Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet

Igor Leturia, Iñaki San Vicente, Xabier Saralegi
Elhuyar Fundazioa R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea
20170 Usurbil. Basque Country
{igor, inaki, xabiers}@elhuyar.com

Abstract

In this paper we propose using search engine queries for collecting bilingual specialized comparable corpora from the Internet, an alternative to using news agencies or focused crawling which will supposedly obtain more varied corpora. The method we propose for obtaining specialized corpora on a language is based on the BootCaT method (querying search engines for random combinations of a list of seed words representative of the domain or topic and retrieving the pages returned) but, instead of the seed words, a sample mini-corpus is used as a basis for the process: most representative words are automatically extracted from it, and a final domain-filtering step is performed using document-similarity measures with this sample corpus. For obtaining the bilingual comparable corpora, two different variants of this method are proposed. One of them uses a sample mini-corpus for each language and launches the corpus-collecting processes for each language independently. The other uses only a sample mini-corpus in one of the languages, and uses dictionaries for translating the extracted seed words and performing the topic filtering for the other language. We have collected two domain-specific Basque-English comparable corpora with each of the methods, and evaluated them using corpus similarity measures.

1 Motivation

Corpora of any type are a very valuable resource for linguistic research, for language standardization and for the development of language technologies. This is more so in the case of Basque, since its standardization and normalization process begun only very recently and since language technologies for Basque are not as advanced as for other languages. But being a small language in terms of number of speakers and economic resources dedicated to it, the Basque language is not exactly rich in corpora.

So far, most of the corpora-building effort for Basque has been put into general monolingual corpora, which is completely logical, since the first step for the normalization of the language was the standardization of the general lexicon. Nowadays, although few and small compared to other languages (25 million words at most), there exist some general corpora in Basque: XX. mendeko Euskararen Corpusa[1], Ereduzko prosa gaur[2] and Klasikoen gordailua[3] are the most significant.

Now that the Academy of the Basque Language has finished with the general lexicon, and that Basque has entered universities and the labor world, there is a great need for specialized corpora, in order to normalize terminology. So far there have been two specialized corpora projects: Zientzia eta Teknologiaren Corpusa[4] (Areta *et al.*, 2007) and a project for an automatic collector of Basque specialized corpora from the Internet (Leturia *et al.*, 2008a).

Over the last years, the development of language technologies has also brought about a need for multilingual corpora, whether general or specialized, for their use in automatic terminology extraction, statistical machine translation training, etc. The Basque language has hardly any resources of this kind, except for some translation memories from public bodies,

[1] http://www.euskaracorpusa.net
[2] http://www.ehu.es/euskara-orria/euskara/ereduzkoa/
[3] http://klasikoak.armiarma.com/
[4] http://www.ztcorpusa.net

the majority of which are small and Basque-Spanish only.

However, other languages encounter this problem too, particularly for specialized areas. That is why comparable corpora are becoming increasingly popular. Although more difficult to exploit for the mentioned tasks than parallel corpora (because of their smaller alignment level, there is less explicit knowledge to extract), they are easier to obtain in large sizes, and so they also have the potential to overcome the limitations of parallel corpora, as recent research in fields such as machine translation (Munteanu and Marcu, 2005), bilingual terminology extraction (Fung and Yee, 1998; Rapp, 1999) or cross-language information retrieval (Talvensaari *et al.*, 2007) has shown. Systems that make use of this kind of corpora have also been developed for Basque (Saralegi *et al.*, 2008a; Saralegi *et al.*, 2008b). Thus the interest of an automatic tool for gathering comparable specialized corpora for Basque from the Internet.

Comparable corpora have traditionally been obtained on a supervised or directed way: out of news agencies, established research corpora (e.g. TREC or CLEF collections), by crawling certain web sites, etc. Both these approaches present some problems for our case. First, both of them need a human choice of the sources, which makes the corpora at least biased and often not very diverse. Besides, for small languages like Basque, in many domains, it would not be easy to identify good sources that would contain a significant amount of documents on the topic. Also, competition corpora do not usually include such languages. Finally, focused crawling for specialized corpora often requires domain filtering, usually based on machine learning, which needs special training for each topic, so building a generic tool for any domain is not possible. Therefore, our comparable corpora collection method is based on search engine querying.

2 Related work

2.1 Obtaining comparable corpora

Surprisingly, there is not much literature about the process of collecting comparable corpora. Most of the literature concerning comparable corpora deal with the exploitation of such resources, and simply mention that the comparable corpus has been obtained, as we have already mentioned, from news agencies (Barzilay and Lee, 2003; Munteanu and Marcu, 2005) or by crawling certain sites.

Talvensaari *et al.* (2008) do describe a system for obtaining comparable corpora based on focused crawling –the idea of focused crawling for monolingual specialized corpora was first introduced by Chakrabarti *et al.* (1999).

Some other works deal with converting comparable corpora from 'light' to 'hard' (Sheridan and Ballerini, 1996; Braschler and Schäuble, 1998; Bekavac *et al.*, 2004; Talvensaari *et al.*, 2008). The 'light' and 'hard' comparability levels for corpora were first introduced by Bekavac *et al.* (2004). A light comparable corpus would be composed of corpora from two (or more) languages composed according to the same principles (i.e. corpora parameters) which are defined by features such as domain, size, time-span, genre, gender and/or age of the authors, etc. The hard type comparability is dependent on already collected and established light comparable corpora. It is derived from them by applying certain language technology tools/techniques and/or document meta-descriptors to find out which documents in lightly comparable corpora really deal with the same or similar topic. A subset of lightly comparable corpora which has been selected by those tools/techniques, whether document-level aligned or not, can be regarded as a hard comparable corpora. Our interest, for the moment, relies on obtaining the light corpora, which again the aforementioned studies treat very superficially.

The approach most closely related to ours is that used by the BootCaT tool (Baroni and Bernardini, 2004), which introduced a new methodology for collecting monolingual domain-specific corpora from the Internet: give a list of words as input, query APIs of search engines for random combinations of these seed words and download the pages. This methodology has in some cases been used to build big general corpora (Sharoff, 2006), but for collecting smaller specialized corpora, it has become the *de facto* standard, replacing focused crawling. Although BootCaT is a monolingual corpora collector, we can expect that, by applying it to word lists on the same subject but in different languages, one could obtain light multilingual comparable corpora.

2.2 Measuring the quality of comparable corpora

The work described in this paper tries two different search engine based approaches for collecting comparable corpora from the Internet, and carries out an evaluation to see which performs best. In order to evaluate these performances, we need some way to measure the degree of comparability of a comparable corpus. However, the criteria to define comparability are not universal and depend on the type of comparable corpus we want and the task we want to use the corpus for. In our case, the comparability measure should somehow reflect domain or topic similarity and suitability for terminological extraction.

Again, the literature on this is scarce. Kilgarriff (1997) and Kilgarriff and Rose (1998) experiment with various methods for measuring corpus similarity based on word-frequency lists, and Rayson and Garside (2000) use also POS and semantic tag frequencies. But these methods are to be applied to monolingual corpora, not to multilingual comparable corpora.

Morin *et al.* (2007) suggest that, for the task of terminology extraction, the quality of a comparable corpus might be more important than its size, and show that they obtain better results with a smaller corpus if both subcorpora belong to the same register. So the genre or register could be another criterion to weigh the comparability. But word-frequency lists are not valid features for genre identification; punctuation marks and POS trigrams should be used for this task (Sharoff, 2007; Argamon *et al.*, 1998). Anyway, domain similarity is more important for terminology extraction than genre or size, so at the moment we are more interested in the former kind of comparability.

Finally, Saralegi *et al.* (2008b) propose measuring the comparability of a corpus by computing the semantic similarities at the document level. The hypothesis behind this is that the containment of many document pairs with a fairly high semantic similarity improves terminology extraction based on context similarity. So this method somehow measures the 'hardness' of 'light' comparable corpora.

3 Our approach

The aim of our research project is to develop a methodology to collect domain-specific comparable corpora from the Internet, using a search engine based approach similar to that of BootCaT. For the moment, our interest is in Basque-English corpora, but the method should work for any language pair.

The first condition, necessary but not sufficient, for two corpora to be considered domain-comparable is, obviously, that they belong to the same domain. The BootCaT tool and method can be used to obtain two such domain-specific corpora in different languages. But any loss or non-perfection in the domain-precision obtained in each of them affects the quality of the comparable corpus. The few studies that the authors have found on the topic precision obtained by BootCaT's word-list method show that this is not at all perfect (Baroni and Bernardini, 2004; Leturia *et al.*, 2008a). Thus, maximizing the domain-precision of each of the corpora obtained is one of our goals.

Then, even if both corpora were 100% domain-specific, this is not enough to guarantee a good comparability. Out of two corpora strictly on computer sciences, one could be mostly made out of texts on hardware and databases and the other on programming and networks; they could not be considered very comparable, and they would most surely not be very practical for any of the aforementioned tasks. Therefore, we are also interested in obtaining corpora as comparable as possible.

3.1 Maximizing domain precision in monolingual corpus collection

In order to try to improve the domain-precision of the BootCaT method, our approach takes, as a starting point, a sample mini-corpus of documents on the topic, instead of a list of words. This mini-corpus has two uses: first, the list of keywords to be used in the queries is automatically extracted from it; second, it is used to filter the downloaded documents according to the domain by using document-similarity techniques (Lee *et al.*, 2005).

Apart from this main contribution, we have also added some other improvements, some of them general and some others that are applied only for obtaining a better performance when the Basque language is involved.

Next we will describe the whole process we use for obtaining monolingual domain-specific corpora, which is the same as in the work of Leturia *et al.* (2008a), step by step and in more detail:

- Sample mini-corpus collection: The sample mini-corpus of documents on the

target domain, which is the basis of our system, has to be collected manually. The criteria when collecting the sample is that it should be as heterogeneous as possible and cover as many different subjects of the domain as possible.

- Automatic keyword extraction: The seed words to be used in the queries are automatically extracted from the sample corpus, with the same method as used by Saralegi and Alegria (2007). The mini-corpus is automatically lemmatised and POS-tagged, and then the most significant nouns, proper nouns, adjectives, verbs, entities and multiword terms are extracted by means of Relative Frequency Ratio or RFR (Damerau, 1993) and applying an empirically determined threshold. In order to maximize the performance of the queries, the extracted list can be revised manually, to remove too specific or too local proper nouns, words that are too general and polysemous words that have other meanings in other areas.

- Querying search engines and downloading: Random combinations of the extracted seed words are sent to the APIs of search engines and the pages returned are downloaded, just as in the BootCaT method. But some changes are introduced in the method when we want a corpus in Basque, because performance of search engines for Basque is very poor, mostly due to the rich morphology of the language and to the fact that no search engine can restrict its results to pages in Basque alone. We try to solve the former by means of morphological query expansion, which consists of querying for different word forms of the lemma, obtained by morphological generation, within an OR operator. In order to maximize recall, the most frequent word forms are used, and recall is improved by up to 60% in some cases. For the latter, we use the language-filtering words method, consisting of adding the four most frequent Basque words to the queries within an AND operator, which improves language precision from 15% to over 90% (Leturia *et al.*, 2008b). These techniques are common use in IR or web-as-corpus tools for Basque (Leturia *et al.*, 2007a; Leturia *et al.*, 2007b).

- Language filter: For filtering content that is not in the target language out of bilingual documents, we use LangId, a language identifier based on character and word trigram frequencies specialized in Basque, applied at paragraph level.

- Length filter: Filtering documents by length is an effective way of reducing noise (Fletcher, 2004). In our case, we reject documents the length of which after conversion to plain text is under 1,000 characters or over 100,000 characters.

- Boilerplate removal: This is another key issue in this project, not only because boilerplate (site headers, navigation menus, copyright notices, etc.) adds noise and redundancy to corpora, but also because it can affect subsequent stages (near-duplicate detection, domain filtering, etc.). For boilerplate removal, we use Kimatu (Saralegi and Leturia, 2007), a system developed by our team that scored very well (74.3%) in the Cleaneval competition (Baroni *et al.*, 2008).

- Near-duplicates and containment detection: We have also included a near-duplicate detection module based on Broder's shingling and fingerprinting algorithm (Broder, 2000), and a containment detection method also based on Broder's works (1997).

- Domain filtering: As we have said before, we perform a final domain filtering stage. We represent both the downloaded documents and each of the documents of the sample corpus with a vector of the most significant keywords, i.e. nouns, proper nouns, adjectives and verbs. These were extracted using Eustagger, a POS-tagger for Basque (Aduriz *et al.*, 1996). The keywords are selected and weighed by some frequency measure, such as Log Likelihood Ratio or the aforementioned RFR. For measuring the similarity we use the cosine, one of the most widely used ways to measure the similarity between documents represented in the vector space model. A document is accepted in the corpus if the maximum of its cosine measures with each of the documents in the sample mini-corpus reaches an empirically defined threshold, and rejected otherwise.

3.2 Collecting multilingual corpora

With the method described above and a topic-filtering threshold that is high enough, we can obtain monolingual specialized corpora with a very high domain precision (Leturia *et al.*, 2008a). For obtaining a specialized bilingual comparable corpus, we have tried two different variants of applying this method to two different languages, which are explained below.

Different sample corpora method

The most obvious way is to use a sample mini-corpus for each language and launch the corpus collecting process independently for each of them. If the sample mini-corpora used are comparable or similar enough (ideally, a parallel corpus would be best), the corpora obtained will be comparable to some extent too (Fig. 1).

Dictionary method

The other method uses only a sample mini-corpus in one of the languages, and uses dictionaries for translating the extracted seed words (this is manually revised) and the domain-filtering vectors for the other language (Fig. 2).

This method, theoretically, presents two clear advantages: first, the sample mini-corpora are as similar as can be (it is only one), thus we can expect a greater comparability in the end; and second, we need only collect one sample corpus.

But in reality, it presents some problems too, mainly the following two: first, because dictionaries do not cover all existing terminology, we can have some OOV (Out Of Vocabulary) words and the method may not work so well –in our case, there are quite a few, although we use a combination of a general dictionary and a specialized one to maximize translation coverage –; second, we have to deal with the ambiguity derived from dictionaries, and selecting the right translation of a word is not so easy. These not at all trivial difficulties lead us to expect worse results from this method; nevertheless, we have also tried and evaluated it. To reduce the amount of OOV words, the ones that have been POS-tagged as proper nouns are included as they are in the translated lists, since most of them are named entities. And for resolving ambiguity, for the moment, we have used a naïve "first translation" approach, widely used as a baseline in NLP tasks that involve translation based on dictionaries. The basic idea this relies on is that many dictionaries order their translations according to the frequency of use.

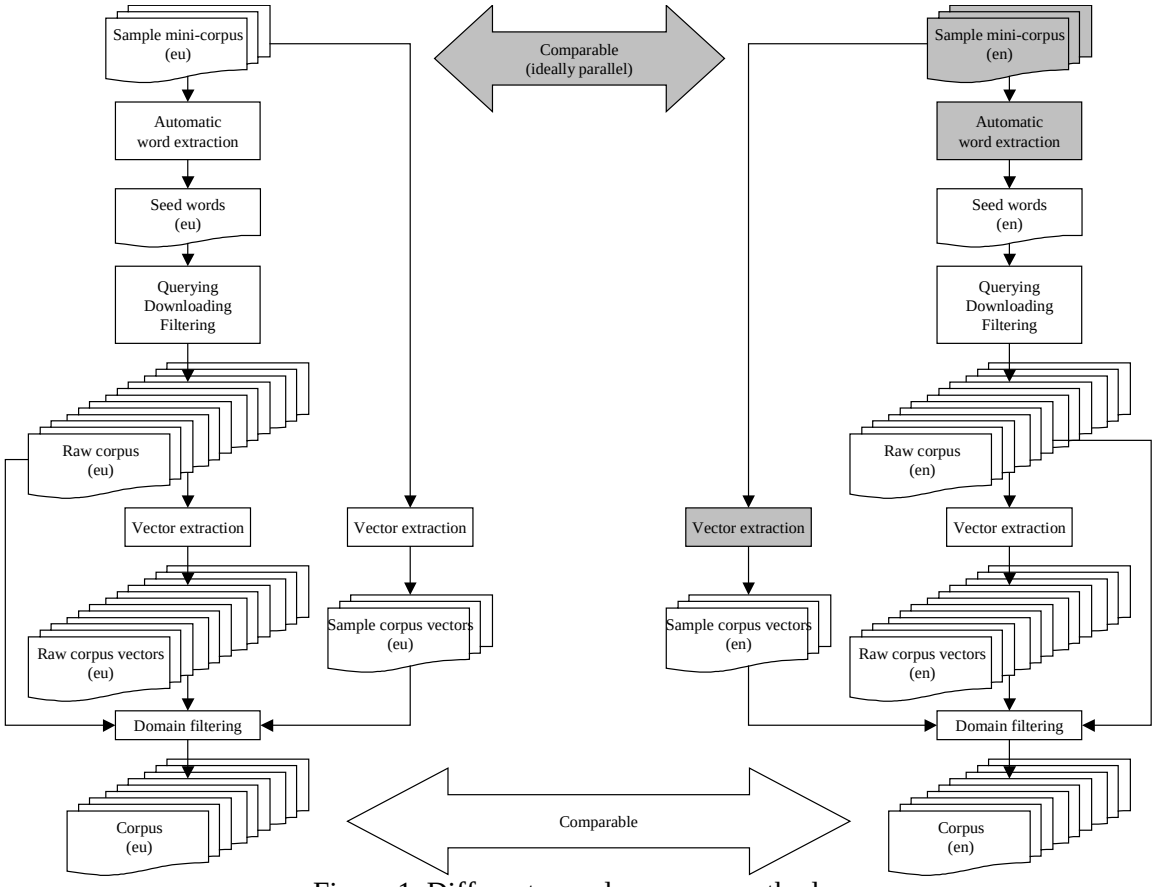

Figure 1. Different sample corpora method

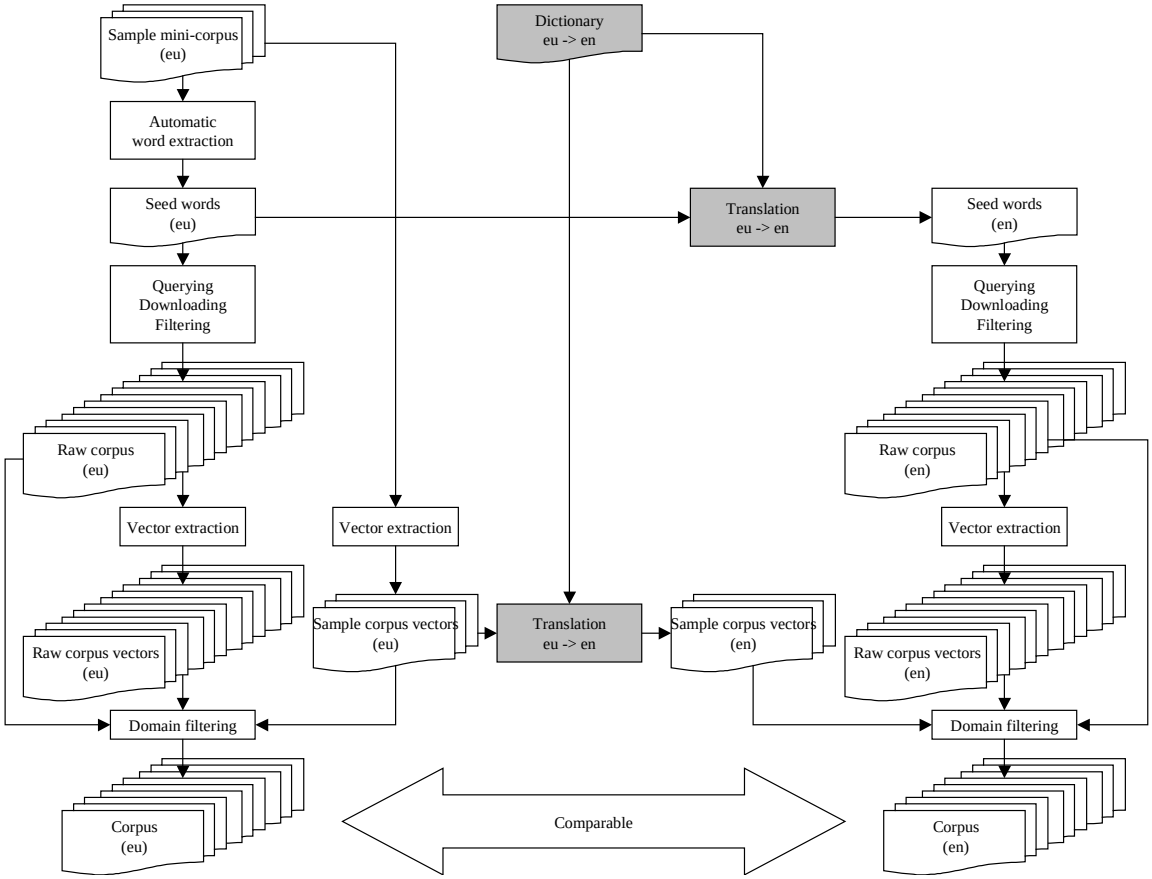

Figure 2. Dictionary method

4 Evaluation

In order to see which of the two methods obtains a higher degree of comparability, we collected two Basque-English comparable corpora, one on computer sciences and the other on tourism, with each of the two methods mentioned above. The sample mini-corpora used for computer sciences are 33 short articles (about 40,000 words) obtained from popular science magazines, and for tourism 10 short articles (about 7,000 words) obtained from tourist office websites. The English versions of the sample mini-corpora are comparable in the case of computer sciences, and parallel in the case of tourism. The final size of the computer sciences corpora amounts to 2.5 million words in each language, and in the case of tourism, 1.5 million words.

Then, for evaluating the two methods, we used two different ways to measure the comparability of the four corpora obtained: one is by calculating the cosine distance between the vectors containing all the keywords of each corpora weighted by LLR; the other is by calculating the Chi Square (χ^2) statistic for the n most frequent keywords, as described by Kilgarriff and Rose (1998). But it must be taken into account that, unlike any other corpora similarity measures mentioned in the literature, the corpora we compare are in different languages, so our measurement necessarily uses dictionaries; again, we resolve ambiguities with a first-translation approach for simplicity.

The results of the evaluation are shown in Table 1. For the cosine, higher values are better; for χ^2, a lower value indicates greater similarity. Best results are shown in bold.

Corpus	Method	Cosine, LLR, all keywords	χ^2, n most frequent keywords				
			500	1,000	5,000	50,000	All
Computer sciences	Different sample corpora	0.4102	700.61	481.57	148.70	17.60	16.55
	Dictionary	**0.4396**	**685.95**	**471.64**	**145.20**	**17.25**	**15.51**
Tourism	Different sample corpora	0.1164	382.80	**256.29**	**83.23**	**12.82**	**12.82**
	Dictionary	**0.1511**	**380.62**	261.78	86.35	13.00	13.00

Table 1. Evaluation results

5 Conclusions and future work

This paper has presented a search engine-based method for collecting specialized comparable corpora from the Internet, by obtaining two specialized, high domain-precision, monolingual corpora out of two sample mini-corpora. We tried a variant of this method that uses only one sample mini-corpus and dictionaries, to see if we could obtain similar or better comparability with less initial effort.

Although the dictionary method might *a priori* appear to be a worse method –owing to OOV words and ambiguity–, the evaluation does not confirm this. In fact, the dictionary method proved to be better in most of the measures. However, this evaluation cannot be considered conclusive, for the following reasons:

- The evaluation was done with only two corpora, which show different results for some of the measures. Besides, we now believe that tourism might not have been a good domain choice for the evaluation, because it does not completely fit into what we know as a specialized domain (interdisciplinary terminology, etc.). Evaluations with more corpora and more domains are needed before stating anything definite.
- There is not much literature on corpora similarity methods. Some measures have been proposed –mostly based on word frequency measures–, but they have not been sufficiently evaluated and indeed there is no standard measure. And regarding corpora in different languages, there is no precedent for measuring similarity. We have employed some of the proposed measures using dictionaries, and they show different results. We believe there is an urgent need for research on and standardization of multilingual corpora similarity methods.
- There might be a bias towards the dictionary method since we are using a dictionary to measure the similarity, too. To illustrate this we can imagine an extreme case, in which using the dictionary method all the seed words have been disambiguated incorrectly and the corpora obtained has nothing to see with the desired topic, but since the same dictionary and disambiguation method is applied to the keyword vectors when evaluating the similarity, the measure obtained might still be high. However, we do not see a solution for this.

For future work, we want to try to improve the dictionary-based approach; as we have already mentioned, the preliminary work needed to obtain a comparable corpus with this method is considerably reduced (only one sample mini-corpus needs to be collected); besides, there is still much room for improvement. One of the things to be tried is to see whether manual revision of the translated vectors to be used in the domain filtering yields a better performance. Another one is to try more complex translation selection techniques –instead of the first-translation approach–, and also synonymy expansion.

Furthermore, for monitoring the improvements in the methodology, we intend to make tests with more corpora and to perform further research on multilingual corpora similarity methods.

We also plan to apply the terminology extraction tool of Saralegi *et al.* (2008b) to corpora obtained with both methods and evaluate the results manually to see if our results on comparability correlate with terminological extraction tasks.

Finally, it would also be very interesting to implement a focused crawling method, download some corpora and compare the results of our method with those, to check whether the extra effort needed in focused crawling is compensated by the results.

References

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza and Ruben Urizar. 1996. EUSLEM: A Lemmatiser/Tagger for Basque. *Proceedings of EURALEX'96*, vol. I, 17-26. Euralex, Göteborg, Sweden.

Nerea Areta, Antton Gurrutxaga, Igor Leturia, Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza and Aitor Sologaistoa. 2007. ZT Corpus: Annotation and tools for Basque corpora. *Proceedings of Corpus Linguistics 2007*. University of Birmingham, Birmingham, UK.

Shlomo Argamon, Moshe Koppel and Galit Avneri. 1998. Routing documents according to style. *Proceedings of the International workshop on Innovative Internet Information Systems (IIIS-98)*, Pisa, Italy.

Marco Baroni, Francis Chantree, Adam Kilgarriff and Serge Sharoff. 2008. Cleaneval: a competition for

cleaning web pages. *Proceedings of LREC 2008*. ELRA, Marrakech, Morocco.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*, 1313-1316. ELRA, Lisbon, Portugal.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *Proceedings of HLT/NAACL 2003*, 16-23. NAACL, Edmonton, USA.

Božo Bekavac, Petya Osenova, Kiril Simov and Marko Tadić. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian & Croatian. *Proceedings of LREC 2004*, 1187-1190. ELRA, Lisbon, Portugal.

Martin Braschler and Peter Schäuble. 1998. Multilingual information retrieval based on document alignment techniques. *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, 183-197. Springer, Heraklion, Greece.

Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium*, 1-10. Springer, Montreal, Canada.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. *Proceedings of Compression and Complexity of Sequences 1997*, 21-29. IEEE Computer Society, Los Alamitos, California, USA.

Soumen Chakrabarti, Martin van der Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific web resource discovery. *Proceedings of the 8th International WWW Conference*, 545-562. W3C, Toronto, Canada.

Fred J. Damerau. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29:433-447.

William H. Fletcher. 2004. Making the web more useful as a source for linguistic corpora. *Corpus Linguistics in North America 2002*. Rodopi, Amsterdam, The Netherlands.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *Proceedings of COLING-ACL 1998*, 414-420. ACL, Montreal, Canada.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. *Proceedings of EMNLP-3*, 46-52. ACL SIGDAT, Granada, Spain.

Adam Kilgarriff. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings of workshop on very large corpora*, 231-245. ACL SIGDAT, Beijing and Hong Kong, China.

Michael D. Lee, Brandon Pincombe and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. *Proceedings of CogSci2005*, 1254-1259. Earlbaum, Stresa, Italy.

Igor Leturia, Iñaki San Vicente, Xabier Saralegi, Maddalen Lopez de Lacalle. 2008. Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. *Proceedings of the 4th Web as Corpus Workshop*, 40-46. ACL SIGWAC, Marrakech, Morocco.

Igor Leturia, Antton Gurrutxaga, Nerea areta, Eli Pociello. 2008. Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. *Proceedings of LREC 2008*. ELRA, Marrakech, Morocco.

Igor Leturia, Antton Gurrutxaga, Iñaki Alegria and Aitzol Ezeiza. 2007. CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. *Proceedings of the 3rd Web as Corpus workshop*, 69-81. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria and Aitzol Ezeiza. 2007. EusBila, a search service designed for the agglutinative nature of Basque. *Proceedings of Improving non-English web searching (iNEWS'07) workshop*, 47-54. SIGIR, Amsterdam, The Netherlands.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi and Kyo Kageura. 2007. Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 664-671. ACL, Prague, Czech Republic.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477-504.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 519-526. ACL, College Park, Maryland, USA.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora*, 1-6. ACL, Hong Kong, China.

Xabier Saralegi and Iñaki Alegria. 2007. Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39:71-78.

Xabier Saralegi, Iñaki San Vicente and Maddalen López de Lacalle, 2008. Mining Term Translations from Domain Restricted Comparable Corpora. *Proceedings of SEPLN 2008*, 273-280. SEPLN, Madrid, Spain.

Xabier Saralegi, Iñaki San Vicente, Antton Gurrutxaga. 2008. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. *Proceedings of Building and using Comparable Corpora workshop*, 27-32. ELRA, Marrakech, Morocco.

Xabier Saralegi and Igor Leturia. 2007. Kimatu, a tool for cleaning non-content text parts from HTML docs. *Proceedings of the 3rd Web as Corpus workshop*, 163-167. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*, 63-98. Gedit, Bologna, Italy.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification.. *Proceedings of the 3rd Web as Corpus Workshop*, 83-94. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Páraic Sheridan and Jean Paul Ballerini. 1996. Experiments in multilingual information retrieval using the SPIDER system. *Proceedings of the 19th Annual International ACM SIGIR Conference*, 58-65. ACM, Zurich, Switzerland.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems*, 25(1):4.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola and Jorma Laurikkala. 2008. Focused web crawling in acquisition of comparable corpora. *Information Retrieval*, 11:427-445.