Document Expansion for Cross-Lingual Passage Retrieval

Eneko Agirre[1], Olatz Ansa[1], Xabier Arregi[1], Maddalen Lopez de Lacalle[2],
Arantxa Otegi[1], Xabier Saralegi[2]

[1] IXA NLP Group, University of the Basque Country. Donostia, Basque Country
{e.agirre,olatz.ansa,xabier.arregi,arantza.otegi}@ehu.es

[2] R&D, Elhuyar Foundation. Usurbil, Basque Country
{m.lopezdelacalle,x.saralegi}@elhuyar.com

Abstract. This article describes the participation of the joint Elhuyar-IXA group in the ResPubliQA exercise at QA&CLEF 2010. In particular, we participated in the English–English monolingual task and in the Basque–English cross-lingual one. Our focus was threefold: (1) to check to what extent information retrieval (IR) can achieve good results in passage retrieval without question analysis and answer validation, (2) to check dictionary techniques for Basque to English retrieval when faced with the lack of parallel corpora for Basque in this domain, and (3) to check the contribution of semantic relatedness based on WordNet to expand the passages to related words. Our results show that IR provides good results in the monolingual task, that our performance drop in the cross-lingual system was much greater than in previous CLIR experiments, and that expansion improves the results in the monolingual task.

Keywords: Cross-lingual passage retrieval, semantic relatedness, word co-occurrences.

1 Introduction

Like last year, the team consisted of two different groups: the Elhuyar Foundation, and the IXA NLP group. Last year we participated in the CLEF 2009 ResPubliQA task by submitting two English-English monolingual runs and two Basque-English cross-lingual runs. It should be mentioned that we were the only team who participated in a cross-lingual task.

Following the positive experience of last year's participation it seemed interesting to continue sharing our experience and knowledge on QA-oriented (CL)IR. Like last year, we participated in the English-English monolingual task and Basque-English cross-lingual task.

With respect to the Basque-English task, we met the challenge of retrieving English passages for Basque questions. We tackled this problem by translating the lexical units of the questions into English. The main setback is that no parallel corpus was available for this pair of languages, given that there is no Basque version of the JRC-Acquis and the Europarl collections. So we explored an approach which does not use parallel corpora when translating queries, which could also be interesting for other less resourced languages. In our opinion, bearing in mind the idiosyncrasy of the European Union, it is worthwhile tackling the search for passages that answer questions formulated in non-official languages.

Question answering systems typically rely on a passage retrieval system. Given that passages are shorter than documents, vocabulary mismatch problems are more significant than in full document retrieval. Most of the previous work on expansion techniques has focused on pseudo-relevance feedback and other query expansion techniques. In particular, WordNet has been used previously to expand the terms in the query with little success [1, 2, 3]. The main problem is ambiguity, and the limited context available to disambiguate the word in the query effectively. As an alternative, we felt intuitively that passages would provide sufficient context to disambiguate and expand the terms in the passage. In fact, we did not do explicit word sense disambiguation, but rather applied a state-of-the-art semantic relatedness method [4] in order to select the best terms to expand the documents.

2 System Overview

2.1 Question pre-processing

We analysed the Basque questions by re-using the linguistic processors included in the *Ihardetsi* question-answering system [5]. This system uses two general linguistic processors: the lemmatizer/tagger named *Morfeus* [6], and the Named Entity Recognition and Classification (NERC) processor called *Eihera* [7]. The use of the lemmatizer/tagger is particularly suited to Basque, as it is an agglutinative language. It provides the corresponding lemma and part of speech of each lexical unit, which also includes both single words and multiword units (MWU). The numerical and temporal expressions are also captured by the lemmatizer/tagger. The NERC processor, Eihera, captures entities such as persons, organizations and locations. The questions thus analyzed are passed to the translation module once the function words are removed. In the case of English, queries were just tokenized without further analysis.

2.2 Translation of the query terms (Basque-English runs)

Once the questions had been linguistically processed, they were translated into English using a dictionary-based method. According to the literature, parallel corpora-based translation methods provide the best translation quality, but these are scarce for small languages like Basque or even for major languages in certain domains. So, a dictionary-based translation approach was chosen. To tackle translation ambiguity produced by the dictionary translation, some techniques have been proposed in the literature, such as structured query-based techniques [8, 9] and co-occurrences-based techniques [10, 11, 12]. According to previous pieces of work [13], structured queries offer better MAP than co-occurrences-based methods on Basque-English CLIR experiments only when dealing with long queries [13]. However, the questions to evaluate in ResPubliQA are short, and structured queries were not supported in the retrieval algorithm used (see Section 2.4), so we adopted a co-occurrences-based translation selection strategy. The dictionary-based translation process designed comprises two main steps, taking the keywords (named entities, MWU and single words tagged as noun, adjective or verb) of the question as source words:

1. Obtaining translation candidates: In the first step the translation candidates of each source word are obtained from a bilingual eu-en dictionary comprising the Basque-English Morris dictionary[1], and the Euskalterm terminology bank[2] which includes 38,184 MWUs. After that, Out-Of-Vocabulary words are solved by searching for their cognates in the target collection. The cognate detection is done in two phases. First, several transliteration rules are applied to the source word. Then, the Longest Common Subsequence Ratio is calculated with respect to all the words from the target collection. Those that reach a previously established threshold (0.9) are selected as translation candidates.

2. Solving ambiguous candidates: The selection of the best translation for each source keyword is performed by an algorithm based on the maximum association degree, explained on detail in [14]. The association degree is computed by calculating co-occurrences of word pairs in the target collection. The algorithm obtains the set of translation candidates that maximizes the association degree between each other in the target collection. This maximization problem is solved by an Expectation Maximization-type greedy algorithm made up of initialization, iteration and normalization steps:

Initially, all the translation candidates provided by the dictionary are equally likely.

In the **iteration** step, the weight of each translation candidate is iteratively updated according to the association degree it has regarding the rest of the source word translation candidates. This association degree is pondered using the weights obtained on the previous iteration. The association degree between two translation candidates is measured by the Log-likelihood ratio using the target collection as a corpus. A factor is included in order to increase the association degree between translation candidates whose source words are near each other in the source query, and whose source words belong to the same MWU.

Finally, after re-computing each term weight, all of them are normalized. The algorithm stops when the difference between the term weights corresponding to previous and current iteration become lower than a predefined threshold.

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

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

2.3 Document Pre-processing and Expansion

Given that the aim of the task was to retrieve a paragraph that contains an answer for each question, we first split the document collection into paragraphs.

One of the main features of our system is that the passages are expanded based on their related concepts according to the background information in WordNet [15]. We selected those concepts that are most closely related to the passage as a whole. For this purpose, we used a technique based on random walks over the graph representation of WordNet 3.0 concepts and relations [4], whose implementation is publicly available[3].

Given a passage and the graph-based representation of WordNet, we obtained a ranked list of WordNet concepts as follows:

1. We first pre-processed the passage to obtain the lemmas and parts of speech of the open category words using the OpenNLP open source software[4]. It should be noted that the lemmatizer/tagger Morfeus used for Basque questions works only with the Basque language.
2. We then assigned a uniform probability distribution to the terms found in the passage. The rest of the nodes were initialized to zero.
3. We computed personalized PageRank [16] over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts. The higher the probability for a concept, the more related it is to the given passage.

In order to select the expansion terms, we chose the 100 highest scoring concepts, and got all the words that lexicalize the given concept. An example of a document expansion is shown in Fig. 1.

We applied the expansion strategy only to passages which had more than 10 words, for two reasons: the first one was that most of the shorter passages were found not to contain relevant information for the task (e.g. “Article 2” or “Having regard to the proposal from the Commission”), and the second was that we thus saved some computation time.

The same expansion strategy has been used in some of our previous work with promising results [17].

2.4 Including Expansions in a Retrieval System

Once we had the list of words for document expansion, we created one index for the words in the original documents and another index with the expansion terms. We used the MG4J search engine [18] as it enables several indices over the same document collection to be combined. This way, we were able to use the original words only, or to include the expansion words during retrieval as well.

We used the BM25 ranking function, which has two free parameters (b and k_1) [19]. In the implementation of BM25 of the MG4J search engine, the two indices are

[3] http://ixa2.si.ehu.es/ukb/

[4] http://opennlp.sourceforge.net/

combined linearly, where the relative weight of the expanded index can be specified setting up the free λ parameter. Further information about the scoring function and the combination of the index we used can be found in [17].

3 Experimental Setup

We participated in the English-English monolingual task and the Basque-English cross-lingual task. For the monolingual run, we did not analyze the English questions, we carried out the passage retrieval only after expanding the documents, as explained in Sections 2.3 and 2.4. For the bilingual runs, we first analyzed the questions (see Section 2.1), then we translated the question terms from Basque to English (see Section 2.2), and, finally, we retrieved the relevant passages for the translated query terms (see Sections 2.3 and 2.4). For both languages, stop words were removed from the queries and a stemming pre-process based on the Porter algorithm was applied to the query and document words.

As we were interested in the performance of passage retrieval on its own, we did not carry out any answer validation, and we just chose the first passage returned by the passage retrieval module as the response. We did not leave any question unanswered.

For both tasks, the only difference between the two runs submitted is the use (or not) of the expansion in the passage retrieval phase. In other words, in the first run (referenced as “run 1” in the tables throughout this paper), apart from the original words that were in the passages, we also used the expanded words during the retrieval. In the second run (referenced as “run 2” in the tables throughout this paper), we only used the original words that were in the passages.

The BM25 parameters and the λ parameter (see Section 2.4) for both languages were fixed after a training phase with the question set from the previous edition of ResPubliQA [20]. Table 1 lists the parameter values used for each run.

Table 1. Free parameters described in Section 2.4. λ is not used in run 2.

Submitted runs		b	k_1	λ
English - English	run 1	0.17	0.30	0.22
	run 2	0.09	0.53	-
Basque - English	run 1	0.35	0.34	0.57
	run 2	0.71	0.23	-

4 Results

This section describes the results obtained in our ResPubliQA 2010 participation and discusses the performance of our document expansion approach and the translation of query terms approach.

Table 2 shows the official results of the four runs we submitted. The Mean Reciprocal Rank (MRR) measure is also shown in the table. We use * to indicate statistical significance at 99% confidence level, based on the Paired Randomization Test [21].

Table 2. Results for submitted runs

Submitted runs		#answered correctly	#answered incorrectly	c@1	MRR
English - English	run 1	130	70	**0.65**	**0.6067***
	run 2	123	77	0.62	0.5658
Basque - English	run 1	66	134	0.33	0.2742
	run 2	72	128	**0.36**	**0.2958**

Table 3 lists, for each language pair, the number of questions answered correctly in run 1 alone (i.e. using expansions), in run 2 alone (i.e. not using expansions) and in both runs, respectively.

Table 3. Comparison between the two runs per language pair

Language pairs	#answered correctly only in run 1	#answered correctly only in run 2	#answered correctly in both runs
English - English	9	2	121
Basque - English	5	11	61

4.1 Analysis of the Document Expansion Approach

Regarding monolingual results (“English-English” row in Table 2), we can see that the number of correct answers is higher in run 1 than in run 2. Since the only difference between the two runs was that run 1 used expanded words of the passages, the results indicate that the use of document expansion is beneficial. It should be noted that the improvement in MRR in run 1 compared with run 2 is statistically significant. To be precise, the correct answer set in run 1 was 130, and 123 in run 2, where the intersection of both sets was 121 (see Table 3).

The results of cross-lingual runs (“Basque-English” row in Table 2) show that the use of the expanded words did not improve the results, but the differences between both runs are not statistically significant. To our surprise, 72 questions were correctly answered without expansion, 6 more than when it was used. However, the answers to 5 questions were only found by the run enriched with expansions (see Table 3). As we obtained improvements using expansions in the training phase and also at ResPubliQA 2009 [14], further analysis of our cross-lingual approach is needed in order to determine why the use of expanded words is favourable only for some settings.

Fig. 1 shows an example of a document expansion which was effective for answering the English question number 32 of the training set: “*Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*”

In the second part of the example we can see some words that we obtained after applying the expansion process explained in Section 2.3 to the original passage also shown in the example. As we can see, there are some new words among the expanded words that are not in the original passage, such as *unfavourable* or *consequence*. Those two words were in the question referred to above (number 32). That could be why our system answered that question correctly when using the expanded words, but not when using the original words alone.

original passage: *Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any adverse effects on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;*

some expanded words: *cistron factor gene coding cryptography secret_writing ... acetyl acetyl_group acetyl_radical ethanoyl_group ethanoyl_radical beta_lactamase penicillinase ... ec eec eu europe european_community european_economic_community european_union ... directive directing directional guiding citizens_committee committee environment environs surround surroundings corn ... maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward gamboge ... unfavorable* ***unfavourable*** *... set_up expostulation objection remonstrance remonstration dissent protest believe light lightly belief feeling impression notion opinion ... reason reason_out argue jurisprudence law* ***consequence*** *effect event issue outcome result upshot ...*

Fig. 1. Example of a document expansion (doc_id: *jrc31998D0293-en.xml,* p_id*: 17).*

4.2 Analysis of the Query Terms Translation Approach

Compared with the monolingual run, the cross-lingual task yielded worse results. 50% of the monolingual performance was achieved for run 1, and 58% for run 2 (see table 3). This drop in performance for the cross-lingual task is worse than the one reported in a similar CLIR experiment [22] with the same cross-lingual method, where 74% of monolingual results were achieved. In that work, the drop in performance in our system was produced mainly because of the lack of recall of the dictionary. The source word appeared on the dictionary, but translations for the corresponding sense did not. This case falls between ambiguity and Out-Of-Vocabulary word. In the experiment carried out in this paper, in addition to the dictionary recall problem, many Out-Of-Vocabulary words corresponding to acronyms were detected. This adversely affects the retrieval performance since the cognate-based method does not solve them. Irrespective of the translation method, the accumulation of errors (i.e. question analysis, automatic lemmatization and entities detection) is another factor which explains the deterioration in the system performance in the cross-lingual task.

Despite this difference between the monolingual and cross-lingual task, some questions were answered correctly only in the cross-lingual runs (see Table 4).

Table 4. Number of questions answered correctly in the monolingual run alone, in the cross-lingual run alone, and in both runs

	Number of questions answered correctly		
	Only in the Monolingual Run	Only in the Cross-lingual Run	In both runs
run 1	75	11	55
run 2	64	13	59

We compared the translations of the test questions provided by our system with the source English questions. Our system translations helped to retrieve the correct passage in those cases because of the following isolated reasons:

a) Some relevant Out-Of-Vocabulary words are translated by cognate detection as they appear spelled in the correct passage (e.g. "*Zimmerman*" was translated to "*Zimmermann*" instead of "*Zimmerman*" as in the source English question).

b) Some words are translated as they appear in the correct passage, but different from spelling in the source English question (e.g. in question number 42, the Basque keyword "*zuzendari*" was translated by our system into "*manager*" which appears in the correct passage, instead of "*director*" as in the source English question).

c) The wrong translation of a word helps to retrieve the appropriate passage because it appears accidentally in the passage.

d) The translations provided by our system give a better distribution of weights by allowing the chance retrieval of the appropriate passage.

5 Conclusions

This paper describes the participation of the joint Elhuyar-IXA team at ResPubliQA

2010. For that purpose we used a system which works with passage retrieval alone, without any question analysis and answer validation steps.

Our English-English results show that good results can be achieved by means of this simple strategy. After expanding the passages based on semantic relatedness and tuning the retrieval system parameters, we obtained improvements for the English-English task. The drop in performance in the Basque-English bilingual runs is significant, and is caused by the accumulation of errors in the analysis and translation of the query. The use of expanded words was not effective for the cross-lingual task. A possible reason is the following: the co-occurrence-based translation selection algorithm uses as the target collection the one without expanded words to calculate the association degree between translation candidates, and consequently, the final translations are adapted to the original collection. Then, when expanded words are added to the passages, instead of helping the retrieval, they could add noise.

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01), imFUTOURnet (IE08-233) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government.

References

1. Kim, S., Seo, H., Rim, H.: Information retrieval using word senses: root sense tagging approach. In: Proceedings of SIGIR. (2004)
2. Liu, S., Yu, C., Meng, W.: Word Sense Disambiguation in Queries. In: Proceedings of the 14th ACM Conference on Information and Knowledge Management, CIKM. (2005)
3. Pérez-Agüera, J.R., Zaragoza, H.: Query Clauses and Term Independence. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
4. Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M.: A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of the annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA (2009)
5. Ansa, O., Arregi, X., Otegi, A., Soraluze. A.: Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
6. Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J.M., Urizar, R.: Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In: COLING-ACL, pp.380–384. (1998)
7. Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., Urizar. R.: Development of a Named Entity Recognizer for an Agglutinative Language. In: IJCNLP, (2004)
8. Darwish, K., Oard, D.W.: Probabilistic Structured Query Methods. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in

Information Retrieval, pp. 338–344. (2003)

9. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63. (1998)
10. Ballesteros, L., Bruce Croft, W.: Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71. (1998)
11. Gao, J., Nie, J.Y., Xun, E., Zhang, J., Zhou, M., Huang, C.: Improving Query Translation for Cross-language Information Retrieval Using Statistical Models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 96-104. (2001)
12. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 520-527. (2005)
13. Saralegi, X., López de Lacalle, M.: Comparing Different Approaches to Treat Translation Ambiguity in CLIR: Structured Queries v. Target Co-occurrence-Based Selection. In: 6th TIR workshop. (2009)
14. Agirre, E., Ansa, O., Arregi, X., Lopez de Lacalle, M., Otegi, A., Saralegi, X., Zaragoza. H.: Elhuyar-IXA: semantic relatedness and cross-lingual passage retrieval. Working Notes of the Cross-Lingual Evaluation Forum, Corfu, Greece. (2009)
15. Fellbaum, C.: WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, Cambridge, Mass. (1998)
16. Haveliwala, T. H.: Topic-sensitive PageRank. In: Proceedings of WWW'02, pages 517-526. (2002)
17. Agirre, E., Arregi, X., Otegi, A.: Document Expansion Based on WordNet for Robust IR. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING). To appear (2010)
18. Boldi, P., Vigna, S.: MG4J at TREC 2005. In: Voorhees, E.M., Buckland, L.P. (eds.) The Fourteenth Text Retrieval Conference (TREC 2005) Proceedings, number SP 500-266 in Special Publications. NIST. http://mg4j.dsi.unimi.it/. (2005)
19. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4):333-389. (2009)
20. Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P.: Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. Working Notes for the CLEF 2009 Workshop. (2009)
21. Smucker, M. D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of CIKM 2007, Lisbon, Portugal. (2007)
22. Saralegi, X., Lopez de Lacalle, M.: Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In the 7th International Conference on Language Resources and Evaluations (LREC). Malta. (2010)