Elhuyar-IXA: semantic relatedness and cross-lingual passage retrieval

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

Arantxa Otegi[1], Xabier Saralegi[2], Hugo Zaragoza[3]

[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

{maddalen,xabiers}@elhuyar.com

[3] Yahoo! Research. Barcelona, Spain
hugoz@yahoo-inc.com

Abstract

This article describes the participation of the joint Elhuyar-IXA group in the RespubliQA exercise at QA&CLEF. We put together tools developed separately and we combined and shared knowledge and technology between the two groups. In particular, we participated in the English–English monolingual task and in the Basque–English cross-lingual one. Our focus has been threefold: (1) to check to what extent IR can achieve good results in passage retrieval without question analysis and answer validation, (2) to check Machine Readable Dictionary techniques for the 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 crosslingual system lowers the performance compared to the monolingual runs, and that semantic relatedness improves the results in both tasks (by 6 and 2 points, respectively).

1 Introduction

The joint team was formed by two different groups, on the one hand the Elhuyar Foundation, and on the other hand the IXA NLP group. The Elhuyar Foundation is a non-profit making organization located in the Basque Country. The Elhuyar Foundation's mission is to popularize science and technology and promote the development of the Basque language. To achieve this, it offers the Basque public at large quality services, tools and resources that are a reference, with innovation being pivotal and with a commitment to operate in the area of education. Related to these objectives it deals with many activities, such as R&D in Natural Language Processing and Information Retrieval fields. The IXA NLP group of the University of the Basque Country has previously participated at CLEF, specifically, in the CLEF 2008 Basque to Basque monolingual QA task [3] and in the CLEF 2008 Robust-WSD task. IXA has participated in the CLEF 2009 Robust-WSD Task this year too.

Both Elhuyar and IXA considered that it would be interesting to share experience and knowledge on QA oriented (CL)IR. We decided to form a single team for participating in the ResPubliQA track. This collaboration allowed us to tackle the English–English monolingual task and the Basque–English cross-lingual one.

Question answering systems typically rely on a passage retrieval system. Given that passages are shorter than documents, vocabulary mismatch problems are more important 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 [9, 10, 11, 13]. The main problem is ambiguity, and the limited context available to disambiguate the word in the query effectively. As an alternative, we felt that passages would provide sufficient context to disambiguate and

expand the terms in the passage. In fact, we do not do explicit WSD, but rather apply a state-of-the-art semantic relatedness method [1] in order to select the best terms to expand the documents.

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 is available for Basque, given that there is no Basque version of the JRC-Acquis collection. So we have explored a corpus parallel free approach for translating queries which could also be interesting for other less resourced languages. Even so, we regarded the cross-lingual exercise as interesting. In our opinion, bearing in mind the idiosyncrasy of the European Union, it is worthwhile dealing with the search of passages that answer questions formulated in unofficial languages.

2 System overview

2.1 Question pre-processing

We analysed the Basque questions by re-using the linguistic processors of the *Ihardetsi* question-answering system [3]. This module uses two general linguistic processors: the lemmatizer/tagger named *Morfeus* [7], and the Named Entity Recognition and Classification (NERC) processor called *Eihera* [2]. The use of the lemmatizer/tagger is particularly suited to Basque, as it is an agglutinative language. It returns only one lemma and one part of speech for each lexical unit, which includes single word terms and multiword terms (MWT) (those included in the MRD introduced in the next subsection). The NERC processor, *Eihera*, captures entities such as *person*, *organization* and *location*. The numerical and temporal expressions are captured by the lemmatizer/tagger. The questions thus analyzed are passed to the translation module.

English queries were tokenized without further analysis.

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

Once the questions had been linguistically processed, we translated them into English. Among the main strategies and methods proposed in the literature to deal with language barriers in IR problems we adopted a Machine Readable Dictionary (MRD)-based method. Due to the scarcity of parallel corpora for a small language or even for big languages in certain domains we have explored a MRD-based method. However, MRD-based approaches have inherent problems, such as the presence of ambiguous translations and out-of-vocabulary (OOV) words. To tackle these problems, both translation ambiguity and OOV words, some techniques have been proposed such as structured query-based techniques [6, 14] and concurrences-based techniques [4, 8, 12]. These approaches have been compared for Basque by obtaining best MAP results with structured queries [15]. However, structured queries were not supported in the retrieval algorithm used (see Section 2.3), so we adopted a concurrences-based translation selection strategy.

The translation process designed comprises two steps and takes the keywords (Name Entities, MWT and singles words tagged as noun, adjective or verb) of the question as source words.

In the first step the translation candidates of each source word are obtained. The translation candidates for the lemmas of the source words are taken from a bilingual eu-en MRD composed from the Basque-English *Morris* dictionary[1], and the *Euskalterm* terminology bank[2] which includes 38,184 MWTs. After that, OOV words and ambiguous translations are dealt with. The number of OOV words quantified out of a total of 421 keywords for the 77 questions of the development set was 42 (10%). These 77 questions were translated by hand from English to Basque in order to carry out the development phase. Nevertheless, it must be said that many of these OOV words were wrongly tagged lemmas and entities. We deal with OOV words by searching for their cognates in the target collection. The cognate detection is done in two phases. Firstly, we apply several transliteration rules to the source word. Then we calculate the Longest Common Subsequence Ratio (LCSR) among words with a similar length (+-10%) from the target collection (see Figure 1). The ones which reach a previously established threshold (0.9) are selected as translation candidates. The attempt to select the best translation candidate will be held in the translation selection phase. The MWT terms that are not found in the dictionary are translated word

[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.

by word, as we realized that most of the MWT could be translated correctly in that way, exactly 91% of the total MWTs identified by hand in the 77 development questions.

err- ---> r- *erradioterapeutiko=radioterapeutiko*
k ---> *c* *radioterapeutiko=radioterapeutico*
LCSR(radioterapeutico, radioterapeutic) = 0.9375

Figure 1: Example of cognate detection

In the second step of the translation process we perform a translation selection step. In the translation selection step, we select the best translation of each source keyword according to an algorithm based on target collection concurrences. This algorithm sets out to obtain the translation candidate combination that maximizes their global association degree. We take the algorithm proposed by Monz and Dorr [12].

Initially, all the translation candidates are equally likely. Assuming that t is a translation candidate of the set of all candidates $tr(s_i)$ for a query term s_i given by the MRD, then:

Initialization step:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|}$$

In the iteration step, each translation candidate is iteratively updated using the weights of the rest of the candidates and the weight of the link connecting them.

Iteration step:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} w_L(t,t') \cdot w_T(t'|s_i)$$

where $inlink(t)$ is the set of translation candidates that are linked to *t*, and $w_L(t,t')$ is the association degree between *t* and *t'* on the target passages measured by Log-likelihood ratio. These concurrences were calculated by taking the passages of the documents of the target collection as window.

After re-computing each term weight they are normalized.

Normalization step:

$$w_L^n(t|s_i) = \frac{w_L^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w_L^n(t_{i,m}|s_i)}$$

The iteration stops when the variations of the term weights become smaller than a predefined threshold.

We have modified the iteration step by adding a factor $w_T(t,t')$ to increase the association degree $w_L(t,t')$ between translation candidates *t* and *t'* whose source words $w_T(t,t')$ are near each other (distance *dis* in words is low) in the source query *Q*, and whose source words $so(t), so(t')$ belong to the same Multi-Word Unit (MWU) *Z* $smw(t,t')$. As the global association degree between translation candidates is estimated from the association degree of pairs of candidates, we score positively these two characteristics when the association degree for a pair of candidates is calculated. Thus, the modified association degree $w'_L(t,t')$ between t and t' will be calculated in this way in the iteration step:

$$w'_L(t,t') = w_L(t,t') * w_T(t,t')$$

$$w_T(t,t') = \frac{\max_{s_i,s_j \in Q}(dis(s_i,s_j))}{dis(so(t),so(t'))} * 2^{smw(t,t')}$$

$$smw(t,t') = \begin{cases} 1 & \{so(t),so(t')\} \subseteq Z \quad \text{where } Z \in MWU \\ 0 & \end{cases}$$

2.3 Passage retrieval

The purpose of the passage retrieval module is to retrieve passages from the document collection which are likely to contain an answer. The main feature of this module is that the passages are expanded based on their related concepts, as explained in the following sections.

2.3.1 Document preprocessing and application of semantic relatedness

Given that the system needs to return paragraphs, we first split the document collection into paragraphs, which are delimited by the mark <*p*> in the documents. Then we lemmatized and POS tagged those passages using the OpenNLP open source software[3].

After preprocessing the documents, we expanded the passages based on semantic relatedness. To this end, we used UKB[4], a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base [1], in this case WordNet 3.0.

Given a passage, UKB returns a vector of scores for concepts in WordNet. Each of these concepts has a score, and the higher the score, the more related the concept is to the given passage, where we represent the passage using the lemmas of all nouns, verbs, adjectives and adverbs in the passage.

Given the list of related concepts, we took the highest-scoring 100 concepts and expanded them to all variants (words that lexicalize the concepts) in WordNet. An example of a document expansion is shown in Figure 2.

The variants for those expanded concepts were included in a new field of the passage representation. This way, we were able to use the original words only, or alternatively, to also include the variants for the most related 100 concepts, as we will be explaining in Section 2.3.2 and Section 2.3.3.

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

2.3.2 Indexing

We indexed the new expanded documents using the MG4J search-engine [5]. MG4J makes it possible to combine several indices over the same document collection. We created one index for each field: one for the original words and one for the expanded words. Porter stemmer was used as per usual.

2.3.3 Retrieval

We used the BM25 ranking function with the following parameters: 1.0 for *k1* and 0.6 for *b*. We did not tune these parameters. MG4J allows multi-index queries, where one can specify which of the indices one wants to search in, and assign different weights to each index. We conducted different experiments, by using the original words alone (the index made of original words) and also by using the index with the expansion of concepts, giving different weights to the original words and the expanded concepts. The weight of the index which was created using the original words from the passages was 1.00 for all the runs. 1.00 was also the weight of the index that included the expanded words for the monolingual run, but it was 1.78 for the bilingual run. These weights were fixed following a training phase with the English development questions provided by the organization, and after the Basque questions had been translated by hand (as no development Basque data was released).The submitted runs are described in the next section.

3 Description of runs

We participated in the English-English monolingual task and the Basque-English cross-lingual task, with two runs per language pair. We did not analyze the English queries for the monolingual run, and we just removed the stopwords.

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

[4] The algorithm is publicly available at http://ixa2.si.ehu.es/ukb/

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 Section 2.3).

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 submitted two runs is the use (or not) of the expansion in the passage retrieval module. That is, in the first run ("run 1" in Table 1), during the retrieval we only used the original words that were in the passage. In the second run ("run 2" in Table 1), apart from the original words, we also used the expanded words.

4 Results

Table 1 summarizes the results of our submitted runs, explained in Section 3.

		#answered correctly	#answered incorrectly	c@1
English - English	**run 1**	211	289	0.42
	run 2	240	260	**0.48**
Basque - English	**run 1**	78	422	0.16
	run 2	91	409	**0.18**

Table 1: Results for submitted runs

The results show that the use of the expanded words (run 2) was effective for both tasks, improving the final result by 6 % in the monolingual task.

Figure 2 shows an example of a document expansion which was effective for answering the English question number 32: "*Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*"

doc_id: *jrc31998D0293-en.xml*
p_id: *17*

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 common_market ec eec eu europe european_community european_economic_community european_union ... directive directing directional guiding citizens_committee committee environment environs surround surroundings corn indian_corn maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward gamboge lemon lemon_yellow ... 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 ...*

Figure 2: Example of a document expansion

In the last part of the example we can see some words that we obtained after applying the expansion process explained in Section 2.3.1 to the original passage showed in the example too. 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 we mentioned before (number 32). That could be why we answered that question correctly when using the expanded words (in run 2), but not when using the original words only (without using the expanded words, in run 1).

As expected, the best results were obtained in the monolingual task. With the intention of finding reasons to explain the significant performance drop in the bilingual run, we analyzed manually 100 query translations obtained in the query translation process of the 500 test queries, and detected several types of errors arising from both the question analysis process and from the query translation process. In the question analysis process, some lemmas were not correctly identified by the lemmatizer/tagger, and in other cases some entities were not returned by the lemmatizer/tagger causing us to lose important information for the subsequent translation and retrieval processes. In the query translation process, leaving aside the incorrect translation selections, the words appearing in the source questions were not exactly the ones that figured in many queries that had been correctly translated. In most cases this happened because the English source query word was not a translation candidate in the MRD. If we assume that the answers contain words that appear in the questions and therefore in the passage that we must return, this will negatively affect the final retrieval process.

5 Conclusions

The joint Elhuyar-Ixa team has presented a system which works on 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. We experimented with applying semantic relatedness in order to expand passages prior to indexing, and the results are highly positive, especially for English-English. The performance drop in the Basque-English bilingual runs is significant, and is caused by the accumulation of errors in the analysis and translation of the query mentioned.

Acknowledgments

This work has been supported by KNOW (TIN2006-15049-C03-01), imFUTOURnet (IE08-233) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government.

References

[1] E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA*, June 2009.

[2] I. Alegria, O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, and R. Urizar. Development of a Named Entity Recognizer for an Agglutinative Language. In *IJCNLP*, 2004.

[3] O. Ansa, X. Arregi, A. Otegi, A. Soraluze. Ihardetsi question answering system at QA@CLEF 2008. *Working Notes of the Cross-Lingual Evaluation Forum, Aarhus, Denmark*. 2008.

[4] L. Ballesteros and W. Bruce Croft, Resolving Ambiguity for Cross-language Retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.64–71*. 2008

[5] P. Boldi and S. Vigna. MG4J at TREC 2005. In Ellen M. Voorhees and Lori P. Buckland, editors, *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, number SP 500-266 in Special Publications. NIST*, 2005. http://mg4j.dsi.unimi.it/.

[6] K. Darwish and D. W. Oard. Probabilistic structured Query Methods. *Proceedings of the 26th annual*

international ACM SIGIR conference on Research and development in information retrieval. P.338–344. 2003.

[7] N. Ezeiza, I. Aduriz, I. Alegria, J.M. Arriola, and R. Urizar. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *COLING-ACL, pp.380–384*, 1998.

[8] J. Gao, J.Y. Nie, E. Xun, J. Zhang, M. Zhou, C. Huang. Improving Query Translation for Cross-language Information Retrieval using Statistcal Models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research an development in information retrieval, p. 96-104*. 2001

[9] S. Kim, H. Seo, H. Rim. Information retrieval using word senses: Root sense tagging approach. *Proceedings of SIGIR 2004*.

[10] S. Liu, F. Liu, C. Yu, W. Meng. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *Proceedings of SIGIR 2004*.

[11] S. Liu, C. Yu, W. Meng. Word Sense Disambiguation in Queries. *Proceedings of ACM Conference on Information and Knowledge Managment, CIKM,* 2005.

[12] C. Monz and B.J. Dorr. Iterative translation disambiguation for cross-language Information Retrieval. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. Pages 520-527, 2005

[13] J.R. Pérez-Agüera, H. Zaragoza. UCM-Y!R at CLEF 2008 Robust and WSD tasks. *Working Notes of the Cross-Lingual Evaluation Forum, Aarhus, Denmark*. 2008.

[14] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p*. 55-63. 1998

[15] X. Saralegi, M. López de Lacalle. Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries v. Target Co-occurrence Based Selection. *6th TIR workshop.* To appear, 2009.