Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online at *http://link.springer.de/series/lncs/*.

Detailed information on LNCS can be found at the series home page *http://www.springer.de/comp/lncs/*.

Proposals for publication should be sent to the

LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany

E-mail: lncs@springer.de

ISSN 0302-9743

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence

Peters · Braschler
Gonzalo · Kluck (Eds.)

LNCS
2406

Evaluation of Cross-Language Information Retrieval Systems

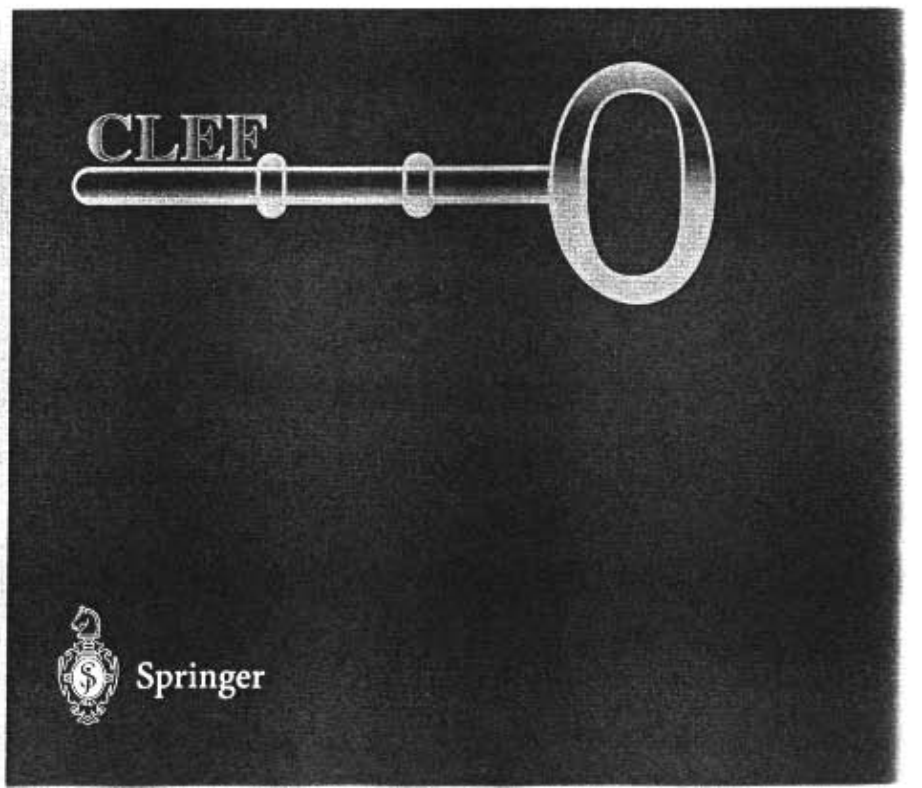
CLEF
2001

LNCS 2406

Carol Peters · Martin Braschler
Julio Gonzalo · Michael Kluck (Eds.)

Evaluation of Cross-Language Information Retrieval Systems

Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001
Darmstadt, Germany, September 2001
Revised Papers



Springer

Table of Contents

Introduction

Carol Peters

IEI-CNR, Area di Ricerca, 56124 Pisa, Italy
carol@iei.pi.cnr.it

The objective of the Cross Language Evaluation Forum (CLEF) is to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes. The second CLEF evaluation campaign was held from January to September 2001 and ended with a workshop held in Darmstadt, Germany, 3-4 September, in which the participants in the campaign reported their experiments and discussed their results.

These Proceedings consist of the revised, extended versions of the preliminary papers presented by the participants at the Workshop. In many cases, the participating groups not only describe and analyse their first results but report additional experiments made subsequent to the workshop. The volume consists of two parts and an appendix. The first part provides an exhaustive overview of the CLEF 2001 experiments whereas the second describes the framework against which these experiments were held. Readers who have never participated in CLEF or in similar evaluation campaigns may well prefer to begin with the second part in order to acquire the necessary background information on the organisation of this type of campaign, before entering into the details of the different cross-language and monolingual retrieval experiments. The appendix presents the results of all the participating groups for each track and task, run by run.

1 CLEF 2001 Experiments

Part I of this volume contains papers from the individual participating groups and provides a complete record of the CLEF 2001 experiments. The first paper by Martin Braschler introduces the experiments by giving a description of the various tracks and tasks and a summary of the main results. The remainder of Part I has been divided into three sections, reflecting to some extent the organisation of the Workshop.

In the first section, we have grouped all those papers that describe cross-language system testing activities: both multilingual, bilingual and domain-specific experiments are included here. The name of the section, "Mainly Cross-language" is determined by the fact that many of the authors also mention strategies implemented in monolingual runs, however, the main focus is on the cross-language aspect of their work. Twenty-two groups submitted runs for this kind of cross-language task, twenty of these groups have contributed a paper to these Proceedings.

Nine groups preferred to remain with the monolingual track at CLEF 2001; most of these were newcomers to CLEF activities. We expect many of them to move onto a

C.A. Peters et al. (Eds.): CLEF 2001, LNCS 2406, pp. 1–5, 2002.

Intelligent Information Access Systems (SINAI) at CLEF 2001: Calculating Translation Probabilities with SemCor

Fernando Martínez-Santiago, L. Alfonso Ureña-López, M. Carlos Díaz-Galiano, Maite Martín Valdivia, and Manuel García-Vega

Dpto. Computer Science. University of Jaén. Avda. Madrid 35. 23071 Jaén, Spain
{dofer,laurena,mcdiaz,maite,mgarcia}@ujaen.es

Abstract. The aim of this paper is to present an approach for bilingual Spanish-English information retrieval based on EuroWordNet but also using another linguistic source known as SemCor. SemCor is used to calculate the translation probabilities of words that share the same meaning in EuroWordNet. The focus of the paper is thus on evaluating the performance of SEMCOR, previously used with success in traditional IR, in a bilingual context.

1 Introduction

CLIR (Cross Language Information Retrieval) is a particular application of Information Retrieval, which aims at building systems capable of retrieving relevant documents in a language that is not necessarily that of the query. This requirement creates a lot of additional problems to those of monolingual IR [1, 2], almost all stemming from the need to overcome the existing linguistic barrier. In our case, the barrier is that between English and Spanish. Namely, we are retrieving texts in English from queries in Spanish. The approach we used is an electronic dictionary (ED) based system. Thus, we start with a query in Spanish and translate it word by word through the ED into English, and use this new query with a traditional IR system. We have used EuroWordNet [3] as if it were an ED. The choice of EuroWordNet has been determined by the main purpose of this study, which is not so much to test a new method for CLIR but rather to highlight the quality of the SemCor linguistic resource in calculating translation probabilities. Whilst there are studies that propose the implementation of CLIR systems using EuroWordNet [4,3], we have focused in this paper on the specific study of the calculation of translation probabilities.

2 SemCor

SemCor [5] is a subset of the Brown Corpus [6] containing documents dealing with several topics such as politics, sports, music, films, philosophy, etc. SemCor is tagged at different syntactic and semantic levels. Every word in SemCor

C. Peters et al. (Eds.): CLEF 2001, LNCS 2406, pp. 185–192, 2002.

```
<contextfile concordance=brown>
<context filename=br a01 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1
lexsn=1:03:00:: pn=group>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1
lexsn=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1
lexsn=1:09:00::>investigation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1
lexsn=1:15:00::>Atlanta</wf>
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2
lexsn=3:00:00:past:00>recent</wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1
lexsn=1:04:00::>primary_election</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4
lexsn=2:39:01::>produced</wf>
<punc> </punc>
<wf cmd=ignore pos=DT>no</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1
lexsn=1:09:00::>evidence</wf>
<punc>''</punc>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=ignore pos=DT>any</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1
lexsn=1:04:00::>irregularities</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1
lexsn=2:30:00::>took_place</wf>
<punc>.</punc>
</s>
</p>
```

Fig. 1. Fragment of a document in SEMCOR

(nouns, verbs, adjectives and adverbs), is annotated with its correct sense in the WORDNET lexical database, when possible.

Figure 1 shows a text fragment of SEMCOR, corresponding to the following sentence from the Brown Corpus:

The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced "no evidence" that any irregularities took place

The figure shows the labels used to indicate the sense of every term that composes the sentence. For example, the term *investigation* appears labelled under the first sense in WORDNET (wnsn=1).

3 EuroWordNet

The EUROWORDNET project implemented a multilingual database, in which the languages present are represented and structured in the style of WORDNET 1.5 [7]. The languages are linked cross-linguistically through English, which acts as an "inter-language" or pivot language, for want of a better word. As in WORDNET, in EUROWORDNET the words are linked by meaning in sets of synonyms (*synsets*). Thus, within one *synset* we will find all those words from a particular language that share a common sense. The *synsets* are linked by a number of relations such as hypernym, holonym, etc. In addition, *synsets* of different languages are also linked by relationships of synonymy or of near synonymy, what we could call "words with close meaning", i.e. words which, without being synonymous over languages, do share a similarity in meaning.

Table 1. Query Translation using EUROWORDNET as ED

Original Spanish	Consequences of Chernobil	
Lematized With MACO+ RELAX	Consecuencias de Chernobil	
No empty words	Consecuencia Chernobil	
Word and meaning, translated according to the relationship of synonym from EUROWORDNET	Consequence (3 meanings)	implication#1 entailment#1 deduction#4
		consequence#2 aftermath#1
		upshot#1 result#3 outcome#2 effect#4 consequence#3
	chernobil (No translation)	chernobil#1

In our experiments with EUROWORDNET we have used the relationship of synonymy for the translation of the words. There are other studies which also make use of the "similar meaning" relationship in the translation [8]. We have preferred to use only the synonymy relationship in a more restrictive approach. Once the query is constructed in this way in Spanish, we eliminate stop words, and lemmatize each word using MACO + RELAX [9] and, for each meaning of the lemma, extract the set of words that make up the corresponding *synsets* in the target language - English in this particular case.

4 Filtering of Queries

This simple approach evidences several problems, already manifest in WordNet: the great amount of "noise" that the *synsets* introduce due to the fine distinction of meanings existing for each word. For instance, the word "capacidad" (capacity) has up to twelve possible translations into English, shared among the five meanings of the source word.

Table 2. Weights for the 3 meanings of the word *absolute*

Word	Meaning	Frequency	Weight
absolute	1	10	0,6665
	2	4	0,2667
	3	1	0,0667

One way of solving this problem may be by classifying, i.e. identifying where the difference is irrelevant to the needs of Information Retrieval [10]. The difficulty of this approach lies in knowing when two or more meanings must be joined into just one. Our approach differs considerably from the idea of grouping according to meaning, although the two methods are not incompatible. The method suggested here attempts to filter the query obtained through a word by word translation using EuroWordNet, disposing of those words we consider to be very rare translations of the Spanish word. It is important to point out that no disambiguation of the original word in Spanish is being made as all the possible meanings of the word are taken into account. All we are trying to achieve is to get rid of all those words in English which are highly unlikely as translations of the original word in Spanish. In short, what we are tying to establish is, for a given word T in Spanish and its corresponding translation into English {S1... Sn}., how probable it is that Si be a translation of T. Although there are lexical databases, such as VLIS [11,12], that make this calculation of translation probabilities process easier, we have decided to calculate this fact from the SemCor corpus.

The idea is simple: SemCor labels every word with its sense, so it is possible to calculate how many times a term is used with a given sense. Thus the probability or weight of every sense for a given term is known automatically: this way, it is easy to build up a frequency table of senses which shows how often a particular sense is assigned to each term. Table 2 shows an example of this process corresponding to the word "absolute". This term receives three senses in SemCor, the third one being the most unusual.

For the translation, we make use of the frequency table of senses as follows. EuroWordNet gives us the translation of every term into different languages by means of the synonymy relationship. For instance, the Spanish word "sanatorio" may be translated into English as *"sanatorium"* or *"home"* because, both pairs < *sanatorio, sanatorium* > and < *sanatorium, home* > share a particular meaning. However the problem is: should we translate the term "sanatorio" with both meanings?

The solution to this problem depends on the strength of the synonymy relationship. The second sense of "home" is synonymous with the first sense of "sanatorio". So, we conclude that the probability of translating "sanatorio" by "home" is exactly that of when "home" is used in its second sense. This frequency information is that stored in our table of senses. Moreover, according to this table, we see that the word "home" usually appears with its first sense rather than the sense shared with "sanatorio". So we consider "home" with the meaning of "sanatorio" as irrelevant.

The use of this method, in addition to its availability, shows another clear advantage of an ED with translation probabilities, that of being readily gradable in word-pairs. It provides information on the probability of translating the words *T1* and *T2* by *S1* and *S2*, assuming we find *S1*, *S2* in the text, each with its specific meaning. The relationship between *S1* and *S2* can be calculated through SemCor according to criteria such as co relation indexes [2] and more complex techniques such as the use of trees of dependence in micro-contexts [13].

Another feature of this approach is that it is very appropriate for the application of disambiguation techniques over the original query, written in Spanish in this instance. Since we are translating T by S, due to the fact that they share a certain meaning, it would be important to know whether T is really acting with the same meaning that it shares with S. Although this approach could almost certainly improve our levels of precision, its use is beyond the scope of this current study.

However, SemCor has two serious drawbacks. The first is its relatively small size (SemCor 1.6 has approximately 31.600 word, meaning pairs)and the second is that it is only available for English.

5 Description of the Experiment

In our experiment we used the ZPrise [14] Information Retrieval System. This choice was determined by its availability and because this system has been recommended in the evaluation of linguistic resources in CLIR tasks such as the one presented here [15]. For our corpus, we used the "Los Angeles Times, 1994" made available by CLEF. This collection has 113.005 documents from the 1994 editions of the "Los Angeles Times". The title, heading and article text were extracted. The official experiments carried out were as follows:

1. *sinai-eng* run: original set of queries in English. This is taken as the best case and used as reference for the rest of runs. For our runs, we used the set of queries provided in Spanish and translated them word by word, using the relationship of cross-language synonymy in EuroWordNet. We then performed three different experiments using this translation:
2. *sinai-esa* run: using the query obtained through the word by word EuroWordNet translation.

3. *sinai-ewn2* run. a filtering was applied, based on the probabilities of translation obtained with SEMCOR, to the set of queries obtained in 1. The aim was to eliminate all those target term candidates below a threshold of 0,25 in their probability of translation. It is important to point out that those words that do not appear in SEMCOR in any of it meanings are retained in the original query, as we have no information for them.

6 Results Obtained

The 11-pt precision we obtained for each of the following experiments is shown in Figure 2, together with the average precision.

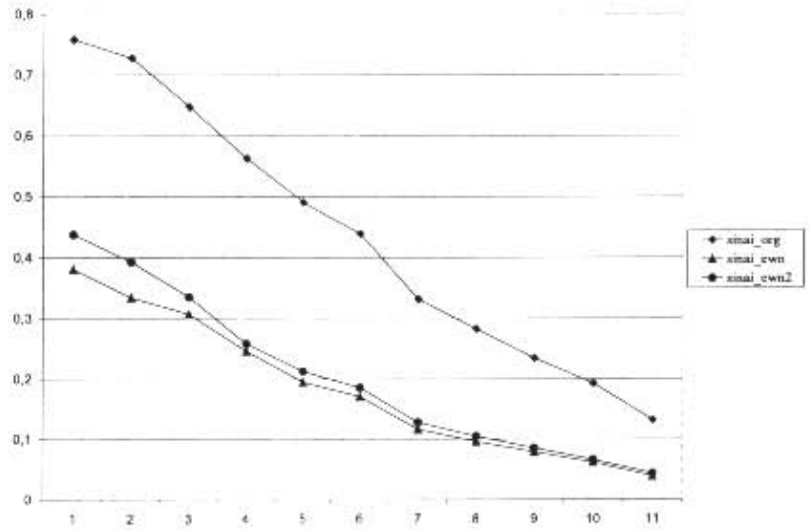


Fig. 2. 11 pt-precision obtained

Table 3. Avg precision obtained

Official run	Average Precision
sinai_org	0.4208
sinai_ewn	0.1701
sinai_ewn2	0.1941

If we take *sinai-org*, as the reference experiment we notice that the loss of precision in the *sinai-ewn* experiment is 59,5% compared with a 53,8% loss in the *sinai-ewn2* (EUROWORDNET+SEMCOR) run. Therefore the use of translation probabilities calculated on SEMCOR reduces the lack of precision by 6,3% compared to that obtained using EUROWORDNET without filtering (*sinai-ewn* experiment). It is likely this percentage would improve if we had a corpus containing all the meanings from EUROWORDNET with a number of words far superior to that of SEMCOR.

Table 4. Breakdown of words in the queries translated from EUROWORDNET. PT = Probability of Translation

	cons_exp		cons_exp + multiwords	
Appear in SEMCOR	PT> 0,25	344	PT> 0,25	42
	PT< 0,25	295	PT< 0,25	12
	Sum:	639	Sum:	54
Do not appear in SEMCOR	196		137	
Sum	735		191	

Table 4 shows how many times SEMCOR provided information that helped to eliminate noise. Thus, we note that of a total of 735 words, which is the sum of words in the *cons-exp* queries, on 196 occasions we do not obtain any information from SEMCOR. This means that 27% of times we cannot decide whether the word is a good translation or not. This situation becomes considerably worse when we consider the multi-words. The percentage of indecision in this case rises to 72%. However, for those multi-words not found in SEMCOR, we note that 77,8% are assigned a probability of translation, PT, superior to 0,25 compared to 53,8% of simple words. This could be read as meaning that multi-words tend to be a more precise translation of the original word, as in general a multi-word tends to be monosemous or have very few meanings.

7 Conclusions and Future Work

We have presented a CLIR system based on an ED. In future work, we will study the effect of multi-words in indexes that can handle these lexical units, rather than just simple words.

Along these lines, in the search for evidence that could indicate the existence of multi-words not registered in EUROWORDNET, studies of the queries as well as of the text retrieved appear promising.

In addition, we have also mentioned a possible solution to the excessively fine granularity of the EUROWORDNET sense disambiguation for Information Retrieval. This solution is based on using the translation probabilities calculated from the frequency of meanings of the words listed in SEMCOR. However, although this approach shows a gain in terms of precision and appears useful

it is far from being completely satisfactory. Our next steps must be directed towards improving the calculation of translation probabilities, through the use of linguistic resources of larger dimensions than SEMCOR, such as large parallel corpus or similar. Another strategy that could be worth studying is the combination of the approach proposed with techniques for lexical disambiguation [16]; we consider these approaches as two sides of the same coin.

References

1. G. Grefenstette. *Cross-Language Information Retrieval*. E. by G. Grefenstette Kluwer Academic Publishers, 1998.
2. D. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *The 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1996.
3. P. Vossen. EuroWordNet: A multilingual database for information retrieval query expansion using lexical-semantic relations. In *THIRD DELOS WORKSHOP Cross-Language Information Retrieval*, 1997.
4. J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying EuroWordNet to cross-language text retrieval. *Computers and the Humanities*, (2/3), 1998.
5. C. Fellbaum. WORDNET: *An Electronic Lexical Database*. The MIT Press, 1998.
6. W. Francis. *Problems of Assembling and Computerizing Large Corpora*. 1982.
7. G. Miller. WORDNET: An on-line lexical database. In *An International Journal of Lexicography*, 1990.
8. T. Gollins and M. Sanderson. CLEF 2000 submission(bilingual track - German to English). In *Working Notes for CLEF 2000 Workshop*, 2000.
9. S. Acebo, A. Ageno, S. Climent, J. Farreres, L. Padró, F. Ribas, H. Rodríguez, and O. Soler. EMACO: Morphological analyzer corpus-oriented. In *In ESPRIT BRA-7315 Acquilex II*, 1994.
10. J. Gonzalo, I. Chugur, and F. Verdejo. Sense clusters for information retrieval: Evidence from SemCor and the interlingual index. In *Proceedings of the ACL'2000 workshop on Word Senses and Multilinguality*, 2000.
11. D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. Twenty-one at CLEF-2000: Translation resources, merging strategies and relevance feedback. In *Working Notes for CLEF 2000 Workshop*, 2000.
12. D. Hiemstra and W. Kraaij. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *Proceedings of the seventh Text Retrieval Conference TREC 7*, 2000.
13. M. Holub and A. Böhmová. Use of dependency tree structures for the microcontext extraction. In *ACL'2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 2000.
14. D. Dimmick. ZPrise NIST. *Available on demand at* `http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html` *[2/6/2001]*, 2001.
15. J. Gonzalo. Language resources in cross-language information retrieval: a CLEF perspective. In *Cross-Language Information Retrieval and Evaluation: Proceedings of the First Cross-Language Evaluation Forum*, 2001.
16. L.A. Ureña, M. Buenaga, and J.M. Gómez. Integrating linguistic resources in TC through WSD. *Computers and the Humanities*, 35(2):215–230, May 2001.