SemEval-2007 Task 16: Evaluation of Wide Coverage Knowledge Resources

Montse Cuadros
TALP Research Center
Universitat Politécnica de Catalunya
Barcelona, Spain
cuadros@lsi.upc.edu

German Rigau
IXA NLP Group
Euskal Herriko Unibersitatea
Donostia, Spain
german.rigau@ehu.es

Abstract

This task tries to establish the relative quality of available semantic resources (derived by manual or automatic means). The quality of each large-scale knowledge resource is indirectly evaluated on a Word Sense Disambiguation task. In particular, we use Senseval-3 and SemEval-2007 English Lexical Sample tasks as evaluation bechmarks to evaluate the relative quality of each resource. Furthermore, trying to be as neutral as possible with respect the knowledge bases studied, we apply systematically the same disambiguation method to all the resources. A completely different behaviour is observed on both lexical data sets (Senseval-3 and SemEval-2007).

1 Introduction

Using large-scale knowledge bases, such as WordNet (Fellbaum, 1998), has become a usual, often necessary, practice for most current Natural Language Processing (NLP) systems. Even now, building large and rich enough knowledge bases for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, dozens of person-years have been invested in the development of wordnets for various languages (Vossen, 1998). For example, in more than ten years of manual construction (from version 1.5 to 2.1), WordNet passed from 103,445 semantic relations to 245,509 semantic relations[1]. That is, around one thousand new relations per month. But this data does not seems to be rich enough to support advanced concept-based NLP applications directly. It seems that applications will not scale up to working in open domains without more detailed and rich general-purpose (and also domain-specific) semantic knowledge built by automatic means.

Fortunately, during the last years, the research community has devised a large set of innovative methods and tools for large-scale automatic acquisition of lexical knowledge from structured and unstructured corpora. Among others we can mention eXtended WordNet (Mihalcea and Moldovan, 2001), large collections of semantic preferences acquired from SemCor (Agirre and Martinez, 2001; Agirre and Martinez, 2002) or acquired from British National Corpus (BNC) (McCarthy, 2001), large-scale Topic Signatures for each synset acquired from the web (Agirre and de la Calle, 2004) or acquired from the BNC (Cuadros et al., 2005). Obviously, these semantic resources have been acquired using a very different set of methods, tools and corpora, resulting on a different set of new semantic relations between synsets (or between synsets and words).

Many international research groups are working on knowledge-based WSD using a wide range of approaches (Mihalcea, 2006). However, less attention has been devoted on analysing the quality of each semantic resource. In fact, each resource presents different volume and accuracy figures (Cuadros et al., 2006).

In this paper, we evaluate those resources on the

[1] Symmetric relations are counted only once.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 81–86,
Prague, June 2007. ©2007 Association for Computational Linguistics

SemEval-2007 English Lexical Sample task. For comparison purposes, we also include the results of the same resources on the Senseval-3 English Lexical sample task. In both cases, we used only the nominal part of both data sets and we also included some basic baselines.

2 Evaluation Framework

In order to compare the knowledge resources, all the resources are evaluated as Topic Signatures (TS). That is, word vectors with weights associated to a particular synset. Normally, these word vectors are obtained by collecting from the resource under study the word senses appearing as direct relatives. This simple representation tries to be as neutral as possible with respect to the resources studied.

A common WSD method has been applied to all knowledge resources on the test examples of Senseval-3 and SemEval-2007 English lexical sample tasks. A simple word overlapping counting is performed between the Topic Signature and the test example. The synset having higher overlapping word counts is selected. In fact, this is a very simple WSD method which only considers the topical information around the word to be disambiguated. Finally, we should remark that the results are not skewed (for instance, for resolving ties) by the most frequent sense in WN or any other statistically predicted knowledge.

As an example, table 1 shows a test example of SemEval-2007 corresponding to the first sense of the noun capital. In bold there are the words that appear in its corresponding Topic Signature acquired from the web.

Note that although there are several important related words, the WSD process implements exact word form matching (no preprocessing is performed).

2.1 Basic Baselines

We have designed a number of basic baselines in order to establish a complete evaluation framework for comparing the performance of each semantic resource on the English WSD tasks.

RANDOM: For each target word, this method selects a random sense. This baseline can be considered as a lower-bound.

Baselines	P	R	F1
TRAIN	65.1	65.1	65.1
TRAIN-MFS	54.5	54.5	54.5
WN-MFS	53.0	53.0	53.0
SEMCOR-MFS	49.0	49.1	49.0
RANDOM	19.1	19.1	19.1

Table 2: P, R and F1 results for English Lexical Sample Baselines of Senseval-3

SemCor MFS (SEMCOR-MFS): This method selects the most frequent sense of the target word in SemCor.

WordNet MFS (WN-MFS): This method selects the first sense in WN1.6 of the target word.

TRAIN-MFS: This method selects the most frequent sense in the training corpus of the target word.

Train Topic Signatures (TRAIN): This baseline uses the training corpus to directly build a Topic Signature using TFIDF measure for each word sense. Note that this baseline can be considered as an upper-bound of our evaluation.

Table 2 presents the precision (P), recall (R) and F1 measure (harmonic mean of recall and precision) of the different baselines in the English Lexical Sample exercise of Senseval-3. In this table, TRAIN has been calculated with a vector size of at maximum 450 words. As expected, RANDOM baseline obtains the poorest result. The most frequent senses obtained from SemCor (SEMCOR-MFS) and WN (WN-MFS) are both below the most frequent sense of the training corpus (TRAIN-MFS). However, all of them are far below the Topic Signatures acquired using the training corpus (TRAIN).

Table 3 presents the precision (P), recall (R) and F1 measure (harmonic mean of recall and precision) of the different baselines in the English Lexical Sample exercise of SemEval-2007. Again, TRAIN has been calculated with a vector size of at maximum 450 words. As before, RANDOM baseline obtains the poorest result. The most frequent senses obtained from SemCor (SEMCOR-MFS) and WN (WN-MFS) are both far below the most frequent sense of the training corpus (TRAIN-MFS), and all of them are below the Topic Signatures acquired using the training corpus (TRAIN).

Comparing both lexical sample sets, SemEval-2007 data appears to be more skewed and simple for WSD systems than the data set from Senseval-3: less

<instance id="19:0@11@wsj/01/wsj_0128@wsj@en@on" docsrc="wsj"> <context>
" A sweeping restructuring of the industry is possible . " Standard & Poor 's Corp. says First Boston , Shearson and Drexel Burnham Lambert Inc. , in particular , are likely to have difficulty shoring up their **credit** standing in months ahead . What worries credit-rating concerns the most is that Wall Street firms are taking long-term **risks** with their own <head> **capital** </head> via leveraged buy-out and junk bond financings . That 's a departure from their traditional practice of transferring almost all **financing** risks to **investors** . Whereas conventional securities financings are structured to be sold quickly , Wall Street 's new penchant for leveraged buy-outs and junk bonds is resulting in long-term lending commitments that stretch out for months or years .
</context> </instance>

Table 1: Example of test id for capital#n which its correct sense is 1

Baselines	P	R	F1
TRAIN	87.6	87.6	87.6
TRAIN-MFS	81.2	79.6	80.4
WN-MFS	66.2	59.9	62.9
SEMCOR-MFS	42.4	38.4	40.3
RANDOM	27.4	27.4	27.4

Table 3: P, R and F1 results for English Lexical Sample Baselines of SemEval-2007

polysemous (as shown by the RANDOM baseline), less similar than SemCor word sense frequency distributions (as shown by SemCor-MFS), more similar to the first sense of WN (as shown by WN-MFS), much more skewed to the first sense of the training corpus (as shown by TRAIN-MFS), and much more easy to be learned (as shown by TRAIN).

3 Large scale knowledge Resources

The evaluation presented here covers a wide range of large-scale semantic resources: WordNet (WN) (Fellbaum, 1998), eXtended WordNet (Mihalcea and Moldovan, 2001), large collections of semantic preferences acquired from SemCor (Agirre and Martinez, 2001; Agirre and Martinez, 2002) or acquired from the BNC (McCarthy, 2001), large-scale Topic Signatures for each synset acquired from the web (Agirre and de la Calle, 2004) or SemCor (Landes et al., 2006).

Although these resources have been derived using different WN versions, using the technology for the automatic alignment of wordnets (Daudé et al., 2003), most of these resources have been integrated into a common resource called Multilingual Central Repository (MCR) (Atserias et al., 2004) maintaining the compatibility among all the knowledge resources which use a particular WN version as a sense repository. Furthermore, these mappings allow to port the knowledge associated to a particular WN version to the rest of WN versions.

The current version of the MCR contains 934,771 semantic relations between synsets, most of them acquired by automatic means. This represents almost four times larger than the Princeton WordNet (245,509 unique semantic relations in WordNet 2.1).

Hereinafter we will refer to each semantic resource as follows:

WN (Fellbaum, 1998): This resource uses the direct relations encoded in WN1.6 or WN2.0 (for instance, tree#n#1–hyponym–>teak#n#2). We also tested WN^2 (using relations at distances 1 and 2), WN^3 (using relations at distances 1 to 3) and WN^4 (using relations at distances 1 to 4).

XWN (Mihalcea and Moldovan, 2001): This resource uses the direct relations encoded in eXtended WN (for instance, teak#n#2–gloss–>wood#n#1).

WN+XWN: This resource uses the direct relations included in WN and XWN. We also tested $(WN+XWN)^2$ (using either WN or XWN relations at distances 1 and 2, for instance, tree#n#1–related–>wood#n#1).

spBNC (McCarthy, 2001): This resource contains 707,618 selectional preferences acquired for subjects and objects from BNC.

spSemCor (Agirre and Martinez, 2002): This resource contains the selectional preferences acquired for subjects and objects from SemCor (for instance, read#v#1–tobj–>book#n#1).

MCR (Atserias et al., 2004): This resource uses the direct relations included in MCR but excluding spBNC because of its poor performance. Thus, MCR contains the direct relations from WN (as tree#n#1–hyponym–>teak#n#2), XWN (as teak#n#2–gloss–>wood#n#1), and spSemCor (as read#v#1–tobj–>book#n#1) but not the indi-

Source	#relations
Princeton WN1.6	138,091
Selectional Preferences from SemCor	203,546
New relations from Princeton WN2.0	42,212
Gold relations from eXtended WN	17,185
Silver relations from eXtended WN	239,249
Normal relations from eXtended WN	294,488
Total	934,771

Table 4: Semantic relations uploaded in the MCR

rect relations of (WN+XWN)^2 (tree#n#1–related–>wood#n#1). We also tested MCR^2 (using relations at distances 1 and 2), which also integrates (WN+XWN)^2 relations.

Table 4 shows the number of semantic relations between synset pairs in the MCR.

3.1 Topic Signatures

Topic Signatures (TS) are word vectors related to a particular topic (Lin and Hovy, 2000). Topic Signatures are built by retrieving context words of a target topic from large corpora. In our case, we consider word senses as topics.

For this study, we use two different large-scale Topic Signatures. The first constitutes one of the largest available semantic resource with around 100 million relations (between synsets and words) acquired from the web (Agirre and de la Calle, 2004). The second has been derived directly from SemCor.

TSWEB[2]: Inspired by the work of (Leacock et al., 1998), these Topic Signatures were constructed using monosemous relatives from WordNet (synonyms, hypernyms, direct and indirect hyponyms, and siblings), querying Google and retrieving up to one thousand snippets per query (that is, a word sense), extracting the words with distinctive frequency using TFIDF. For these experiments, we used at maximum the first 700 words of each TS.

TSSEM: These Topic Signatures have been constructed using the part of SemCor having all words tagged by PoS, lemmatized and sense tagged according to WN1.6 totalizing 192,639 words. For each word-sense appearing in SemCor, we gather all sentences for that word sense, building a TS using TFIDF for all word-senses co-occurring in those sentences.

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

political_party#n#1	2.3219
party#n#1	2.3219
election#n#1	1.0926
nominee#n#1	0.4780
candidate#n#1	0.4780
campaigner#n#1	0.4780
regime#n#1	0.3414
identification#n#1	0.3414
government#n#1	0.3414
designation#n#3	0.3414
authorities#n#1	0.3414

Table 5: Topic Signatures for party#n#1 obtained from Semcor (11 out of 719 total word senses)

In table 5, there is an example of the first word-senses we calculate from party#n#1.

The total number of relations between WN synsets acquired from SemCor is 932,008.

4 Evaluating each resource

Table 6 presents ordered by F1 measure, the performance of each knowledge resource on Senseval-3 and the average size of the TS per word-sense. The average size of the TS per word-sense is the number of words associated to a synset on average. Obviously, the best resources would be those obtaining better performances with a smaller number of associated words per synset. The best results for precision, recall and F1 measures are shown in bold. We also mark in italics those resources using non-direct relations.

Surprisingly, the best results are obtained by TSSEM (with F1 of 52.4). The lowest result is obtained by the knowledge directly gathered from WN mainly because of its poor coverage (R of 18.4 and F1 of 26.1). Also interesting, is that the knowledge integrated in the MCR although partly derived by automatic means performs much better in terms of precision, recall and F1 measures than using them separately (F1 with 18.4 points higher than WN, 9.1 than XWN and 3.7 than spSemCor).

Despite its small size, the resources derived from SemCor obtain better results than its counterparts using much larger corpora (TSSEM vs. TSWEB and spSemCor vs. spBNC).

Regarding the basic baselines, all knowledge resources surpass RANDOM, but none achieves neither WN-MFS, TRAIN-MFS nor TRAIN. Only

KB	P	R	F1	Av. Size
TSSEM	**52.5**	**52.4**	**52.4**	103
MCR^2	45.1	45.1	45.1	26,429
MCR	45.3	43.7	44.5	129
spSemCor	43.1	38.7	40.8	56
(WN+XWN)^2	38.5	38.0	38.3	5,730
WN+XWN	40.0	34.2	36.8	74
TSWEB	36.1	35.9	36.0	1,721
XWN	38.8	32.5	35.4	69
WN^3	35.0	34.7	34.8	503
WN^4	33.2	33.1	33.2	2,346
WN^2	33.1	27.5	30.0	105
spBNC	36.3	25.4	29.9	128
WN	44.9	18.4	26.1	14

Table 6: P, R and F1 fine-grained results for the resources evaluated individually at Senseval-03 English Lexical Sample Task.

KB	P	R	F1	Av. Size
(WN+XWN)^2	**54.9**	**51.1**	**52.9**	5,153
TSWEB	54.8	47.8	51.0	700
XWN	50.1	39.8	44.4	96
WN+XWN	45.4	36.8	40.7	101
MCR	40.2	35.5	37.7	149
TSSEM	35.1	32.7	33.9	428
MCR^2	32.4	29.5	30.9	24,896
WN^3	29.3	26.3	27.7	584
WN^2	25.9	27.4	26.6	72
spSemCor	31.4	23.0	26.5	51.0
WN^4	26.1	23.9	24.9	2,710
WN	36.8	16.1	22.4	13
spBNC	24.4	18.1	20.8	290

Table 7: P, R and F1 fine-grained results for the resources evaluated individually at SemEval-2007, English Lexical Sample Task .

TSSEM obtains better results than SEMCOR-MFS and is very close to the most frequent sense of WN (WN-MFS) and the training (TRAIN-MFS).

Table 7 presents ordered by F1 measure, the performance of each knowledge resource on SemEval-2007 and its average size of the TS per word-sense[3]. The best results for precision, recall and F1 measures are shown in bold. We also mark in italics those resources using non-direct relations.

Interestingly, on SemEval-2007, all the knowledge resources behave differently. Now, the best results are obtained by (WN+XWN)^2 (with F1 of 52.9), followed by TSWEB (with F1 of 51.0). The lowest result is obtained by the knowledge encoded in spBNC mainly because of its poor precision (P of 24.4 and F1 of 20.8).

Regarding the basic baselines, spBNC, WN (and also WN^2 and WN^4) and spSemCor do not surpass RANDOM, and none achieves neither WN-MFS, TRAIN-MFS nor TRAIN. Now, WN+XWN, XWN, TSWEB and (WN+XWN)^2 obtain better results than SEMCOR-MFS but far below the most frequent sense of WN (WN-MFS) and the training (TRAIN-MFS).

5 Combination of Knowledge Resources

In order to evaluate deeply the contribution of each knowledge resource, we also provide some results of the combined outcomes of several resources. The combinations are performed following a very basic strategy (Brody et al., 2006).

[3]The average size is different with respect Senseval-3 because the words selected for this task are different

KB	Rank
MCR+(WN+XWN)^2+TSWEB+TSSEM	**55.5**

Table 8: F1 fine-grained results for the 4 system-combinations on Senseval-3

Rank-Based Combination (Rank): Each semantic resource provides a ranking of senses of the word to be disambiguated. For each sense, its placements according to each of the methods are summed and the sense with the lowest total placement (closest to first place) is selected.

Table 8 presents the F1 measure result with respect this method when combining four different semantic resources on the Senseval-3 test set.

Regarding the basic baselines, this combination outperforms the most frequent sense of SemCor (SEMCOR-MFS with F1 of 49.1), WN (WN-MFS with F1 of 53.0) and, the training data (TRAIN-MFS with F1 of 54.5).

Table 9 presents the F1 measure result with respect the rank mthod when combining the same four different semantic resources on the SemEval-2007 test set.

KB	Rank
MCR+(WN+XWN)^2+TSWEB+TSSEM	38.9

Table 9: F1 fine-grained results for the 4 system-combinations on SemEval-2007

In this case, the combination of the four resources obtains much lower result. Regarding the baselines, this combination performs lower than the most frequent senses from SEMCOR, WN or the training data. This could be due to the poor individual performance of the knowledge derived from SemCor (spSemCor, TSSEM and MCR, which integrates spSemCor). Possibly, in this case, the knowledge comming from SemCor is counterproductive. Interestingly, the knowledge derived from other sources (XWN from WN glosses and TSWEB from the web) seems to be more robust with respect corpus changes.

6 Conclusions

Although this task had no participants, we provide the performances of a large set of knowledge resources on two different test sets: Senseval-3 and SemEval-2007 English Lexical Sample task. We also provide the results of a system combination of four large-scale semantic resources. When evaluated on Senseval-3, the combination of knowledge sources surpass the most-frequent classifiers. However, a completely different behaviour is observed on SemEval-2007 data test. In fact, both corpora present very different characteristics. The results show that some resources seems to be less dependant than others to corpus changes.

Obviously, these results suggest that much more research on acquiring, evaluating and using large-scale semantic resources should be addressed.

7 Acknowledgements

We want to thank the valuable comments of the anonymous reviewers. This work has been partially supported by the projects KNOW (TIN2006-15049-C03-01) and ADIMEN (EHU06/113).

References

E. Agirre and O. Lopez de la Calle. 2004. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of LREC*, Lisbon, Portugal.

E. Agirre and D. Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of CoNLL*, Toulouse, France.

E. Agirre and D. Martinez. 2002. Integrating selectional preferences in wordnet. In *Proceedings of GWC*, Mysore, India.

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of GWC*, Brno, Czech Republic.

S. Brody, R. Navigli, and M. Lapata. 2006. Ensemble methods for unsupervised wsd. In *Proceedings of COLING-ACL*, pages 97–104.

M. Cuadros, L. Padró, and G. Rigau. 2005. Comparing methods for automatic acquisition of topic signatures. In *Proceedings of RANLP*, Borovets, Bulgaria.

M. Cuadros, L. Padró, and G. Rigau. 2006. An empirical study for automatic acquisition of topic signatures. In *Proceedings of GWC*, pages 51–59.

J. Daudé, L. Padró, and G. Rigau. 2003. Validation and Tuning of Wordnet Mapping Techniques. In *Proceedings of RANLP*, Borovets, Bulgaria.

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

S. Landes, C. Leacock, and R. Tengi. 2006. Building a semantic concordance of english. In *WordNet: An electronic lexical database and some applications. MIT Press, Cambridge,MA., 1998*, pages 97–104.

C. Leacock, M. Chodorow, and G. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166.

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*. Strasbourg, France.

D. McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

R. Mihalcea and D. Moldovan. 2001. extended wordnet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

R. Mihalcea. 2006. Knowledge based methods for word sense disambiguation. In *E. Agirre and P. Edmonds (Eds.) Word Sense Disambiguation: Algorithms and applications.*, volume 33 of *Text, Speech and Language Technology*. Springer.

P. Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .