

# BabelDomains: Large-Scale Domain Labeling of Lexical Resources

Jose Camacho-Collados and Roberto Navigli

Department of Computer Science

Sapienza University of Rome

{collados,navigli}@di.uniroma1.it

## Abstract

In this paper we present BabelDomains, a unified resource which provides lexical items with information about domains of knowledge. We propose an automatic method that uses knowledge from various lexical resources, exploiting both distributional and graph-based clues, to accurately propagate domain information. We evaluate our methodology intrinsically on two lexical resources (WordNet and BabelNet), achieving a precision over 80% in both cases. Finally, we show the potential of BabelDomains in a supervised learning setting, clustering training data by domain for hypernym discovery.

## 1 Introduction

Since the early days of Natural Language Processing (NLP) and Machine Learning, generalizing a given algorithm or technique has been extremely challenging. One of the main factors that has led to this issue in NLP has been the wide variety of domains for which data are available (Jiang and Zhai, 2007). Algorithms trained on the business domain are not to be expected to work well in biology, for example. Moreover, even if we manage to obtain a balanced training set across domains, our algorithm may not be as effective on some specific domain as if it had been trained on that same target domain. This issue has become even more challenging and significant with the rise of supervised learning techniques. These techniques are fed with large amounts of data and ought to be able generalize to various target domains. Several studies have proposed regularization frameworks for domain adaptation in NLP (Daumé III and Marcu, 2006; Daumé III, 2007; Lu et al., 2016). In this paper we tackle this problem but approach it from

a different angle. Our main goal is to integrate domain information into lexical resources, which, in turn, could enable a semantic clusterization of training data by domain, a procedure known as multi-source domain adaptation (Crammer et al., 2008). In fact, adapting algorithms to a particular domain has already proved essential in standard NLP tasks such as Word Sense Disambiguation (Magnini et al., 2002; Agirre et al., 2009; Faralli and Navigli, 2012), Text Categorization (Navigli et al., 2011), Sentiment Analysis (Glorot et al., 2011; Hamilton et al., 2016), or Hypernym Discovery (Espinosa-Anke et al., 2016), *inter alia*.

The domain annotation of WordNet (Miller et al., 1990) has already been carried out in previous studies (Magnini and Cavaglià, 2000; Bentivogli et al., 2004; Tufiş et al., 2008). Domain information is also available in IATE<sup>1</sup>, a European Union inter-institutional terminology database. The domain labels of IATE are based on the Eurovoc thesaurus<sup>2</sup> and were introduced manually. The fact that each of these approaches involves manual curation/intervention limits their extension to other resources, and therefore to downstream applications.

We, instead, have developed an automatic hybrid distributional and graph-based method for encoding domain information into lexical resources. In this work we aim at annotating BabelNet (Navigli and Ponzetto, 2012), a large unified lexical resource which integrates WordNet and other resources<sup>3</sup> such as Wikipedia and Wiktionary, augmenting the initial coverage of WordNet by two orders of magnitude.

<sup>1</sup><http://iate.europa.eu/>

<sup>2</sup><http://eurovoc.europa.eu/drupal/?q=navigation&cl=en>

<sup>3</sup>See <http://babelnet.org/about> for a complete list of the resources integrated in BabelNet.

Animals	Engineering and technology	Language and linguistics	Philosophy and psychology
Art, architecture and archaeology	Food and drink	Law and Crime	Physics and astronomy
Biology	Games and video games	Literature and theatre	Politics and government
Business, economics and finance	Geography and places	Mathematics	Religion, mysticism and mythology
Chemistry and mineralogy	Geology and geophysics	Media	Royalty and nobility
Computing	Health and medicine	Meteorology	Sport and recreation
Culture and society	Heraldry, honors and vexillology	Music	Transport and travel
Education	History	Numismatics and currencies	Warfare and defense

Table 1: The set of thirty-two domains.

## 2 Methodology

Our goal is to enrich lexical resources with domain information. To this end, we rely on BabelNet 3.0, which merges both encyclopedic (e.g. Wikipedia) and lexicographic resources (e.g. WordNet). The main unit in BabelNet, similarly to WordNet, is the synset, which is a set of synonymous words corresponding to the same meaning (e.g.,  $\{midday, noon, noontide\}$ ). In contrast to WordNet, a BabelNet synset may contain lexicalizations coming from different resources and languages. Therefore, the annotation of a BabelNet synset could directly be expanded to all its associated resources.

As domains of knowledge, we opted for domains from the *Wikipedia featured articles page*<sup>4</sup>. This page contains a set of thirty-two domains of knowledge.<sup>5</sup> Table 1 shows the set of thirty-two domains. For each domain, there is a set of Wikipedia pages associated (127 on average). For instance, the Wikipedia pages *Kolkata* and *Oklahoma* belong to the *Geography* domain<sup>6</sup>. Our methodology for annotating BabelNet synsets with domains is divided into two steps: (1) we apply a distributional approach to obtain an extensive distribution of domain labels in BabelNet (Section 2.1), and (2) we complement this first step with a set of heuristics to improve the coverage and correctness of the domain annotations (Section 2.2).

### 2.1 Distributional similarity

We exploit the distributional approach of Camacho-Collados et al. (2016, NASARI). NASARI<sup>7</sup> provides lexical vector representations for BabelNet synsets. In order to obtain a full distribution for each BabelNet synset, i.e. a list

<sup>4</sup>[https://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](https://en.wikipedia.org/wiki/Wikipedia:Featured_articles)

<sup>5</sup>Biography domains are not considered.

<sup>6</sup>For simplicity we refer to each domain with its first word (e.g., *Geography* to refer to *Geography and Places*).

<sup>7</sup><http://lcl.uniroma1.it/nasari/>

of ranked domains associated, each domain is first associated with a given vector. Then, the Wikipedia pages from the featured articles page are leveraged as follows. First, all Wikipedia pages associated with a given domain are concatenated into a single text. Second, a lexical vector is constructed for each text as in Camacho-Collados et al. (2016), by applying lexical specificity over the bag-of-word representation of the text. Finally, given a BabelNet synset  $s$ , the similarity between its respective NASARI lexical vector and the lexical vector of each domain is calculated using the Weighted Overlap comparison measure (Pilehvar et al., 2013).<sup>8</sup>

This enables us to obtain, for each BabelNet synset, scores for each domain label denoting their importance. For notational brevity, we will refer to the domain whose similarity score is highest across all domains as its *top domain*. For instance, the top domain of the BabelNet synset corresponding to *rifle* is *Warfare*, while its second domain is *Engineering*. In order to increase precision, initially we only tag those BabelNet synsets whose maximum score is higher than 0.35.<sup>9</sup>

### 2.2 Heuristics

We additionally propose three heterogeneous heuristics to improve the quality and coverage of domain annotations. These heuristics are applied in cascade (in the same order as they appear on the text) over the labels provided on the previous step.

**Taxonomy.** This first heuristic is based on the BabelNet hypernymy structure, which is an integration of various taxonomies: WikiData, WordNet and MultiWiBi (Flati et al., 2016). The main intuition is that, in general, synsets connected by a hypernymy relation tend to share the same domain

<sup>8</sup>Weighted Overlap has been proved to suit interpretable vectors better than cosine (Camacho-Collados et al., 2015).

<sup>9</sup>This value was set through observation to increase precision but without drastically decreasing recall.

(Magnini and Cavaglià, 2000).<sup>10</sup> This taxonomy-based heuristic is intended to both increase coverage and refine the quality of synsets annotated by the distributional approach. First, if all the hypernyms (at least two) of a given synset share the same top domain, this synset is annotated (or re-annotated) with that domain. Second, if the top domain of an annotated synset is different from at least two of its hypernyms, this domain tag is removed.

**Labels.** Some Wikipedia page titles include general information about the page between parentheses. This text between parentheses is known as a label. For example, the Wikipedia page *Orange (telecommunications)* has *telecommunications* as its label. In BabelNet these labels are kept in the main senses of many synsets, information which is valuable for deciding their domain. For those synsets sharing the same label, we create a distribution of domains, i.e. each label is associated with its corresponding synsets and their domains. Then, we tag (or retag) all the synsets containing the given label provided that the most frequent domain for that label gets a number of instances higher than 80% of the total of instances containing the same label.<sup>11</sup> As an example, before applying this heuristic the label *album* contained 14192 synsets which were pre-tagged with a given domain. From those 14192 synsets, 14166 were pre-tagged with the *Music* domain (99.8%). Therefore, the remaining 26 synsets and all the rest containing the *album* label were tagged or re-tagged with the *Music* domain.

**Propagation.** In this last step we propagate the domain annotations over the BabelNet semantic network. First, given an unannotated input synset, we gather a set with all its neighbours in the BabelNet semantic network. Then we retrieve the domain with the highest number of synsets associated among all annotated synsets in the set. Similarly to the previous heuristic, if the number of synsets of such domain amounts to 80% of the whole set, we tag the input synset with that domain. Otherwise, we repeat the process with the

<sup>10</sup>In WordNet this property is satisfied most of the times. However, in Wikipedia, especially given its large amount of entities, this is not always the case. For instance, *Microsoft* is a *company* (tagged with the *Business* domain) but it would arguably better have *Computing* as its top domain.

<sup>11</sup>This threshold is set in order to improve the precision of the system, as there are labels which might be ambiguous within a domain (e.g., country names).

	New	Re-ann.	Removed
Distributional	1.31M	-	-
Taxonomy	164K	32K	7K
Labels	94K	4K	-
Propagation	1.11M	-	-
Total	2.68M	-	-

Table 2: Number of tagged synsets (*new*, *re-annotated* and *removed*) in each of the domain annotation steps.

second-level neighbours and, if still not found, with its third-level neighbours.

### 3 BabelDomains: Statistics and Release

We applied the methodology described in Section 2 on BabelNet 3.0. This led to a total of 2.68M synsets tagged with a domain. Note that this number greatly improves on the number given in previous studies for WordNet. In our approach, in addition to WordNet, we provide annotations for other lexical resources such as Wikipedia or Wiktionary. Table 2 shows some statistics of the synsets tagged in each step of the whole domain annotation process. The largest number of annotated synsets were obtained in the first distributional step (1.31M) and the final propagation (1.11M), while the taxonomy and labels heuristics contributed to not only increasing the coverage, but also to refining potentially dubious annotations.

BabelDomains is available for download at [lcl.uniroma1.it/babeldomains](http://lcl.uniroma1.it/babeldomains). In the release we include a confidence score<sup>12</sup> for each domain label. Additionally, the domain labels have been integrated into BabelNet<sup>13</sup>, both in the API and in the online interface<sup>14</sup>.

### 4 Evaluation

We evaluated BabelDomains both intrinsically (Section 4.1) and extrinsically on the hypernym discovery task (Section 4.2).

<sup>12</sup>The confidence score for each synset’s domain label is computed as the relative number of neighbours in the BabelNet semantic network sharing the same domain.

<sup>13</sup>In its current 3.7 release version we have included two additional domains to the ones included in Table 1: *Farming and Textile and Clothing*

<sup>14</sup>See <http://babelnet.org/search?word=house&lang=EN> for an example of the domain annotations of all senses of *house* in BabelNet.

	WordNet			BabelNet		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
BabelDomains	81.7	68.7	<b>74.6</b>	<b>85.1</b>	32.0	<b>46.5</b>
Distributional	<b>84.0</b>	59.8	69.9	78.1	16.0	26.6
Wikipedia-idf	45.9	29.7	36.1	8.8	6.5	7.5
WN-Taxonomy Prop.	71.3	70.7	71.0	-	-	-
BN-Taxonomy-Prop.	73.5	<b>73.5</b>	73.5	48.3	<b>37.2</b>	42.0
WN-Domains-3.2	93.6	64.4	76.3	-	-	-

Table 3: Precision, Recall and F-Measure percentages of different systems on the gold standard WordNet and BabelNet domain-labeled datasets.

#### 4.1 Intrinsic Evaluation

In this section we describe the evaluation of our domain annotations on two different lexical resources: BabelNet and WordNet. To this end, we used the domain-labeled datasets released by Camacho-Collados et al. (2016). The WordNet dataset is composed of 1540 synsets tagged with a domain. These domain labels were taken from WordNet 3.0 and manually mapped to the domains of the Wikipedia featured articles page. The BabelNet dataset is composed of 200 synsets randomly extracted from BabelNet 3.0 which were manually annotated with domains.

As comparison systems we included a baseline based on Wikipedia (Wikipedia-idf). This baseline first constructs a *tf-idf*-weighted bag-of-word vector representation of Wikipedia pages and, similarly to our distributional approach, calculates its similarity with the concatenation of all Wikipedia pages associated with a domain in the Wikipedia featured articles page.<sup>15</sup> We additionally compared with WN-Domains-3.2 (Magnini and Cavaglia, 2000; Bentivogli et al., 2004), which is the latest released version of WordNet Domains<sup>16</sup>. However, this approach involves manual curation, both in the selection of seeds and correction of errors. In order to enable a fair comparison, we report the results of a system based on its main automatic component. This baseline takes annotated synsets as input and propagates them through the WordNet taxonomy (WN-Taxonomy Prop.). Likewise, we report the results of the same baseline by propagating through the BabelNet taxonomy (BN-Taxonomy Prop.). These two systems were evaluated by 10-fold cross validation on the

<sup>15</sup>For the annotation of WordNet we used the direct Wikipedia-WordNet mapping from BabelNet.

<sup>16</sup><http://wndomains.fbk.eu/>

corresponding datasets. Finally, we include the results of the distributional approach performed in the first step of our methodology (Section 2.1).

Table 3 shows the results of our system and four comparison systems. Our system achieves the best overall F-Measure results, with precision figures above 80% on both WordNet and BabelNet datasets. These results clearly improve the results achieved by applying the first step of distributional similarity only, highlighting that the inclusion of the heuristics was beneficial. These precision figures are especially relevant considering the large set of domains (32) used in our methodology. By analyzing the errors, we realized that our system tends to provide domains close to the gold standard. For instance, the synset referring to *entitlement*<sup>17</sup> was tagged with the Business domain instead of the gold Law. Other domains which produced imperfect choices due to their close proximity were Mathematics-Computing and Animals-Biology. As regards the generally low recall on the BabelNet dataset, we found that it was mainly due to the nature of the dataset, including many isolated synsets which are hardly used in practice.

#### 4.2 Extrinsic Evaluation

One of the main applications of including domain information in sense inventories is to be able to cluster textual data by domain. Supervised systems may be particularly sensitive to this issue (Daumé III, 2007), and therefore training data should be clustered accordingly. In particular, two recent studies found that clustering training data was essential for distributional hypernym discovery systems to perform accurately (Fu et al., 2014; Espinosa-Anke et al., 2016). They discovered that

<sup>17</sup>Defined as *right granted by law or contract (especially a right to benefits)*.

	<b>Art</b>	<b>Bio</b>	<b>Edu</b>	<b>Geo</b>	<b>Hea</b>	<b>Med</b>	<b>Mus</b>	<b>Phy</b>	<b>Tra</b>	<b>War</b>
BabelDomains	<b>0.30</b>	<b>0.87</b>	<b>0.39</b>	<b>0.43</b>	<b>0.12</b>	<b>0.71</b>	0.42	<b>0.20</b>	<b>0.63</b>	<b>0.13</b>
Distributional	0.18	0.41	0.30	0.26	0.10	0.46	<b>0.43</b>	0.08	0.56	0.11
Non-filtered	0.00	0.68	0.00	0.10	0.05	0.25	0.11	0.00	0.34	0.00

Table 4: MRR (Mean Reciprocal Rank) performance of TaxoEmbed in the hypernym discovery task by filtering (BabelDomains and Distributional) or not filtering training data by domains.

hypernymy information is not encoded equally in different regions of distributional vector spaces, as it is stored differently depending on the domain.

The hypernym discovery task consists of, given a term as input, finding its most appropriate hypernym. In this evaluation we followed the approach of Espinosa-Anke et al. (2016, TaxoEmbed), who provides a framework to train a domain-wise transformation matrix (Mikolov et al., 2013) between the vector spaces of terms and hypernyms. As in the original work, we used the sense-level vector space of Iacobacci et al. (2015) and training data from Wikidata.<sup>18</sup> We used the domain annotations of BabelDomains for clustering the training data by domain, and compared it with the domains obtained through the distributional step, as used in Espinosa-Anke et al. (2016). We additionally included a baseline which did not filter the training data by domain. The training data<sup>19</sup> was composed of 20K term-hypernym pairs for the domain-filtered systems and 200K for the baseline, while the test data was composed of 250 randomly-extracted terms with their corresponding hypernyms in Wikidata.

Table 4 shows the results of TaxoEmbed in the hypernym discovery task on the same ten domains<sup>20</sup> evaluated in Espinosa-Anke et al. (2016). Our domain clusterization achieves the best overall results, outperforming the clusterization based solely on distributional information in nine of the ten domains. The results clearly show the need for a pre-clusterization of the training data, confirming the findings of Espinosa-Anke et al. (2016) and Fu et al. (2014). Training directly without pre-clusterization leads to very poor results, despite being trained on a larger sample. This baseline

<sup>18</sup>We used the code and data available at <http://www.taln.upf.edu/taxoembed>

<sup>19</sup>Training data was extracted randomly from Wikidata, excluding the terms of the test data.

<sup>20</sup>Domains are represented by their three initial letters. From left to right in the table: Art, Biology, Education, Geography, Health, Media, Music, Physics, Transport, and Warfare.



provides competitive results on `Biology` only, arguably due to the distribution of Wikidata where biology items are over-represented.

## 5 Conclusion

In this paper we presented BabelDomains, a resource that provides unified domain information in lexical resources. Our method exploits at best the knowledge available in these resources by combining distributional and graph-based approaches. We evaluated the accuracy of our approach on two resources, BabelNet and WordNet. The results showed that our unified resource provides reliable annotations, improving over various competitive baselines. In the future we plan to extend our set of domains with more fine-grained information, providing a hierarchical structure following the line of Bentivogli et al. (2004).

As an extrinsic evaluation we used BabelDomains to cluster training data by domain prior to applying a supervised hypernym discovery system. This pre-clustering proved crucial for finding accurate hypernyms in a distributional vector space. We are planning to further use our resource for multi-source domain adaptation on other NLP supervised tasks. Additionally, since BabelNet and most of its underlying resources are multilingual, we plan to use our resource in languages other than English.

## Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.  

Jose Camacho-Collados is supported by a Google Doctoral Fellowship in Natural Language Processing. We would also like to thank Jim McManus for his comments on the manuscript.

## References

- Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2009. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1501–1506, Pasadena, California.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing. In *Proceedings of the COLING Workshop on Multilingual Linguistic Resources*, pages 101–108, Geneva, Switzerland.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577, Denver, USA.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.
- Koby Crammer, Michael Kearns, and Jennifer Wortman. 2008. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774.
- Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.
- Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263, Prague, Czech Republic.
- Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP*, pages 424–435.
- Stefano Faralli and Roberto Navigli. 2012. A New Minimally-supervised Framework for Domain Word Sense Disambiguation. In *Proceedings of EMNLP*, pages 1411–1422, Jeju, Korea.
- Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artificial Intelligence*, 241:66–102.
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, pages 1199–1209, Baltimore, USA.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520, Bellevue, Washington, USA.
- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of EMNLP*, pages 595–605, Austin, Texas.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105, Beijing, China.
- Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of ACL*, pages 264–271, Prague, Czech Republic.
- Wei Lu, Hai Leong Chieu, and Jonathan Löfgren. 2016. A general regularization framework for domain adaptation. In *Proceedings of EMNLP*, pages 950–954, Austin, Texas.
- Bernardo Magnini and Gabriella Cavaglia. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC*, pages 1413–1418, Athens, Greece.
- Bernardo Magnini, Carlo Strapparava, Giovanni Pezulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(04):359–373.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
- Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier de Lacalle, and Eneko Agirre. 2011. Two birds with one stone: Learning semantic models for text categorization and Word Sense Disambiguation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, pages 2317–2320, Glasgow, UK.
- Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351, Sofia, Bulgaria.
- Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceaşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of 4th Global WordNet Conference, GWC*, pages 441–452, Bucharest, Romania.