Ihardetsi: a Basque Question Answering System at QA@CLEF 2008

Olatz Ansa, Xabier Arregi, Arantxa Otegi, and Ander Soraluze

IXA Group, University of the Basque Country
olatz.ansa@ehu.es

Abstract. This paper describes *Ihardetsi*, a question answering system for Basque. We present the results of our first participation in the QA@CLEF 2008 evaluation task. We participated in three subtasks using Basque, English and Spanish as source languages, and Basque as target language. We approached the Spanish-Basque and English-Basque cross-lingual tasks with a machine translation system that first processes a question in the source language (i.e. Spanish, English), then translates it into the target language (i.e. Basque) and, finally, sends the obtained Basque question as input to the monolingual module.

1 Introduction

In the QA@CLEF 2008 edition Basque language was incorporated for the first time both as source language and as target language. In this context a new monolingual task, Basque-Basque, and two cross-lingual tasks, English-Basque and Spanish-Basque, were organised.

The main goal of our first participation in QA@CLEF for Basque was to evaluate our basic system by comparing it with the state of the art of non-English question answering systems. Besides, the analysis of the results could reveal a number of future system improvements. We took part in Basque-Basque, Spanish-Basque and English-Basque tasks. Our system, Ihardetsi, is a Basque monolingual system, and we use a Spanish-Basque and a English-Basque machine translation systems [1] for the cross-lingual tasks in order to translate the questions into Basque.

This paper is structured as follows. The next section presents the corpus processing. Section 3 describes the system architecture. Section 4 introduces the results and a preliminary analysis of the kind of errors that the system made. Conclusions and directions of future work follow in section 5.

2 Corpus processing

The QA@CLEF 2008 campaign establishes two different document collections for each language: a newswire collection and the Wikipedia. In the case of Basque a dump of the *Wikipedia 2006* and the *Euskaldunon Egunkaria* newspaper collection (from 2000 until 2002) were provided.

The document collection has been lemmatized before indexing it. Due to the fact that Basque is an agglutinative language, a given lemma makes many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite) for nouns and adjectives, and the person (me, he, etc.) and the tense (present, past, etc.) for verbs. For example, the lemma *lan* (”work”) forms the inflections *lana* (”the work”), *lanak* (”works” or ”the works”), *lanari* (”to the work”), etc. This means that looking only for the exact word given or the word plus an ”s” for the plural is not enough for Basque. And the use of wildcards, which some search engines allow, is not an adequate solution, as these can return occurrences of not only conjugations or inflections of the word, but also derivatives, unrelated words, etc. For example, looking for *lan** would also return all the forms of the words *lanabes* (”tool”), *lanbide* (”job”), *lanbro* (”fog”), and many more.

Before analysing the Wikipedia it needed to be parsed to clean the text, e.g. to get rid of HTML tags. So, we created a XML parser that extracts page title, paragraphs, and lists and then creates a simple XML document, which is very similar to the XML of the newspaper collection.

The entire document collection was lemmatized, part-of-speech tagged and named entity recognised. The named entity recogniser for Basque captures entities such as PERSON, ORGANISATION and LOCATION. The numerical and temporal expressions are captured by the lemmatizer/tagger.

Finally, the document collection was indexed by lemma; Swish-e[1] search engine was used and the retrieval unit is the passage.

3 System overview

A XML configuration file governs the running of the system. The configuration file is a declarative document where all the features involved in a run are described. The set of features is divided into two categories:

1. General requirements. It includes specifications such as the corpus to be used, the location of the list of questions to be answered, and the metrics and conditions for the evaluation.
2. Descriptors of the QA process itself. This subset of features represents the characteristics of the answering process. Mainly, it determines which modules act during the answering process, describes them and specifies the parameters of each module. In that way, the process is controlled by means of the configuration file, and different processing options, techniques, and resources can be easily activated/deactivated and adapted.

The principles of versatility and adaptability have guided the development of the system. It is based on web services, integrated using SOAP communication protocol. Some tools previously developed in the IXA group are used as

[1] http://swish-e.org

autonomous web services. This distributed model allows to parameterise the linguistic tools and to adjust the behaviour of the system during the development and testing phases.

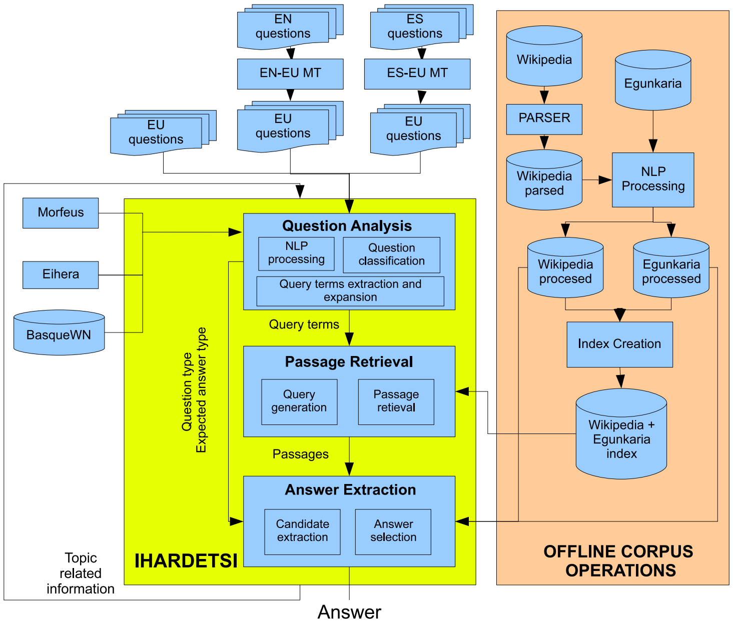

Fig. 1. The system general architecture

The communication between the web services is done using XML documents. This model has been adopted by some other systems ([2], [3]). The current monolingual version has three main modules: question analysis, passage retrieval and answer extraction. A complementary module, the question translation, has been added for cross-lingual versions. Fig. 1 shows the architecture of the system.

3.1 Question translation

A machine translation engine named *Matxin*[2] [4] has been used for question translations. This engine has been developed for translation from Spanish to Basque and it is rule-based. Due to the different structures of the languages the quality of the translations is not enough for dissemination, but it can be used for assimilation. It has been developed for a general domain and tested with texts from newspapers, but not with questions. A shallow test was carried out on factoid questions from previous trails of CLEF and we considered that the

[2] A free version of the MT engine is in a public repository (*matxin.sourceforge.net*)

results were good enough for using it in this task. Anyway a wider evaluation is required.

In the English to Basque translation we have used an early version of the English to Basque engine based on the same technology. The quality was poor and in a similar shallow test with factoid questions we detected that the translation of some question types were wrong, specially when the question marker was composed of two words that appeared as non-contiguous (i.e. *Where is he from?*). To face this problem a heuristic was applied after the translation process in order to repair bad translations of question markers. The heuristic was implemented using a few number of conditional rules, which work on the original and on the translated sentences.

3.2 Question analysis

The main goal of this module is to analyse the question and to generate the information needed for the next tasks. Concretely, a set of search terms are extracted for the passage retrieval module, and both the question type (factoid, list or definition) and the expected answer type along with some lexical information are passed to the answer extraction module. To achieve this goal, our question analyser performs the following steps:

Linguistic processing: The question analysis uses a set of general purpose tools like the morphological analyser, *Morfeus* [5], and the Name Entity recogniser and classifier, *Eihera* [6].

Question classification: In order to identify the question type, the question focus and the expected answer type, a set of rules has been defined after the examination of a Basque question set.

The question focus is the word or the word sequence that defines or disambiguates the question.For example, in the question *Which river is in the south of this country?*, the focus is *river* and in question *What is the North Pole?*, the focus is *North Pole*.

The next step is to identify the expected answer type. Our system's answer type taxonomy distinguishes the following classes: PERSON, ORGANISATION, DESCRIPTION, LOCATION, QUANTITY, TEMPORAL, ENTITY and OTHER. The assignment of a class to the analysed question is performed using the question stem, the syntactic construction and the type of the question focus. The question focus type is used to detect the expected answer type using BasqueWN [7] semantic file to the categories PERSON, ORGANISATION, LOCATION, QUANTITY, TEMPORAL.

Query terms extraction and expansion: All nouns, verbs, adjectives and abbreviations of the question constitute the set of search terms. They are lemmatized and arranged by their *Inverse Document Frequency* (IDF) value in the corpora in descending order.

Optionally, the search terms can be expanded by using synonymy, hyponymy and hypernymy information. To do this, the system uses a service which consults the lexical-semantic database BasqueWN.

3.3 Passage retrieval

The retrieval unit is a passage and not the entire document. The corpus is indexed by lemma using swish-e search engine. The corpus is batch-processed (see section 2): all words are lemmatized, and complex lexical units and entities are marked.

This module produces a set of queries taking as input a) the search terms selected by the question analysis module, b) the search terms selected by the question analysis module for the first question of a topic (if the question is not the first) and c) the first three answers of the first question of a topic (if the question is not the first). For each group a set of queries are created using relaxation techniques [8], and then they are combined to generate the set of final queries. Finally, they are executed until one of the queries retrieves a passage.

3.4 Answer extraction

Two tasks are performed in sequence: Candidate Extraction and Answer Selection. The candidate extraction consists of extracting all the candidate answers from the highest scoring passages. The answer selection consists of choosing the best three answers.

Candidate Extraction The process is carried out on the set of passages obtained in the previous step. First, all candidate answers are detected from each retrieved passage and a set of windows are defined around them. The selected window for each candidate answer is the smaller one which has all the query terms, or taxonomically related terms, in.

Then the candidate answers extraction process addresses each question type in a different manner, as follows:

- **Question type is Factoid:** the answer selection depends on entities in the most of the cases except when the expected answer type is *Entity* or *Other*. In such cases, all the entities and nouns near the question focus are selected.
- **Question type is Definition:** a set of rules have been defined to extract definition from retrieved text passages.
- **Question type is List:** we followed an heuristic looking for lists of candidate answers in the same passage, but it did not get the expected results.

Answer Selection: In order to select the best answers from the set of candidates, the same answers that appear in different passages must be combined. We try to map as identical those answers that refer to the same entity.

4 Results

This section describes the results obtained in our CLEF-2008 participation. We submitted four runs: one Basque monolingual, one English-Basque cross-lingual, and two Spanish-Basque cross-linguals. The methodology we employed targeted precision at the cost of recall, therefore we always choose NIL answers for those questions in which we could not reliably locate a candidate answer in the retrieved passage.

4.1 Monolingual system

As it was expected the best results were obtained for the monolingual task. Table 1 illustrates the results achieved by our system in the monolingual run.

It is clear that the best results were achieved for factoid questions. It is due to the fact that we focused on this type of questions in the development of the system. There were 145 factoid questions and 50 had a correct or inexact answer in the proposed three answers, 22 had a NIL answer (incorrect) and 73 had an incorrect answer. Analysing these 73 questions we detected that for 17 the correct passage was detected, but the system did not extract the correct answer.

Table 1. Results obtained in Basque to Basque monolingual run at QA@CLEF 2008.

	OVERALL	FACTOID	DEFINITION	LIST	TEMPORALLY RESTRICTED
RIGHT	26	23	3	0	2
WRONG	163	113	36	14	19
INEXACT	11	9	0	2	2
UNSUPPORTED	0	0	0	0	0
TOTAL	200	145	39	16	23
ACCURACY	13%	15.862%	7.692%	0	8.696%

The system answered NIL for 57 questions but only 4 of them were correct. Analysing the reasons we can group them in 5 groups:

- The expected answer type detection failed: 6 questions
- No passage was retrieved: 14 questions
- The passage had the answer but the system could not extract it: 13 questions
- Retrieved passages had not the answer: 16 questions
- Some other reasons: 4 questions

It is remarkable that no other system took part in the Basque as target task, so obtained results could not be directly compared with another Basque system. Anyway, it is interesting to contrast our results with some other languages, specially with those that, like Basque, have lack of resources, tools and experience in the QA field. For that purpose, we choose QA@CLEF2007 [9] results as a reference because it was the first time that topic-related questions and Wikipedia corpus were included. Although our results are faraway from the best ones, with overall accuracy of 54%, we realized that almost 40% of the runs got worse results than those of our system.

4.2 Cross-lingual systems

Three cross-lingual runs, two for Spanish-Basque and one for English-Basque, have been performed. The aim of the second run for Spanish-Basque was to test if the semantic expansion (see 3.2 section) of the question could compensate the lost of precision in the translation process. The results of the three runs are shown in Table 2.

Table 2. Results obtained in cross-lingual runs at QA@CLEF 2008.

	EN-EU					ES-EU					ES-EU with synonymy				
	R	W	X	U	ACC.	R	W	X	U	ACC.	R	W	X	U	ACC.
OVERALL	11	182	7	0	5.5%	11	182	7	0	5.5%	7	185	8	0	4.5%
FACTOID	8	130	7	0	5.517%	10	129	6	0	6.897%	7	130	8	0	4.828%
DEFINITION	3	36	0	0	7.692%	1	37	1	0	2.564%	0	39	0	0	0%
LIST	0	16	0	0	0%	0	16	0	0	0%	0	16	0	0	0%
TEMPORAL RESTRICTED	1	22	0	0	4.348%	2	21	0	0	8.696%	1	22	0	0	4.384%

The main conclusions we want to remark are:

- The results are quite poor. The loss of precision compared to the results of the monolingual system is more than 50%.
- Very similar results are obtained for the basic Spanish-Basque and for the English-Spanish runs (in both there are 11 right answers, 7 right answers in 2nd or 3rd place and 7 inexact in the first place). Due to the better quality of the Spanish-Basque translator we expected to improve the results for this pair of languages, but all runs have similar accuracy. Although the right results do not correspond always to the same questions. Only five of the eleven right answers are common.
- The semantic expansion in the second run for Spanish-Basque did not achieve better results. A slight smaller precision is observed, because some right answers are lost. However, new right or inexact answers appear but not in the first place. These figures might lead us to think that a higher number of *passages* are recovered, but it is not true, because the number of recovered *passages* remains at the same level (about 40 of 200).

5 Conclusions and Future work

The development stage of our monolingual Basque to Basque QA system has been described in this paper, as well as our participation in the QA@CLEF campaign. Thanks to this track we have had the opportunity of testing our system. Although the results might look not so good, our general conclusion is very positive taking into account that it was our first participation. In order to compare our results with other systems, it must be taking into account the particularities of Basque language. A shallow analysis reveals that our system is comparable with those that were relatively novels in the QA@CLEF track, especially when topics and Wikipedia had to be managed, as in the 2007 and 2008 editions. Moreover, we have been able to extract some of the strengths and weakness of each module of the system, which will be considered for future improvements.

6 Acknowledgements

We would like to thank Gorka Labaka and Mikel Lersundi who collaborated in this research. Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was funded by the Basque Government (Department of Language Policy and Department of Industry Anhitz project, Etortek IE06-185) and the Education Ministry (KNOW, TIN2006-15049-C03-01).

References

1. Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source. In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. (2008) 59–64
2. Tomás, D., Vicedo, J., Saiz, M., Izquierdo, R.: Building an XML framework for Question Answering. In: CLEF. (2005)
3. Hiyakumoto, L.: Planning in the JAVELIN QA System. In: CMU-CS-04-132. (2004)
4. Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: Transfer-based MT from Spanish into Basque: reusability, standardization and open source. In: Cicling 2007
5. Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J., Urizar, R.: Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In: COLING-ACL. (1998) 380–384
6. Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., Urizar, R.: Design and Development of a Named Entity Recognizer for an Agglutinative Language. In: IJCNLP. (2004)
7. Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. In: Proc. of the 2nd Global WordNet Conference. (2004)
8. Bilotti, M.: Query Expansion Techniques for Question Answering. Master's thesis, Massachusetts institute of technology (2004)
9. Giampiccolo, D., Peñas, A., Ayache, C., Cristea, D., Forner, P., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 multilingual question answering track. In: CLEF 2007 Working Notes, Online-Proceedings. (2007)