Using Wikipedia for Named-Entity Translation

Izaskun Fernandez
Tekniker-IK4
Eibar, Basque Country
ifernandez@tekniker.es

Iñaki Alegria
IXA Group, EHU
Donostia, Basque Country
i.alegria@ehu.es

Nerea Ezeiza
IXA Group, EHU
Donostia, Basque Country
n.ezeiza@ehu.es

Abstract: In this paper we present a system for translating named-entities from Basque to English using Wikipedia's knowledge. We can exploit interlingual links from Wikipedia (WIL) to get named-entity translation, but entities without interlingual links can be translated using the Wikipedia as a corpus, suggesting new interlingual links. In this second case the interlingual links can be used as a test corpus in order to evaluate the translation process. We just need Wikipedia articles in both languages (specially in the target language) and a bilingual dictionary to apply this methodology to other language pairs.
Keywords: Named-Entity Translation, exploiting Wikipedia

1 Introduction

Person, location and organization names, are the main types of named entities (NEs), and they are expressions commonly used in all kinds of written texts. Recently, these expressions have become indispensable units for many applications in the area of information extraction, as well as for many searching engines. Named-entity translation task also has an increasing interest in the NLP community, since this kind of systems might help in the improvement of multilingual systems, such as machine translation or question answering systems. The proper processing of named-entities might not only improve numerical results in machine translation but also comprehensiveness of translations. Most systems dealing with NE translation are based on parallel corpora, which are aligned to extract the necessary information about different kinds of phrases, including NEs. However, and as it is widely known, obtaining parallel corpora is not an easy task, and it is even harder when one of the languages in the language pair is a minority language, as it is the case of Basque.

Our main goal is to build a multilingual NE database that would be used in multilingual or cross-lingual systems in general. Since getting the information for that multilingual NE database was a complex task, we decided to work in the field of NEs translation, designing a system for translating those expressions between different language pairs.

Wikipedia[1] is a free on-line multilingual encyclopedia written collaboratively by volunteers, where anyone can add and change articles. Each article in Wikipedia is uniquely identified by its title. Normally, the title is the most common name for the entity described in the article. Those forms that refer to an entity but are not the common forms, are represented in the Wikipedia through redirect and disambiguation pages.

Since Wikipedia is a multilingual resource, we can find entries in the Wikipedias for different languages, representing the same entity in each corresponding language. For instance, the Basque entry *Euskal Herria* and the English *Basque Country* represent the same entity in different languages. Wikipedia uses interlingual links (WIL)[2] in order to relate those forms in different languages. So

[1] http://en.wikipedia.org

[2] Links for each Wikipedia entry that connect them to the corresponding entries in the Wikipedias for other languages.

an exhaustive translation proccess may be avoided if we exploit WILs. For those entities without WILs, we propose a translation system based on the contents of Wikipedia in two different languages, following similar steps described in (Alegria *et al.*, 2008) for translation based on comparable corpora.

The paper is structured as follows. Section 2 presents the related works. Section 3 presents how to exploit Wikipedia for named-entity translation task. In section 4 we describe the development of the NE translation system using a limited amount of linguistic knowledge. In section 5, we present the results of the experiments, and finally, section 6 presents some conclusions and future work.

2 Related Works

Considerable research effort has been recently focused on machine translation systems (MT). Even though, most of the MT systems will translate the Spanish form *escuela de derecho de Harvard* into *school of the right of Harvard* instead of *Harvard Law School* which is the correct English form (Reeder, 2001). So, besides being a good way to obtain multilingual NE information, NE translation can be also considered a helpful task for MT improvement.

Concerning the resources, despite the difficulty to get bilingual parallel corpora for many languages, most NE translation systems work with parallel datasets. Those bilingual corpora are aligned not only at the paragraph level but also at the sentence level. For example, Moore's work (Moore, 2003) uses bilingual parallel English–French aligned corpora, and he obtains a French form for each English entity applying different statistical techniques.

Although comparable corpora have been less studied, there are some known systems designed to work with them as well, such as the system that translates entity names from Arabic to English (Al-Onaizan and Knight, 2002a) (Al-Onaizan and Knight, 2002b), and the Chinese–English translation tool presented in ACL 2003 (Chen *et al.*, 2003).

The main goal of both systems is to obtain the equivalent English form, taking Chinese and Arabic respectively as source language. Two kinds of translations can be distinguished in both systems: direct/simple translations and transliterations[3]. However, each tool uses a different technique. Frequency-based methods are used in Chinese–English translations, while in the Arabic–English language pair, a more complex combination of techniques is applied.

Similar techniques are applied in (Sproat *et al.*, 2006) and (Tao *et al.*, 2006), which transliterate English–Chinese NEs using comparable corpora. The former combines a supervised phonetic transliteration technique and a phonetic frequency correlation approach, while the latter combines those techniques, but applying the phonetic approach in an unsupervised way, where the distance is determined by means of combining the substitution, insertion and deletion of characters.

Not only approaches to named-entity transliteration have been presented in this area. The system presented in (Poliquen *et al.*, 2005) integrated at the news analysis system NewsExplorer[4], tries to extract person names from multilingual news collections to match name variants referring to the same person, and to infer relationships between people based on the co-occurrence information in related news.

WIL links are used to try to enrich the German–English pair (Sorg and Cimiano, 2008). They show that roughly 50% of the articles in German are linked to their corresponding English version and only 14% from English to German. They present a classification-based approach based on text-based features and/or graph-based features for that enrichment. The experiments show that their approach has a recall of 70% with a precision of 94%.

Multilingual named-entity recognition based on Wikipedia is faced on (Richman and Schone, 2008), showing that English language data can be used to bootstrap the NER process in other languages. For multilingual categorization they use links among languages when possible, and categories with their English equivalents in the remaining cases.

Concerning Basque, in our previous work we have found two different approaches to translate Basque NEs into Spanish (Alegria *et al.*, 2006). The first one was a language de-

[3]Transliteration is the proccess of replacing words in the source language with their approximate phonetic or spelling equivalents in the target language.

[4]`http://press.jrc.it/NewsExplorer/entities/en/1.html`

pendent tool for translating NEs from Basque to Spanish using comparable corpora. That system used linguistic information for both transliteration and entity element rearrangement. This system was tested using a set of the most common entities, and it obtained an f-score of 78.7% in named-entity translation task.

However, as the development of a language dependent system for each language pair was very expensive, we tried a relatively language semi-independent tool following a similar strategy, and using comparable corpora and bilingual dictionaries. This tool was tested first in the Basque–Spanish language pair and it shows that the performance was quite close to the language dependent tool, obtaining an f-score of 77.5%.

To confirm that the methodology was general enough, we tried using the translation methodology for the Spanish–English language pair in (Alegria *et al.*, 2008) and we obtained almost 65% f-score, which is a considerable lower performance. After observing the errors in detail, it was detected that due to the bad quality of the comparable corpora the 13% of the English entity forms where badly defined in the target corpus, so by correcting them the system would get results that are as good as the ones got for Basque–Spanish language pair. So the methodology based on comparable corpora seemed to be a good choice for developing systems to translate NE for different language pairs.

Thus we have obtained the language semi-independent approach to develop the Basque–English NE translation tool but using Wikipedia as a corpus. We will describe in more detail the system architecture in the following section.

3 Exploiting Wikipedia

Many articles of Wikipedia can be found in different languages, and that is why this encyclopedia can be considered an interesting resource to get named-entity translations between different language pairs, specially if the target language is English.

In this work, we exploit Wikipedia contents in many different ways. As we have mentioned before, the encyclopedia has interconnected entries from different languages by means of interlingual links (WIL), which means that both entry forms represent the same entity in their corresponding language. So this is the most effective and cheapest way to get named-entity translations from Wikipedia. Unfortunately, some named-entity pairs lack an interlingual link, which means that other techniques must be used in order to translate them.

Most of the translation systems use a target language lexicon, and as we will see in the next section, we also use an English lexicon in our system. Since the English Wikipedia is very rich and a resource that is continuously growing, we considered it an interesting source for our target lexicon generation. But as we are dealing with NEs, we are only interested in words appearing in this kind of expressions and not just in any kind of words. Because of the wide coverage of the target Wikipedia we assume that most of the source words would have their corresponding translation in the target lexicon if we use this resource.

Yahoo! has a semantically annotated Wikipedia[5] version presented in (Atserias *et al.*, 2008), where NER task has been applied. We took this version and extracted all the tagged NEs. We constructed the target lexicon by means of including words in those NEs and excluding grammatical words such as prepositions, articles, etc. using a stoplist[6].

Combining the lexicon with some techniques that we will see in the following section, we constructed a system that proposes some translation candidates for each given Basque entity. In order to make sure if they are suitable proposals, we can use Wikipedia again, this time for searching if proposals have an entry in the on-line encyclopedia. So, the system only gives revised named-entity translation proposals.

4 System Description

The system proposed for named-entity translation task in this work uses three main modules: 1) a searching module with different resources for searching, 2) an entity elements translation module using a transliteration grammar combined with a bilingual dictionary for those words that cannot be translated only by applying transliteration but still need some translation and 3) an el-

[5] `http://barcelona.research.yahoo.net/dokuwiki`

[6] `http://www.lc.leidenuniv.nl/awcourse/oracle/text.920/a96518/astopsup.htm`

ement rearranging module for the construction of the whole entity from components, which will treat the possible differences in syntactic structures.

As Figure 1 shows these three modules are applied following four main steps when a Basque NE is given for translation:

- Step 0: Searching for WIL between the Basque entity and an English Wikipedia entry (*Searching Module*)
- Step 1: Searching for a translation for the entire Basque NE as a multiword lemma in the bilingual dictionary (*Searching Module*)
- Step 2: Searching the Basque entity in the English Wikipedia (*Searching Module*)
- Step 3: Translation/transliteration of entity elements themselves, finally constructing the entire translation proposal using the individual translations and searching these entire proposals in the English Wikipedia.

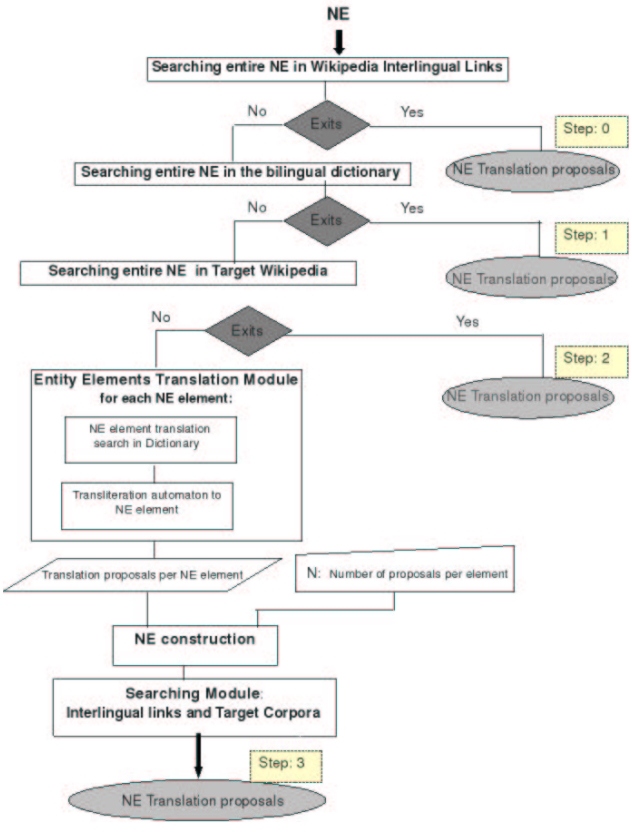

Figure 1: System Architecture

In the following subsections we will present each module in detail.

4.1 Searching Module

This module contains three main functionalities which are used in different steps of the system architecture: searching in Wikipedia's interlingual links, in the target Wikipedia, and in the bilingual dictionary.

As the previous step (0) the system exploits the interlingual links of Wikipedia and if a link exists between the input Basque entity and an English Wikipedia entry, the system suggests this English form as translation for the input expression. If no link is found, step (1) is applied.

In the step (1) the whole NE is searched in the bilingual dictionary. If the form is found the translation is obtained. This step is applied as a baseline in the experiments. For example the translation *Euskal Herria–Basque Country* is resolved using the dictionary. If no translation is found, step (2) is applied.

In the step (2) the system verifies if the input Basque form exists with the same form in the English Wikipedia. If it does, then the system will propose the entity in the target language as translation proposal.

When no translation proposal is obtained in the previous steps, and the system has to translate each element and construct then the entire translation posible forms, the last functionality of this module is applied in order to reduce the amount of suggested proposals and to improve the quality of them: in this step the system checks for every translation proposal if the proposed entry exists in the English Wikipedia, and only forms with entries will be given as possible translations.

As an additional functionality the system exploits the redirection links of Wikipedia. As it has been explained previously, these links connect the different forms in Wikipedia to refer to an entity in the same language. This way, the system returns not only those entity translations found or trusted to exist in the Wikipedia, but also all their connected forms.

The system uses the MediaWiki API [7] to exploit both the interlingual and the redirection links. For example the translation (*Alpeak*, *Alps*) can be obtained exploiting interlingual links and the result can be enriched with the pair *Alps–The Alps* using the redirection links.

[7] http://www.mediawiki.org/wiki/API

4.2 Entity element translation module

When no entity translation proposal is obtained from step (0), step (1) and step (2) the system applies the word-by-word translation process. The bilingual dictionary and the finite-state transducer, combined with the English lexicon, will be used in order to obtain translation proposals for each entity element.

As it is explained in (Alegria *et al.*, 2008), edit distance (Kukich, 1992) based on a finite-state grammar and a lexicon of the target language are neccesary for constructing trasliteration rules. Since each rule can be applied n times for each word, the set of all translated words that we obtain after applying rules independently and combining them, is too large. In order to reduce the size of the set of proposals, the system combines the grammar with the lexicon of the target language obtained from the Wikipedia, and it restricts the transformation rules to at most two applications per word, avoiding the generation of words with more than two transformations, as it is shown in the top of Figure 2.

With this transliteration automaton, the system will be able to translate *Txina* into *China*.

However, there are some translations that cannot be obtained applying only transliteration/edition rules. The system uses a source-target bilingual dictionary, converted into an transducer for this aim. The module strategy is shown in the bottom part of Figure 2 and is applied in the following order:

- get translation looking up the bilingual dictionary.
- suggest an identical word if it is in the target lexicon.
- propose words in the target lexicon with distance 1 from the source word.
- suggest words in the target lexicon with distance 2 from the source word.

So this module is able to translate not only the transliterated words such as *Kuba–Cuba*, but also, words that cannot be translated using transformation knowledge and need information from a bilingual dictionary, such as *Erakunde–Organization*.

For both transliteration and bilingual dictionary based automata, the system uses

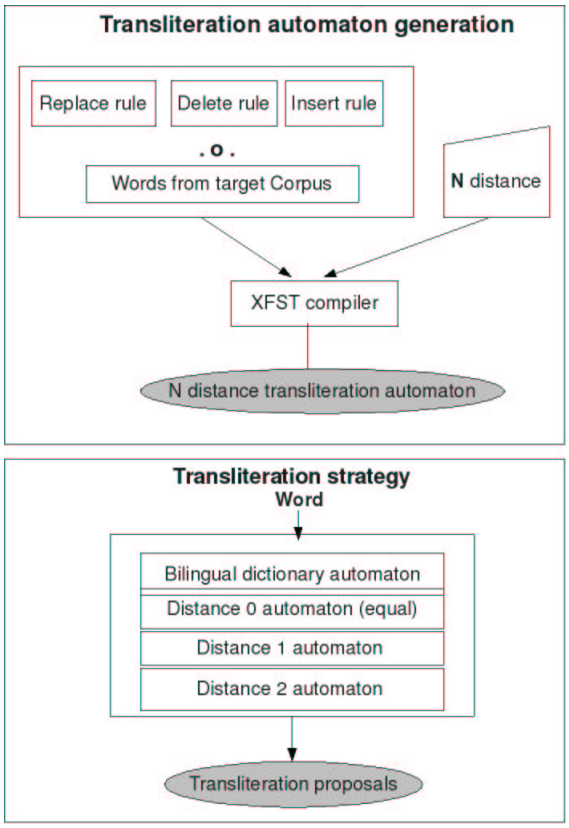

Figure 2: Transliteration automaton and strategy

the lemmatized form of the entity elements, applying Eustagger, the Basque lemmatizer/tagger (Ezeiza *et al.*, 1998) developed by IXA group .

Both kinds of automata are combined for translating entities like *Ipar Katalunia* into *North Catalonia*, using the dictionary for converting *Ipar* into *north*, *northern*, *north wind* and transliteration for transforming *Katalunia* into *Catalonia*, *Catalunya* and *Katatonia*. After looking up in the Wikipedia *North Catalonia* and *Northern Catalonia* forms are suggested.

4.3 Entire Entity Construction

Once each element is translated, the entire entity construction must be performed. For this work we cannot ignore the different syntactic patterns between languages, and this makes necessary to include some treatment for element rearrangement. This module is applied before searching for translation candidates in the target Wikipedia. As mentioned in (Alegria *et al.*, 2008), this module combines each proposed element with the rest, considering that each proposal can ap-

pear in any position within the entity.

Although in some cases prepositions and articles are needed to obtain the correct target form, the translation candidates for the whole entity will not contain any element apart from the translated words of the original entity. So, we will take into account the lack of these elements in the following step.

For reducing the amount of translation proposals, only the N most suitable translations for each word will be considered for the entire construction. For instance, when the system tries to translate *Itsaso Gorria* , the system gets candidate *Sea* for *Itsaso* and *Red* for *Gorria*. And this module generates the following entire candidates, considering that both elements can appear in any position: *Sea Red* maintaining the original position of words, and *Red Sea* inverting the positions. In this case, the correct form is the one obtained changing word order.

5 Experiments

We had no evaluation corpus for the system, so we considered convenient to generate an evaluation corpus in a semi-automatic way. We used two resources for this task: Wikipedia and the CLEF evaluation set. For the Wikipedia set we exploited the interlingual links in order to obtain the gold standard for testing; so, in this case the (0) step will be not applied.

For both evaluation sets we have used the same three measures:

- $Precision = \frac{correctly\_translated\_NEs}{Translated\_NEs}$
- $Recall = \frac{correctly\_translated\_NEs}{All\_NEs}$
- $f - score = \frac{2*Precision*Recall}{Precision+Recall}$

5.1 Evaluation with Wikipedia based corpus

For the construction of the first evaluation corpus, we have used a Basque article collection borrowed from *Euskaldunon Egunkaria*[8], which is a newspaper entirely written in Basque closed since February 20th 2003, and the WILs. The Basque corpus has 40,648 articles with 9,655,559 words and 142,464 NEs tagged in the Hermes project[9] (news databases: cross-lingual information retrieval and semantic extraction).

We selected the most frequent NEs from the Basque collection and searched the WILs in Basque–English direction to find linked English forms, we got a collection of 575 entity pairs interlinked in the Basque–English Wikipedia languages. Since interlingual links are used for the corpus generation, we will not use them for suggesting translations (Step 0 in Figure 1 is not carried out).

Steps	Total	Correct
Step (1)	17	11
Step (2)	391	375
Step (3)	59	48
No-Translation	108	0

Table 1: Translations distribution

Table 1 shows the number of translated entities in each step of the system, together with the amount of well translated entities. In the first row, we can see the number of entities that have been found in the dictionary. The second row shows how many Basque entities have been found in the English Wikipedia, thus they have no need to be translated element by element. In the third row we can see the entities that have been translated using the language semi-independent system[10]. Finally, we can see that around 19% of times the system did not obtain a translation.

	Pr.	R.	fs
Baseline	59.82%	59.82%	59.82%
Our system	93.36%	75.82%	83.68%

Table 2: Results for Wikipedia-based test set

In Table 2 we present the evaluation, and in the first row a baseline is shown. The baseline is calculated considering correct translations when Basque and English forms are identical.

The results are very encouraging, since we have obtained 83.68% f-score.

Analysing the errors in the development corpora we observed that sometimes WILs do not link the same entity form. For instance, if *Dorre Bikiak* is searched in the Basque Wikipedia and the interlingual link is used to obtain the English translation, the same way it has been used to build the test corpus, the Basque form found is *World Trade*

[8] http://www.unibertsitatea.net/blogak/ixa/egunkaria-hizkuntza-teknologiako-baliabideen-sortzailea

[9] http://nlp.uned.es/hermes/

[10] Translating each entity, constructing the English proposals with elements' translation and finding the best proposals with the searching module

Center, instead of *Twin Towers* which must be the interlingual linked form.

With the proposed translation system this kind of new links could be good suggestion to be added to the Wikipedia.

5.2 Evaluation with CLEF based corpus

Since Wikipedia has been used for constructing the named-entity translation system presented in this work, it can be considered that the evaluation corpus is biased in favor of the system. So, we considered interesting to evaluate the system using another NE set, independent from this encyclopodedia.

For that purpose, we used the ResPubliQA CLEF-2009[11] test set. This test set has 500 questions translated into Bulgarian, Basque, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish. We used Basque and English versions to construct the new evaluation set for the NE translation system.

Exploiting the questions set, we obtained 72 Basque–English NEs pairs, where 9 of them has no entry in the target Wikipedia. Since our system only proposes english translations trusted in the English Wikipedia, even if it gets correct English forms for that 9 entities, it would never propose them because they are not in the English Wikipedia. So, we can say that for this test set, our system's topline recall is 87.5%.

We have tested the system in two different ways for this evaluation corpus: When the system does not find out any of its proposals in the target Wikipedia,

- no translation is returned (silence-mode).
- the original Basque form is returned as translation proposal (talkative-mode).

	Pr.	**R.**	**fs**
Baseline	23.61%	23.61%	23.61%
Silence-mode	92.68%	52.77%	67.25%
Talkative-mode	55.5%	55.5%	55.5%

Table 3: Results for CLEF test set

Since CLEF test set does not belong to the Wikipedia, for this evaluation, we exploit the WILs between Basque and English Wikipedias(step0 in the system architecture), in order to evalaute the complete system architecture. Exploiting those links, we obtain translations proposals for 26 of the 72 entity colection, and all but one agree with the test set English forms.

As you can see in Table 3, this time the results are not as encouranging as the previous ones, but we want to highlight the improvement that the system gets respect to the baseline. So it would be pretty good to evaluate the system with a bigger and more extensive corpus.

6 System Improvement

Analysing the errors occurred with both evaluation sets, we detect that our biligual dictionary was not very suitable in many cases for translating words appearing in NEs. For example, if we try to translate *Nazio Batuak* into *United Nations* and we use our bilingual dictionary for it. First we will lemmatize the basque elements, then look up them into the dicitonary and finally we will get *Nation* and *Union* respectively. With this forms we will never get the correct English entire entity form.

But if we were able to enrich the dictionary with *Nazio–Nations* and *Batuak–United* pairs, the system will be able to obtain *United Nations* as translation candidate.

So, we decided to carry out an automatic dictionary lexical enrichment expoiting a small Basque–English WIL set, and then check if that enrichment improves our system performance, evaluating the system with CLEF test corpus.

We take as Basque–English WIL input set the wrong translated entities from the Wikipedia Based test corpus, concretly 84 entity pairs. For each entity pair, we try to match each Basque entity element with their corresponding English entity element mantaining the source Basque form, or translating with the existing bilingual dictionary. When every element in the Basque entity is parsed, if only one basque element has not been matched and in the target entity there is only one element too that has not been assigned, we will consider those elements as translations, and we will enrich the bilingual dictionary with them. This metholodogy will be applied iteratively, enriching the dicitonary in each step and using it in the following one, until no new elements' translation proposals are obtained.

[11] `http://celct.isti.cnr.it/ResPubliQA/`

For instance supose that in the input WILs set we have *Europako Parlamentua–European Parliament*. Carrying out the previously explained matching, we will not get any matching for *Europako* because it does not exist in bilingual dictionary, and it does not match identically with any of the English elements. Even though, from the bilingual dictionary we get that *Parlamentua* matches with *Parliament*. So *Europako* is the unique element in the Basque form without matching, and *European* in the English form. Applying the previous assumption, we will consider *European* a possible translation form for *Europako*, and we will entrich the dictionary with it.

	Pr.	**R.**	**Fs**
Silence-mode	93%	55.5%	69.51%
Talkative-mode	58.33%	58.33%	58.33%

Table 4: Results for CLEF test set with lexical enrichment

As you can see in Table 4, observing a very small set of entity pairs, we obtain a sligth improvement in the system. So it could be interesting to consider the entire Basque–English pairs linked with WILs to get a better lexical enrichment.

7 Conclusions and Further Work

We have presented an approach to translate NEs using the Wikipedia encyclopedia as main resource. It has been shown that exploiting Wikipedia might benefit in two directions: on the one hand it may help in building a good quality named-entity translation system; on the other hand, new interlingual links for Wikipedia might be suggested.

The evaluation gives us promising results but a deeper evaluation and error analysis is needed, for studying solutions for entities with different number of elements in each language. It would be also interesting to test this technique in other languages.

As further work we want to disambiguate NE written in minority languages such as Basque. Since the resources for that kind of languages are very limited, we are intending to use the translation system proposed in this paper for exploiting the information of the languages with much more resources like English, and the Wikipedia's disambiguation links.

Acknowledgment

This was partially supported by the Spanish Ministry of Education and Science (FIT-340000-2007-157 carried out at Tekniker and TIN2006-15307-C03-01)

References

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98. Pgs.380–384 Vol 1. Montreal(Canada). August 10-14,1998.

Alegria I., Ezeiza N., Fernandez I. 2006. *Named Entities Translation Based on Comparable Corpora*. Proceedings of Multi-Word-Expressions in a Multilingual Context Workshop in EACL 2006. W06–2401.

Alegria I., Ezeiza N., Fernandez I. 2008. *Translating Named Entities using Comparable Corpora*. Proceedings of Building and Using Comparable Corpora Workshop in LREC 2008.

Al-Onaizan Y., Knight K. 2002. *Translating Named Entities Using Monolingual and Bilingual Resources*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002. Pgs. 400–408.

Al-Onaizan Y., Knight K. 2002. *Machine Transliteration of Names in Arabic Text*. Proceedings of the ACL-02 workshop on Computational approaches to semitic languages. Pgs. 1–13.

Atserias J., Zaragoza H., Ciaramita M., Attardi G. 2008. *Semantically Annotated Snapshot of the English Wikipedia*. Proceedings of LREC 2008. L08–1165.

Beesley K.R., Karttunen L. 2003. *Finite State Morphology: Xerox Tools and Techniques.* CSLI Publications

Chen H., Yang C., Lin Y. 2003. *Learning Formulation and Transformation Rules for Multilingual Named Entities*. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition. Vol. 15 Pgs. 1–8.

Kukich K., 1992. *Techniques for automatically correcting word in text. ACM Computing Surveys* Vol. 24 No. 4 377-439

Moore R. C., 2003. *Learning Translations of Named-Entity Phrases from Parallel Corpora*. Proceedings of EACL 2003. Vol. 1 Pgs. 259–266.

Poliquen B., Steinberger R., Ignat C., Temnikova I., Widiger A., Zaghouani W., Žižka J. 2005. *Multilingual person name recognition and transliteration*. CORELA - COgnition, REpresentation, LAnguage, Poitiers, France, CERLICO. ISSN 1638-5748, 2005, vol. 3/3, no. 2, pp. 115-123.

Reeder F. 2001. *The Naming of Things and the Confusion of Tongues*. MT Evaluation: Who Did What To Whom Workshop on MT Summit VIII. Pgs. 55–59.

Richman A. E., Schone P. 2008. *Mining Wiki Resources for Multilingual Named Entity Recognition* Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008. Pgs. 1–9.

Sorg P., Cimiano T. 2008. *Enriching the Crosslingual Link Structure of Wikipedia - A Classification Based Approach*. Proceedings of the AAII 2008, Workshop on Wikipedia and Artificial Intelligence, 2008.

Sproat R., Tao T., Zhai C. 2006. *Named Entity Translation with Comparable Corpora*. Proceedings of the 21st International Conference on Computational Linguistic and 44th Annual Meeting of the ACL 2006. Pgs. 73–80.

Tao T., Yoon S., Fister A., Sproat R., Zhai C. 2006. *Unsupervised Named Entity Translation Using Temporal and Phonetic Correlation*. Proceedings of the 2006 EMNLP. Pgs. 250–257.

Ihardetsi: A Question Answering system for Basque built on reused linguistic processors

Iñaki Alegria, Olatz Ansa, Xabier Arregi, Arantxa Otegi, Ander Soraluze
IXA Group. University of the Basque Country
xabier.arregi@ehu.es

Abstract: This paper presents *Ihardetsi*, a question answering system oriented to Basque. We describe the main architecture of the system, paying special attention to the use of linguistic resources and tools. The system has been built reusing such tools, on the basis that general linguistic processors can be adapted to satisfy the requirements of the question answering task. This methodology can be suitable for other projects, specially when lesser resourced languages are involved. Along with the description of the system, we outline its performance presenting some experiments and the obtained results.

Keywords: QA, Question Answering, Basque

1 Introduction

Question answering systems tackle the task of finding a precise and concrete answer for a natural language question on a document collection.

This task involves the use and adaptation of IR (Information Retrieval) and NLP (Natural Language Processing) resources, techniques and tools.

The current version of *Ihardetsi*,[1] a Basque question answering system, takes Basque questions as input and the corpora on which the answers are searched are written in Basque too.

The system incorporates tools and resources developed previously in the IXA group, like a lemmatizer/tagger, *Morfeus* (Ezeiza et al., 1998), and a recognizer and classifier of named entities (NERC) for Basque, *Eihera* (Alegria et al., 2004). Additionally the Basque Wordnet (Agirre et al., 2006) has been used in order to improve the results of the system.

The remainder of the paper is organized as follows. Section 2 is devoted to introduce the general architecture of the system. In section 3 we describe the main modules of the QA system. Then, in section 4 it is explained how the tool has been adapted to a new domain and to a multilingual environment. Finally, evaluation issues are discussed and some conclusions and suggestions for future research are pointed.

2 General Architecture

The principles of versatility and adaptability have guided the development of the system. It is based on web services, integrated by the SOAP (Simple Object Access Protocol)[2] communication protocol. As we have already remarked, some tools previously developed in the IXA group are reused as autonomous web services, and the QA system becomes a client that calls these services when it needs them. This distributed model allows to parametrize the linguistic tools, and to adjust the behaviour of the system during the development and testing phases.

The communication between the web services is done using XML documents. This model has been adopted by some other systems (Tomás et al. 2005, Hiyakumoto 2004).

The global features of each run are described in a XML configuration file. The set of features is divided into two categories:

1. General requirements. It includes specifications such as the corpus to be used, the processing model of the corpus,

[1] The name of the system, *Ihardetsi,* comes from a Basque word, generally used in the North dialect, which means "to answer".

[2] *www.w3.org/TR/soap/*

the location of the list of questions to be answered, and the description of the type of questions.

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.

As it is common in the question answering systems *Ihardetsi* is founded on three main modules: the question analysis module, the passage retrieval module and the answer extraction module.

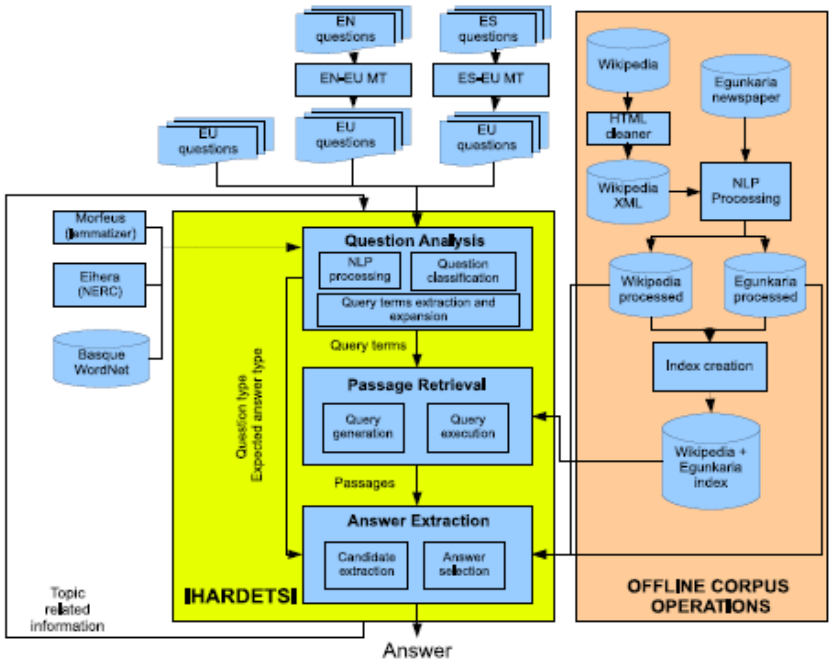

Figure 1: General architecture of the system.

3 Main modules of the system

The main modules in Fig. 1 are explained in the next subsections. Some of the off-line corpus operations are explained beside the Passage Retrieval section.

3.1 Question Analysis

The main goal of this module is to analyse the question and to generate the information needed for the next tasks. On the one hand, a set of search terms are extracted for the passage retrieval module (see section 3.2); on the other, the question type (*factoid*, *list* or *definition* mainly) and the expected answer type, along with some lexical information is passed to the answer extraction module.

The question analyser performs the following steps:

1. *Linguistic processing of the question*: The question analysis reuses a set of general purpose tools like the lemmatizer/tagger named *Morfeus*, and the NERC processor named *Eihera*.
2. *Question classification*: For identifying the main features of a question we attend to the question type, the question focus and the expected answer type. This process is carried out by means of a set of pattern rules that have been defined according to the Basque questions' structure.

   The taxonomy of answer types in our system distinguishes the following classes: *person*, *organisation*, *description*, *location*, *quantity*, *time*, *entity* and *other*. The assignment of a class to the analysed question is performed using the interrogative word, some heuristics of syntactic nature and the type of the question focus. The question focus is mapped with the semantic file characteristic of the Basque WordNet, and in this way we can determine more precisely the expected answer type.
3. *Extraction and expansion of query terms*: 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)[3] value in the corpora.

[3] *Inverse Document Frequency*, a factor in the TF-IDF principle (http://en.wikipedia.org/wiki/Tf-idf)

Optionally, the search terms can be expanded using synonymy, hyponymy and hypernymy information. This expansion is carried out just in one level, without applying any word-sense disambiguation. The synonyms, hyponyms and/or hypernyms are selected by reusing a service that consults the *Basque WordNet* lexical-semantic database. This resource is the Basque version of *EuroWordNet*, and is integrated in the *Multilingual Central Repository* (MCR), which is a multilingual lexical database developed in the *Meaning* project (Atserias, 2004).

3.2 Passage retrieval

Basically an information retrieval task is performed, but in this case the retrieved units are passages and not entire documents.

This module receives as input the selected query terms and produces a set of queries that are passed to a search engine. We have tested different search engines, like *Swish-e*[4], *Jirs*[5] and *Indri* (Strohman et al., 2005). As long as none of them is equipped to deal with Basque, we need to process the corpus before indexing it, as will be explained in the following paragraph.

Since Basque is an agglutinative language, a given lemma takes 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 given exact word, 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 it can be returned occurrences of not only inflections of the lemma, 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. So, a stemmer or a lemmatizer/tagger is almost indispensable for this kind of languages. *Snowball*[6] can be a good option for stemming for some languages since it is open-source. We made an attempt for Basque, but it was not successful.

In our case, the entire document collection was lemmatized/tagged with part-of-speech and named entities. In this task we reused the lemmatizer/tagger named *Morfeus*, that returns only one lemma and one part-of-speech for each lexical unit. The NERC processor, *Eihera*, captures entities such as *person*, *organization* and *location*. The numerical and temporal expressions are captured by the lemmatizer/ tagger. Up to date, no semantic pre-processing has been performed.

3.3 Answer extraction

In this module two tasks are performed in sequence: the candidate extraction and the answer selection. Basically, the candidate extraction consists of extracting all the candidate answers from the retrieved passages, and the answer selection consists of choosing the best answers among the considered as candidates.

Candidate extraction: Firstly, 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 smallest one which has all the query terms.

In order to extract the candidate answers the system addresses each question type in a different manner, as follows:

- Question type is *factoid*:[7] the answer selection depends on the named entities.
- Question type is *definition*: a set of rules have been defined to extract definitions from retrieved text passages.
- Question type is *list*: we followed a heuristic looking for lists of candidate answers in the same passage.

Answer selection: In order to select the best answers from the set of candidates, the same answers that appear in different passages must

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

[5] http://sourceforge.net/projects/jirs/

[6] http://snowball.tartarus.org/

[7] As far as we know, there is no formal definition of *factoid question*. Intuitively, a *factoid* is asking about a simple *fact* or relationship, and the answer is easily expressed, usually by means of a named entity.

be combined. We try to map as identical those answers that refer to the same entity. For instance, "Miguel Indurain" and "Indurain" are different strings, but the system must detect that both refer to the same person. This point is relevant when assigning weights to the candidate answers. The formula used to compute the final score of each answer is as follows:

$$S(C) = \frac{\sum_{i=1}^{p} w_i}{N}$$

where *S(C)* is the score of the candidate *C*, *p* is the number of answers identical to *C*, w_i is the confidence score of the *i*-th identical answer, and *N* is the number of all candidate answers.

4 Applications

In this section we present several scenarios where *Ihardetsi* has been used and tested. The nature and the conditions of these scenarios are quite heterogeneous. The evaluation of some of them is reported in the next section.

The first version of the tool was carried out using a news corpus (*Euskaldunon Egunkaria*), which was previously processed in a project on IR.[8] *Euskaldunon Egunkaria* was the only Basque language newspaper in the world for years, but it was closed in 2003. We use a corpora from 2000, 2001 and 2002 years, with about 7 million words.

In 2008 we were involved in the CLEF-QA 2008 evaluation,[9] using the previous corpus and the Basque Wikipedia as evidence for the questions. The architecture of the system remained the same, and only a small change was introduced when the Wikipedia was preprocessed: the headword of an entry was inserted at the beginning of every paragraph of the entry, with the aim of considering all the paragraphs when answering questions about such entry. We participated in the cross-lingual QA evaluation too, answering questions in Spanish and English based on the Basque repository. We just translated the questions by reusing the *Matxin* technology (Alegria et al., 2007) with a small adaptation in order to improve the translation of the questions. Taking into account the specific structure of some questions, we performed an automatic post-edition (based on regular expressions) process for repairing some translations.

After, during the *Anhitz* project we had to adapt *Ihardetsi* to new features: cross-language, in the same way as in CLEF, new domain, science and technology, and multimodal process, question were processed by a speech recognizer.

The aim of the *AnHitz* project (Arrieta et al., 2008), whose participants are research groups with very different backgrounds, is to carry out research into language, speech and visual technologies for Basque. Several resources, tools and applications were integrated into a prototype of a 3D virtual expert on science and technology. It includes our QA system.

5 Evaluation

We have combined quantitative evaluations (mainly those that are framed in the CLEF campaigns), with more qualitative ones (specifically we tackle them in the *Anhitz* project).

5.1 CLEF 2008

This section describes the results we obtained in our first participation in the CLEF 2008 campaign, specifically in the Basque to Basque monolingual QA task (Ansa et al., 2008).

The exercise of the main QA task consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and contained co-references between one question and the others. Moreover, besides the usual news collections provided by ELRA/ELDA, articles from Wikipedia were considered as an answer source. Some questions could have answers only in one collection, i.e. either only in the news corpus or in Wikipedia.

A Basque corpus and a set of Basque questions were offered for the first time in this edition.

The methodology we employed targeted precision at the expense of recall, therefore we always choose NIL answers for those questions

[8] HERMES project: News databases. cross-lingual information retrieval and semantic extraction (TIC-2000-0335-C03-03), founded by the Spanish Government.

[9] www.clef-campaign.org/2008/working_notes/

we could not reliably locate a candidate answer in the retrieved passage. Table 1 illustrates the results achieved by *Ihardetsi* in the monolingual run.

It is clear that the best results were achieved for factoid questions. This is due to the fact that we focused on this type of questions in the development of the system. A set of 145 factoid questions was processed and, taking into account the first three answers, we obtained the following results: 50 questions had a correct or inexact[10] answer in the proposed three answers, 22 had a wrong NIL answer[11] and 73 had a wrong answer. Analysing these 73 questions we detected that for 17 the correct passage was detected but the system did not extract the correct answer.

	R	W	I	ACC
OVERALL	26	163	11	13.0%
FACTOID	23	113	9	15.9%
DEFINITION	3	36	0	7.7%
LIST	0	14	2	0.0%

(R: Right, W: Wrong, I: Inexact, ACC: Accuracy)

Table 1: Results for the first answer obtained in the monolingual run (Ansa et al., 2008).

There are not correct answers for LIST questions because at the time of sending the runs we had not yet implemented the heuristics for answering such questions.

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

It is remarkable that no other system took part in the Basque target task, so the obtained results could not be directly compared with another Basque system. Nonetheless, it is interesting to contrast our results with some other languages. For that purpose, we choose QA@CLEF 2007 (Giampiccolo et al., 2008) results as a reference because that was the first time that topic-related questions and the Wikipedia corpus were included. Although our results are far from the best ones, with overall accuracy of 54%, we realized that almost 40% of all the runs got worse results than those of our system.

5.2 The *Anhitz* Project

The *Ihardetsi* system, included as the QA functionality in the demo prototype developed in *AnHitz,* has been evaluated in order to measure its performance and weigh the impression of potential users about it. A group of 50 users formulated 3 questions and 3 cross-lingual searches each, making 300 tests in total. During the interaction of the testers with the system, some objective observations were noted down, such as the number of failures and successes of the QA system.

As it can be observed in Table 2, *Ihardetsi* answered correctly 30.61% of the times, and in another 15.30% the correct answer was among the first five possible answers given. 54.08% of the times the system did not give a correct solution or did not answer at all. We could not evaluate whether the correct answer was in the corpus or not.

Correct answer	**%**
In the 1st place	30,61
In the 2nd place	8,16
In the 3rd place	1,02
In the 4th place	3,06
In the 5th place	3,06
The right answer was not among the possible answers	36,73
The system did not answer at all	17,35

Table 2: Results for qualitative evaluation in *Anhitz*-QA

[10] An answer is incorrect if it contains less or more information than that required by the query.

[11] A NIL answer can be correct if really the question has no answer in the corpus.

5.3 Addressing cross-linguality

Although the main aim of *Ihardetsi* is to deal with Basque questions and Basque documents, we have carried out some cross-lingual experiments, such as the Spanish-Basque and English-Basque bilingual tasks at QA@CLEF 2008, and the Basque-English task in the RespubliQA exercise at QA@CLEF 2009.

5.3.1 Bilingual tasks at QA@CLEF

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 of the question could compensate the lost of precision in the translation process. The results of the three runs are quite poor. The loss of precision respect to the monolingual system is more than 50% (Ansa et al., 2008).

The main conclusions we want to remark are:

- Very similar results are obtained for the basic Spanish–Basque and for the English–Basque 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 hoped better results for this run.
- Although the results are similar in average, 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 does not achieve better results. A slightly smaller precision is observed, because some right answers are lost. In compensation to this, new right or inexact answers appear, but not in the first place. In view of these figures, one might 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).

5.3.2 ResPubliQA at QA@CLEF 2009

ResPubliQA[12] has been presented as a new task at CLEF 2009, and it consists of retrieving a passage string (small snippet of text) containing the answer to a question in natural language. *JRC-Acquis*[13] is the reference corpus to generate the questions and search for answers. In this corpus aligned documents are available in Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish.

Although no Basque documents are available, organizers have arranged a Basque–English task, so that, given a pool of Basque questions, systems must retrieve passages from the English document collection. This is a new cross-lingual experience for us, given that Basque is the source language.

Our system analyses the Basque questions, translates and disambiguates the query terms and searches for the passages that are relevant to the query. The techniques that have been used in the translation/disambiguation of the query terms are described in (Saralegi et al., 2008), and the approach for passage retrieval is based on the ideas of (Otegi et al., 2008).

We submitted two runs, but we have not received yet the evaluations.

6 Conclusions

This article shows a general architecture for Question-Answering, where general linguistic processors are reused and integrated. The underlying idea is that QA systems can be built, even for languages with fewer resources, reusing existing linguistic tools.

We present different evaluations of the system, which have been carried out in different scenarios.

In the QA@CLEF 2008 campaign we compared the performance of our system with other monolingual and cross-lingual systems on a heterogeneous document collection (news articles and Wikipedia). Although the results might look not so good, our general conclusion is positive considering that it was our first participation and taking into account the particularities of the Basque language. Moreover, we have been able to identify some of the strengths and weaknesses of each module of the system.

In the context of the *Anhitz* project, our QA system has been integrated and tested in a prototype that has received ample media

[12] http://celct.isti.cnr.it/ResPubliQA/

[13] JRC-Acquis is the total body of European Union (EU) law applicable in the EU Member States.

coverage and has been welcomed by Basque society. The system has been evaluated by 50 users who have completed a total of 300 tests, showing good performance and acceptance. We consider that this kind of qualitative evaluations are quite interesting.

The ResPubliQA experience at the QA@CLEF 2009 campaign has showed us that we can deal with cross-lingual tasks even if the target language is not Basque.

All these experiences constitute the background of future improvements. It would be useful to apply other techniques, such as the syntactic pattern matching and the anaphora resolution. As these tools are developed for Basque, *Ihardetsi* will use them.

Acknowledgements

This research was supported in part by the Spanish Ministry of Education and Science (Know TIN2006-15049-C03-01) and the Basque Government (AnHITZ 2006IE06-185).

References

Agirre E., Aldezabal I., Etxeberria J., Iruskieta M., Izagirre E., Mendizabal K., Pociello E.. Improving the Basque WordNet by corpus annotation. Proceedings of Third International WordNet Conference. pp. 287-290. Jeju Island (Korea). 2006.

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

Alegria I., Diaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and Sarasola K.. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS. Springer. Vol. 4394/2009. pp. 374-384. 2007.

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

Arrieta K., Diaz de Ilarraza A., Hernáez I., Iturraspe U., Leturia I., Navas E., Sarasola K.. AnHitz, development and integration of language, speech and visual technologies for Basque. Second International Symposium on Universal Communication Osaka. pp.338-343. 2008.

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, pp. 23-30. 2004.

Bilotti M. Query Expansion Techniques for Question Answering. Master's thesis, Massachusetts Institute of technology, 2004.

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, pp.380–384, 1998.

Giampiccolo G., Herrera J., Peñas A., Ayache C., Forascu C., Jijkoun V., Osenova P., Rocha P., Sacaleanu B., and Sutcliffe R. Overview of the CLEF 2007 Multilingual Question Answering Track. LNCS. Springer. Volume 5152/2008. pp. 200-236. 2008.

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

Otegi A., Agirre E., Rigau G.. IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. Working Notes of the Cross-Lingual Evaluation Forum, Aarhus. 2008.

Saralegi X., San Vicente I., Gurrutxaga A.. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. 6th International Conference on Language Resources and Evaluations (LREC) - Building and using Comparable Corpora workshop. Marrakech. 2008.

Strohman H. T., Metzler D. and Croft W.B. Indri: A language model-based search engine for complex queries. Proceedings of the International Conference on Intelligence Analysis. Poster. 2005.

Tomás D., Vicedo J.L., Saiz M., and Izquierdo R.. Building an XML framework for Question Answering. Working Notes of the Cross-Lingual Evaluation Forum. Alacant. 2005.