Language Resources in Cross-Language Text Retrieval: a CLEF perspective

Julio Gonzalo

Departamento de Lenguajes y Sistemas Informáticos de la UNED
Ciudad Universitaria s/n, 28040 Madrid, Spain
julio@lsi.uned.es
http://sensei.lsi.uned.es/NLP

Abstract. Language resources such as machine dictionaries and lexical databases, aligned parallel corpora or even complete machine translation systems are essential in Cross-Language Text Retrieval (CLTR), although not standard tools for the Information Retrieval task in general. We outline the current use and adequacy for CLTR of such resources, focusing on the participants and experiments performed in the CLEF 2000 evaluation. Our discussion is based on a survey conducted on the CLEF participants, as well as the descriptions of their systems that can be found in the present volume. We also discuss how the usefulness of the CLEF evaluation campaign could be enhanced by including additional tasks which would make it possible to distinguish between the effect on the results of the resources used by the participating systems, on the one hand, and the retrieval strategies employed, on the other.

1 Introduction

Broadly speaking, traditional Information Retrieval (IR) has paid little attention to the linguistic nature of texts, keeping the task closer to a *string processing* approach rather than a *Natural Language Processing* (NLP) one. Tokenization, removal of non-content words and crude stemming are the most "language-oriented" IR tasks. So far, more sophisticated approaches to indexing and retrieval (e.g. phrase indexing, semantic expansion of queries, etc.) have generally failed to produce the improvements that would compensate for their higher computational cost. As a consequence, the role of language resources in standard text retrieval systems has remained marginal.

The Cross-Language Information Retrieval (CLIR) challenge - in which queries and documents are stated in different languages- is changing this landscape: the indexing spaces of queries and documents are different, and the relationships between them cannot be captured without reference to cross-linguality. This means that Language Engineering becomes an essential part of the retrieval process. As the present volume attests, research activities in CLIR include the development, adaptation and merging of translation resources; the study of methods to restrict candidate terms in query translation; the use of Machine Translation (MT) systems, in isolation or (more commonly) in combination with other strategies, etc.

In this paper, we will study the use of Language Resources by groups participating in CLEF 2000, assuming that this provides a representative snapshot of the research being conducted in CLIR as a whole. We will use "language resources" in its broadest sense to include not only dictionaries and corpora but also Natural Language Processing tools (stemmers, morphological analyzers and compound splitters, MT systems, etc.).

The next section summarizes the language resources, and their current capabilities and shortcomings, used in the first CLEF campaign. In Section 3 we propose possible ways to complement the current CLEF evaluation activity to take into account the balance between the quality of language resources, on one hand, and cross-language retrieval techniques, on the other. The final section briefly extracts some conclusions.

2 Language Resources in CLEF 2000

We have collected information about the language resources and tools used in the first CLEF campaign, using two sources of information: a survey conducted on the CLEF participants, and the papers contained in the present volume.

The survey was sent to all participants in CLEF, and we received 14 responses. The teams were asked to list the resources used (or tested) in their CLTR system, specifying the provider, the availability and the approximate size/coverage of the resource. They were also asked a) whether the resources were adapted/enriched for the experiment, and how; b) what were the strengths and limitations of the resources employed; and c) their opinion about key issues for future CLTR resources. Finally, we scanned the descriptions of systems contained in the present volume to complete the information obtained in the responses to the survey.

We have organized language resources into three groups: dictionaries (from bilingual word pair lists to lexical knowledge bases), aligned corpora (from the Hansard corpus to data mined from the web) and NLP software (mainly MT systems, stemmers and morphological analyzers). Before discussing in more depth each of these three categories, some general observations can be made:

- More than 40% of the resources listed have been developed by the participants in the CLIR evaluation. This is a strong indication that CLEF is not just evaluating CLIR strategies built on top of standard resources, but also evaluating resources themselves.
- Only 5 out of 34 resources are used by more than one group: a free dictionary (*Freedict*[5]), a web-mined corpus (*WAC*[21]), an online MT service (*Babelfish*[1]), a set of stemmers (*Muscat*[8]) and an automatic morphology system (*Automorphology*[14]). This is partially explained by the fact that many participants use their own resources, and there are only two cases of effective resource sharing: the web-mined corpus developed by U Montreal/RALI (three users including the developers) and the Automorphology system developed by the U. of Chicago (used also by the U. Maryland group [22]).

Languages	developer/provider	size	teams
EN,GE,FR,IT	IAI	EN 40K, GE 42K FR 33K, IT 28K	IAI
EN-GE,FR,IT	IAI	EN/FR 39K, EN/GE 46K, EN/IT 28K	IAI
NL-EN	Canadian web company	?	Syracuse U
NL-EN,GE,FR,IT	www.travlang.com/Ergane	NL 56K, EN 16K, FR 10K GE 14K, IT 4K	CWI, U Montreal/RALI
EN-GE	www.quickdic.de	EN 99K, GE 130K	U. Maryland
EN-FR	www.freedict.com	EN 20K, FR 35K,	U. Maryland, U. Glasgow
EN-IT	www.freedict.com	EN 13K, IT 17K	U. Maryland, U. Glasgow
EN-GE	www.freedict.com	88K	IRIT
EN-GE	www.leo.online	224K	U Dortmund
FI,SW,GE→EN	?	100K	U Tampere
GE-EN	?	?	Eurospider
EN-FR	Termium	1M per lang.	U Montreal/RALI
GE-FR,IT	Eurospider sim. thesauri	?	Eurospider
GE-EN-SP-NL IT-FR-CZ-ET	EuroWordNet/ELRA	EN 168K, IT 48K, GE 20K, FR 32K	U Sheffield
EN/GE/NL	CELEX/LDC	51K lemmas	U Sheffield
NL-GE,FR, EN,SP	VLIS/Van Dale	100K lemmas	TNO/Twente

Table 1. Dictionaries and lexical databases

- The coverage and quality of the resources are very different. In general, the participating teams found that good resources (in coverage, consistency, markup reliability, translation precision, richness of contextual information) are expensive, and free resources are of poor quality. With a few (remarkable) exceptions, better resources seem to lead to better results.
- Of all the "key issues for the future", the one quoted most often by CLEF participants was simply "availability" and sharing of lexical resources. This is partially explained by the points mentioned above:
  - many resources used in CLEF are developed by the participants themselves, and it is not clear whether they are accessible to other researchers or not, except for a few cases.
  - a general claim is that good resources (especially dictionaries) are expensive, and freely available dictionaries are poor.
  - the diversity and minimal overlapping of the resources used by CLEF participants indicate lack of awareness of which resources are available and what is their cost/benefit for CLIR tasks. Hopefully, the CLEF activities should provide an excellent forum to overcome many of these difficulties.
- Two trends seem to be consolidating:
  - The lack of parallel corpora is being overcome, in corpus-based approaches, either by mining the web (U Montreal/RALI [18]) or by using comparable corpora (Eurospider [12]).
  - The distinction between corpus-based and dictionary-based approaches is becoming less useful to classify CLIR systems, as they tend to merge whatever resources are available. U Montreal/RALI, Eurospider, TNO/Twente [18], IRIT [11] systems are examples of this tendency.

2.1 Dictionaries

It is easy to imagine the features of an ideal dictionary for CLIR: wide coverage and high quality, extensive information to translate phrasal terms, translation probabilities, domain labels, rich examples of usage to permit contextual disambiguation, domain-specific extensions with coverage of named entities, semantically-related terms, clean markup ... In general, such properties are listed by CLEF participants as features that are lacking in present resources and desirable features for future CLIR resources.

In practice, 14 different lexical resources were used by the 18 groups participating in CLEF this year (see Table 1). They are easier to obtain and use than aligned corpora and thus their use is more generalized. The distinctive feature of the dictionaries used in CLEF is their variety:

- Under the term "dictionary" we find a whole range of lexical resources, from simple lists of bilingual word pairs to multilingual semantic databases such as EuroWordNet.

- In most cases, however, the lexical knowledge effectively used by the CLEF systems is quite simple. Definitions, domain labels, examples of usage, semantically related terms, are examples of lexical information that are hardly used by CLEF participants. Information on translation probabilities, on the other hand, is something that the dictionaries did not provide and would have been used by many teams, according to the survey.
- The size of the dictionaries used also covers a wide spectrum: from the 4000 terms in the Italian part of the Ergane dictionary [3] to the 1 million terms per language in the Termium database [9] used by the U Montreal/RALI group. Sizes that differ by more than two orders of magnitude!
- Some of them (four at least) are freely available in the web; two are obtainable via ELRA [4] (European Language Resources Association) or LDC (Linguistic Data Consortium) [7]; one is distributed by a publishing company (Van Dale) and at least three have a restricted distribution.
- Only one dictionary is used by more than one group (*Freedict* in its English-French and English-Italian versions). As has already been pointed out, this is a strong indication that sharing resources/knowledge about resources is not yet a standard practice in the CLIR community.
- As could be expected, the more expensive the resource, the higher its quality and coverage and the better the results, in the opinion of the participants. Freely available dictionaries tend to be the most simple and noisy, and have lower coverage.

Table 1 does not include the GIRT thesaurus, which was provided to all participants in the specific-domain retrieval task. UC Berkeley [13], for instance, used this social sciences bilingual thesaurus to produce a domain specific translation list; the list was used, together with a generic bilingual dictionary for uncovered words, to produce better results than an MT approach. This is an interesting result that shows that, although thesauri are not considered as lexical resources per se, they can be successfully adapted for translation purposes. The similarity thesaurus included in Table 1 was derived automatically from comparable corpora (see below).

2.2 Aligned Corpora

Only 5 aligned corpora were used by CLEF participants, mainly by the JHU/APL group (see Table 2). Most of them are domain-specific (e.g. the Hansard corpus [6] or the United Nations corpus[16]) and not particularly well suited to the CLEF data. Obviously the lack of aligned corpora is a major problem for corpus-based approaches. However, the possibility of mining parallel web pages seems a promising research direction, and the corpora and the mining software developed by U Montreal/RALI and made freely available to CLEF participants have been used by more groups than any other resource (U Montreal/RALI, JHU/APL [19], IRIT, TNO/Twente).

Besides parallel corpora, a German/Italian/French comparable corpus consisting on Swiss national news wire, provided by SDA (Schweizerische Depeschenagen-

Resource	Languages	developer/provider	size	teams
WAC (web corpus)	FR,EN, IT,GE	U Montreal/RALI	100MB per lang.	U Montreal/RALI, JHU/APL, IRIT
web corpus	EN/NL	TNO/Twente	3K pages	TNO/Twente
Hansard	EN-FR	LDC	3M sentence pairs	JHU/APL
UN	EN-SP-FR	LDC	50K EN-SP-FR docs	JHU/APL
JOC	EN-FR-SP-IT-GE	ELRA	10K sentences	JHU/APL

Table 2. Aligned Corpora

tur) was used to produce a multilingual similarity thesaurus [12]. The performance of this thesaurus and the availability of comparable corpora (much easier to obtain, in theory, than parallel corpora) makes such techniques worth pursuing.

Overall, it becomes clear that corpus-based approaches offer two advantages over dictionaries: a) they make it possible to obtain translation probabilities and contextual information, which are rarely present in dictionaries, and b) they would provide translations adapted to the searching domain, if adequate corpora were available. The practical situation, however, is that aligned translation equivalent corpora are not widely available, and are very costly to produce. Mining the web to construct bilingual corpora and using comparable corpora appear to be promising ways to overcome such difficulties, according to CLEF results.

2.3 NLP Software

Stemmers, morphological analyzers and MT systems have been widely used by the participants. The list of tools can be seen in Table 3. Some results are worth pointing out:

- The best groups in the German monolingual retrieval task all did some kind of compound analysis, confirming that morphological information (beyond crude stemming) may be crucial for languages with a rich morphology. Variants of the Porter stemmer for languages other than English are, according to CLEF participants, much less reliable than the original English stemmer.
- The best monolingual results for the other languages in the monolingual task, Italian and French, are obtained by two groups that concentrated on monolingual retrieval (IRST [10] and West Group [20]) and applied extensive lexical knowledge: lexical analysis and part-of-speech tagging in the case of IRST, and lexicon-based stemming in the case of West Group.

Resource	Languages	developer/provider	teams
babelfish.altavista.com	EN,FR,GE,IT,SP	Altavista/Systran	U Dortmund, U Salamanca, UC Berkeley
Systran MT system	EN-FR,GE,IT	Systran	JHU/APL
L&H Power Translator Pro 7.0	EN-FR,GE,IT	Lernout & Hauspie	UC Berkeley
stemmers	EN,GE,FR IT,NL	open.muscat.com	CWI, West Group
stemmers (from assoc. dic.)	IT,FR,GE	UC Berkeley	UC Berkeley
ZPRISE stemmers	FR,GE	NIST	U. Glasgow
stat. stemmer	FR,GE, IT,EN	U. Chicago, U. Maryland	U. Maryland
Spider stemmers	FR,IT,GE	Eurospider	Eurospider
Automorphology	EN,GE, IT,FR	U. Chicago	U.Chicago, U. Maryland
morph. analyser	FIN,GE, SWE,EN	LINGSOFT	U Tampere
compound splitter	NL	Twente	CWI/Twente
MPRO morph. anal.	GE	IAI	IAI
stemmers based on morph. anal.	FR,GE	?	West Group
morph. analyser/ POS tagger	IT	ITC-IRST	ITC-IRST
grammars	EN,IT, GE,FR	IAI	IAI

Table 3. NLP software

– Automatic stemming learned from corpora and association dictionaries appears as a promising alternative to stemmers a la Porter. Three groups (Chicago, UC Berkeley and Maryland) tested such techniques in CLEF 2000.
– MT systems are the only language resources that are not mainly developed by the same groups that participate in the CLEF evaluation. All the MT systems used are commercial systems: the free, online version of Systran software (babelfish), a Systran MT package and a Lernout & Hauspie version of the Power Translator.

3 Language Resources in CLIR evaluation

Competing systems in CLEF and TREC multilingual tracks usually make two kinds of contributions: the creation/adaptation/combination of language resources, on one hand, and the development of retrieval strategies making use of such resources, on the other hand. A problem of CLEF tasks is that they are designed to measure overall system performance. While the results indicate promising research directions, it is harder to discern which language resources worked better (because they were tested with different retrieval strategies) and it is also unclear what were the best retrieval strategies (as they were tested using different language resources). Of course, the main evaluation task should always be an overall task, because a good resource together with a good retrieval strategy will not guarantee a good overall system (for instance, the resource may not be compatible with the kind of information demanded by the retrieval algorithm). But CLEF could perhaps benefit from additional tracks measuring resources and retrieval strategies in isolation. In the rest of this section, we list some possibilities:

3.1 Task with a Fixed Monolingual IR System

A frequent approach to CLIR by CLEF participants is to translate the queries and/or documents and then perform a monolingual search with an IR system of their own. A wide range of IR systems are used in CLEF, from vector model systems to n-gram language models and database systems. This produces a different monolingual retrieval baseline for each individual group, making it hard to compare the cross-language components of each system.

A possible complementary task would be to ask participants to generate queries and/or document translations, and then feed a standard system (e.g. the Zprise system provided on demand by NIST to participants) with monolingual runs. A substantial number of participants would probably be able to provide such translations, and the results would shed some additional light on CLEF results with respect to the translation components used.

3.2 Task with Fixed Resources

A track in which all participants use the same set of language resources, provided by the CLEF organization, would make it possible to compare retrieval algorithms that participate in the main tracks with different resources. Ideally, CLEF

could cooperate with the European Language Resources Association (ELRA) to provide a standard set of resources covering (at least) the languages included in the multilingual track. We see some obvious benefits:

- Such standard resources would enormously facilitate the participation in the multilingual track for groups that need to scale up from systems working on a specific pair of languages.
- A track of this type would highlight promising retrieval strategies that are ranked low simply because they are tested with poor resources.

What kind of resources should be made available? There is no obvious answer, in our opinion, to this question. Fixing a particular type of language resource will restrict the potential number of participating systems, while providing all kinds of resources will again make the comparison of results problematic.

From the experience of CLEF 2000, it seems reasonable to start with a multilingual dictionary covering all languages in the multilingual track, or a set of bilingual dictionaries/translation lists covering a similar functionality. In its catalogue, ELRA offers at least two resources that would fit the requirements for the CLEF 2001 multilingual track (which will include English, Spanish, German, Italian and French): One is a basic multilingual lexicon with 30000 entries per language, covering the five languages in the multilingual track [2]. This dictionary has already been evaluated for CLIR purposes in [17]. The other one is the EuroWordNet lexical database, which offers interconnected wordnets for 8 European languages in a size range (for the five languages in the multilingual task) between 20000 word meanings for German and 168000 for English [23]. EuroWordNet has already been used in CLEF 2000 by the Sheffield group [15].

3.3 Task with a Large Set of Queries

In a real world application, the coverage of query terms by the language resources is essential for the response of a system. Coverage, however, is poorly measured in CLEF for a majority of systems that do query translation: the whole set of queries (summing up title, description and narrative) contain only a couple of thousand term occurrences (including stop words), and the results are quite sensitive to the ability to provide translations for a few critical terms. In addition, many relevant problems in cross-language retrieval systems are under represented in current queries.

As an example, let us consider a system that makes a special effort to provide adequate translations for proper nouns. This tends to be a critical issue in the newspapers domain, where a high percentage of queries include, or even consist of, this type of terms. Figure 1 gives a snapshot of queries to the EFE newswire database that reflects the importance of proper nouns [1]. However, the set of 40 queries in CLEF 2000 only contains three names of people ("Pierre Bérégovoy", "Luc Jouret" and "Joseph di Mambro") with a total of five occurrences, less than 0.1 occurrences per query.

[1] EFE is the provider of Spanish data for the CLEF 2001 campaign

...
Jul 26 08:33:49 2000; (joaquin garrigues walker)
Jul 26 08:34:34 2000; (descenso and moritz)
Jul 26 08:34:52 2000; (convencion republicana)
Jul 26 08:38:32 2000; (baloncesto real-madrid)
Jul 26 08:38:37 2000; (caricom)
Jul 26 08:38:41 2000; SHA REZA PAHLEVI
Jul 26 08:38:43 2000; SHA REZA PAHLEVI
Jul 26 08:38:45 2000; SHA REZA PAHLEVI
Jul 26 08:38:54 2000; (noticias internacional)
Jul 26 08:40:18 2000; (CONCORDE)
Jul 26 08:40:34 2000; (DOC) AND (CONCORDE)
Jul 26 08:42:31 2000; (MANUEL FERNANDEZ ALVAREZ)
...

Fig. 1. A 9 minute snapshot of EFE news archive search service

Another example is a system that tries to find variants and translations for named entities in general. In the CLEF 2000 queries, there are approximately 31 terms (excluding geographical names) that can be associated with named entities, such as "Electroweak Theory" or "Deutsche Bundesbahn". This represents only around 0.1% of the total number of terms.

A final example can be the ability of the resources to translate certain acronyms, such as "GATT". There are 5 acronyms in the collection (excluding country names); its coverage may affect the final results, but this variation will not be representative as to how well the resources used cover acronym translation.

It is impractical to think of a substantially larger set of queries for CLEF that is representative of every possible query style or cross-language issue. However, a practical, compromise would be to use a multilingual aligned corpora (such as the UN corpus) with documents containing a summary or a descriptive title. The titles or the summaries could be used as queries to retrieve the corresponding document in a known-item retrieval task. Obviously, such a task is no closer to real world IR than CLEF or TREC ad-hoc queries, but it would produce useful complementary information on the performance consistency of systems on a large query vocabulary, and would probably leave room to test particular issues such as proper noun translation or recognition of named entities.

4 Conclusions

The systems participating in CLEF 2000 provide a representative snapshot on language resources for CLIR tasks. From the reported use of such resources in CLEF, together with the results of a survey conducted on the participant groups, some interesting conclusions can be drawn:

– There is a wide variety (in type, coverage and quality) of resources used in CLIR systems, but little reuse or resource sharing. CLEF campaigns could provide a key role in improving availability, dissemination and sharing of resources.
– Corpus-based approaches, which were less popular due to the lack of parallel corpora, are successfully employing web-mined parallel corpora and comparable corpora.
– The distinction between corpus-based and dictionary-based approaches is becoming less useful to classify CLIR systems, as they tend to merge whatever resources are available.
– Richer lexical analysis seems to lead to better monolingual results in languages other than English, although the difference is only significant for German, where decompounding is essential.
– System builders devote a significant part of their efforts to resource building. Indirectly, CLEF campaigns are also evaluating such resources. We have proposed three complementary tasks that would reflect the systems/resources duality in CLIR better than a single, overall retrieval task: a) a task with a fixed monolingual IR system, fed with query translations provided by participants; b) a task with fixed resources provided by CLEF; c) a task with a large set of queries to provide a significant number of cases for relevant CLIR problems (e.g. proper nouns or vocabulary coverage).

Acknowledgements

I am indebted to Carol Peters and Felisa Verdejo for their valuable comments on earlier versions of this paper, and to Manuel Fuentes for providing the EFE query log.

References

[1] Babel Fish Corporation. http://www.babelfish.com.
[2] Basic multilingual lexicon (MEMODATA). http://www.icp.grenet.fr/ELRA/cata /text_det.html#basmullex.
[3] Ergane multi-lingual dictionary. www.travlang.com/Ergane.
[4] European Language Resources Association. http://www.icp.grenet.fr/ELRA/home.html.
[5] Freedict dictionaries. http://www.freedict.com.
[6] Hansard French/English. http://www.ldc.upenn.edu/Catalog/LDC95T20.html.
[7] Linguistic Data Consortium (LDC). http://www.ldc.upenn.edu/.
[8] Omsee - the open source search engine (formerly Muscat). http://www.omsee.com.
[9] Termium. http://www.termiumplus.bureaudelatraduction.gc.ca.
[10] Nicola Bertoldi and Marcello Federico. Italian text retrieval for CLEF 2000 at ITC-IRST. In *this volume*. 2001.
[11] Mohand Boughanem and Nawel Nassr. Mercure at CLEF 1. In *this volume*. 2001.
[12] Martin Braschler and Peter Schäuble. Experiments with the Eurospider retrieval system for clef 2000. In *this volume*. 2001.

[13] Fredric C. Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the CLEF collections - comparing multiple methods of retrieval. In *this volume*. 2001.
[14] John Goldsmith, Derrick Higgins, and Svetlana Soglasnova. Automatic language-specific stemming in information retrieval. In *this volume*. 2001.
[15] Tim Gollins and Mark Sanderson. Sheffield University: CLEF 2000. In *this volume*. 2001.
[16] David Graff. Overview of the UN parallel text corpus. http://www.ldc.upenn.edu/readme_files/un.readme.html, 1994.
[17] G. Grefenstette. The problem of cross-language information retrieval. In *Cross-Language Information Retrieval*. Kluwer AP, 1998.
[18] Djoerd Hiemstra, Wessel Kraaij, Renée Pohlmann, and Thijs Westerveld. Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In *this volume*. 2001.
[19] Paul McNamee, James Mayfield, and Christine Piatko. A language-independent approach to european text retrieval. In *this volume*. 2001.
[20] Isabelle Moulinier, J. Andrew McCulloh, and Elizabeth Lund. West Group at CLEF 2000: Non-english monolingual retrieval. In *this volume*. 2001.
[21] Jian-Yun Nie, Michel Simard, and George Foster. Multilingual information retrieval based on parallel texts from the web. In *this volume*. 2001.
[22] Douglas W. Oard, Gina-Anne Levow, and Clara Cabezas. CLEF experiments at the University of Maryland: statistical stemming and back-off translation strategies. In *this volume*. 2001.
[23] P. Vossen. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, 1998.