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

Fernando Martínez Santiago
L. Alfonso Ureña López
Manuel Carlos Díaz Galiano
Manuel García Vega
Maite Martín Valdivia
Computer Science Department. University of Jaén. Spain
{dofer, laurena, mcdiaz, mgarcia, maite}@ujaen.es

Summary: This work aims to present an approach for the retrieval of bilingual Spanish-English information based on EuroWordNet and basing itself on another linguistic source such as SemCor, the latter used to calculate the translation probability in words that share same meaning in EuroWordNet. It is, therefore, about evaluating the linguistic aid SemCor, of long-standing tradition in IR tasks, in a bilingual ambience.

Key Words: Cross-language Information Retrieval, EuroWordNet, SemCor, Multiwords, Translation Probabilities.

1. Introduction

CLIR (Cross Language Information Retrieval) is a task within Information Retrieval, whose end is the setting up of systems capable of retrieval relevant documents in a language not necessarily that use in the consultation. This situation creates a lot of additional problems[1,6], almost all stemming from the need to improve the existing linguistic barrier. In one case, the barrier is that between English and Spanish. Namely, here we retrieve texts in English from queries in Spanish. The approach used is that within the systems based on electronic dictionaries (ED). Thus, we start with a query in Spanish which must be translated word by word through the ED into the English language, using this new query with a traditional IR system. We have used EuroWordNet [3] as if it were a DE. The choice for EuroWordNet is due to the final end of this study, not so much to show a new method within those already existent in CLIR, than to highlight the quality of SemCor linguistic resource in the calculus of translation probabilities. Whilst there are studies that propose possible implementation of CLIR systems from EuroWorNet [4,5], we have focused on the specific study of calculus of translation probabilities.

2. EuroWordNet

The EuroWordNet project is about the development of a multilingual database, in a way that the languages present are represented and linked in the style of WordNet 1.5 [5,7]. The link between anyone of these languages though English, which acts as an “inter-language” or pivot language, for want of a better word. As in WordNet, in EuroWordNet the words link by meaning in sets of synonyms (*synsets).* Thus, within one synset we will find all those words from a particular language which share a common meaning. Among these synsets there are certain linguistic links such as hypernym, holonym, etc. In addition to this, among synsets of different languages a relationship of synonym develops and what we could also call “words with close meaning”. Words which, without being synonymous in one language and the other, do share a similarity in meaning

Table I: 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 we have used the relationship of synonymy for the translation of the words. There are works which also make use of the "similar meaning" in the translation [7]. We have preferred to use only the synonymy relationship for a more restrictive approach.

In that way, having consulted in Spanish, you do away with empty words, you lemmatized each word using MACO + RELAX [8] , and you extract for each meaning of the lemma, the set of words that make up the corresponding synsets in the target language, English being this particular case.

3. Filtering of Queries

This simple approach shows several problems, already manifest in WordNet, the great amount of "noise" that the synsets bring, 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 which the original word has.

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

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

One way of solving this problem could be trying to put meanings into groups, whose difference is irrelevant to all needs of Information Retrieval [9]. The difficulty in this approach is precisely knowing when to join 2 or more meanings into just one. Our approach differs considerably from the idea of grouping according to meaning, although they are not incompatible. The method suggested here, attempts to filter the consultation obtained through translating word by word carried out on EuroWordNet, disposing of those words we consider to be a translation of the word in Spanish very rarely. It is important to point out that no ambiguity is being carried out over the original word in Spanish, for all the possible meanings of the word are taken into account. All we are trying to achieve is get rid of all those words in English which are a translation of the original word in Spanish

though highly unlikely. In short, what we are tying to establish is the probability that a given word T in Spanish, and its corresponding translation into English { S1,....Sn}., how probable Si be a translation of T. There are lexical databases, such a VLIS [10.11] that make this calculation of translation probabilities process easier, although we have decided to calculate this fact from corpus SemCor. The idea is simple: each –T, S1- couple share a particular meaning. For instance, the word "sanatorio" in its different meanings can be translated as "sanatorium" with meaning 1 and meaning 2 (sanatorium#1, sanatorium#2), or as "home" with meaning 2 (home#2). However, it is unusual for "home" to appear with the meaning "sanatorium" and so we expect the translation probabilities P("sanatorio"/"home") to be low. In other words, the probability of "home#2" must be low.

To obtain the translation probabilities, each word is jotted down with its meaning. We can calculate the frequency of each meaning of a given word and, from the table of frequency of meanings, the probabilities are that, give a set word, this latter is behaving in a particular meaning. Table II shows an example of this process.

Once the probability of each meaning is calculated for a given word S, we can learn that the probability that S be a translation of a specific Spanish word T, shows a relationship of synonymy with some of the meanings of S, then it will be precisely the probability of this meaning of S that we will suppose to be probability of translating T for S. In other words, if we consider a word T in Spanish, can be translated into some of its meanings for the word S with the meaning "j" in English, then we may conclude that the probability of translating T for S is precisely that where S acts with the j meaning for in that case it would be obvious that S and S' share the meaning j. That is why, indirectly, we are expanding our original consultation by adding all those synonyms of T. The use of this method, in addition to its availability, shows another clear advantage with regard to a ED with probability of translation, of being readily gradable in word-pairs. That is: how probable is translating the words T1 and T2 for S1 and S2, assuming we find S1, S2 in the text, each one with its specific meaning. The relationship between S1 and S2 can be calculated through SemCor according to criteria such as co-relation indexes [6] and more complex techniques such as the use of tress of dependence in micro-contexts [12].

Another peculiarity of this approach is that it is very suitable for applying disambiguity techniques over the original consultation, written in Spanish in this instance. Since we are translating T for S, due to the fact that they share a certain meaning, it would be very worthwhile 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 exactness, its use is beyond the reach of this current study.

Finally, SemCor shows two serious drawbacks. The first being its relatively small size (SemCor 1.6 has approximately 31600 pair –word, meaning-)and the second is that it is only available for the English language.

4. Description of the experiment

In our experiment we have opted for the ZPrise [13] Information Retrieval System. This choice has been determined by the availability and for being a system recommended in the evaluation of linguistic resources in CLIR tasks like that presented here [16]. As a Corpus, we have used the "Los Angeles Times, 1994" Documents (LAT94), used in the CLIF conferences for the evaluation of IR systems. This collection has 113.005 documents from the "Los Angeles Times", in its 1994 editions. The tittle, heading and article core have been extracted. On that basis, the official experiments carried out were as follows:

i. *sinai_org* run: original consulting game in English. This is taken as our best case and the reference for the rest of runs

Following that, we have considered the original consultations in Spanish, to later translate them word by word, using to this end the existing relationship of synonymy in EuroWordNet. On this translation we have done three more experiments:

ii. *sinai-ewn* run: carrying out of the consultation we obtained through the translationword by word with EuroWordNet.

iii. *sinai-ewn2* run: to the set of consultations obtained in (ii), you apply a filtering based on the probabilities of translation obtained with SemCor: eliminate all those that do not surpass the 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 remain in the original consultation, as they are words of which we have no information.

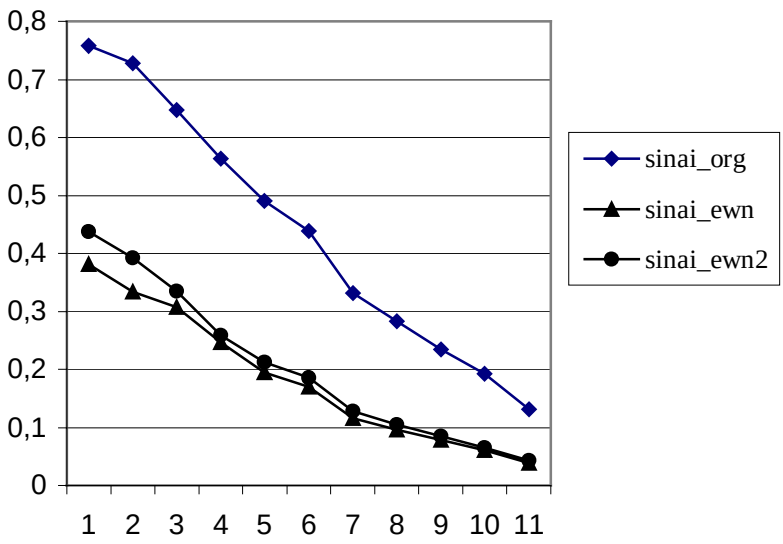
5. Results obtained

The 11-pt precision we have obtained for every one of the following experiments is detailed in the next graph, together with the average precision

Table IV. Avg precision obtained.

official run	Avg. Prec
sinai_org	0,4208
sinai_ewn	0,1701
sinai_ewn2	0,1941

Figure I. 11pt-precision obtained.



In relative terms, if we consider the *sinai-org*, as the best experiment we notice that the loss of precision in the *sinai-ewn* experiment is 59,5% compared to a 53,8% loss in the *sinai-ewn2* (EuroWordNet + SemCor) one. Therefore the use of probabilities of translation calculated on SemCor reduces the lack of precision to 6,3% compared to that obtained using EuroWordNet without filtering (*sinai-ewn* experiment).

It is likely this percentage would improve if you could count with a corpus filled with the meanings from EuroWordNet with a number of words for superior to that of SemCor.

Table V. Breakdown of words in the consultations 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 app ear in SemCor		196		137
Sum		735		191

In Table V you can see how many times SemCor has lent some information for the elimination of noise. Thus, we notice that a total of 735 words, which is the balance of *cons-exp* consultations, on 196 occasions we get no information at all from SemCor. That is on a 27% of times we cannot decide whether the word is a good translation or not. This situation become acutely worse if we consider the multi-words. Then the percentage of indecision rises to 72%. However, for those multi-words we do not find on SemCor, we notice that 77,8% turn out to have a probability of translation PT superior to 0,25 compared to 53,8% of simple words. This could be read as that because multi-words tend to be a more precise translation of the original word, as in general a multi-word tends to be monosemous or with very few meanings, it is therefore, more likely that if we find a multi-word in a determined text, this normally happens with the same sense as the word from which it is translated.

6. Conclusions and future works

We have presented a CLIR system based on ED. In future works, we will study the effect of such multi-words in indexes capable of working with lexical units of this kind, and not just with simple words. Along those lines, research into the exploration of the consultation as well as the recovered text looks very promising, in the search for evidence that could indicate the existence of multi-words not registered on EuroWordNet.

In addition to that, we have also mentioned a possible solution to the excessively tine grain that appears on EuroWordNet, for Information Retrieval tasks, based on the probability of translation calculated from the frequency of meanings of the words listed in SemCor. Whilst you benefit in terms of precision, it is far from adequate, although it shows that an approach of this kind could be useful. Our next steps must be headed towards improving these calculation of translation probabilities, through the use of linguistic resources of greater reach than SemCor, such as large parallel corpus or similar. Another aspect worth taking into consideration is combining the “peaks” of the queries here carried out with techniques of lexical disambiguity because they are in a certain sense, two sides of the same coin.

References

[1] Gregory Grefenstette (1998). "The Problem Of Cross-Language Information Retrieval”. In: *Cross-Language Information Retrieval*, Capítulo 1. Kluwer Academic Publishers.

[2] Schauble, P. & Sheridan, P. (1997). "Cross-Language Information Retrieval (CLIR) Track Overview". In: Voorhees, E. M. & Harman, D. K. (eds.), *NIST Special Publication 500-226: The Sixth Text REtrieval Conference (TREC-6)*. NIST. Available at http://trec.nist.gov/pubs/trec6/t6_proceedings.html [15/02/2000].

[3] Vossen, P. (1997). "EuroWordNet: A Multilingual Database for Information Retrieval". In: *THIRD DELOS WORKSHOP Cross-Language Information Retrieval*, pp. 85-94. *European Research Consortium For Informatics and Mathematics*. Available at: http://www.ercim.org/publication/ws-proceedings/DELOS3/Vossen.pdf [01/03/2000].

[4] Gonzalo, J., Verdejo, F., Peters, C. & Calzolari, N. (1998). "Applying EuroWordNet to cross-language text retrieval", In *Computers and the Humanities*, 32(2-3)**,** pp 185-207

[5] Vossen, P. (1997). "EuroWordNet: A Multilingual Database for Information Retrieval". In: *THIRD DELOS WORKSHOP Cross-Language Information Retrieval*, pp. 85-94. European Research Consortium For Informatics and Mathematics. [Online]. Available at: http://www.ercim.org/publication/ws-proceedings/DELOS3/Vossen.pdf [01/03/2000].

[6] David A. Hull, Gregory Grefenstette (1996). "Experiments in Multlingual Information Retrieval". In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Available at: http://www.xerox.fr/people/grenoble/ hull/papers/sigir96.ps.

[7] Tim Gollins, Mark Sanderson (2000). "CLEF 200 Submission(Bilingual Track – German to English)". In *Working Notes for CLEF 2000 Workshop* . Available at: http://www.iei.pi.cnr.it/DELOS/CLEF/sheffield.doc [1/4/2001].

[8] S. Acebo, A. Ageno, S. Climent, J. Farreres, L. Padró, F. Ribas, H. Rodríguez, O. Soler (1994) "E*MACO: Morphological Analyzer Corpus-Oriented"*. In *ESPRIT BRA-7315 Acquilex II*, Working Paper 31.

[9] Gonzalo, J., Chugur, I. and Verdejo, F. (2000), "Sense Clusters for Information Retrieval: Evidence from Semcor and the InterLingual Index". In *Proceedings of the ACL'2000 workshop on Word Senses and Multilinguality*, Hong-Kong. Available at http://sensei.ieec.uned.es/~julio/publications.html [12/4/2001]

[10] Djoerd Hiemstra, Wessel Kraaij, Renée Pohlmann, and Thijs Westerveld (2000), "Twenty-One at CLEF-2000: Translation resources, merging strategies and relevance feedback", In *Working Notes for CLEF 2000 Workshop* . Available at: http://www.iei.pi.cnr.it/DELOS/CLEF/sheffield.doc [1/4/2001].

[11] D. Hiemstra and W. Kraaij (1999). "Twenty-One at TREC-7: Ad-hoc and cross-language track". In *Proceedings of the seventh Text Retrieval Conference TREC-7*, NIST Special Publication 500-242, pages 227–238.

[12] Martin Holub and Alena Böhmová (2000). "Use of Dependency Tree Structures for the Microcontext Extraction". In *ACL'2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*., Hong-Kong.

[13] ZPrise, developed by Darrin Dimmick (NIST) . Available on demand at http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html [2/6/2001]

[14] L.A. Ureña, M. Buenaga y J.M. Gómez (2001). "Integrating linguistic resources in TC through WSD. In *Computers and the Humanities*, 35/2, pp. 215-230. May 2001.

[15] CLEF, Cross Language Evaluation Forum, http://galileo.iei.pi.cnr.it/DELOS/CLEF/index.html [2/6/2001]

[16] Gonzalo, J. (2001). "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*, Lecture Notes in Computer Science, Springer-Verlag.