Automatic Domain Assignment for Word Sense Alignment

Tommaso Caselli
TrentoRISE / Via Sommarive, 18
38123 Povo, Italy
t.caselli@gmail.com

Carlo Strapparava
FBK / Via Sommarive, 18
38123 Povo, Italy
strappa@fbk.eu

Abstract

This paper reports on the development of a hybrid and simple method based on a machine learning classifier (Naive Bayes), Word Sense Disambiguation and rules, for the automatic assignment of WordNet Domains to nominal entries of a lexicographic dictionary, the Senso Comune De Mauro Lexicon. The system obtained an F1 score of 0.58, with a Precision of 0.70. We further used the automatically assigned domains to filter out word sense alignments between MultiWordNet and Senso Comune. This has led to an improvement in the quality of the sense alignments showing the validity of the approach for domain assignment and the importance of domain information for achieving good sense alignments.

1 Introduction and Problem Statement

Lexical knowledge, i.e. how words are used and express meaning, plays a key role in Natural Language Processing. Lexical knowledge is available in many different forms, ranging from unstructured terminologies (i.e. word list), to full fledged computational lexica and ontologies (e.g. WordNet (Fellbaum, 1998)). The process of creation of lexical resources is costly both in terms of money and time. To overcome these limits, semi-automatic approaches have been developed (e.g. MultiWordNet (Pianta et al., 2002)) with different levels of success. Furthermore, important information is scattered in different resources and difficult to use. Semantic interoperability between resources could represent a viable solution to allow reusability and develop more robust and powerful resources. Word sense alignment (WSA) qualifies as the preliminary requirement for achieving this goal (Matuschek and Gurevych, 2013).

WSA aims at creating lists of pairs of senses from two, or more, (lexical-semantic) resources which denote the same meaning. Different approaches to WSA have been proposed and they all share some common elements, namely: i.) the extensive use of sense descriptions of the words (e.g. WordNet glosses); and ii.) the extension of the basic sense descriptions with additional information such as hypernyms, synonyms and domain or category labels.

The purpose of this work is two folded: first, we experiment on the automatic assignment of domain labels to sense descriptions, and then, evaluate the impact of this information for improving an existing sense aligned dataset for nouns. Previous works has demonstrated that domain labels are a good feature for obtaining high quality alignments of entries (Navigli, 2006; Toral et al., 2009; Navigli and Ponzetto, 2012). The WordNet (WN) Domains (Magnini and Cavaglia, 2000; Bentivogli et al., 2004) have been selected as reference domain labels. We will use as candidate lexico-semantic resources to be aligned two Italian lexica, namely, MultiWordNet (MWN) and the Senso Comune De Mauro Lexicon (SCDM) (Vetere et al., 2011).

The two resources differ in terms of modelization: the former, MWN, is an Italian version of WN obtained through the "expand model" (Vossen, 1996) and perfectly aligned to Princeton WN 1.6, while the latter, SCDM, is a machine readable dictionary obtained from a paper-based reference lexicographic dictionary, De Mauro GRADIT. Major issues for WSA of the lexica concern the following aspects:

- SCMD has no structure of word senses (i.e. no taxonomy, no synonymy relations, no distinction between core senses and subsenses for polysemous entries) unlike MWN;
- SCDM has no domain or category labels associated to senses (with the exception of specific terminological entries) unlike MWN;
- the Italian section of MWN has only 2,481 glosses in Italian over 28,517 synsets for nouns (i.e. 8.7%).

The remainder of this paper is organized as follows: Section 2 will report on the methodology and experiments implemented for the automatic assignment of the WN Domains to the SCDM entries. Section 3 will describe the dataset used for the evaluation of the WSA experiments and the use of the WN Domains for filtering the sense alignments. Finally, Section 4 illustrates conclusion and future work.

2 Methodology and Experiments

The WN Domains consist of a set of 166 hierarchically organized labels which have been associated to each

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 414–418,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

Classifiers	P	R	F1	10-Fold F1
NaiveBayes$_{lemma}$	0.77	0.58	0.66	0.66
MaxEnt$_{lemma}$	0.70	0.49	0.58	0.63
NaiveBayes$_{wsd}$	0.77	0.58	0.66	0.69
MaxEnt$_{wsd}$	0.74	0.54	0.62	0.67

Table 2: Results for the Naive Bayes and Maximum Entropy binary classifiers.

synset[1] and express a subject field label (e.g. SPORT, MEDICINE). A special label, FACTOTUM, has been used for those synsets which can appear in almost all subject fields.

The identification of a domain label to the nominal entries in the SCDM Lexicon is based the "One Domain per Discourse" (ODD) hypothesis applied to the sense descriptions. We have used a reduced set of domains labels (45 normalized domains) following (Magnini et al., 2001).

To assign the WN domain label to the SCDM entries, we have developed a hybrid method: first a binary classifier is applied to the SCDM sense descriptions to discriminate between two domain values, FACTOTUM and OTHER, where the OTHER value includes all remaining 44 normalized domains. After this, all entries classified with the OTHER value are analyzed by a rule based system and associated with a specific domain label (i.e. SPORT, MEDICINE, FOOD . . .).

2.1 Classifier and feature selection

We have developed a training set by manually aligning noun senses between the two lexica. The sense alignment allows us to associate all the information of a synset to a corresponding entry in the SCDM lexicon, including the WN Domain label. Concerning the test set, we have used an existing dataset of aligned noun pairs as in (Caselli et al., 2014). We report in Table 1 the figures for the training and test sets. Multiple alignments with the same domain label have been excluded from the training set.

Characteristics	Training Set	Test Set
# lemmas	131	46
# of aligned pairs	369	166
# of SCDM senses	747	216
# of MWN synsets	675	229
# SCDM with WN Domain label	350	118

Table 1: Training and test sets for the classifier.

In order for the classifier to predict the binary domain labels (FACTOTUM and OTHER), each sense description of the SCDM Lexicon has been represented by means of a two-dimensional feature vector (e.g. for training data: BINARY_DOMAIN_LABEL GENERIC:val SPECIFIC:val). Feature values have been obtained through two strategies:

- lemma label: we extract all normalized domain labels associated to each sense of each lemma in the sense description from MWN. The value of the feature GENERIC corresponds to the sum of the FACTOTUM labels. The value of the feature SPECIFIC corresponds to the sum of all other specific domain labels (e.g. MEDICINE, SPORT etc.) after they have been collapsed into a single value (i.e. NOT-FACTOTUM).
- word sense label: for each sense description, we have first performed Word Sense Disambiguation by means of an adapted version to Italian of the UKB package[2] (Agirre et al., 2010; Agirre et al., 2014)[3]. Only the highest ranked synset, and associated WN Domain(s), was retained as good. Similarly to the lemma label strategy, the sum of the domain label FACTOTUM is assigned to the feature GENERIC, while the sum of all other domain labels collapsed into the single value NOT-FACTOTUM is assigned to the feature SPECIFIC.

We experimented with two classifiers: Naive Bayes and Maximum Entropy as implemented in the MALLET package (McCallum, 2002). We illustrate the results in Table 2. The classifiers have been evaluated with respect to standard Precision (P), Recall (R) and F1 against the test set. Ten-fold cross validation has been performed on the training set as well. Classifiers trained with the first strategy will be associated with the label *lemma*, while those trained with the second strategy with the label *wsd*.

Both classifiers obtains good results with respect to the test data in terms of Precision and Recall. The Naive Bayes classifier outperforms the Maximum Entropy one in both training approaches, suggesting better generalization capabilities even in presence of a small training set and basic features. The role of WSD has a positive impact, namely for the Maximum Entropy classifier (Precision +4 points, Recall +5 points with respect to the lemma label). Although such a positive effect of the WSD does not emerge for the Naive Bayes classifier with respect to the test set, we can still observe an improvement over the ten-fold cross validation (F1= 0.69 *vs.* F1=0.66). We finally selected the

[1] The full set of labels and hierarchy is available at http://wndomains.fbk.eu/hierarchy.html

[2] Available at http://ixa2.si.ehu.es/ukb/

[3] We used the WN Multilingual Central Repository as knowledge base and the MWN entries as dictionary

predictions of Naive Bayes$_{wsd}$ classifier as input to the rule-based system as it provides the highest scores.

2.2 Rules for WN Domain assignment

The rule based classifier for final WN Domain assignment works as follows:

- lemmatized and word sense disambiguated lemmas in the sense descriptions are associated with the corresponding WN Domains from MWN;
- frequency counts on the WN Domain labels is applied; the most frequent WN Domain is assigned as the correct WN Domain of the nominal entry;
- in case two or more WN Domains have same frequency, the following assignment strategy is applied: if the frequency scores of the WN Domains is equal to 1, the value FACTOTUM is selected; on the contrary, if the frequency score is higher than 1, all WN Domain labels are retained as good.

We report the results on final domain assignment in Table 3. The final system, NaiveBayes+Rules, has been compared to two baselines. Both baselines apply frequency counts over the WN Domains labels of the lemmas of the sense descriptions for the entire set of the 45 normalized domain values, including the FACTOTUM label, as explained in Section2. The Baseline$_{lemma}$ assigns the domain by taking into account every WN Domain associated to each lemma. On the other hand, the Baseline$_{wsd}$ selects only the WN Domain of sense disambiguated lemmas. WSD for the second baseline has been performed by applying the same method described in Section 2.1. The results of both baselines have high values for Precision (0.58 for Baseline$_{lemma}$, 0.70 for Baseline$_{wsd}$). We consider this as a further support to the validity of the ODD hypothesis which seems to hold even for text descriptions like dictionary glosses which normally use generic lexical items to illustrate word senses. It is also interesting to notice that WSD on its own has a positive impact in Baseline$_{wsd}$ system for the assignment of specific domain labels (F1=0.53).

The hybrid system performs better than both baselines in terms of F1 scores (F1=0.58 *vs.* F1=0.45 for Baseline$_{lemma}$ *vs.* F1=0.53 for Baseline$_{wsd}$). However, both the hybrid system and the Baseline$_{wsd}$ obtain the same Precision. To better evaluate the performance of our hybrid approach, we computed the paired t-test. The results of the hybrid system are statistically significant with respect to the Baseline$_{lemma}$ ($p < 0.05$) and for Recall only when compared to the Baseline$_{wsd}$.

To further analyze the difference between the hybrid system and the Baseline$_{wsd}$, we performed an error analysis on their outputs. We have identified that the hybrid system is more accurate in the prediction of the FACTOTUM class with respect to the baseline. In particular, the accuracy of the hybrid system on this class is 79% while that of the baseline is only 65%. In addition to this, the hybrid system provides better results in terms of Recall (R=0.50 *vs.* R=0.43). Although comparable, the hybrid system provides more accurate results with respect to the baseline.

System	P	R	F1
NaiveBayes$_{wsd}$+Rules	0.70†	0.50†∗	0.58†
Baseline$_{lemma}$	0.58	0.36	0.45
Baseline$_{wsd}$	0.70	0.43	0.53

Table 3: Results of WN Domain Assignment over the SDCM entries. Statistical significance of the NaiveBayes+Rules system has been marked with a † for the Baseline$_{lemma}$ and with a ∗ for the Baseline$_{wsd}$

3 Domain Filtering for WSA

This section reports on the experiments for improving existing WSA for nouns between SDCM and MWN. In this work we have used the same dataset and alignment methods as in (Caselli et al., 2014), shortly described here:

- Lexical Match: for each word *w* and for each sense *s* in the given resources $R \in$ {MWN, SCDM}, we constructed a sense descriptions d_R(s) as a bag of words in Italian. The alignment is based on counting the number of overlapping tokens between the two strings, normalized by the length of the strings;
- Cosine Similarity: we used the Personalized Page Rank (PPR) algorithm (Agirre et al., 2010) with WN 3.0 as knowledge base extended with the "Princeton Annotated Gloss Corpus". Once the PPR vector pairs are obtained, the alignment is obtained on the basis of the cosine score for each pair[4].

The dataset consists of 166 pairs of aligned senses from MWN and SCDM for 46 nominal lemmas (see also column "*Test set*" in Table 1). Overall, SCDM covers 53.71The main difference with respect to (Caselli et al., 2014) is that the proposed alignments have been additionally filtered on the basis of the output of the WN domain system (NaiveBayes$_{wsd}$+Rules). In particular, for each aligned pair which was considered as good in (Caselli et al., 2014), we have applied a further filtering based on the WN domain system results as follows: if two senses are aligned but do not have the same domain, they are excluded from the WSA results, otherwise they are retained. Table 4 illustrates

[4] The vectors for the SCDM entries were obtained by, first, applying Google Translate API to get the English translations and, then, PPR over WN 3.0.

System	P	R	F1
LexicalMatch	0.76 (0.69)	0.27 (0.44)	0.40 (0.55)
Cosine_noThreshold	0.27 (0.12)	**0.47 (0.94)**	0.35 (0.21)
Cosine > 0.1	0.77 (0.52)	0.21 (0.32)	0.33 (0.40)
Cosine > 0.2	**0.87 (0.77)**	0.14 (0.21)	0.24 (0.33)
LexicalMatch+Cosine > 0.1	0.73 (na)	0.40 (na)	**0.51 (na)**
LexicalMatch+Cosine > 0.2	**0.77 (0.67)**	0.37 (0.61)	0.50 (0.64)

Table 4: Results for WSA of nouns with domain filtering.

the results of the WSA approaches with domain filters. We report in brackets the results from (Caselli et al., 2014). The filtering based on WN Domains has a big impact on Precision and contributes to increase the quality of the aligned senses. Although, in general, we have a downgrading of the performance with respect to Recall, the increase in Precision will reduce the manual post-processing effort to fully aligned the two resources[5]. Furthermore, it is interesting to notice that, when merging together the results of the pre-filtered alignments from the two alignment approaches (LexicalMatch+Cosine > 0.1 and LexicalMatch+Cosine > 0.2), we still have a very high Precision (> 0.70) and an increase in Recall (> 0.40) with respect to the results of each approach. Finally, we want to point out that what was reported as the best alignment results in (Caselli et al., 2014), namely LexicalMatch+Cosine > 0.2, can be obtained, at least for Precision, with a lower filtering cut-off threshold on the Cosine Similarity approach (i.e cut-off threshold at or higher than 0.1)

4 Conclusions and Future Work

This work describes a hybrid approach based on a Naive Bayes classifier, Word Sense Disambiguation and rules for assigning WN Domains to nominal sense descriptions of a lexicographic dictionary, the Senso Comune De Mauro Lexicon. The assignment of domain labels has been used to improve WSA results on nouns between the Senso Comune Lexicon and MultiWordNet. The results support some observations, namely: i.) domain filtering plays an important role in WSA, namely as a strategy to exclude wrong alignments (false positives) and improve the quality of the aligned pairs; ii.) the method we have proposed is a viable approach for automatically enriching existing lexical resources in a reliable way; and iii.) the ODD hypothesis also apply to sense descriptions.

An advantage of our approach is its simplicity. We have used features based on frequency counts and obtained good results, with a Precision of 0.70 for automatic WN Domain assignment. Nevertheless, an important role is played by Word Sense Disambiguation. The use of domain labels obtained from sense disambiguated lemmas improves both the results of the classifier and those of the rules. The absence of statistical significance with respect to the Baseline$_{wsd}$ is not to be considered as a negative result. As the error analysis has showed, the classifier mostly contributes to the identification of the FACTOTUM value, which tends to be overestimated even with sense disambiguated lemmas, and to Recall. We are planning to extend this work to include domain clusters to improve the domain assignment results, namely in terms of Recall.

Acknowledgments

One of the author wants to thank Vrije Univerisiteit Amsterdam for sponsoring the attendance to the EMNLP conference.

References

Eneko Agirre, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring knowledge bases for similarity. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 101–108. Association for Computational Linguistics.

Tommaso Caselli, Carlo Strapparava, Laure Vieu, and Guido Vetere. 2014. Aligning an italianwordnet with a lexicographic dictionary: Coping with limited data. In *Proceedings of the Seventh Global WordNet Conference*, pages 290–298.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press.

[5] The F1 of 0.64 in (Caselli et al., 2014) is obtained with a Precision of 0.67, suggesting that some alignments are false positives

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into wordnet. In *Proceedings of the conference on International Language Resources and Evaluation (LREC 2000)*.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2001. Using domain information for word sense disambiguation. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 111–114. Association for Computational Linguistics.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 2:to appear.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Rada Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York.

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. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the* 44^{th} *Annual Meeting of the Association for Computational Linguistics joint with the* 21^{st} *International Conference on Computational Linguistics (COLING-ACL)*, Sydney, Australia.

Elisabeth Niemann and Iryna Gurevych. 2011. The peoples web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Singapore, January.

Emanuele Pianta, Luisa Bentivogli, and Cristian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

German Rigau and Agirre Eneko. 1995. Disambiguating bilingual nominal entries against WordNet. In *Proceedings of workshop The Computational Lexicon, 7th European Summer School in Logic, Language and Information*, Barcelona, Spain.

Adriana Roventini, Nilda Ruimy, Rita Marinelli, Marisa Ulivieri, and Michele Mammini. 2007. Mapping concrete entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and results. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, June.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Proceedings of the Third international conference on Advances in Web Intelligence*, AWIC'05, Berlin, Heidelberg. Springer-Verlag.

Antonio Toral, Oscar Ferrández, Eneko Aguirre, and Rafael Munoz. 2009. A study on linking and disambiguating wikipedia categories to wordnet using text similarity. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*.

Guido Vetere, Alessandro Oltramari, Isabella Chiari, Elisabetta Jezek, Laure Vieu, and Fabio Massimo Zanzotto. 2011. Senso Comune, an open knowledge base for italian. *JTraitement Automatique des Langues*, 53(3):217–243.

Piek Vossen. 1996. Right or wrong: Combining lexical resources in the eurowordnet project. In *Euralex*, volume 96, pages 715–728. Citeseer.