Ihardetsi question answering system at QA@CLEF 2008

Olatz Ansa, Xabier Arregi, Arantxa Otegi, Ander Soraluce
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 process a question in the source language (i.e. Spanish, English), translates into the target language (i.e. Basque) and, finally, the obtained Basque question is sent to Ihardetsi system. We submitted four runs, one for Basque-Basque subtask, one for English-Basque subtask and two for Spanish-Basque subtask.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Cross-lingual Question answering, Natural Language Processing

1 Introduction

Question answering (QA) systems tackle the task of finding a precise and exact answer to a question formulated in a natural language. Cross-lingual QA capabilities enable systems to retrieve the answer in one language (the target language) to a question posed in a different language (the source language).

This year, a new Basque-to-Basque monolingual task and English-to-Basque, Spanish-to-Basque cross-lingual QA tasks were organised for the first time within the context of the CLEF campaign.

The main goal of our first participation in QA@CLEF for Basque with Ihardetsi system was to evaluate our basic system in order to compare with other systems dealing with Basque and the state of the art with non-English question answering systems. Besides, the results analysis will reveal a number of future system improvements directions. We took part in Basque-Basque, Spanish-Basque and English-Basque tasks. Ihardetsi system is a Basque monolingual system, and we use a Spanish-Basque and English-Basque machine translation systems [4] for the Cross-lingual tasks 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 to solve main problems follow in section 5.

2 Corpus processing

This year's document collection consists of two different collections: a dump of the Wikipedia 2006 articles and Egunkaria newswire collection from 2000 until 2002.

The document collection has been lemmatized before indexing it. Due to 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"), *lanei* ("to the works"), *lanaren* ("of the work"), *lanen* ("of the works"), 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 the Wikipedia was analysed it needed to be parsed to clean the text, getting out html tags. So, we created a XML parser that extracts page title, paragraphs, and lists creating 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, ORGANIZATION, LOCATION and the numerical and temporal expressions are captured by the lemmatizer/tagger.

And finally, the document collection was indexed by lemma using Swish-e search engine and the retrieval unit is the passage.

3 System overview

The system relies on NLP tools, which perform a linguistic analysis, both on the question and on the corpus.

A XML configuration file governs the running of these components. 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. These descriptors constitute the documentation support of the system.

The principles of versatility and adaptability have guided the development of the system. The system is based on web services, integrated using SOAP communication protocol. Some tools previously developed in the IXA group are used as autonomous web services, and the QA system becomes a client that calls these services when it needs them. This distributed model allows to parameterize the linguistic tools, and to adjust the behaviour of the system during the development and testing phases.

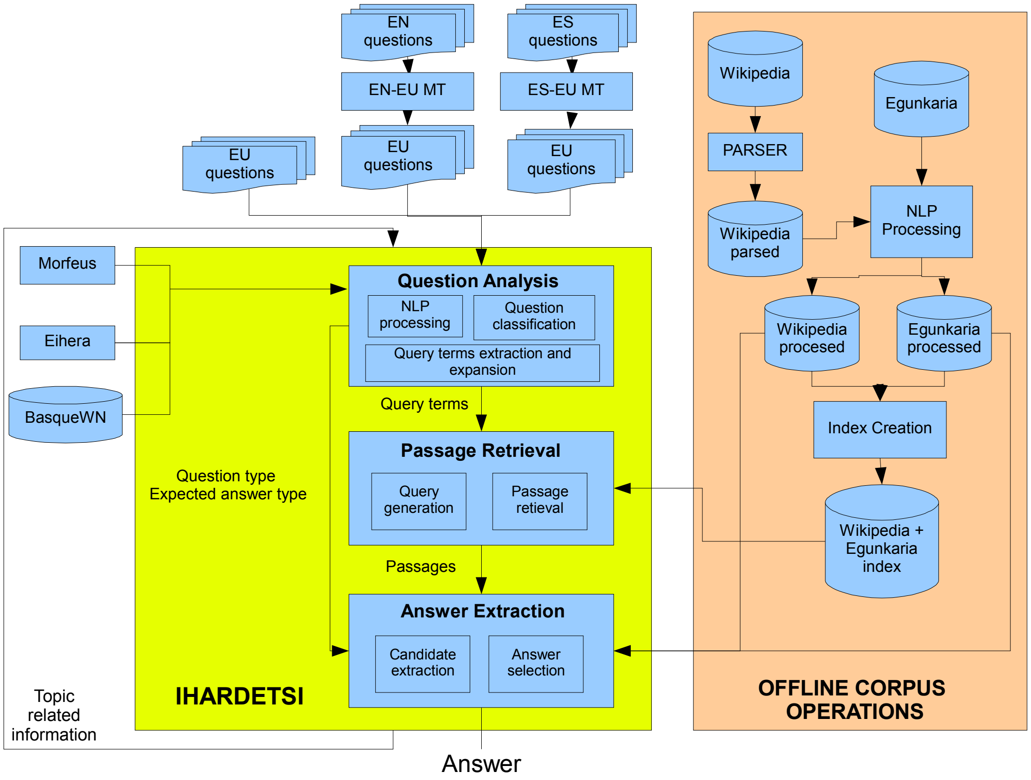

Figure 1: The system general architecture

The communication between the web services is done using XML documents. This model has been adopted by other systems ([1], [8]). Each service receives the input data in XML documents, and consults the general configuration file for specific information about execution parameters. The current version has three main modules, as it is very common in the question answering systems: Question analysis, Passage Retrieval and Answer extraction.

3.1 Question translation

A Machine Translation engine named *Matxin* [2] 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 it 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 the results were enough good for using it in this task. Anyway a wider evaluation is necessary. A free version of the MT engine is in a public repository (*matxin.sourceforge.net*).

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 the translated sentences.

In the near future we want to evaluate the quality of the translation of questions and to improve it, using if it is possible a corpus-based approach.

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, the question type (factoid, list or definition) and the expected answer type along with some lexical information is 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* [7], and the Name Entity recogniser and classifier, *Eihera* [3].

- **Question classification:** For identifying 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, i.e. it pinpoints what the question is searching for or what it is about. 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*.

  Next step is to identify the expected answer type. Our system's answer type taxonomy distinguishes the following classes: PERSON, ORGANIZATION, DESCRIPTION, LOCATION, QUANTITY, TEMPORAL, ENTITY, and OTHER. The assignment of a class to 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[1] semantic file to the categories PERSON / ORGANIZATION / 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 in descending order by their *Inverse Document Frequency* (IDF) value in the corpora.

  Optionally, the search terms can be expanded 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[2] search engine. The corpus is batch-processed (see section 2): all words are lemmatized, and complex lexical units and entities are marked.

This module takes as input:

1. the search terms selected by the question analysis module
2. the search terms selected by the question analysis module for the first question of a topic (if the question is not the first)
3. the first three answers of the first question of a topic (if the question is not the first)

[1] It is the Basque version of EuroWordNet. This resource is integrated in the Multilingual Central Repository (MCR), which is a multilingual lexical database developed in the Meaning project [5].

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

and produces a set of queries. For each group a set of queries are created using relaxation techniques [6], and then they are combined to generate the set of final queries. Finally, they are executed until one of them 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 answer score is computed like this:

  $$score_{CA} = \frac{\sum_{i=1}^{n} w_i}{n} \tag{1}$$

  where n is the window size and w_i is the i word weight. w_i is 1 for search terms, 0.8 for the synonyms of the search terms, 0.5 for hyponyms and hypernyms, and 0.3 for other question terms.

  The candidate answers extraction process then 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. For Entity and Other types we select all the entities and nouns near the question focus. Although the numerical and temporal expressions are marked in the processed corpus (see section 2), a grammar has been applied to mark even more.
  - **Question type is Definition:** a set of rules have been defined to extract definition from retrieved text passages.
  - **Question type is List:** an attempt has been done similar to Factoid questions but it was asked to be a list of answers being in the same sentence.

- 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. The formula used to compute the final score of each answer is as follows:

  $$final\_score_{CA} = \frac{\sum_{i=1}^{p} w_i}{N} \tag{2}$$

  where p is the number of identical answers and N is the number of candidate answers.

4 Results

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

	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%

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

4.1 Monolingual system

At 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.

The system answered NIL for 57 questions but only 4 of them were correct. Analysing the reasons for this 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 the answer: 13 question
- Retrieved passage had not the answer: 16 questions
- Some other reasons: 4 questions

After an analysis of the results of the NIL questions, we realized that some questions did not get any documents due to a bug in the system. Once it was corrected we performed a new run for Basque questions; 10 more questions were answered (9 DEFINITION questions and 1 FACTOID question) and for 8 questions the answer was changed (losing two correct answers). The answered new factoid question (i.e. *Where is Ocotal?*) was answered correctly and for the nine definitional questions 4 were answered correctly and 2 more had the correct answer in the second place.

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.

	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%

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

The main conclusions we want to remark are:

- The results are quite poor. The loss of precision respect to 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 better quality of the Spanish-Basque translator we hoped better results for this run. Anyway, it is necessary a wider evaluation of each MT engine when translating questions.
- Although the results are similar in average, the right results do not correspond always to the same questions. Only five of the eleven right answer are common.
- The semantic expansion in the second run for Spanish-Basque do not achieve better results. A slight smaller precision is observed, because some right answer are lost. In compensation to this, new right or inexact answer appear but not in the first place. With this figures we can think that at least a higher number of "passages" are recovered, but it is not true, because the number of recovered "passages" remains at 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 no good, our general conclusion is very positive taking into account that it was our first participation. We can not directly compare our system results with the results of other languages systems due to the particularities of Basque language. However, we have been able to extract some of the strengths and weakness of each module of the system, which we will take into account for future improvements. Besides we study the possibility of adding a fourth module to deal with topic-related questions.

6 Acknowledgements

We would like to thank Gorka Labaka and Mikel Lersundi who have collaborated in this research. Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work has been 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] Tomás D., Vicedo J.L., Saiz M., and Izquierdo R. Building an XML framework for Question Answering. In *CLEF*, 2005.

[2] Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and Sarasola K. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. In *Cicling 2007*.

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

[4] Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and 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*, pages 59–64, 2008.

[5] Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., and Vossen P. The MEANING Multilingual Central Repository. In *Proc. of the 2nd Global WordNet Conference*, 2004.

[6] Bilotti M. Query Expansion Techniques for Question Answering. Master's thesis, Massachusetts institute of technology, 2004.

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

[8] Hiyakumoto L. S. Planning in the JAVELIN QA System. In *CMU-CS-04-132*, 2004.