SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval

Eneko Agirre
IXA NLP group
University of the Basque Country
Donostia, Basque Counntry
e.agirre@ehu.es

Bernardo Magnini
ITC-IRST
Trento, Italy
magnini@itc.it

Oier Lopez de Lacalle
IXA NLP group
University of the Basque Country
Donostia, Basque Country
jibloleo@ehu.es

Arantxa Otegi
IXA NLP group
University of the Basque Country
Donostia, Basque Country
jibotusa@ehu.es

German Rigau
IXA NLP group
University of the Basque Country
Donostia, Basque Country
german.rigau@ehu.es

Piek Vossen
Irion Technologies
Delftechpark 26
2628XH Delft, Netherlands
Piek.Vossen@irion.nl

Abstract

This paper presents a first attempt of an application-driven evaluation exercise of WSD. We used a CLIR testbed from the Cross Lingual Evaluation Forum. The expansion, indexing and retrieval strategies where fixed by the organizers. The participants had to return both the topics and documents tagged with WordNet 1.6 word senses. The organization provided training data in the form of a pre-processed Semcor which could be readily used by participants. The task had two participants, and the organizer also provide an in-house WSD system for comparison.

1 Introduction

Since the start of Senseval, the evaluation of Word Sense Disambiguation (WSD) as a separate task is a mature field, with both lexical-sample and all-words tasks. In the first case the participants need to tag the occurrences of a few words, for which hand-tagged data has already been provided. In the all-words task all the occurrences of open-class words occurring in two or three documents (a few thousand words) need to be disambiguated.

The community has long mentioned the necessity of evaluating WSD in an application, in order to check which WSD strategy is best, and more important, to try to show that WSD can make a difference in applications. The use of WSD in Machine Translation has been the subject of some recent papers, but less attention has been paid to Information Retrieval (IR).

With this proposal we want to make a first try to define a task where WSD is evaluated with respect to an Information Retrieval and Cross-Lingual Information Retrieval (CLIR) exercise. From the WSD perspective, this task will evaluate all-words WSD systems indirectly on a real task. From the CLIR perspective, this task will evaluate which WSD systems and strategies work best.

We are conscious that the number of possible configurations for such an exercise is very large (including sense inventory choice, using word sense induction instead of disambiguation, query expansion, WSD strategies, IR strategies, etc.), so this first edition focuses on the following:

- The IR/CLIR system is fixed.
- The expansion / translation strategy is fixed.
- The participants can choose the best WSD strategy.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 1–6,
Prague, June 2007.

- The IR system is used as the upperbound for the CLIR systems.

We think that it is important to start doing this kind of application-driven evaluations, which might shed light to the intricacies in the interaction between WSD and IR strategies. We see this as the first of a series of exercises, and one outcome of this task should be that both WSD and CLIR communities discuss together future evaluation possibilities.

This task has been organized in collaboration with the Cross-Language Evaluation Forum (CLEF[1]). The results will be analyzed in the CLEF-2007 workshop, and a special track will be proposed for CLEF-2008, where CLIR systems will have the opportunity to use the annotated data produced as a result of the Semeval-2007 task. The task has a webpage with all the details at `http://ixa2.si.ehu.es/semeval-clir`.

This paper is organized as follows. Section 2 describes the task with all the details regarding datasets, expansion/translation, the IR/CLIR system used, and steps for participation. Section 3 presents the evaluation performed and the results obtained by the participants. Finally, Section 4 draws the conclusions and mention the future work.

2 Description of the task

This is an application-driven task, where the application is a fixed CLIR system. Participants disambiguate text by assigning WordNet 1.6 synsets and the system will do the expansion to other languages, index the expanded documents and run the retrieval for all the languages in batch. The retrieval results are taken as the measure for fitness of the disambiguation. The modules and rules for the expansion and the retrieval will be exactly the same for all participants.

We proposed two specific subtasks:

1. Participants disambiguate the corpus, the corpus is expanded to synonyms/translations and we measure the effects on IR/CLIR. Topics[2] are not processed.
2. Participants disambiguate the topics per language, we expand the queries to synonyms/translations and we measure the effects on IR/CLIR. Documents are not processed

The corpora and topics were obtained from the ad-hoc CLEF tasks. The supported languages in the topics are English and Spanish, but in order to limit the scope of the exercise we decided to only use English documents. The participants only had to disambiguate the English topics and documents. Note that most WSD systems only run on English text.

Due to these limitations, we had the following evaluation settings:

IR with WSD of topics , where the participants disambiguate the documents, the disambiguated documents are expanded to synonyms, and the original topics are used for querying. All documents and topics are in English.

IR with WSD of documents , where the participants disambiguate the topics, the disambiguated topics are expanded and used for querying the original documents. All documents and topics are in English.

CLIR with WSD of documents , where the participants disambiguate the documents, the disambiguated documents are translated, and the original topics in Spanish are used for querying. The documents are in English and the topics are in Spanish.

We decided to focus on CLIR for evaluation, given the difficulty of improving IR. The IR results are given as illustration, and as an upperbound of the CLIR task. This use of IR results as a reference for CLIR systems is customary in the CLIR community (Harman, 2005).

2.1 Datasets

The English CLEF data from years 2000-2005 comprises corpora from 'Los Angeles Times' (year 1994) and 'Glasgow Herald' (year 1995) amounting to 169,477 documents (579 MB of raw text, 4.8GB in the XML format provided to participants, see Section 2.3) and 300 topics in English and Spanish (the topics are human translations of each other). The relevance judgments were taken from CLEF. This

[1] `http://www.clef-campaign.org`

[2] In IR topics are the short texts which are used by the systems to produce the queries. They usually provide extensive information about the text to be searched, which can be used both by the search engine and the human evaluators.

might have the disadvantage of having been produced by pooling the results of CLEF participants, and might bias the results towards systems not using WSD, specially for monolingual English retrieval. We are considering the realization of a post-hoc analysis of the participants results in order to analyze the effect on the lack of pooling.

Due to the size of the document collection, we decided that the limited time available in the competition was too short to disambiguate the whole collection. We thus chose to take a sixth part of the corpus at random, comprising 29,375 documents (874MB in the XML format distributed to participants). Not all topics had relevant documents in this 17% sample, and therefore only 201 topics were effectively used for evaluation. All in all, we reused 21,797 relevance judgements that contained one of the documents in the 17% sample, from which 923 are positive[3]. For the future we would like to use the whole collection.

2.2 Expansion and translation

For expansion and translation we used the publicly available Multilingual Central Repository (MCR) from the MEANING project (Atserias et al., 2004). The MCR follows the EuroWordNet design, and currently includes English, Spanish, Italian, Basque and Catalan wordnets tightly connected through the Interlingual Index (based on WordNet 1.6, but linked to all other WordNet versions).

We only expanded (translated) the senses returned by the WSD systems. That is, given a word like 'car', it will be expanded to 'automobile' or 'railcar' (and translated to 'auto' or 'vagón' respectively) depending on the sense in WN 1.6. If the systems returns more than one sense, we choose the sense with maximum weight. In case of ties, we expand (translate) all. The participants could thus implicitly affect the expansion results, for instance, when no sense could be selected for a target noun, the participants could either return nothing (or NOSENSE, which would be equivalent), or all senses with 0 score. In the first case no expansion would be performed, in the second all senses would be expanded, which is equivalent to full expansion. This fact will be mentioned again in Section 3.5.

[3]The overall figures are 125,556 relevance judgements for the 300 topics, from which 5700 are positive

Note that in all cases we never delete any of the words in the original text.

In addition to the expansion strategy used with the participants, we tested other expansion strategies as baselines:

noexp no expansion, original text

fullexp expansion (translation in the case of English to Spanish expansion) to all synonyms of all senses

wsd50 expansion to the best 50% senses as returned by the WSD system. This expansion was tried over the in-house WSD system of the organizer only.

2.3 IR/CLIR system

The retrieval engine is an adaptation of the Twenty-One search system (Hiemstra and Kraaij, 1998) that was developed during the 90's by the TNO research institute at Delft (The Netherlands) getting good results on IR and CLIR exercises in TREC (Harman, 2005). It is now further developed by Irion technologies as a cross-lingual retrieval system (Vossen et al.,). For indexing, the TwentyOne system takes Noun Phrases as an input. Noun Phases (NPs) are detected using a chunker and a word form with POS lexicon. Phrases outside the NPs are not indexed, as well as non-content words (determiners, prepositions, etc.) within the phrase.

The Irion TwentyOne system uses a two-stage retrieval process where relevant documents are first extracted using a vector space matching and secondly phrases are matched with specific queries. Likewise, the system is optimized for high-precision phrase retrieval with short queries (1 up 5 words with a phrasal structure as well). The system can be stripped down to a basic vector space retrieval system with an tf.idf metrics that returns documents for topics up to a length of 30 words. The stripped-down version was used for this task to make the retrieval results compatible with the TREC/CLEF system.

The Irion system was also used for preprocessing. The CLEF corpus and topics were converted to the TwentyOne XML format, normalized, and named-entities and phrasal structured detected. Each of the target tokens was identified by an unique identifier.

2.4 Participation

The participants were provided with the following:

1. the document collection in Irion XML format
2. the topics in Irion XML format

In addition, the organizers also provided some of the widely used WSD features in a word-to-word fashion[4] (Agirre et al., 2006) in order to make participation easier. These features were available for both topics and documents as well as for all the words with frequency above 10 in SemCor 1.6 (which can be taken as the training data for supervised WSD systems). The Semcor data is publicly available [5]. For the rest of the data, participants had to sign and end user agreement.

The participants had to return the input files enriched with WordNet 1.6 sense tags in the required XML format:

1. for all the documents in the collection
2. for all the topics

Scripts to produce the desired output from word-to-word files and the input files were provided by organizers, as well as DTD's and software to check that the results were conformant to the respective DTD's.

3 Evaluation and results

For each of the settings presented in Section 2 we present the results of the participants, as well as those of an in-house system presented by the organizers. Please refer to the system description papers for a more complete description. We also provide some baselines and alternative expansion (translation) strategies. All systems are evaluated according to their Mean Average Precision [6] (MAP) as computed by the `trec_eval` software on the pre-existing CLEF relevance-assessments.

3.1 Participants

The two systems that registered sent the results on time.

PUTOP They extend on McCarthy's predominant sense method to create an unsupervised method of word sense disambiguation that uses automatically derived topics using Latent Dirichlet Allocation. Using topic-specific synset similarity measures, they create predictions for each word in each document using only word frequency information. The disambiguation process took aprox. 12 hours on a cluster of 48 machines (dual Xeons with 4GB of RAM). Note that contrary to the specifications, this team returned WordNet 2.1 senses, so we had to map automatically to 1.6 senses (Daude et al., 2000).

UNIBA This team uses a a knowledge-based WSD system that attempts to disambiguate all words in a text by exploiting WordNet relations. The main assumption is that a specific strategy for each Part-Of-Speech (POS) is better than a single strategy. Nouns are disambiguated basically using hypernymy links. Verbs are disambiguated according to the nouns surrounding them, and adjectives and adverbs use glosses.

ORGANIZERS In addition to the regular participants, and out of the competition, the organizers run a regular supervised WSD system trained on Semcor. The system is based on a single k-NN classifier using the features described in (Agirre et al., 2006) and made available at the task website (cf. Section 2.4).

In addition to those we also present some common IR/CLIR baselines, baseline WSD systems, and an alternative expansion:

noexp a non-expansion IR/CLIR baseline of the documents or topics.

fullexp a full-expansion IR/CLIR baseline of the documents or topics.

wsdrand a WSD baseline system which chooses a sense at random. The usual expansion is applied.

1st a WSD baseline system which returns the sense numbered as 1 in WordNet. The usual expansion is applied.

wsd50 the organizer's WSD system, where the 50% senses of the word ranking according to the WSD system are expanded. That is, instead of expanding the single best sense, it expands the best 50% senses.

3.2 IR Results

This section present the results obtained by the participants and baselines in the two IR settings. The

[4] Each target word gets a file with all the occurrences, and each occurrence gets the occurrence identifier, the sense tag (if in training), and the list of features that apply to the occurrence.

[5] `http://ixa2.si.ehu.es/semeval-clir/`

[6] `http://en.wikipedia.org/wiki/Information_retrieval`

	IRtops	IRdocs	CLIR
no expansion	0.3599	0.3599	0.1446
full expansion	0.1610	0.1410	0.2676
UNIBA	0.3030	0.1521	0.1373
PUTOP	0.3036	0.1482	0.1734
wsdrand	0.2673	0.1482	0.2617
1st sense	0.2862	0.1172	0.2637
ORGANIZERS	0.2886	0.1587	0.2664
wsd50	0.2651	0.1479	0.2640

Table 1: Retrieval results given as MAP. IRtops stands for English IR with topic expansion. IRdocs stands for English IR with document expansion. CLIR stands for CLIR results for translated documents.

Senseval-2 all words			
	precision	recall	coverage
ORGANIZERS	0.584	0.577	93.61%
UNIBA	0.498	0.375	75.39%
PUTOP	0.388	0.240	61.92%
Senseval-3 all words			
	precision	recall	coverage
ORGANIZERS	0.591	0.566	95.76%
UNIBA	0.484	0.338	69.98%
PUTOP	0.334	0.186	55.68%

Table 2: English WSD results in the Senseval-2 and Senseval-3 all-words datasets.

second and third columns of Table 1 present the results when disambiguating the topics and the documents respectively. Non of the expansion techniques improves over the baseline (no expansion).

Note that due to the limitation of the search engine, long queries were truncated at 50 words, which might explain the very low results of the full expansion.

3.3 CLIR results

The last column of Table 1 shows the CLIR results when expanding (translating) the disambiguated documents. None of the WSD systems attains the performance of full expansion, which would be the baseline CLIR system, but the WSD of the organizer gets close.

3.4 WSD results

In addition to the IR and CLIR results we also provide the WSD performance of the participants on the Senseval 2 and 3 all-words task. The documents from those tasks were included alongside the CLEF documents, in the same formats, so they are treated as any other document. In order to evaluate, we had to map automatically all WSD results to the respective WordNet version (using the mappings in (Daude et al., 2000) which are publicly available).

The results are presented in Table 2, where we can see that the best results are attained by the organizers WSD system.

3.5 Discussion

First of all, we would like to mention that the WSD and expansion strategy, which is very simplistic, degrades the IR performance. This was rather expected, as the IR experiments had an illustration goal, and are used for comparison with the CLIR experiments. In monolingual IR, expanding the topics is much less harmful than expanding the documents. Unfortunately the limitation to 50 words in the queries might have limited the expansion of the topics, which make the results rather unreliable. We plan to fix this for future evaluations.

Regarding CLIR results, even if none of the WSD systems were able to beat the full-expansion baseline, the organizers system was very close, which is quite encouraging due to the very simplistic expansion, indexing and retrieval strategies used.

In order to better interpret the results, Table 3 shows the amount of words after the expansion in each case. This data is very important in order to understand the behavior of each of the systems. Note that UNIBA returns 3 synsets at most, and therefore the wsd50 strategy (select the 50% senses with best score) leaves a single synset, which is the same as taking the single best system (wsdbest). Regarding PUTOP, this system returned a single synset, and therefore the wsd50 figures are the same as the wsdbest figures.

Comparing the amount of words for the two participant systems, we see that UNIBA has the least words, closely followed by PUTOP. The organizers WSD system gets far more expanded words. The explanation is that when the synsets returned by a WSD system all have 0 weights, the wsdbest expansion strategy expands them all. This was not explicit in the rules for participation, and might have affected the results.

A cross analysis of the result tables and the number of words is interesting. For instance, in the IR exercise, when we expand documents, the results in

		English	Spanish
No WSD	noexp	9,900,818	9,900,818
	fullexp	93,551,450	58,491,767
UNIBA	wsdbest	19,436,374	17,226,104
	wsd50	19,436,374	17,226,104
PUTOP	wsdbest	20,101,627	16,591,485
	wsd50	20,101,627	16,591,485
Baseline	1st	24,842,800	20,261,081
WSD	wsdrand	24,904,717	19,137,981
ORG.	wsdbest	26,403,913	21,086,649
	wsd50	36,128,121	27,528,723

Table 3: Number of words in the document collection after expansion for the WSD system and all baselines. wsdbest stands for the expansion strategy used with participants.

the third column of Table 1 show that the ranking for the non-informed baselines is the following: best for no expansion, second for random WSD, and third for full expansion. These results can be explained because of the amount of expansion: the more expansion the worst results. When more informed WSD is performed, documents with more expansion can get better results, and in fact the WSD system of the organizers is the second best result from all system and baselines, and has more words than the rest (with exception of wsd50 and full expansion). Still, the no expansion baseline is far from the WSD results.

Regarding the CLIR result, the situation is inverted, with the best results for the most productive expansions (full expansion, random WSD and no expansion, in this order). For the more informed WSD methods, the best results are again for the organizers WSD system, which is very close to the full expansion baseline. Even if wsd50 has more expanded words wsdbest is more effective. Note the very high results attained by random. These high results can be explained by the fact that many senses get the same translation, and thus for many words with few translation, the random translation might be valid. Still the wsdbest, 1st sense and wsd50 results get better results.

4 Conclusions and future work

This paper presents the results of a preliminary attempt of an application-driven evaluation exercise of WSD in CLIR. The expansion, indexing and retrieval strategies proved too simplistic, and none of the two participant systems and the organizers system were able to beat the full-expansion baseline. Due to efficiency reasons, the IRION system had some of its features turned off. Still the results are encouraging, as the organizers system was able to get very close to the full expansion strategy with much less expansion (translation).

For the future, a special track of CLEF-2008 will leave the avenue open for more sophisticated CLIR techniques. We plan to extend the WSD annotation to all words in the CLEF English document collection, and we also plan to contact the best performing systems of the SemEval all-words tasks to have better quality annotations.

Acknowledgements

We wish to thank CLEF for allowing us to use their data, and the CLEF coordinator, Carol Peters, for her help and collaboration. This work has been partially funded by the Spanish education ministry (project KNOW)

References

E. Agirre, O. Lopez de Lacalle, and D. Martinez. 2006. Exploring feature set combinations for WSD. In *Proc. of the SEPLN*.

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The MEANING Multilingual Central Repository. In *Proceedings of the 2.nd Global WordNet Conference, GWC 2004*, pages 23–30. Masaryk University, Brno, Czech Republic.

J. Daude, L. Padro, and G. Rigau. 2000. Mapping WordNets Using Structural Information. In *Proc. of ACL*, Hong Kong.

D. Harman. 2005. Beyond English. In E. M. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 153–181. MIT press.

D. Hiemstra and W. Kraaij. 1998. Twenty-One in ad-hoc and CLIR. In E.M. Voorhees and D. K. Harman, editors, *Proc. of TREC-7*, pages 500–540. NIST Special Publication.

P. Vossen, G. Rigau, I. Alegria, E. Agirre, D. Farwell, and M. Fuentes. Meaningful results for Information Retrieval in the MEANING project. In *Proc. of the 3rd Global Wordnet Conference*.