Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain

X. Saralegi, I. San Vicente, A. Gurrutxaga

Elhuyar R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country
{xabiers, inaki, agurrutxaga}@elhuyar.com

Abstract

In the literature several approaches have been proposed for extracting word translations from comparable corpora, almost all of them based on the idea of context similarity. This work addresses the aforementioned issue for the English-Basque pair in a popular science domain. The main tasks our experiments focus on include: designing a method to combine some of the existing approaches, adapting this method to a popular science domain for the English-Basque pair, and analyzing the effect the comparability of the corpora has on the results. Finally, we evaluate the different prototypes by calculating the precision for different cutoffs.

1. Introduction

In the literature several strategies have been proposed for extracting lexical equivalences from corpora. Most of them are designed to be used with parallel corpora. Although these kinds of corpora give the best results, they are a scarce resource, especially when we want to deal with certain language pairs and certain domains and genres. As a solution to this limitation the first algorithms (Rapp 1995, Fung 1995) were developed for automatic extraction of translation pairs from comparable corpora. These kinds of corpora can be easily built from the Internet.

The techniques proposed for the extraction task are mainly based on the idea that translation equivalents tend to co-occur within similar contexts. An alternative is to detect translation equivalents by means of string similarity (cognates). Nevertheless, none of these techniques achieve the precision and recall obtained with the parallel corpora techniques.

This work focuses on the Basque-English pair and popular-science domain. Taking this scenario as the starting point, we channeled our efforts towards designing a hybrid approach to the methods proposed in the literature, adapting it to the scenario, and designing a measure to compute the comparability of a corpus. The results of the techniques applied to comparable corpora depend on the degree of comparability of a corpus. Hence, a proper measure is a determining factor to evaluate the adequacy of a corpora for terminology extraction.

2. Comparable Corpora

Comparable corpora are defined as collections of documents sharing certain similar characteristics and written in more than one language. In bilingual lexicon extraction some of these characteristics depend on the lexicon type we aim to extract. Thus, achieving a high degree of comparability with regard to these characteristics is very important, since context similarity techniques will be more effective. The more similar the corpora are, the higher the comparability between the collocated words of the equivalent translations (Morin et al. 2007).

In order to guarantee this comparability fully, we believe a global measure that takes different aspects relating to global comparability into account needs to be designed.

This work focuses on bilingual comparable corpora in popular science, that is, the domain is 'science' and the type of discourse is 'news for non-specialized readers'. Besides these two main aspects, there are other characteristics that are related to the degree of comparability, such as distribution of topics and publication dates. All of them can be measured in order to estimate the global comparability of the corpora. Our hypothesis is that the comparability correlates with both the presence of word translations and the comparability of their contexts or collocates.

We introduce a method to compute the similarity between corpora, based on the Earth Movers Distance (EMD) (Rubner et al. 1997). This measure has been used to compute document similarity (Wang and Peng 2005). Section 4.1 further explains our strategy behind using this measure.

3. Identification of Equivalents

3.1. Context Similarity

The main method is based on the idea that the same concept tends to appear with the same context words in both languages, that is, it maintains many collocates. It is the same hypothesis that is used for the identification of synonyms. There are various approaches for implementing this technique. Problems arise with low frequency words, polysemous words and very general words, because they are difficult to represent. The representativity of the context vectors depends on the representativity of the corpus. However, we are only interested in the comparability of the context vectors, so while the representativity of the corpus is a significant problem, it is nevertheless a secondary one. The methods based on context similarity consist of two steps: modeling of the contexts, and calculation of the degree of similarity using a seed bilingual lexicon (Rapp 1999, Fung 1998).

The majority of the methods for modeling are based on the "bag-of-words" paradigm. Thus, the contexts are represented by weighted collections of words. There are several techniques for determining which words make up the context of a word: distance-based window, syntactic based-window (Gamallo 2007). Different measures have been proposed for establishing the weight of the context words with regard to a word: Log-likelihood ratio (LLR), Mutual Information, Dice coefficient, Jaccard measure,

frequency, tf-idf, etc. Another way of representing the contexts is by using language models (Shao et al. 2004).

After representing word contexts in both languages, the proposed algorithms compute for each word the similarity between its context vector and all the context vectors in the other language by means of measures such as Cosine, Jaccard or Dice. According to the hypothesis, the correct translation should be ranked in the first positions. To be able to compute the similarity, the context vectors are put in the same space by translating one of them. This translation can be done by using dictionaries or statistical translation models.

3.2. Cognates

Another technique proposed in the literature is the identification of translations by means of cognates (Al-Onaizan and K.Night 2002). This method could be appropriate in a science domain where the presence of cognates is high. In fact, using a Basque-English technical dictionary we were able to calculate automatically that around 30% of the translation pairs were cognates. Dice coefficient or LCSR (Longest Common Subsequence Ratio) measures are proposed for computing string similarity.

4. Experiments

4.1. Measuring the Comparability Degree of Corpora

The degree of comparability between two corpora depends on several features of their texts (document topics, publication dates, genre, corpus size, etc.), and certain criteria must be adopted to tackle the problem of measuring comparability. Besides, the criteria depend on the target of the task and the methodology used to achieve that target. Our objective is to extract bilingual terminology from popular science texts by using a method based on comparing contexts of words. Therefore, we need a method to guarantee a minimum amount of comparable contexts of translation equivalents.

There are few works in the literature on this topic, and they do not deal with the impact of comparability on terminology extraction. Among them, (Kilgarriff 1998) evaluates certain measures and concludes that techniques based on word frequency information perform better. These techniques extract lists of the most frequent *n* words appearing in both corpora, and then these are compared by means of Hypothesis Tests. While (Kilgarriff 1998) uses raw word lists, (Rayson & Garside 2000) also tests POS tag lists and semantic tag lists.

We aim to find a measure which can tell how similar two corpora are; what is meant by *similar* is that the corpora are semantically alike on a document level. The more similar the documents are, the more similar the contexts of the words should be and hence, the performance of the term extraction process is expected to improve.

The method we propose in order to obtain a degree of comparability between two corpora takes the document as a unit for comparison. Let us say that the corpus C_1 (Basque) has *m* documents eu_i (where $i \in 0..m$) and the corpus C_2 (English) has *n* documents en_j (where $j \in 0..n$). Document similarity is computed for all of the inter-corpora document pairs, using *Dokusare, a* tool for cross-lingual similarity measuring described in (Saralegi and Alegria 2007). As a result, we obtain a *nxm* matrix (*DM*), where each $d_{i,j}$ entry corresponds to the content similarity between eu_i and en_j. This matrix is passed as a parameter to the EMD, which calculates the global similarity score.

$$DM = \begin{array}{c} \begin{array}{ccccc} en_1 & .. & en_j & .. & en_m \end{array} \\ \begin{pmatrix} d_{11} & .. & d_{1j} & .. & d_{1m} \\ .. & .. & .. & .. & .. \\ d_{i1} & .. & d_{ij} & .. & d_{3} \\ .. & .. & .. & .. & .. \\ d_{n1} & .. & d_{nj} & .. & d_{nm} \end{pmatrix} \end{array} \begin{array}{c} eu_1 \\ .. \\ eu_i \\ .. \\ eu_n \end{array}$$

Where DM is the matrix storing distance between documents computed using Dokusare.

$$p_j = en_j$$
$$q_i = eu_i$$
$$P = \{(p_1; w_{p_1}), \ldots, (p_m; w_{p_m})\} = \{(en_1; 1/m), \ldots, (en_m; 1/m)\}$$
$$Q = \{(q_1; w_{q_1}), \ldots, (q_n; w_{q_n})\} = \{(eu_1; 1/n), \ldots, (eu_n; 1/n)\}$$

We want to find a flow $F = [f_{ij}]$ with f_{ij} being the flow between p_i and q_j, which minimizes the overall cost

$$WORK(P; Q; F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

constraints:

$$f_{ij} \geqslant 0, 1 \leqslant i \leqslant m; 1 \leqslant j \leqslant n$$
$$\sum_{j=1}^{n} f_{ij} \leqslant w_{pi}; 1 \leqslant i \leqslant m$$
$$\sum_{i=1}^{m} f_{ij} \leqslant w_{qi}; 1 \leqslant j \leqslant n$$
$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(\sum_{i=1}^{m} w_{pi}; \sum_{j=1}^{n} w_{qj})$$

The EMD is defined as the work normalized by the total flow:

$$EMD(P; Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}$$

4.2. Term Extraction from Comparable Corpora

4.2.1. Preprocess

We needed to identify the words we considered to be meaningful for our process, that is, content-words. POS tags were used for this task. Treetagger is the tagger we chose to tag the English corpus and Eustagger in the case of the Basque corpus. Only nouns, adjectives and verbs are regarded as content words. In our experiments, adverbs were found to produce noise. Proper nouns also

produced noise due to a cultural bias effect. Both were removed.

4.2.2. Vector-contexts Construction

We established a window depending on the POS of the word being focused on. The window size was determined empirically: 10 words for Basque (plus and minus 5 around a given word) and 14 for English (plus and minus 7). Furthermore, our experiments showed that using punctuation marks to delimit the window improved the results. Therefore, we also included this technique in our system.

We calculated the weight of the words within the context by means of the absolute frequency, LLR, Dice coefficient or Jaccard measure, and then contexts were modeled in a vector space. The best results were achieved by using the LLR. In addition, experiments were conducted combining the LLR with a distance factor between the center word *x* and the word *y disfactor(x,y),* for which the weight was being calculated:

$$LLRmod(x,y) = LLR(x,y) * disfactor(x,y)$$

The distance factor increases hyperbolically when the average distance between *x* and *y* decreases. We adopted this strategy to penalize the words farther from the center word, because the farther two words are from each other the weaker their relation is.

4.2.3. Context Vector Translation

To compute the translation of a Basque word, we translated its context vector in order to make it comparable with English context vectors. A bilingual Machine Readable Dictionary (MRD) was used for this purpose. If a word had more than one translation, we included all of them in the translated context vector, since the English equivalents were not sort by frequency of use. Our hypothesis is that the probability of concurrence of wrong translations in an English context-vector is low, and consequently, the first positions of the similarity-ranking are not distorted. In the case of the cosine distance, vectors were normalized before translation in order to prevent the noise produced by hypothetically wrong translations. Otherwise, the recall of the MRD determines the representativity of the context vector. In our experiments with a general dictionary, the average translation recall by vector was 55%. The higher the recall the greater the possibilities of finding the right translation for a word, because context vectors held more detailed information about the word in question.

To increase the recall of our translated vectors, we try to find equivalents not included in the dictionary by means of cognates. For all the Out Of Vocabulary (OOV) words, we looked for cognates among all the context words in the target language. The identification of these cognates is made by calculating the LCSR between the Basque and English context words. Before applying the LCSR, we processed some typographic rules to normalize equal phonology n-grams (e.g., *ph→f phase=fase*) or regular transformation ones (e.g., *-tion→-zio*, *action=akzio*) in both equivalent candidates. The candidates that exceeded a certain threshold (0.8, determined after several tests) were taken as translations.

4.2.4. Context Similarity Calculation

To obtain a ranked list of the translation candidates for a Basque word, we calculated the similarity between its translated context vector and the context vectors of the English words by using different similarity measures (Dice coefficient, Jaccard measure and Cosine). The best results were obtained with cosine. Furthermore, to prevent noise candidates, we pruned those that had a different grammatical category from that of the word to be translated.

4.2.5. Equivalent Similarity Calculation

In addition to context similarity, string similarity between source words and equivalent candidates is also used to rank candidates. LCSR is calculated between each source word and its first 100 translation candidates in the rank obtained after context similarity calculation. LCSR is applied in the same way as in context vector translation.

When used in combination with context similarity, LCSR data is used as the last ranking criteria. The candidates that exceeded the 0.8 threshold are ranked first, the remaining candidates not changing their positions in the rank. A drawback to this method is that cognate translations are promoted over the translations based on context vector similarity.

5. Evaluation

5.1. Building Test Corpora

We built two corpora with different characteristics in order to analyze the effect that comparability has on the results. The sources of the documents were science information web-sites. Zientzia.net (Basque), Sciam.com. AlphaGalileo, BBC News, ESA, EurekAlert!, NASA, New Scientist, news@nature, and ScienceNOW (English).

Zientzia.net and Sciam.com are quite similar with respect to the distribution of topics and register, so we chose them to build the first corpus (test corpus A). A correlation between topic and date was expected and for that reason we downloaded only all news items between 2000 and 2008. Moreover, other types of documents like articles, dossiers, etc. were rejected in order to maintain the same register throughout the corpus. Finally, the HTML documents were cleaned and converted into text using Kimatu (Saralegi & Leturia 2007). The size of this corpus was 1,092 million tokens for Basque and 1,107 for English. The distribution of the documents among the domains was comparable (table 1).

We built a second corpus (test corpus B), aiming for a lower comparability degree. We tried to unbalance important characteristics for the comparability degree like distribution among dates, topics and sources. We took the test corpus A as a starting point and randomly removed 1,000 documents from each language. In order to produce the bias we introduced 1,000 Basque news items from Zientzia.net belonging to the 1985-2000 period, and 1,000 English news items from the sources other than Sciam belonging to the 2007-2008 period. All new HTML documents were also cleaned and converted into text by Kimatu. The size of this corpus was 1,106 million tokens for Basque and 1,319 for English.

Domain	Sciam	Zientzia.net
Health, Mind & Brain	15.99%	14.85%
Space	9.83%	9.17%
Technology & Innovation	8.53%	15.40%
Biology	16.29%	28.35%
Earth & Environment, Archaeology & Paleontology	22.25%	17.88%
Physics, Chemistry, Math	11.15%	5.95%
History of Science, Society & Policy	15.96%	8.41%

Table 1: Domain distribution of documents for test corpus A.

The degree of comparability was computed using the EMD for both corpora. The value obtained for test corpus B was higher than the one obtained for the test corpus A. However, it was not as high as we expected. We are aware that these are only relative values, since there is no reference or threshold to compare them with. Anyway, the EMD value obtained in both cases is far from 0, which would indicate the maximum comparability degree. These high values are partly due to the rigorousness of Dokusare for calculating content similarity.

corpus	#word		#doc		EMD
	eu	en	eu	en	
Test corpus A	1,092K	1,107K	2,521	2,900	0.84
Test corpus B	1,106K	1,319K	2,521	2,900	0.86

Table 2: Characteristics of test-corpora

5.2. Tests

For the automatic evaluation of our system, we need a list of Basque-English equivalent terms occurring in each part of the corpora and which are not included in the dictionary used for the translation of content words in the construction of context vectors. To build that list, firstly we take all the Basque content words obtained in the preprocess step for the two built corpora. Secondly, those words are searched in the Basque-English Morris dictionary[1], and, for all the Basque words not included in that dictionary, we randomly select 200 pairs of words that reached a minimum frequency (10) and which appeared in two terminology Basque-English dictionaries (*Elhuyar Science and Technology Dictionary*[2] and Euskalterm terminology bank[3]).

This enabled us to estimate the precision automatically. In order to analyze the impact the frequency has on the results, we divide this set in two subsets. The first one includes words of high frequency (>50), and the other one, medium-low frequency words (within the 10-30 frequency range).

We also analyze the effect that the dispersion of the source test-words across the domains has on the precision of the system. Some scholars have pointed out the existence of a general academic vocabulary (Coxhead 2000) or a *lexique scientifique transdisciplinaire* (Drouin 2007). Those kinds of words are widely used in science-domain texts but do not belong to a specific domain. Therefore, the contexts of those words are, in principle, more heterogeneous than the contexts of specialized terms, and it is reasonable to suppose that they will be more difficult to extract. To analyze this effect, we calculated the correlation between the position of the target word in the ranking and the dispersion of the source word across the domains. We measured this dispersion by computing the coefficient of variation (CV) of the frequency of the source word across the domains. The reference domain list is the one used in Zientzia.net to classify news:

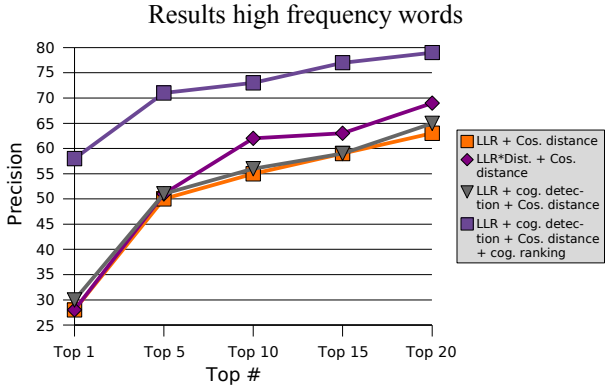
- Biology
- Space
- Physics, Chemistry, Math
- Computer science
- Earth sciences
- Environment
- Health
- Technology
- General

We analyzed different variables: the comparability of the corpus, the modeling of the contexts, and the way to combine the different approaches.

- Comparability: we processed the two test corpora in order to analyze the effect of the degree of comparability has on the results
- Modeling of contexts: Association Measures (AM), techniques to reduce OOVs
- Combining methods: context similarity, cognates

5.3. Results

Figures 1 and 2 show the results for both test corpora.

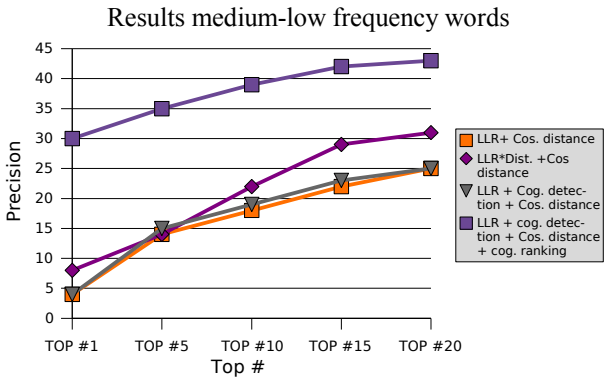


a)

1 English/Basque dictionary including 67,000 entries and 120,000 senses.

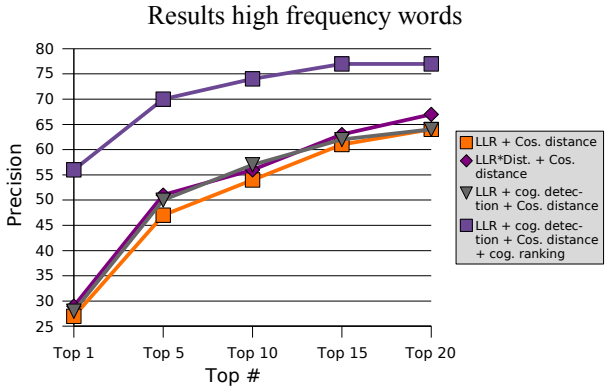
2 Encyclopaedic dictionary of science and technology including 15,000 entries in Basque with equivalences in Spanish, French and English.

3 Terminological dictionary including 100,000 terms in Basque with equivalences in Spanish, French, English and Latin.

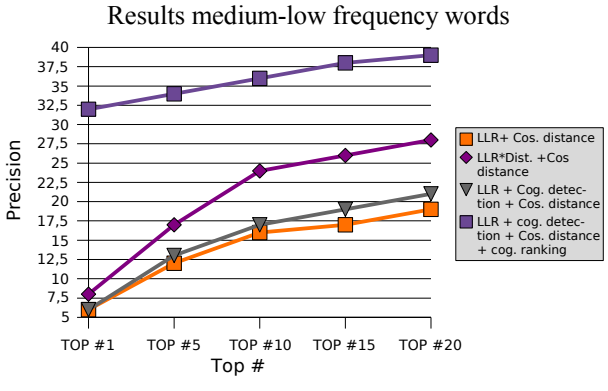


b)

Figure 1: Precision results for test corpus A. Context similarity (cosine) combined with and without cognates detection during the vector translation phase (LCSR>0.8) and/or the ranking phase. Weighting the words in context vectors according to their distance from the centre word is also presented here.



a)



b)

Figure 2: Precision results for test corpus B.

In general, the precision obtained for the test corpus A is slightly better than the one obtained with the test corpus B. Although the difference is small, we can observe the influence of the degree of comparability on the precision. Another aspect that should be evaluated is the relation between the degree of comparability and the recall. As we mentioned in section 1, our hypothesis is that the comparability degree correlates with both the presence of word translations and the comparability of their contexts. In any case, more experiments must be carried out to deeper analyze these relations.

We have observed that combining the identification of cognates in the list of equivalents with context similarity (as proposed in section 4.2.5) improves the precision of the final rank. The high presence of these kinds of translations explains this improvement.

The detection of cognates in the translation of the context-vectors slightly outperforms translation based exclusively on dictionaries. Besides, the use of the distance factor together with the LLR also improves precision slightly, specially in the case of medium-low frequency words. This fact can be explained on the ground that co-occurrence data could not be enough to estimate correct association degree for the context words.

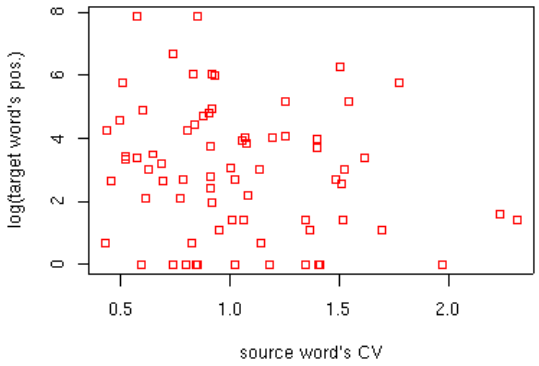


Figure 3: Dispersion diagram for source word's CV and target word's rank position

Figure 3 shows some results of the experiments done to measure the influence of the domain specificness of a source word on the rank position of the target word (corresponding to the LLR+Cos. distance experiment of Figure 1. a). There is no statistically significant correlation, contrary to our initial suspicion. There is no clear relation between the heterogeneity of the context of a word and its domain specificness, and therefore we could conclude that this factor does not have a significant effect on extraction based on context similarity calculation. Nevertheless, we think that a deeper analysis needs to be conducted in order to characterize difficult words, e.g. by analyzing the dispersion of frequency across the senses.

6.Conclusions

We've developed the first experiments towards terminology extraction from comparable corpora integrating different existing techniques and adapted them for a new language pair. The combination of the cognates detection in the final ranking as well as in the translation process of the context vectors seems suitable for corpora of science domain where the presence of cognates is high. On the other hand, our corpora are relatively small by current standards, and this leads to a significant decrease in the recall, since very few words reach the minimum frequency threshold necessary to obtain good precision in context similarity based extraction. In fact, in our test corpora only around 18% of the unknown source words

(Basque) reaches a frequency of 10. So, the maximum recall we could obtain is low.

As for the building of corpora, we have analyzed the importance of taking into account certain criteria in order to build comparable corpora for the terminology extraction task. Specifically, we have analyzed the effect that data and domain distribution also have on the degree of comparability and on the precision of the extraction process. The experiments we carried out showed a small effect. This could be due to the fact that the bias we induced in the test corpus B was not strong enough. Besides, we presented a new measure to quantify the degree of comparability, based on the EMD. Nevertheless, only preliminary experiments were conducted with this measure, and so further tests need to be done in order to tune it and ensure its reliability.

7. Future Work

We plan to build bigger corpora for the next experiments. To tackle the problems less-resourced languages like Basque have, we plan to use the Internet as the source of corpora as SIGWAC[4] suggests. So we are currently designing methods for building comparable corpora from the web.

Otherwise, we plan to extend our experiments to other languages, like Spanish, German and French.

In order to improve the extraction process, on the one hand, techniques for correct translation selection based on monolingual co-occurrences models will be integrated into the context vector translation process. On the other hand, we are planning to experiment with probabilistic models to represent contexts.

References

Coxhead, A. (2000). "A new Academic Word List." In *TESOL Quarterly*, 34.

Déjean, H, Gaussier, E & Sadat, F. (2002) "An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction." In *COLING 2002*.

Druoin, P. (2007). "Identification automatique du lexique scientifique transdisciplinaire." In Tutin, A. (Ed.) *Lexique des écrits scientifiques.* Revue Française de Linguistique Appliqué. Volume XII-2

Fung, P. (1995) "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus." In *Proceedings of the Third Workshop on Very Large Corpora*, p.173-183, Boston, Massachusetts.

4 The Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus

Fung, P. and Lo Yuen Yee (1998) "An IR Approach for Translating New Words from Nonparallel Comparable Texts." In *COLING-ACL* 1998: 414-420.

Gamallo, P. (2007) "Learning Bilingual Lexicons from Comparable English and Spanish Corpora." In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, pp. 191-198.

Kilgarriff, A., Rose, T. (1998) "Measures for corpus similarity and homogeneity." *In Proc. 3rd Conf. on Empirical Methods in Natural Language Processing* (EMNLP-3). Granada, Spain, June: 46-52.

Morin, E., Daille, B., Takeuchi, K. and Kageura, K. (2007) "Bilingual Terminology Mining - Using Brain, not brawn comparable corpora." In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, June 2007, Prague, Czech Republic, ACL p. 664-671.

Rapp, R. (1995) "Identifying word translations in non-parallel texts." In *ACL*, p.320-322.

Rapp, R. (1999) "Automatic identification of word translations from unrelated English and German corpora." In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, p.519-526, June 20-26, 1999, College Park, Maryland.

Saralegi, X. and Alegria, I. (2007) "Similitud entre documentos multilingües de carácter técnico en un entorno Web." In *SEPLN 2007.* Sevilla. p.71-78.

Saralegi, X. and Leturia, I. (2007) "Kimatu, a tool for cleaning non-content text parts from html docs." In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain,* pp. 163—167.

Shao, Li and Ng, Hwee Tou (2004) "Mining New Word Translations from Comparable Corpora." In Proceedings of the 20th *International Conference on Computational Linguistics* (COLING 2004). (pp. 618-624). University of Geneva, Geneva, Switzerland.

Rubner, Y., Guibas, L.J. and Tomasi, C. (1997) "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval." In *Proceedings of the ARPA Image Understanding Workshop,* New Orleans, LA, May 1997, pp. 661-668.

Wan, X. and Peng, Y. (2005) "The Earth Mover's Distance as a Semantic Measure for Document Similarity." In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp.301– 302.

Rayson, P. and Garside, R. (2000) "Comparing corpora using frequency profiling." In *Proceedings of the workshop on Comparing Corpora (38th ACL),* Hong Kong, pp. 1–6.