Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision

I. Leturia, I. San Vicente, X. Saralegi, M. Lopez de Lacalle
Elhuyar Fundazioa, R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country
E-mail: {igor, inaki, xabiers, maddalen}@elhuyar.com

Abstract

The *de facto* standard process for collecting corpora from the Internet (with a given list of words, asking APIs of search engines for random combinations of them and downloading the returned pages) does not give very good precision when searching for texts on a certain topic. And this precision is much worse when searching for corpora in the Basque language, due to certain properties inherent in the language and in the Basque web.

The method proposed in this paper improves topic precision by using a sample mini-corpus as a basis for the process: the words to be used in the queries are automatically extracted from it, and a final topic-filtering step is performed using document-similarity measures with this sample corpus. We also describe the changes made to the usual process to adapt it to the peculiarities of Basque, alongside other adjustments to improve the general performance of the system and quality of the collected corpora.

1. Introduction

1.1 Motivation

Basque needs corpora more than many other bigger languages, as its standardisation began only very recently. And above all it is in need of specialized corpora, because terminology is the area with least *de jure* normalization. The only specialized corpus in Basque is the ZT Corpus (Areta et al., 2007), a corpus on Science and Technology that is a very valuable resource, but which does not fulfil all the needs of Basque for many reasons: it does not include texts on social sciences; it is divided into very general topics, so it is impossible to search texts dealing exclusively with anatomy or computer sciences, for example; and it is not kept up-to-date.

But building specialized corpora the classical way, i.e. out of printed texts, is normally a very costly process, and Basque is not exactly what we would call a language with plenty of economic resources. So we embarked on a project to build a system to collect specialized corpora in Basque, using the Internet as a source.

1.2 Low topic precision

Before BootCaT (Baroni & Bernardini, 2004) came onto the scene, collecting corpora on a certain topic from the web was mainly done by crawling sites related to the topic and subsequently filtering the pages using some sort of topic classifier, as in (Chakrabarti et al., 1999). BootCaT introduced a new methodology: give a list of words as input, query APIs of search engines for 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. Since then, the subsequent topic-filtering stage has been left aside, as it has been assumed that the search for words on a topic suffices for obtaining the corresponding texts on it alone.

And yet there are not many studies on the precision obtained by the word-list method, and the results of the few that have been done give us reason to believe that a topic-filtering stage *is* necessary: in the aforementioned paper on BootCaT, an evaluation was performed on a small sample of 30 texts of each of the two corpora collected, and a third of them proved to be uninformative or unrelated to the topic. Depending on the application, this amount of noise in the corpora can be considered to be unacceptable.

1.3 Problems with Basque

Obtaining an increase in precision is even more important in our case, since some features of the Basque language and the Basque web cause topic precision to fall dramatically when using the standard methodology, as the experiment we describe next shows.

We used BootCaT to gather some small corpora on geology and computer sciences: we made 20 queries with 2, 3 and 4 n-gram combinations and downloaded the first 10 pages. Then we looked at all of the documents to see if they were appropriate for the corpus (desired topic and language, informative, not duplicates, etc.), and the results we obtained are shown in Table 1.

Topic	n	Total		Appropriate			
		Docs	Words	Docs	%	Words	%
Comp. Sci.	2	65	1,282,001	33	50.77	289,259	22.56
	3	60	2,853,710	25	41.67	406,426	14.24
	4	48	2,321,888	22	45.83	355,254	15.30
Geol.	2	85	2,526,820	13	15.29	379,131	15.00
	3	31	1,606,312	8	25.81	184,371	11.48
	4	3	195,246	2	66.67	101,731	52.10
Total		**292**	**10,785,977**	**103**	**35.27**	**1,716,172**	**15.91**

Table 1: BootCaT topic precision results

The percentage of each of the reasons for a document to be considered inappropriate are shown below:

Topic	n	Reason					
		Wrong topic		Wrong language		Other[1]	
		Docs	%	Docs	%	Docs	%
Comp. Sci.	2	21	65.63	5	15.63	6	18.75
	3	17	48.57	11	31.43	7	20.00
	4	16	61.54	4	15.38	6	23.08
Geol.	2	31	43.06	26	36.11	15	20.83
	3	4	17.39	2	8.70	17	73.91
	4	0	0.00	0	0.00	1	100.00
Total		**89**	**47.09**	**48**	**25.40**	**52**	**27.51**

Table 2: Kinds of inappropriate pages

This study is by no means exhaustive, but our objective was not to quantify the loss in precision exactly. We were just aiming to show that topic precision and general quality of a corpus obtained with BootCaT are much worse when looking for corpora in Basque. Besides, we must take into account that in this experiment we did not perform the bootstrapping process of extracting the words out of the downloaded pages to get new ones; if we had done so, the pages downloaded in the next stage would most likely have yielded even worse topic precision.

The reasons for this are diverse. One is that no search engine offers the possibility of returning pages in Basque alone, so when looking for technical words (as is often the case with specialized corpora), it is very probable that they exist in other languages too, and that the queries return many pages that are not in Basque. Another reason is that the Basque web is not as big as those of other languages, and this means that the only pages existing for certain queries with combinations of various words are very long documents (blogs, magazines in PDF format, etc.) where the desired topic is just a small part of the whole document, or where the words searched for are simply found by chance in different parts of the long document. This phenomenon is exacerbated by the fact that Basque is a morphologically rich language and any lemma has many different word forms, so looking for a word's base form alone, as search engines do, brings fewer results.

2. Our approach

2.1 System objectives and description

The objective of our project is to develop a system to obtain specialized corpora in Basque from the Internet, aimed at improving topic precision and solving Basque-specific problems.

In order to try to improve topic precision, our method 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 topic by using document-similarity techniques (Lee et al., 2005).

And considering the inferior quality that is obtained when Basque is involved, we also try to improve this by using techniques and methods known to obtain better performances with Basque IR, as well as other little adjustments to the general process.

2.2 Evaluation corpora

In order to evaluate and measure the improvements of our system, we built some corpora by putting the system into practice. We chose the same two topics with which we evaluated the performance of BootCaT with Basque, i.e. computer sciences and geology. We built three sample corpora of each topic, consisting of 10, 20 and 30 documents, the two smaller ones made up of documents chosen at random out of the bigger one. For each of these six sample mini-corpora, we automatically extracted the word lists and revised them manually. Then out of each of the six lists we built three different corpora using 2-, 3- and 4-word combinations in the queries. These are the final sizes of the 18 corpora collected:

Topic	Sample size	n		
		2	3	4
Computer Sciences	10	758	274	43
	20	745	256	56
	30	674	176	52
Geology	10	97	22	3
	20	125	14	3
	30	146	27	2

Table 3: Sizes of the collected corpora

These are the corpora that have been used for the various evaluations and partial results mentioned in the next sections, which describe the method and system developed.

3. Automatic keyword extraction from a sample mini-corpus

The basis of our system is a sample mini-corpus of documents on the target topic, which will have to be collected manually. This sample will be used for extracting the word list for the queries and in the final topic-filtering stage as well, so the criteria when collecting the sample is that it should be as heterogeneous as possible and cover as many different subjects of the topic as possible. According to our experiments, as few as 10 documents may be enough for a very specialized topic, but more might be needed for more general topics.

The words to be used in the queries are automatically

[1] Duplicate, part of a much bigger text including other topics, spam, etc.

extracted from this sample corpus, thus avoiding the work of finding appropriate words on the topic. This is usually more laborious than finding texts on the topic, at least for Basque, because there are many topics for which there are still no specialised dictionaries or glossaries.

The keyword extraction method is based on the work previously performed in our team in the DokuSare project (Saralegi & 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), which we calculate by dividing the relative frequency of a word in the specialized mini-corpus by the relative frequency of the word in a general corpus, and applying an empirically determined threshold. The general corpus we use is a 450,000-word corpus consisting of newspaper articles.

The extracted list consists of (mostly) topic-specific words, but some of them might be too specific or rare, as the RFR measure tends to promote on excess words that are not present in the general corpus. The usual way to avoid this is to use a raw frequency threshold to choose the candidate words for the RFR measure, but this is not so easy to apply in our case, because the sample mini-corpora are small (on purpose). And in any case, these undesired words are usually removed in the manual revision stage explained in the next paragraph.

In order to maximise the performance of the queries, the extracted list is revised manually. Too specific or too local proper nouns, too general words and polysemous words that have other meanings in other areas are removed. Normally, the total process of obtaining the mini-corpus and manual revision of the word list is still less costly than trying to obtain a word list, because of the absence of specialised dictionaries explained above.

4. Optimizing for Basque

It is a well-known fact that search engines do not work well with many non-English languages (Bar-Ilan & Gutman, 2005). In 2007 there was even a SIGIR workshop on the subject (Lazarinis et al., 2007). Specifically, 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, and these are the main reasons for the poor performance of BootCaT with Basque. But search engines can be made to work much better with Basque by applying the techniques known as morphological query expansion and language-filtering words, as shown in the projects CorpEus (Leturia et al., 2007 a) and EusBila (Leturia et al., 2007 b).

4.1 Methodology description

In Basque, a lemma can form very many different surface forms, so just looking for the exact base form does not return all the pages that actually contain occurrences of a word. This is true, to a greater or lesser extent, for many other languages too, but while search engines usually apply some sort of stemming for major languages, this does not happen in the case of Basque. Morphological query expansion, also called Frequent Case Generation in some other works (Kettunen, 2007), consists of asking the search engine not only for the lemma of a word, but also for various different word forms of the lemma, which are obtained by morphological generation, within an OR operator. In order to maximize recall, the most frequent word forms are used. In the case of Basque, the morphological generation is done using a tool developed by the IXA Group of the University of the Basque Country, and recall is improved by up to 60% in some cases. The anticipated effect of this increase in recall in our project is a smaller percentage of big PDFs in the downloaded documents, and more pages downloaded in some topics with 4-word combinations in the queries.

The other problem is caused by the fact that no search engine offers the possibility of restricting its results to pages in Basque. The result is that when searching for technical words, short words or proper nouns, many non-Basque pages are returned, since those words may be used in other languages too. The language-filtering words method, consisting of adding the most frequent Basque words to the queries within an AND operator, improves language precision from 15% to over 90%. There is also a non-negligible loss in recall, because pages not containing the filtering words may be left out, but these are normally short and so uninteresting for corpora. Besides, the practical effect in a project like ours is actually a gain in recall: where some normal searches would return many non-Basque pages that would afterwards be filtered out in the language- or size-filtering step and yield few or even no results, with the language-filtering method, however, we would obtain pages in Basque.

We are aware that BootCaT does give the option of language-filtering by means of a list of frequent words in the language, but that filtering is done after downloading the pages. If filtering is conducted that way, many searches for words that exist in other languages will bring no results in Basque and all the pages will be filtered out, thereby wasting bandwidth, time and calls to the API of the search engine.

However, the language-filtering words method ensures that almost all of the pages downloaded will have Basque in them, but not that they will be exclusively in Basque. Due to the Basque language being co-official with Spanish in the Basque Autonomous Community and in some parts of the Charter Community of Navarre, there are many bilingual web pages and documents, e.g. many local and regional government gazettes. Including those bilingual documents in the corpora would cause too much noise, but not including them means we could lose many interesting documents.

In order to solve this problem, we use LangId, a language identifier developed by the IXA Group of the University of the Basque Country, applied at paragraph level. This does not mean that we remove every non-Basque paragraph; if we did, we could also remove some short quotes important for the understanding of a text. As our intention is to eliminate sufficiently large amounts of noise, we remove sequences of non-Basque paragraphs that exceed 10% of the length of the document, and individual paragraphs only if the total amount of the language of the paragraph in the document exceeds 40%.

But working with a minority language like Basque does not always mean more difficulties. Spam and porn filtering, for example, turn out to be very easy. Since as big an audience as possible is usually targeted, there is practically no spam or porn in Basque, so language filtering does the job perfectly.

4.2 Evaluation

The effectiveness of the language-filtering words method for obtaining pages exclusively in Basque from the queries had already been proven in the aforementioned CorpEus and EusBila projects, and the results achieved in this experiment confirm it (only 2.46% of documents retrieved by search engines did not contain any Basque).

As to the language identifier that is applied at paragraph level, it removes supposedly non-Basque parts from 28% of the downloaded documents. Due to the amount of work this entails, we did not evaluate the recall of this step (that is, we did not look at all the documents to see how many non-Basque parts had been left out). But we did look at a sample of the cleaned documents to see if the removed parts were really non-Basque, and although we did not measure it quantitatively, the performance can be considered to be very good.

The morphological query expansion method improves recall in Basque IR, so the number of long PDFs should go down when it is used, which in fact turns out to be the case: in the BootCaT experiments, almost 72% of the documents were PDFs, but now only 13% are PDFs in the computer sciences corpus and 41% in the geology corpus; and the average document length also went down by a 25%.

5. Other improvements

5.1 Description

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. That is to say, we remove documents that are roughly shorter than half a page (not enough continuous text to be interesting) or longer than 50 pages (not likely to be on a specialized topic).

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

We have also included a near-duplicate detection module based on Broder's shingling and fingerprinting algorithm (Broder, 2000). We have prioritised non-redundancy over recall and have rejected not only almost equal documents, but all that have a level of coincidence of over 50% with some other one. The reason for this is that nowadays many on-line news sites and blogs have a main page with some news that changes over time with the addition of new items, but at certain times many news items may coincide. Also, they often show the list of posts related to a category or a tag, and these can have many articles in common too.

Broder's earlier works on near-duplicate detection also dealt with containment (Broder, 1997). But while near-duplicate detection was improved enough to require a very small set of features and very fast processing (as much as to be used at web level), containment detection did not attain this level of optimization. However, we think containment detection is important: again, many blogs and news sites have a main page or section where many individual articles that also have their own URL are contained. And near-duplicate detection methods do not detect containment. So we took up again Broder's method for containment detection, which on our scale is perfectly usable.

5.2 Evaluation

31% of the downloaded documents were filtered out because they were too long, and 10% and 3% of the computer sciences and geology corpora, respectively, because they were too short. By taking a look at the rejected ones, we confirmed that the filter achieves its goal, as the great majority were uninteresting, general or multi-topic documents.

The near-duplicates filter removes 5% of the downloaded documents, and the containment filter another 5%. In the small evaluation we made for precision we found no errors; recall was not evaluated.

6. Topic precision obtained with the improvements

All the improvements made to the process, both Basque-specific or general, that have been described above, have already been evaluated individually. But the aim of each and every one of them is to enhance the quality of the corpora obtained, mainly regarding topic precision. So it is imperative to evaluate the collected corpora by looking at topic precision, to see if the performance tweaks for Basque and the other improvements had any effect and actually improved the BootCaT results. We took a random sample of 30 documents out of each of the 18 corpora built for the evaluation, and saw whether they belonged to the desired topic or not. Due to their small size (see Table 3), all the documents of n=4 and of geology n=3 were checked. These were the results we obtained:

Topic	**Sample size**	**n**			
		2	3	4	Avg.
Computer Sciences	10	46.66%	63.33%	82.93%	64.31%
	20	50.00%	66.66%	70.00%	62.22%
	30	53.33%	63.33%	63.89%	60.19%
	Avg.	**50.00%**	**64.44%**	**72.27%**	**62.24%**
Geology	10	53.33%	40.91%	100.00%	64.75%
	20	56.66%	64.29%	100.00%	73.65%
	30	46.66%	56.76%	100.00%	67.81%
	Avg.	**52.22%**	**53.98%**	**100.00%**	**68.74%**
Avg.		**51.11%**	**59.21%**	**86.14%**	**65.49%**

Table 4: Topic precision before topic-filtering stage

In view of these results, we can conclude that our little improvements, all together, do yield much better topic precision results when looking for corpora in Basque, and are not far short of the baseline for other languages.

7. Topic filtering

As we have pointed out already, this topic precision can be considered insufficient in many cases, and another aim of our project was to try to improve it.

7.1 Methodology description

Topic or domain detection is usually approached through machine learning methods. While these can obtain good performances, they also have their drawbacks: they need fairly big training sets and times, they are trained for a fixed set of topics, etc.

Our approach to this matter has been to try to use a small set of sample documents (i.e. the sample mini-corpus out of which the keywords are extracted) and document similarity measures based on keyword frequencies to say whether a document belongs to a topic or not. According to Sebastiani (2002), topic or domain detection can be done using keywords.

These kinds of document similarity measures are usually applied between two documents to see if they deal with the same or a similar subject, as in the aforementioned DokuSare project. But in our case, we have a document and a corpus, which are elements of different scale, and also the level of similarity to be handled is somehow smaller, since we just need to measure whether they coincide on the topic.

They have also been applied to measure similarity between two corpora (Kilgarriff & Rose, 1998), which is also a little different from our case.

However, the general idea of our project is very similar to that of DokuSare: to represent both the documents to be filtered and the sample mini-corpus through a set of features based on keywords, and to use some similarity measure to see if they share the same topic.

But as we said, we are going to measure the similarity between elements of a different scale, i.e. a document and a set of documents. So we have tried by measuring the similarity between a document and the mini-corpus directly, and also by measuring the similarity of a document with each of the documents of the sample mini-corpus, and taking the maximum.

For the representation of both the downloaded documents and the sample corpus or each of the documents of the sample corpus, we use the bag-of-words paradigm, which models the most significant keywords, i.e. nouns, proper nouns, adjectives and verbs, in a vector. The words are selected and weighted by a certain frequency measure. We have tried two: the aforementioned RFR and a new one we have defined as Relative Rank Ratio or RRR.

We felt that this new frequency measure fitted Zipf's law better (Zipf, 1949) and could be better suited for comparing documents of different sizes. It is defined as the ratio between the relative frequency-ranking of a word in the document or corpus involved, and the relative frequency-ranking of a word in a general corpus. This is its exact formula:

$$RRR(w_i, dok) = \frac{1 - \dfrac{Freq.Rank(w_i, dok)}{RankCount(dok) + 1}}{1 - \dfrac{Freq.Rank(w_i, gen.corp.)}{RankCount(gen.corp.) + 1}}$$

We have observed that this measure works better if we apply some sort of smoothing to words that are not found in the general corpus, because otherwise the formula gives them very high values, and they are often rare words or spelling errors that worsen the results.

For measuring the similarity we use the cosine, the most extended way to measure the similarity between documents represented in the vector space model.

So for comparing two documents x and y, w_i ($i \in \{1, n\}$) being the keywords present in any of the two, we prepare the vectors ($x_1, x_2, \ldots x_n$) and ($y_1, y_2, \ldots y_n$), where x_i and y_i are the RFR or RRR ratios of the word w_i in the documents x and y respectively, and then we calculate the cosine between the two, which is specified as follows:

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2} \sqrt{\sum_{i=1}^{n} y^2}}$$

7.2 Evaluation

As an evaluation experiment, we took the corpora collected for the evaluation, and out of each of them we manually chose a sample of appropriate documents and another one of inappropriate ones, each made up of 15 documents (if the corpus was large enough). Then we applied the aforementioned similarity measures to these development datasets in the two ways explained, and for each of the 18 corpora we obtained charts like those shown in figures 1 to 4. More precisely, these correspond to the average of the geology and computer sciences corpora collected using 20-document sample mini-corpora and using 2-word combinations.

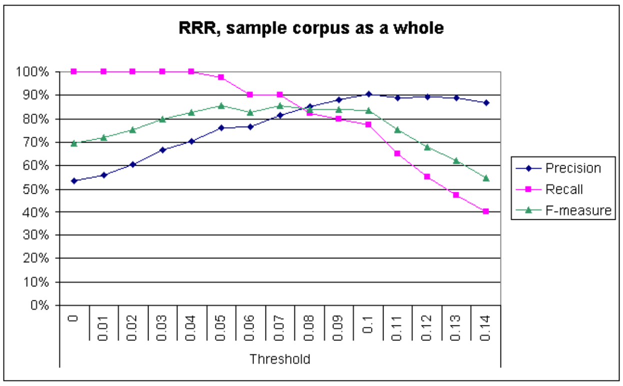

Figure 1: Results with RRR measure, taking the sample mini-corpus as a whole

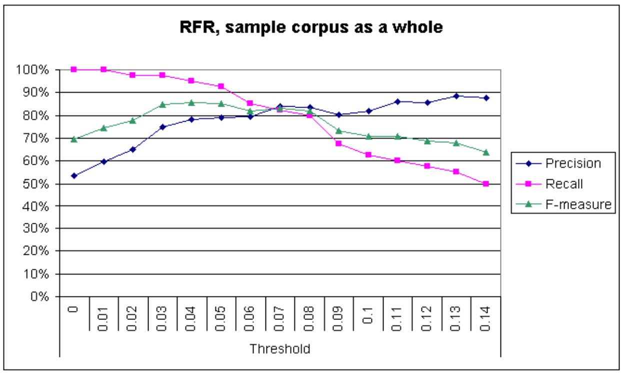

Figure 2: Results with RFR measure, taking the sample mini-corpus as a whole

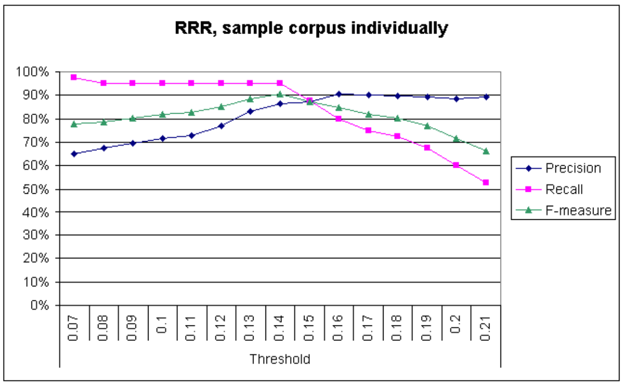

Figure 3: Results with RRR measure, taking each document of the sample mini-corpus individually

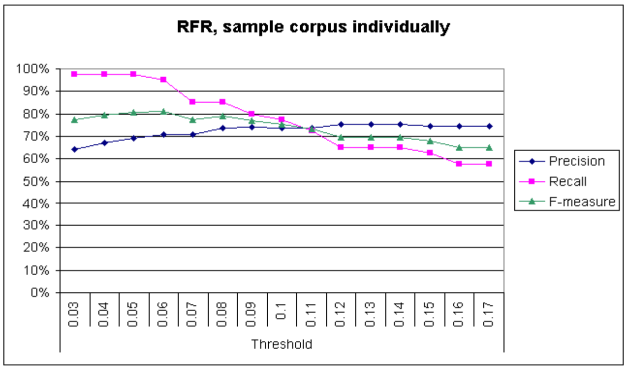

Figure 4: Results with RFR measure, taking each document of the sample mini-corpus individually

It is impossible to show here all the charts for all of the 18 corpora and the different averages. Instead, we will explain the conclusions we have drawn from their observation.

Since our primary objective is to improve topic precision, we are interested in finding a measure and a threshold that will maximise the F-measure but which will prime precision. This is usually obtained somewhere to the right and near the crossing point of the precision and recall series. On average, the highest crossing points are found with the RRR measure when compared with each document of the sample corpus individually.

We have also tried to improve the results by combining more than one of them. For example, we have tried first measuring the similarity with the whole sample mini-corpus and, if the measure is not above the threshold, trying again with the one-by-one comparison. But the only effect of this was that more documents were accepted, both good and bad ones, thus augmenting recall but at the cost of precision.

If we are to significantly improve the baseline of 66% topic precision, we would need a precision of 80% minimum, without a great loss in recall. The RRR-individual method can obtain precision and recall above 80% for most of the corpora, but with different thresholds. In other words, there is no threshold that maximises F-measure and obtains a precision above 80%, and which works for all of the corpora.

In any case, for higher thresholds we usually obtain a higher precision (at least until it falls at some point), so it is possible to assure high precision (80-90%) if recall is not an issue. This might not be the case of Basque, since, as we have observed before, some topics already yield very small corpora and a recall of 60-40% may not be acceptable. But for English or other bigger languages, with the RRR-individual method and a threshold from 0.18 to 0.20 we can obtain a topic precision of 80-90%.

8. Conclusions

The series of improvements to the standard method for collecting specialized corpora from the Internet that we propose, and which are intended to improve the otherwise disastrous performance when looking for documents in Basque, seem to achieve their purpose, since our results are similar to the baseline of other languages. We have also observed that, without any filtering, the best topic precision results are obtained, logically, with 4-word queries, but due to the reduced amount of Basque content on the Internet, corpora obtained on some topics are extremely small with these kinds of queries. And there is no way one can know *a priori* which topics will be affected, so it is better to use 3-word queries, even though the topic precision obtained will be a little lower.

We have also proposed a method for improving the topic precision for any language, based on a sample mini-corpus, automatic extraction of words for the queries and easily computable document similarity measures. In particular, we have shown that it is possible to attain a high precision (80-90%) using the RRR measure and the cosine to compare the documents of the corpus with each document of the sample corpus and taking the maximum, and applying a high enough threshold. But there is also a non-negligible loss in recall, which might be an issue at least for Basque. However, adding the initial word extraction and the final topic filtering as new optional modules to BootCaT could be very interesting.

However, there is an important aspect to point out regarding this method. Obtaining high topic precision does not imply that the corpus obtained will be highly representative of the universe. In fact, since we are filtering by applying similarity measures using the documents of the sample mini-corpus, if this is not wide enough, that is, if not all the subareas of the topic are represented there, we might be missing areas without ever knowing it. So the quality and heterogeneity (and also size) of the sample mini-corpus is a key issue in the method proposed. But it is not easy to say what is a minimum or optimum size of the sample mini-corpus to assure good representativeness, since it greatly depends on whether the topic is very specialised, or quite general, etc. This alone could be a matter for another paper.

9. References

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. In Proceedings of *Corpus Linguistics 2007*. Birmingham, UK: University of Birmingham.

Bar-Ilan, J., Gutman, T. (2005). How the search engines respond to some non-English queries? *Journal of Information Science*, 31(1), pp. 13--28.

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004.* Lisbon, Portugal: ELDA, pp. 1313--1316.

Baroni, M., Chantree, F., Kilgarriff, A., Sharoff, S. (2008). Cleaneval: a competition for cleaning web pages. In *Proceedings of LREC 2008.* Marrakech, Morocco: ELDA.

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

Broder, A.Z. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997.* Los Alamitos, CA: IEEE Computer Society, pp. 21--29.

Chakrabarti, S., van der Berg, M., Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In *Proceedings of the 8th International WWW Conference*. Toronto, Canada: University of Toronto, pp. 545--562.

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

Fletcher, W.H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor & T. Upton (Eds.), *Corpus Linguistics in North America 2002*. Amsterdam, The Netherlands: Rodopi.

Kettunen, K. (2007). Managing keyword variation with frequency based generation of word forms in IR. In *Proceedings of NODALIDA Conference*. Tartu, Estonia: University of Tartu, pp. 318--323.

Kilgarriff, A., Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*. Granada, Spain: ACL SIGDAT, pp. 46--52.

Lazarinis, F., Vilares, J., Tait, J. (2007). Improving non-English web searching (iNEWS07). *ACM SIGIR Forum*, 41(2), pp. 72--76.

Lee, M.D., Pincombe, B., Welsh, M. (2005) An empirical evaluation of models of text document similarity. In *Proceedings of CogSci2005*. Stresa, Italy: Earlbaum, pp. 1254--1259.

Leturia, I., Gurrutxaga, A., Alegria, I., Ezeiza, A. (2007). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, pp. 69--81.

Leturia, I., Gurrutxaga, A., Areta, A., Alegria, I., Ezeiza, A. (2007). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*. Amsterdam, The Netherlands: SIGIR, pp. 47--54.

Saralegi, X., Alegria, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39, pp. 71--78.

Saralegi, X., 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.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), pp. 1--47.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), *WaCky! Working papers on the Web as Corpus*. Bologna, Italy: Gedit.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.