An Overview of the Linguistic Resources used in Cross-Language Question Answering Systems in CLEF Conference

1. Introduction

The development of the Semantic Web requires great economic and human effort. Consequently, it is very useful to create mechanisms and tools that facilitate its expansion. From the standpoint of information retrieval (hereafter IR), access to the contents of the Semantic Web can be favored by the use of natural language, as it is much simpler and faster for the user to engage in his habitual form of expression. The growing popularity of Internet and the wide availability of web informative resources for general audiences are a fairly recent phenomenon, although man´s need to hurdle the language barrier and communicate with others is as old as the history of mankind. The World Wide Web, also known as WWW, together with the growing globalization of companies and organizations, and the increase of the non-English speaking audience, entails the demand for tools allowing users to secure information from a wide range of resources. Yet the underlying linguistic restrictions are often overlooked by researchers and designers. Against this background, a key characteristic to be evaluated in terms of the efficiency of IR systems is its capacity to allow users find a corpus of documents in different languages, and to facilitate the relevant information despite limited linguistic competence regarding the target language.
In the field of Information Retrieval, monolingual and cross-language tools are being created that can greatly assist specialists in their work; as well as helping other users find a wide variety of information. One of the main difficulties facing these cross-language tools is the task of translating queries made by users and the documentary sources found in response (Diekema, 2003). Given the current expansion in research, development, and the creation of cross-language IR systems, it was considered worthwhile analysing and evaluating the resources used by one type of these systems: cross-language question answering systems (hereafter QA systems).

Recent evaluation efforts try to keep their work relevant for the real world and make their results interesting for practical applications. Yet, in order to cope with these new heterogeneous requirements and to account for the changing necessities of different domains and information needs, new approaches and tasks need to be established (Mandl, 2008). A study from the perspective of translation may offer a different focus on the problem of translation and resources. Researchers currently working in the field of cross-language QA systems are searching for new methods to optimise the efficiency of IR without using too many resources for language problems. However, a system cannot easily retrieve relevant information for the user without an optimal solution for translation resources. For this reason, translation is crucial

in this environment and enables problems to be analysed from a fresh point of view. Any progress made in solving problems of cross-language communication can be added to existing information retrieval systems.

The paper is structured as follows: a first section introduces the background of the QA systems. Secondly, the methodology of our research is described, so we analyze the evaluation results obtained with the different approaches. Finally, some conclusions and orientations are presented.

2. Background of QA systems

Recent advances in IR and Web globalization mean that cross-language search systems have been developed in which translation and language resources are as important as the documentary and computer tools. This type of system has opened a new research field that examines the most effective methods for IR, as well as studying which resources are required for a correct translation.

Information overload is felt more strongly on the Web than elsewhere. All too often a query made with a web search tool (search engine, meta-search engine) results in the retrieval of too many pages – many of which are useless or irrelevant to the user. Therefore, professionals from various areas are beginning to recognize the usefulness of other types of systems, such as QA systems, for quickly and effectively finding specialist information [(Crouch et al., 2005) and (Lee et al., 2006)].

Cross-language IR or CLIR (*cross-lingual information retrieval*) involves at least two languages in this process. In a cross-language environment such as the Web, most IR systems (search engines) are limited to finding documents in the language of the query; or alternatively, include machine translation systems, which are only useful once the documents are located and do not effectively cross the language barrier.

Given a particular query, CLIR systems run on a collection of multi-lingual documents and retrieve relevant information regardless of the language used in the query (Grefenstette, 1998). Within the area of multi-lingual IR, the object of our study is multi-lingual QA systems and these systems are opening a new field of research that is becoming increasingly important within CLIR.

Traditionally, CLIR is described as having the problem of offering documents to users which they cannot read (Oard and Gonzalo, 2001). However, that is not all. One of the earlier works in CLIR was conducted by Salton (1970) and compared the effectiveness of English and German queries with that of queries obtained using a bilingual thesaurus for retrieving documents in both languages. Salton empirically showed that CLIR, using a hand-crafted bilingual thesaurus, is comparable with mono-lingual information retrieval in performance. Usually with CLIR a multi-lingual thesaurus of some sort is created to hold a list of descriptors for each document in a collection and the semantic relations between them, and each term in the thesaurus must be translated for each language involved. The descriptors can be added to the thesaurus manually or automatically (if the system can learn which terms are likely to be important from previous indexing) (López Ostenero, 2002).

These circumstances have fuelled academic interest in multi-lingual IR, or CLIR, and the techniques of natural language processing. Although Salton (1970) is considered the "father" of the earliest research initiatives concerning CLIR, the first Workshop geared specifically to CLIR topics was celebrated in Zurich and it was organized by the Association for Computing Machinery (ACM) during the Special Interest Group on Information Retrieval, SIGIR-96 Conference (Grefenstette, 1998). Nowadays, there are four important international forums about the evaluation of IR systems focusing on techniques and proceedings related to CLIR: Text REtrieval Conference (TREC)[1], the Cross-Language Evaluation Forum (CLEF)[2], the NII Text Collection for IR Systems (NTCIR)[3] and the Language Resources and Evaluation Conference (LREC)[4] (Olvera-Lobo, 2009).

In 2000 it is created CLEF, the most important European forum to the evaluation of multi-lingual and multimedia retrieval systems. CLEF is developed to promote research and development in multi-lingual information access by *a)* developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both mono-lingual and multi-lingual contexts, and *b)* creating test-suites of reusable data which can be employed by system developers for benchmarking purposes. The final objective is to boost and encourage information retrieval technologies development in Europe in order to guarantee its competitiveness in a global sphere.

CLEF is divided in different topics (tracks) which research the several aspects of the multi-lingual information retrieval (see figure 1), such as searching for a text (ad-hoc task), geographical information search (GeoCLEF), search of information on the Web (WebCLEF), image retrieval (ImageCLEF), and question answering systems (QA@CLEF), among others. Then, each track offers different tasks about diverse aspects focused specifically on that topic.

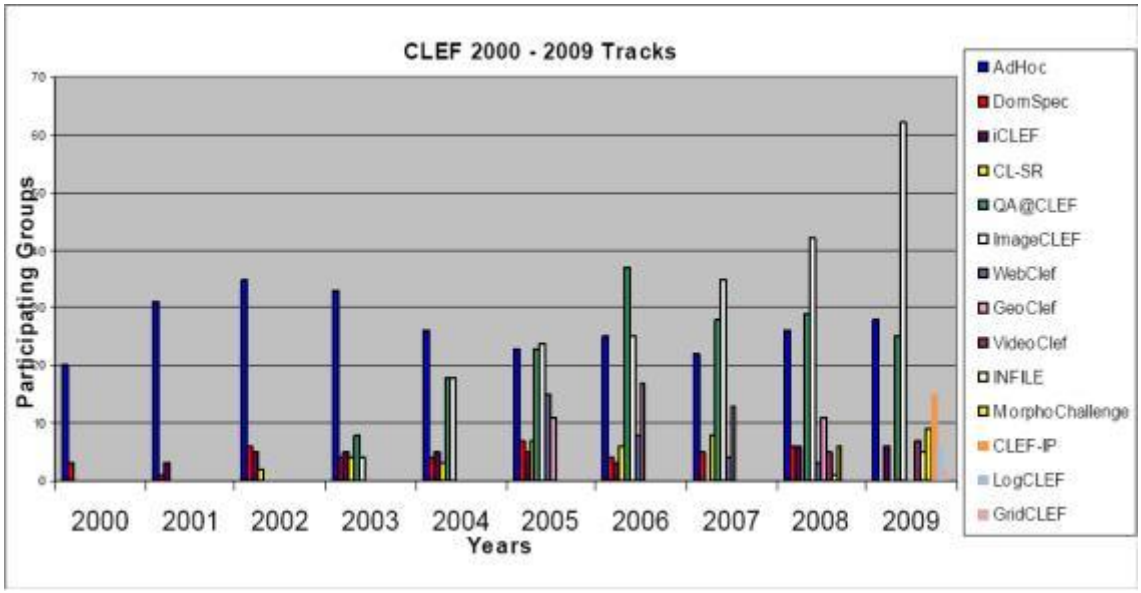

Figure 1. CLEF 2000 – 2009: Participation per Track (Peters, 2009)

[1] Available at: trec.nist.gov/ (Accessed October 27, 2010)
[2] Available at: www.clef-campaign.org/ (Accessed October 27, 2010)
[3] Available at: research.nii.ac.jp/ntcir/index-en.html (Accessed October 27, 2010)
[4] Available at: www.lrec-conf.org/ (Accessed October 27, 2010)

Within the area of cross-language IR, the object of our study is cross-language QA systems and these systems are opening a new field of research that is becoming increasingly important within CLIR. Question-answering systems are an evolutionary improvement in IR systems. As an alternative to traditional IR systems they give correct and understandable answers to factual questions – rather than just offering a list of documents related to the search (Jackson and Schilder, 2005). The benefit is that users do not have to read whole documents to find the desired information. QA systems have attracted major attention since the TREC-8 (Text REtrieval) conference on information retrieval (Vorhees, 1999). TREC conferences have been the major forum for sharing and encouraging international research in information retrieval since 1992.

When a query is entered into the interface, the system proceeds to analyze the question by separating the word or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the specialized area of the question (Olvera-Lobo and Gutierrez-Artacho, 2010). Subsequently, the system evaluates and eliminates redundant information, or information that does not respond correctly to the question, and submits one or more prepared responses to the user [(Tsur, 2003) and (Cui et al., 2004)].

All the QA systems have a very similar architecture, and as described in the literature (Ferrandez et al., 2009), this general architecture is summarized in the following modules: (i) question analysis to extract all the useful information from the question, (ii) document retrieval to obtain a set of relevant documents, (iii) passage retrieval to obtain only the relevant information from retrieved documents and (iv) answer extraction to determine which parts of the selected passages are potential answers.

While the development of QA systems represents progress, the systems nevertheless suffer restrictions. Many were only developed as prototypes, or demonstration versions, and few were marketed. Some researchers have designed and created systems that were presented and discussed at various forums and conferences. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users.

The QA systems were mostly based on the implementation of a set of rules for the question analysis extracting features to extract an answer from structured knowledge (Moreda et al., 2010). These systems usually have a simple interface where users can enter their queries, while some offer a list of recent queries to help users understand how the system works. QA systems handle these queries by applying algorithms and methods of linguistic analysis; as well as using natural language processing to identify the components and determine the expected response (Zweigenbaum, 2005). QA systems may be general domain and so answer questions from diverse fields. Alternatively, they may be domain-specific and focus on a specialized area (Frank et al., 2006). Domain-specific systems use specific linguistic resources that enable more precise answers to be given. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users.

These circumstances have fuelled academic interest in cross-language IR, or CLIR, and the techniques of natural language processing. Given a particular query, CLIR systems run on a collection of cross-language documents and retrieve relevant information regardless of the language used in the query (Grefenstette, 1998).

In cross-language QA systems, the language of the question may differ from the language of the retrieved document. However, QA systems differ from other CLIR systems because they do not retrieve whole documents and instead respond to queries with a short answer. QA systems are a set of coordinated monolingual systems in which each extracts responses from a collection of separate monolingual documents (Aceves Pérez, 2008). Normally, cross-language QA systems are similar to monolingual QA systems, the main difference being the incorporation of a translation module and/or linguistic tool for cross-language recovery.

Translation is crucial in CLIR because queries and documents do not always share the same language. The main translation problems identified are: lexical ambiguity, lack of translation coverage, multi-modal lexemes, and errors in lexical resources (Diekema, 2003). However, translation aspects have been relatively neglected during the development of these systems.

Usually QA systems that deal with multiple languages rely on a translation module. The user enters his specific query, generally including some interrogative adverb (*How? When? Where?)* in a given natural source language. This question is translated by an automatic translator. In the stage of query analysis, the QA system examines the user´s question and determines what type of information is being demanded. The classification of the questions is a key for the system, as this information will be utilized in the search stage, and in the selection and extraction of the potential responses (García Cumbreras *et al*, 2005). The resulting search expression will be, then, the *input*, or the formulation of the query to be used by the search engine of the system for comparing and matching it with the documents in the database. Once the documents that are relevant to the query are located, the system breaks them up into sections, selects the excerpts that include the candidate responses, and selects a final response. This response, along with its location in the corresponding document, is finally delivered to the user (Olvera-Lobo & García-Santiago, 2010).

Five main types of linguistic resources used in cross-language QA systems were identified following an analysis of the literature. The main resource types were databases, corpora, dictionaries, ontologies, and thesauri. There were also two types of linguistic tools used by these systems, namely, machine translation and computational grammars. Nowadays, due to the growing amount of online digital data available, some open-domain QA systems use Internet as corpora to answer questions in even wider domains, like Wikipedia and web pages.

These resources and tools, along with their various types and subtypes, do not run in the same way and use differing methods of processing information. Sometimes, a single resource was insufficient and several resources were used together to achieve better results.

Previous works (Diekema, 2003) identified four major sources of translation in CLIR – ontologies, bilingual dictionaries, machine translation, and corpora (see Figure 2). This study shows that CLIR has grown in popularity in recent years and that some resources are often used. Following an analysis of the literature and after identifying the resources and tools used by cross-language QA systems, a classification was made dividing these resources into two large groups: linguistic resources and linguistic tools.

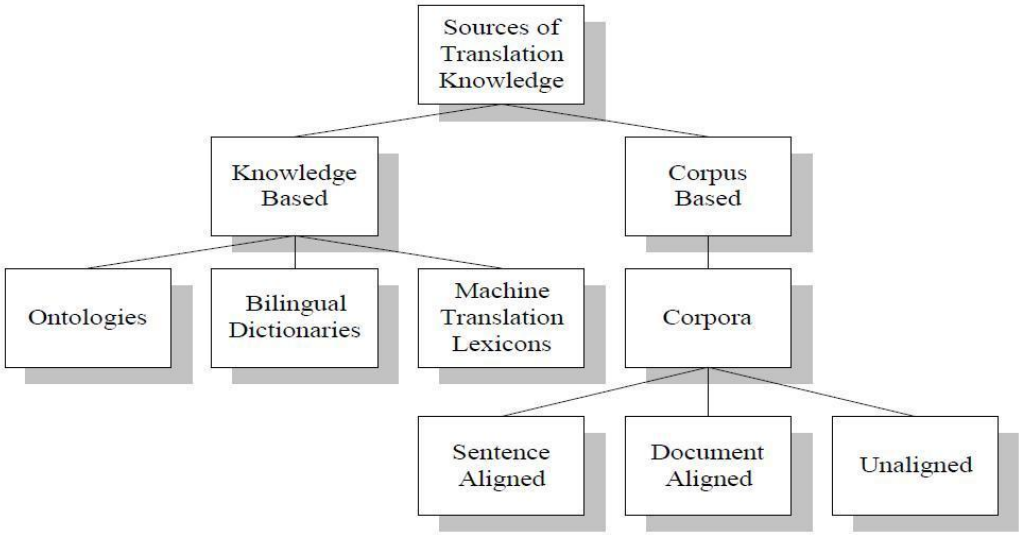

Figure 2. Resources used by cross-language QA systems for translation (Diekema, 2003)

Recent research and advances made in cross-language QA systems relate mainly to the more effective incorporation of new language resources, the creation of faster and more efficient systems, and the production of more transparent results. However, there remains an unsolved challenge: translation.

3. Method Section

An analysis methodology was adopted for this study and the collection of data about the tools and linguistic resources employed by these systems; as well as their use and implementation.

The aim was to find, analyze, and compare the different types of linguistic resources used in the QA systems presented in CLEF during the last ten years. In total, we analysed 947 papers presented in CLEF published between 2000 and 2009. No papers from 2010 were included because CLEF 2010 was celebrated last September and the working notes are not been published yet. We studied the subject discussed in each paper –including the language resources and tools used. Although all the papers discussed the linguistic aspects of IR systems, only some tackled the cross-language QA systems and their resources as the main theme. So, we have analyzed 215 papers of 947 contributions presented in CLEF during the last nine years.

For the studied period, the year with the largest number of papers published on cross-language QA systems was 2005. A growing level of interest peaked from 2004 and it has been continued the following years. Nevertheless, from 2006 interest began shifting to other types of QA systems such as image, voice, and expertise domains.

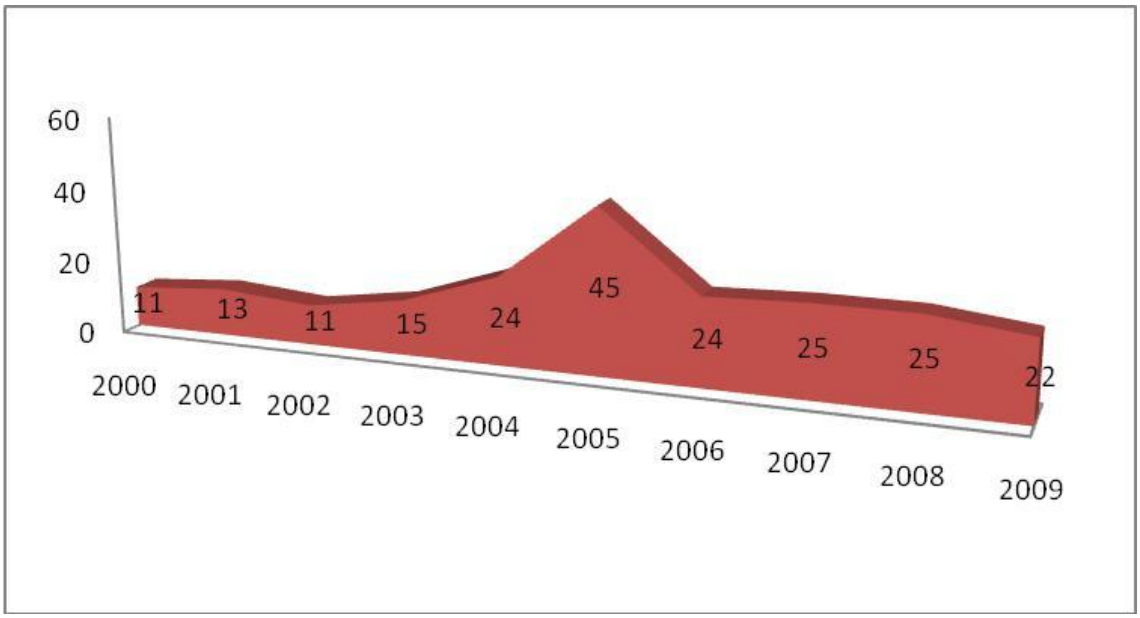

Figure 3. Papers about cross-language QA systems by year

In a second phase we explored the resources used by cross-language QA systems showed in the CLEF conferences. We did an important documentary observation phase, because of analyzing and assessing what are the different linguistic resources and tools used by these systems, so, it enabled us to monitor the progress made by these developers.

4. Results and Discussion

In analyzing the literature, it was found that the resource most used by cross-language QA systems was corpora (mostly parallel), followed by machine translation, and Wikipedia (see Figure 4). We can realise that the traditional trilogy of resources in cross-language QA systems (dictionaries, machine translation and corpora) has changed (Nguyen et al., 2009).

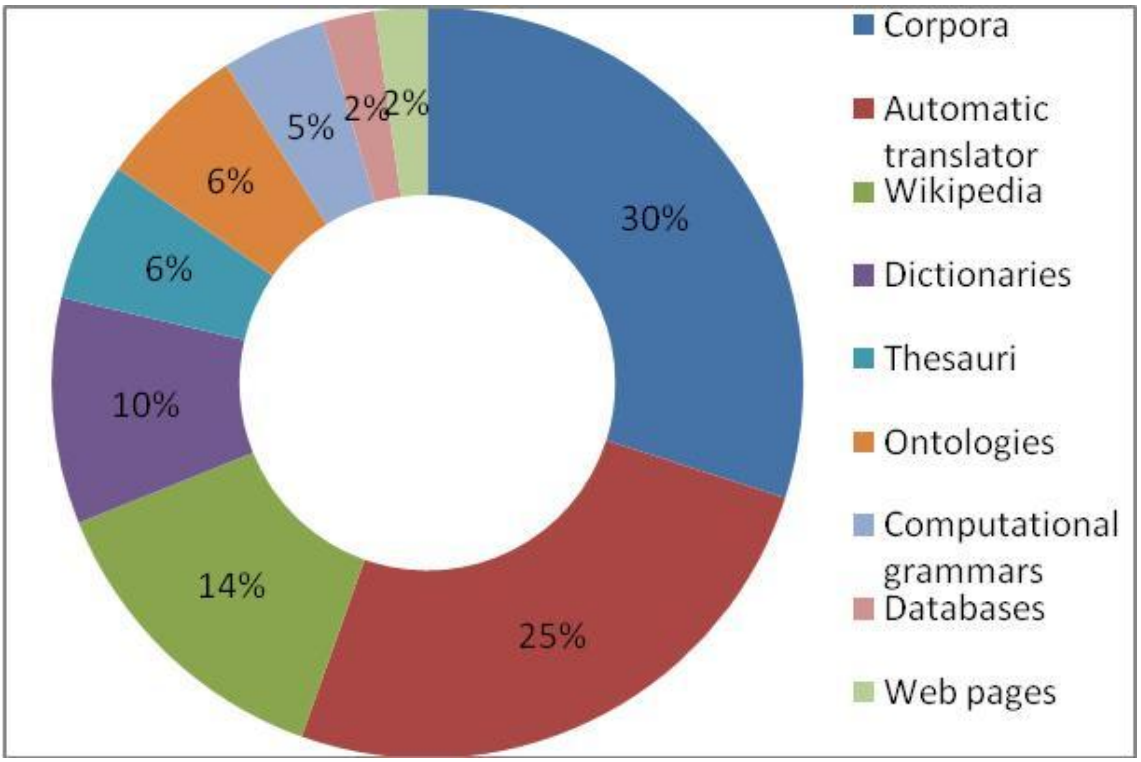

Figure 4. Linguistic resources and tools used in CLEF

The first most commonly used resource is the corpus with 80 occurrences and used in 30 per cent of the systems. The ostensible popularity of corpora is explained by the fact that many variants of corpora are included. The most surprising aspect of this resource was its nearly steady growth in recent years and the peak in 2005 – when corpora appeared in 18 of the 45 papers reviewed (Figure 5). We saw a significant decline in use in 2006, but this may be partially attributed to the fact that only 24 papers on cross-language QA systems were found for the year.

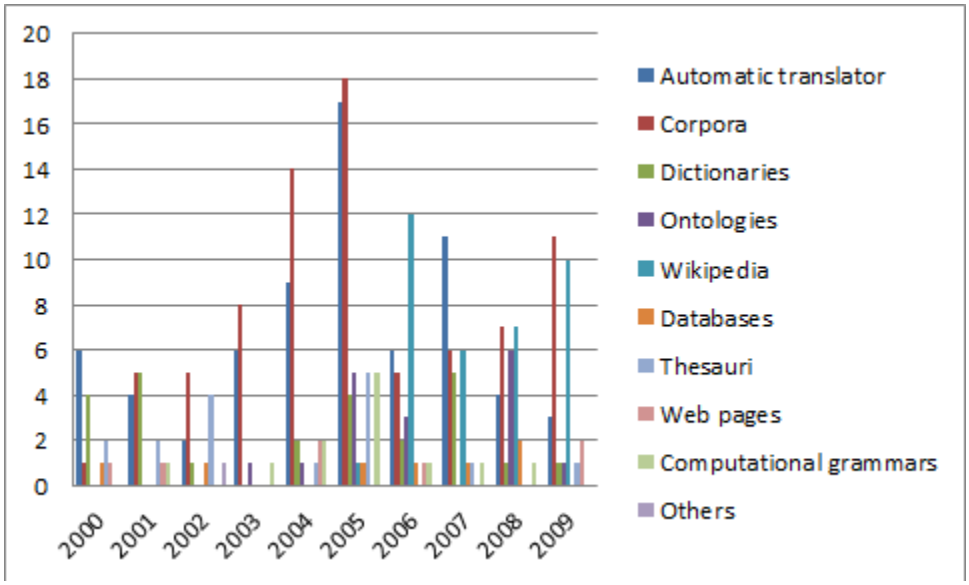

Figure 5. Use of resources and tools per year

Linguistic corpora are very useful resources for specialized domains. This is because the information received by users will be complete and correct when a translation is made or reviewed by professional translators. Existing corpora can be made available on the Web in several languages, so solving two of the main problems raised earlier: computational cost and storage.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Automatic translator	6	4	2	6	9	17	6	11	4	3
Corpora	1	5	5	8	14	18	5	6	7	11
Dictionaries	4	5	1	0	2	4	2	5	1	1
Ontologies	0	0	0	1	1	5	3	0	6	1
Wikipedia	0	0	0	0	0	1	12	6	7	10
Databases	1	0	1	0	0	1	1	1	2	0
Thesauri	2	2	4	0	1	5	0	1	0	1
Web pages	1	1	0	0	2	0	1	0	0	2
Computational grammars	0	1	0	1	2	5	1	1	1	0
Others	0	0	1	0	0	0	0	0	0	0

Table 1. Number of linguistic resources and tools used by QA systems in CLEF per year

Machine translation was used in 68 of the 215 papers reviewed. This tool is often incorporated individually or in combination with other linguistic resources to offer better coverage. The rate and quality of this translation can vary. But even the most sophisticated automatic translators cannot yet produce translations on a large scale that do not need absolutely any revision by a person. The automatic translators also have restrictions about the nature of the texts that they can translate better (Olvera-Lobo & García-Santiago, 2010; García-Santiago & Olvera-Lobo, 2010). Although most authors confirm the problems of ambiguity and the poor quality of texts, they continue to prefer this tool because it is one of the cheapest and easiest to incorporate into systems. Machine translation usually gives better results in general domain QA systems than in specific domains. This is because machine translation cannot identify and correctly translate certain specialized terms. Nor can this tool be recommended for systems that use non-Western languages, or more than two languages. In fact, machine translation is effective in these tasks.

However, the use of machine translation has declined in recent years (see Figure 5). They were used in six of eleven papers reviewed in the year 2000. However, their presence declines substantially over the next two years (2001 and 2002). The number of cross-language QA systems using automatic translation rose again after 2005, yet not individually as in earlier years, but in combination or in support of other language resources. Machine translation has continued to be used in the most recent years – but in a smaller number of systems.

Wikipedia was used on 36 occasions. This is one of the most innovative resources and is growing rapidly in popularity. It was first incorporated in 2005, and its presence grew substantially the following year. Wikipedia is a large document collection and has shown less redundancy than other resources (Roger et al., 2008). Other advantages of Wikipedia are that it uses hyperlinks to avoid information repetition, the similar data appears in several different languages and it is usually highly structured. Nevertheless, the data can be edited by everyone and sometimes it does not happen.

The following most commonly used resources are dictionaries and ontologies, with 26 and 17 appearances respectively. Dictionaries, together with machine translation and corpora, are the resources traditionally used by these systems. However, grammar and ambiguity problems have

recently reduced their popularity, so that only 9 of the 96 systems studied over the past four years used this resource. The results obtained by incorporating Wikipedia into such systems are unclear; some researchers claim it can be considered like one of the most faithful resource, while others stress that it can solve the main problems of management, storage and retrieval of information.

Very different behaviour is seen with ontologies. This resource was not used in the early years, but from the year 2003 has begun to slowly gain acceptance in cross-language QA systems. Ontologies offer many advantages and especially in specialized domain systems. Most systems are composed of texts that have been completely translated into various working languages, and so relationships are easily established. Another advantage is that there are many research teams working closely with cross-language ontologies and studying the various relationships that can be made between terms – and this existing body of work ensures a quality final product.

The use of thesauri, 16 appearances, was very limited and irregular – despite to being used very frequently at the beginning (see table 2). This is surprising given that the architecture of thesauri makes them one of the most suitable resources for these systems; although when used alone they are not used to be very useful in the good information retrieval. However, the situation can vary when specialist domain cross-language QA systems are developed because many well established thesauri exist on a wide range of topics.

	1	2	3
2000	Automatic Translator	Dictionaries	Thesauri
2001	Corpora	Automatic Translator/ Dictionaries	Thesauri
2002	Corpora	Thesauri	Automatic Translator
2003	Corpora	Automatic Translator	Ontologies
2004	Corpora	Automatic Translator	Dictionaries
2005	Corpora	Automatic Translator	Thesauri
2006	Wikipedia	Automatic Translator	Corpora
2007	Automatic Translator	Wikipedia/ Corpora	Dictionaries
2008	Corpora/ Wikipedia	Ontologies	Automatic Translator
2009	Corpora	Wikipedia	Automatic Translator

Table 2. Ranking of the most used resources per year

Computational grammars were used in 12 occasions. This resource is a tool for working with all levels of languages, and to efficient inference systems for performing reasoning. Its use was irregular, but in 2005 it was used by a significant number of QA systems (Figure 5). Finally, the linguistic resources with a slower peak were databases and web pages, with 6 appearances. The use of these resources was very irregular –being entirely absent during some years.

5. Conclusions

This study has analyzed the main publications in all the CLEF conferences– from 2000 to 2009. In total, we were analysed and assessed 215 papers of 947 contributions presented at one of the main conference about cross-language retrieval information and we extracted as much data as possible for an overview of the situation.

Five of most used resources were identified and studied: databases, dictionaries, corpora, ontologies and thesauri. The second group in our study consisted of two linguistic tools: computational grammars and machine translation. The inclusion of grammars in cross-language QA systems is relatively recent, and so the above classifications have not be-en taken into account. Finally, we studied the various other types of translation used on Internet (web pages and Wikipedia).

After reading and analysing 215 papers of our study, we have found corpora remain the most popular option. The second one is machine translation, despite the fact that the authors of the papers recognise the resulting problems of ambiguity (see Figure 5). The low computational cost and ease of storage are two of the main advantages of these two resources. In our opinion, these resources can be adequate for cross-language QA systems when combined with others. However, there have been some changes in the use and incorporation of these resources and tools. The three most popular traditional resources (machine translation, dictionaries, and corpora) are gradually leaving a widening gap for others – such as ontologies and the free encyclopaedia Wikipedia. In addition, other approaches such as computational grammars are slowly attracting more researchers who are experienced in handling the results they produce.

A comparison of the evolution and use of different resources and tools shows that trends favour the traditionally more popular tools (machine translation and corpora). However, Wikipedia shows trends that match, or nearly match, the traditional resources. The remaining tools are timidly growing in popularity and have promising futures. However, the trends for each combination of tools in cross-language QA systems were not studied exhaustively. This data suggests that we may see unexpected changes in the future and this area deserves to be studied and evaluated in future research.

6. References

Aceves Pérez, R. M.: *Búsqueda de Respuestas en Fuentes Documentales Multilingües.* Dissertation. Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico (2008).

Crouch, D., Saurí, R., Fowler., A.: AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines. *Palo Alto Research Center*. (2005) Retrieved October 13, 2010,

http://www2.parc.com/isl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf

Cui, H., Kan, M.Y., Chua, T.S., Xiao, J.: A Comparative Study on Sentence Retrieval for Definitional Question Answering. SIGIR Workshop on Information retrieval for Question Answering, Sheffield (2004)

Diekema, A. R.: *Translation Events in Cross-Language Information Retrieval: Lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations.* Dissertation. University of Syracuse (2003).

Ferrández, O., Izquierdo, R., Ferrández, S., Vicedo, J. L.: Addressing ontology-based question answering with collections of user queries. *Information Processing and Management*, 45, vol. 2, pp. 175-188 (2009).

Frank, A., Kirefer, H.U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., Shäfer, U.: Question answering from structured knowledge sources. *Journal of Applied Logic, Special Issue on Questions and Answers: Theorical and Applied Perspectives,* Vol. 5, pp. 20-48 (2006).

García Cumbreras, M. A., Ureña-López, L. A.; Martínez-Santiago, F.; Montejo Raez, A.: Búsqueda de respuestas multilingüe: clasificación de preguntas en español basadas en aprendizaje, *Procesamiento del lenguaje natural*, 34, 31-40, (2005).

Grefenstette, G.: Cross-Language Information Retrieval. *Kluwer academic publishers*, Vol. 1. (1998).

Jackson, P., Schilder. F.: Natural Language Processing: Overview. In: Brown, K. (eds.) *Encyclopedia of Language & Linguistics*, 2nd., pp. 503–518. Elsevier Press, Amsterdam (2005)

Lee, M., Cimino, J.; Zhu; H.R., Sable; C., Shanker; V., Ely, J., Yu., H.:.Beyond Information Retrieval – Medical Question Answering. *AMIA* (2006)

López-Ostenero, F.: *Un Sistema Interactivo para la Búsqueda de Información en Idiomas Desconocidos por el Usuario*. Unpublished doctoral dissertation. Universidad Nacional de Educación a Distancia, Spain. (2002*).*

Mandl, T.: Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance. *Informatica* 32, pp. 27-38 (2008).

Moreda, P, Llorens, H., Saquete, E., Palomar, M.: Combining semantic information in question answering systems. *Information Processing and Management* (2010, in press).

Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D.R.B., Hiemstra, D. de Jong, F.M.G.: WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia. In Peters et al. (eds.): *CLEF 2008, LNCS 5706*, pp. 58-65 (2009).

Oard, D., and Gonzalo: JThe CLEF 2001 Interactive Track. *Evaluation of Cross-Language Information Systems,* LNCS 2785, Springer, (2001).

Olvera-Lobo, M.D. (2009). Cross-language Information Retrieval on the Web. In Cruz-cunha, M. M. (ed.)**,** Oliveira, E. F. (ed.)**,** Tavares, A. J. V. (ed.) y Ferreira, L. G. (ed.), *Handbook Of Research On Social Dimensions Of Semantic Technologies And Web Services*, Chapter XXXIV, 704–719.

Olvera-Lobo, M. D., & García-Santiago, L. Analysis of errors in the automatic translation of questions for translingual QA systems, *Journal of Documentation*, 66 (3), 434–455, (2010)

Olvera-Lobo, M. D., Gutiérrez-Artacho, J.: Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation. *Health Information and Library Journal* (2010, in press).

Peters, C., What Happened in CLEF 2009: Introduction to the Working Notes. In F. Borri, A. Nardi and Peters, C. (eds.), *Working Notes for the CLEF 2009 Workshop*. Corfu, Greece (2009).

Roger, S., Vila, K., Ferrández, A., Pardiño, M., Gómez, J.M., Puchol-Blasco, M., Peral. J.: Aliqan, Spanish QA System at CLEF-2008. In F. Borri, A. Nardi and Peters, C. (eds.), *Working Notes for the CLEF 2008 Workshop* (2008).

Salton, G: Automatic Processing of Foreign Language Documents. *Journal of American Society for Information Sciences*, 21:187–194, (1970)

Tsur, O.: Definitional Question-Answering Using Trainable Text Classifiers. Dissertation. Institute of Logic Language and Computation (ILLC), University of Amsterdam (2003)

Voorhees, E.M.: The TREC 8 Question Answering Track Report. En Voorhees, E.M. y D.K. Harman (eds.), Proceedings of the 8th Text REtrieval Conference, vol. 500–246, pp. 107–130. NIST, Gaithersburg (1999)

Zweigenbaum, P.: Question answering in biomedicine. In De Rijke, M. y B. Webber (eds.), Proceedings Workshop on Natural Language Processing for Question Answering, EACL 2003, pp. 1-4. ACL, Budapest (2005)