOMAINS. These approaches will be discussed in detail in the full version of the paper. Examples of the mapped WordNet topic synsets to WORDNET DOMAINS is shown in Table 2. 396 of the 400 topic synsets could be mapped to wordnet domains with these techniques. We have also tried to extend these topic domains to other synsets using already existing semantic relations in WordNet such as 'hyponymy'. We have cross-checked the expansion of WordNet topic assignment using the hyponymy relation and the resulting WORDNET DOMAINS mapping using existing synset WORDNET DOMAINS mapping available which is discussed in Section 2.1. We were able to expand WordNet topic synset assignment to 5664 synsets using the hyponymy relation, with a 91.72% matching with existing synset WORDNET DOMAINS assignments.

method	topic synset	domain_label	id
exact_match	photography.n.02	Photography	13536794
LSA	forestry.n.01	Agriculture	6071934

Table 2: Mapping between WORDNET DOMAINS and Princeton WORDNET topics

For the purpose of this work, the mapping between domain labels in WORDNET DOMAINS and WordNet *topic* synsets is important because it produces multilingual domain labels. The domain labels in WORDNET DOMAINS are in English, they are mapped onto synsets in the Princeton WORDNET, and, through the aligned synsets in MultiWordNet, to all the other languages we work with.

2.3. Wikipedia Categories

While the Dewey Decimal Classification was purposefully built to organize printed material, Wikipedia categories have arisen from the contributors' natural propensity for categorization. As such they are varied and they capture various categorization criteria – e.g. BOOKS BY GENRE, BOOKS BY LANGUAGE. Compared with the 200 domains of WORDNET DOMAINS, the English Wikipedia version of May 2013 (20130503) has 1,009,577 categories. Among these varied categories we have noted that many of the domains from the Dewey Decimal Classification appear. A link between them could ensure a coarse mapping of WORDNET DOMAINS onto Wikipedia categories, and, through the cross-language links, to their counterparts in the various available languages. Through this mapping, we could associate to each WordNet domain a domain-specific corpus, gathered from the articles subsumed by the corresponding Wikipedia categories, for each of the desired languages. Cross-language links in Wikipedia are not complete, that is a category that exists in English might not

[1]Some synsets do not belong to a specific domain but rather correspond to general language and may appear in any context. Such senses are tagged in with a Factotum domain label.

WN-Domain	*Wiki Categories*						*Method of mapping*
	English	Italian	German	Portuguese	French	Spanish	
Diplomacy	Diplomacy	Diplomazia	Diplomatie	Diplomacia	Diplomatie	Diplomacia	direct mapping
Athletics	Athletics_(sport)	Atletica_leggera	Leichtathletik	Atletismo	Athlétisme	Atletismo	disambiguation
Graphic Arts	Graphics		Grafik		Arts_graphiques	Artes_gráficas	triangular mapping

Table 3: Sample WORDNET DOMAINS Mappings to Wikipedia categories.

have a mapping linked to corresponding Italian or Spanish category. To fill in these gaps we have used a methodology called *triangulation*. For example, if there is no direct mapping from People(en) to Spanish category, we can exploit that People(en) → Pessoas(pt) and Pessoas(pt) → Biografías(es). We use this indirect mapping to fill in the gaps. We have also cross verified this methodology using the full mappings, i.e. verifying if category(l1) → category(l2) and category(l2) → category(l3) then category (l1) maps to category (l3) where l1, l2 and l3 are three different languages in our set.

3. Domain Specific Resources

Domain specific resources have proved to be beneficial in tasks such as word sense disambiguation, terminology extraction, machine translation (Arcan et al., 2012), sentiment analysis (Choi et al., 2009). In this section we explain a simple methodology used to create a multilingual domain specific corpora from Wikipedia. Steps involved in corpus creation are:

- Map WORDNET DOMAINS to Wikipedia categories in English using simple string matching techniques.
- Extend the mappings above to other languages using cross-language links from the corresponding English Wikipedia category pages to other languages using Dbpedia (Bizer et al., 2009).
- Use triangulation to fill in the mappings which do not have mappings from English.
- Iteratively traverse the category hierarchy and collect clean text from all the articles that fall under the "domain" category until a corpus of a pre-specified size is acquired, or all pages are added.

We have experimented with six languages English, French, German, Italian, Portuguese and Spanish. We observed that at least 85% of WORDNET DOMAINS could be mapped onto Wikipedia categories for all the 6 languages considered. Five domains – Pure Science, Body Care, Artisanship, Factotum, Psychological Features – had no mappings in any of the 6 languages. Examples of Wikipedia category mappings with WORDNET DOMAINS are listed in Table 3. The respective URLs in Wikipedia can be derived using the following patterns:

Diplomacy(en) → http://**en**.wikipedia.org/wiki/**Category**:*Diplomacy*
Diplomazia(it) → http://**it**.wikipedia.org/wiki/**Categoria**:*Diplomazia*.

3.1. Evaluation

Corpus evaluation is not a straight-forward task. One objective evaluation that could be performed is to compare the lexica from two different corpora, for example in terms of word usage. For domain specific corpora for the same domain, similar word usage can be interpreted as similar coverage. We compared the corpora created for the domains Medicine and Finance with the multilingual corpus available from European Medicine Agency (EMEA) and European Central Bank (ECB) Corpus (Tiedemann, 2009). (Kilgarriff, 2001) has proposed a method based on χ^2 statistic to measure the similarity of two corpora. We use that χ^2 statistic over most 500 frequent words in the union of the corpora. Kilgariff's method assumes that the corpora compared are of the same size.

Language	Medicine wiki/(EMEA)	Finance wiki/ECB
English	254	158
Italian	84	38
Spanish	90	57
Portuguese	122	35
German	34	17

Table 4: Pairwise corpus similarity ($\times 10^3$) using χ^2 for EMEA/ECB and wiki corpus from same domain

Language	Medicine wiki/(ECB)	Finance wiki (EMEA)
English	290	286
Italian	89	54
Spanish	105	84
Portuguese	142	46
German	46	17

Table 5: Pairwise corpus similarity ($\times 10^3$) using χ^2 EMEA/ECB and wiki corpus from different domain

To this we have randomly sampled 1 million words from EMEA Medicine and ECB Bank corpus and compared against corresponding wiki domain specific corpora that we created. We measured the similarity of two corpora as the average pairwise χ^2 similarity between their sub-corpora. Lower χ^2 implies that the corpora are similar. We also executed the cross-domain similarity experiments to verify that χ^2 values for same domain are smaller compared to cross-domain scores. From Table 4 and Table 5, in which same domain and cross-domain corpus similarity scores are reported, we can observe that corpus similarity values of same domain are lesser when compared to cross-domain corpus similarity. This shows that the domain specific corpora created using wikipedia category hierarchy is comparable to the manually created domain specific resources.

3.2. Multilingual Domain-Specific Sentiment Lexicon

In this section we sketch the potential of having domain-specific corpora by creating a domain-specific sentiment lexicon from the raw corpora built as described in Section 3. A domain-specific sentiment lexicon could be used to boost the performance of sentiment analysis, the task of understanding the sentiment of texts such as online reviews of products and organisations. To create this resource we have used a simple bootstrapping algorithm proposed by (Weichselbraun et al., 2011). The technique is easily adaptable and applicable to any language and domain having a raw corpus and a seed lexicon. For the experiments reported here, we have used domain corpora of 1 million words from each domain. As a sentiment seed lexicon (for learning domain specific polarity terms) we have used general, well-known lexicons available such as Bing Liu's Opinion lexicon (Hu and Liu, 2004) for English and Spanish sentiment lexicon (Pérez-Rosas et al., 2012) for Spanish.

4. Future Work

Starting form this work, future directions will include exploring the uses of the extracted multilingual domain-specific corpora for the acquisition of multilingual resources. High priority on our list is domain specific terminology acquisition – with particular emphasis on multiword expressions – for several languages, and creating a multilingual resource for domain-specific translation or query expansion, particularly for languages with few resources.

We would also like to explore automatic word sense disambiguation on Wikipedia articles, by exploiting the sense restrictions imposed by the specific domains assigned to the articles according to our mappings between Princeton WordNet topics, WORDNET DOMAINS and the articles' categories.

While working with synset-aligned wordnets gives us redundant mappings to Wikipedia categories for a more confident alignment, the process described can be used without these, to produce domain-specific corpora in other languages represented in Wikipedia, relying solely on cross-language links. We plan to implement and test such an extension of our approach.

5. Conclusions

In this paper, we described the construction of a novel resource, obtained by automatic mapping of Princeton WordNet topics with WORDNET DOMAINS. We used this mapping to extend the assignment of the *topic* relation to other synsets, using also the hyponymy relation in WORDNET. Likewise WORDNET DOMAINS, any language having a synset-aligned wordnet to Princeton WORDNET can take advantage of this resource.

We have also created domain-specific corpora for WORDNET DOMAINS from Wikipedia, by exploiting Wikipedia's category hierarchy and category-article links. We have generated such domain-specific corpora for 6 languages – English, French, German, Italian, Portuguese and Spanish – using the domain to synset mapping, and the synset alignment over the different language versions, as well as cross-language links in Wikipedia (through Dbpedia). To exemplify the usefulness of the created resource, we outlined an application – the extraction of a domain-specific sentiment lexicon.

Acknowledgment

This work was partially supported by the EuroSentiment EU project.

6. References

Arcan, M., Buitelaar, P., and Federmann, C. (2012). Using domain-specific and collaborative resources for term translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 86–94. Association for Computational Linguistics.

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.

Choi, Y., Kim, Y., and Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44. ACM.

Dewey, M. (1876). *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*. Kingsport Press. Project Gutenberg EBook: http://www.gutenberg.org/files/12513/12513-h/12513-h.htm.

Fellbaum, C., editor. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Fernando, S. and Stevenson, M. (2012). Mapping wordnet synsets to wikipedia articles. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, Mayay. European Language Resources Association (ELRA).

Gliozzo, A. and Strapparava, C. (2009). *Semantic Domains in Computational Linguistics*. Springer, August.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Magnini, B. and Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, June.

Milne, D., Medelyan, O., and Witten, I. H. (2006). Mining domain-specific thesauri from wikipedia: A case study. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence*, pages 442–448. IEEE Computer Society.

Pérez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in spanish. In *LREC*, pages 3077–3081.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January.

Ponzetto, S. P. and Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st international joint conference on Artifical Intelligence*, pages 2083–2088, California, USA.

Ponzetto, S. P. and Strube, M. (2007). Deriving a large scale taxonomy from wikipedia. In *AAAI*, volume 7, pages 1440–1445.

Procter, P., editor. (1978). *Longman Dictionary of Contemporary English*. Longman Group Limited.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). YAGO: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Weichselbraun, A., Gindl, S., and Scharl, A. (2011). Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1053–1060. ACM.