Widad MUSTAFA el HADI  
CERSATES/IDIST, Université Charles De Gaulle Lille 3, FRANCE

Dynamics of the Linguistic Paradigm in Information Retrieval

Abstract: In this paper we briefly sketch the dynamics of the linguistic paradigm in Information Retrieval (IR) and its adaptation to the Internet. The emergence of Natural Language Processing (NLP) techniques has been a major factor leading to this adaptation. These techniques and tools try to adapt to the current needs, i.e. retrieving information from documents written and indexed in a foreign language by using a native language query to express the information need. This process, known as cross-language IR (CLIR), is a field at the cross roads of both Machine Translation and IR. This field represents a real challenge to the IR community and will require a solid cooperation with the NLP community.

1. Introduction

In this proposal, we wish to review the dynamics of the linguistic paradigm in information retrieval and its adaptation to the Internet era. The emergence of Natural Language Processing (NLP) techniques has been a major factor leading to this adaptation. NLP applications range from knowledge representation, knowledge acquisition for lexicon construction, automatic term extraction, machine translation, automatic indexing, discourse analysis for automatic text understanding and abstracting, monolingual Information Retrieval (IR) to cross-language IR. The www being a huge repository of multilingual information freely accessible through the various search engines has certainly increased interest in cross-language IR. The use of Machine Translation (MT), such as the “Translate” facility available on the Altavista web service is now a necessity for users wishing to retrieve documents in various languages. But the quality of the translations provided is not satisfactory and a lot of work remains in front of us. The tools and linguistic resources to be developed in order to improve cross-language IR range from part-of-speech taggers, syntactic analysers, monolingual dictionaries, bilingual dictionaries (for language pairs) to lexical databases, terminological databases, monolingual corpora and aligned corpora (for language pairs).

Information retrieved from Web sites by indexing languages can in turn be used to build linguistic resources i.e., lexical/terminological databases, monolingual/bilingual or multilingual corpora on a specific field of knowledge, etc. Some term extraction tools retrieve simple or complex terms from a corpus in a given field of knowledge. The output, i.e., a list of terms, is used to retrieve relevant corpora from other information repositories dealing with the same field. By so doing huge corpora can be built and can further be used in NLP applications and evaluation.

This presentation will be divided in two main parts: first we will focus on the role of terminology and on its major contributions, particularly on the growth of NLP-based term extraction techniques. Then we will trace the development of language engineering techniques and tools in the Internet era and look at the way they adapted to the current needs. We will focus exclusively on the linguistic techniques used in cross-lingual IR and we will review some linguistic techniques and resources involved in this application. We have a full range of techniques starting with full database translation, moving to query translation which leave the documents in their original language; or using parallel corpora which find translation equivalents; possibly accessing an online multilingual dictionary offering translations for
query terms; trying Lexical Conceptual Structures (LCS) or finally reformulation.

2. Terminology: An Application for Natural Language Technology

NLP witnessed a growing interest in automatic handling of terms, linguistic units characterizing specialized domains. Automatic term recognition is much needed because a simple but coherently built terminology is the starting point of many applications such as human or machine translation, indexing, thesaurus construction, knowledge organization, text summarization, etc.

2.1. Approaches to Automatic Term Recognition

As a result of the growing interest in this type of approaches to computational linguistics, a number of products aiming at term recognition and extraction have been developed. Early applications of automatic term extraction and recognition were developed in the context IR. While this research can be traced back to the late 1950s, automatic term extraction aiming at building reference tools for translation, technical writing or knowledge-based systems and CLIR is a recent development.

Although many approaches use both linguistic and statistic techniques the emphasis is usually on either one or the other. As for linguistic-based approaches the primary concern of most of the research done in this field is focus on complex terms. System designers seem to agree on one point, i.e. that terms are linguistically materialized as noun phrases: Lauriston 1994; Dagan et al., 1994; Béguin et al., 1997; Mustafa el Hadi et al.; 1996, Mustafa el Hadi, 1998, among others. Most of the extracting tools identify terms by using frequency distribution and categorial pattern matching (Dagan et al., 1994). All lexical units contained in a given text are analysed and matched to patterns (typical forms of terminological units) previously described by rules. The linguistic approaches are heavily language-dependant, which means that the extension to another language involves a thorough redesigning of programs. L'Homme, 1996; Kageura et al., 1996; Dagnan et al., 1994 describe some of the systems which are tested in the framework the AUF evaluation project.

As for statistically-oriented approaches, studies for recognizing complex terms follow a simple procedure. System designers, for instance, choose to extract and weigh complex terms by simply taking two adjacent words (or two words with one intervening elements) and gives them weights on the basis of their constituent parts. An overview of the statistical techniques for automatic term recognition can be found in Kageura et al., (1996); Béguin et al., (1997).

2.2. Types of Applications which can be Associated with IR

Term extraction deals with two main activities: term acquisition i.e. covering of the domain which includes discovering new terms (cf. Jacquemin, 1997) and indexing. The purpose of term acquisition is the construction of terminological resources, such as controlled vocabularies. Thesauri are used for assisting information access through query expansion (Salton and Lesk, 1971; Voorhees 1993), for cross-lingual information retrieval (Eichmann et al., 1998), or for thesaurus-based querying. In this domain, automatic thesaurus construction is a well established tradition and studies have focused either on the discovery of terms or on the discovery of semantic relations (cf. Jacquemin, 1997 for a detailed review).

Unlike term acquisition, the purpose of indexing is not to built terminological data, but to link texts with terms that are expected to summarize the content of the documents for subsequent querying. Terms are known to represent highly informative parts of technical and scientific documents. Indexing is therefore a type of informationaly sound technique in the event of a text simplification which is expected to store the main contents of a document while replacing the whole text by term occurrences.

Automatic term extraction and recognition though different, because of its various
applications is technically closely related to the field of information retrieval where we generally speak of "automatic indexing" or "automatic key-word extraction". Some researchers on both sides (terminology extraction and automatic indexing) have begun to recognize the relations between the two fields from a variety of perspectives (cf. Kageura et al., 1996).

Bilingual term extraction tools (cf. TERMIGHT, Dagan et al., 1994) based on part-of-speech tagging and word alignment technologies to extract candidate terms and their translations can be used for both tasks. The monolingual part of this type of systems can be used for indexing and for term extraction while the bilingual component can serve for both translation and information retrieval in multilingual text collections (see also infra). Although this type of system had initially been created for term extraction and translation it was later reused and adapted for indexing as well, as reported by (Landauer et al., 1990, quoted by Dagan et al., 1994).

Term extraction designers should keep in mind this new requirement (i.e. translation and information retrieval in multilingual text collections) when designing such systems because its need is going to be ever more evident.

3. Cross-lingual Information Retrieval (CLIR)

3.1. Background

The availability of electronic data in many languages thanks to Internet has created increased interest in searching across languages. NLP techniques and tools try to adapt to the current needs, i.e. retrieving information from documents written and indexed in a foreign language by using a native language query to express the information need. This process, known as cross-language IR (CLIR), involves elements of MT & IR: finding possible translations of the words and terms appearing in the original query (MT); ways of controlling words in order to match stored indexes and approaches to weighing terms in a query (IR), (cf. Grefenstette, 1998 : 523-524).

The ideal strategy is to enable the user to express his query in natural language and in his or mother tongue so as to grant him access to any textual information relevant to his query, independently of the language in which the needed information is presented. Most of the users can read documents written in a foreign language but have difficulties in finding the appropriate words to express their query in an efficient way. The choice of an inaccurate term or of inaccurate expressions can drastically affect a search process and lead to a poor performance. For these reasons it is preferable to give the user the opportunity to express his query in the language he knows best.

3.2. A Review of Some Techniques Used in CLIR

This review is by far not exhaustive. We are just trying to give a brief account of the most popular methods used in CLIR.

3.2.1) Approaches Using Machine Translation (MT)

This approach is based on the use of MT system to translate either the database or the queries:

a) Even at a very poor quality level MT, Altavista Web translation service for instance, is becoming popular as it enables users to read documents in different languages. Many users are able to understand documents in another language, even if they are not fluent enough in that language to query a database.

Translating the complete database required for an information search task is quite difficult because of the amount of translation required. The memory space required for storing language pairs translations prevents using this approach for large document collections.
Moreover, the limitations of this systems are widely known: MT systems provide a word for word translation, incorrect translations for polysemous words, etc.

b) As for query translation, the idea is to leave the documents in their original languages and to transform each query into every language at run time (Dorr et al., 1998). Query translation is efficient when short queries are presented, but unsophisticated techniques based on simple word substitution will negatively affect retrieval efficiency when compared with monolingual information retrieval scenarios. This adverse impact as pointed out by Dorr et al. (1998), seems to derive from translation ambiguity, limited lexical coverage, and a failure to correctly translate noncompositional phrases.

This second approach, is considered to be "faster and straightforward and has demonstrated retrieval performances comparable to the document translation approach", (Dorr et al. 1998).

Grefenstette (1998a : 523) interestingly compared CLIR to MT in the following way:

"CLIR is both easier and harder than MT. It is easier because MT systems must both (1) chose only one translation alternative for each input term and, (2) produce a syntactically correct output for each input sentence; information retrieval systems, on the other hand, function using a bag of words (classical IR systems consider both documents and queries as simple bags of words, and try to match up bags which have the most similar items): the words in a query do not have any syntactic order, and the more the words in the query, the better the results. A CLIR system producing a foreign language query from a native language query then does not have to produce syntactically correct output in the target language, and can retain more than one translation alternative for each original query term."

This technique uses online bilingual and multilingual dictionary to offer translations for queries. Another technique using parallel corpora of texts, i.e. naturally-occurring sets of documents which are not translations of each other to generate necessary translation facility, is documented in Grefenstette (1998a) and Harnann (1998).

Automatic translation works better when domains are well defined and when the system is supplied with a specific terminology. What is particularly binding with CLIR is that IR is not limited to a specific domain so that for the operation to be successful we should have as many specialized terminologies as there are domains about which information is being sought.

3.2.2) Automatic Building of Cross-language Thesaurus Using Comparable Corpus as Input.

Harnann (1998) described an experience conducted by the Swiss Federal Institute of Technology in which they automatically built a cross-language thesaurus by using a comparable corpus as input. This corpus, consisting of Swiss newswire in three languages, was used as a naturally-occurring set of documents not being translations of each other (as in parallel corpora), and were independently-produced articles dealing with the same time period (Harnann, 1998 : 521-522). This experience has been evaluated within the TREC6 cross-language track framework (for more details see Harnann, 1998).

Hudon (1997) has pointed out the relevance of multilingual thesauri as a potential solution to increasing communication difficulties.

3.2.3) Bilingual and Multilingual Reformulation

Another linguistic technique such as "Reformulation" is used by SPIRIT+. It is a technique initially used in expanding the search to other parts of a database by using stemming and semantic relations. For CLIR the approach consists of using a bilingual
reformulation which tries every possible translation for each word. If there is an answer to the query the ambiguities are then solved by using relevant documents as a semantic filter to help choosing the right translation (Fluhr et al., 1999).

The purpose of the mechanisms developed in the EMIR³ (European Multilingual Information Retrieval) is to match the terms expressed in the users' query with the terms used - possibly in another language- by the authors of the documents contained in the database. The matching procedure relies on the lexical semantic knowledge integrated into the reformulation tool. It is based on an automatic linguistic (morphological and syntactic) analysis of both the queries and the documents. A statistical treatment allows a comparison between the query and the documents linked to the proffered concepts.

SPIRIT with its reformulation module exemplifies this type of technology (cf. Fluhr, 1994, Bisson et al., 2000).

4. Drawbacks to CLIR & Improvement Possibilities

As we have all experienced it, the present level of translation when using the Altavista "Translate" facility is not adequate. The main shortcoming of this approach is that MT tools can provide only one translation for each word and that if the wrong translation is provided, something which is very likely in the case of polysemous words, the result is useless.

Efforts have concentrated on the query translation techniques. There are two ways of finding translation terms. The first and most common one is through the use of a bilingual dictionary listing the possible translation terms. The main issue in this type of application is the construction and development of bilingual resources (dictionaries, terminologies, etc.).

The other one is based on alignment technology: word alignment, compound alignment, sentence alignment and whole corpora alignment. The whole spectrum can be used to improve query translations. There is a growing interest for such a type of resource (cf. Langlais et al. 1998) which are regularly growing in quality.

Two important requirements for these bilingual dictionaries designed for CLIR are: 1) the coverage provided by the dictionary for domain independent corpora 2) the efficiency of the translations provided for finding relevant documents in the other language.

There is of course an even more serious problem in properly understanding nuances that are more cross-cultural than cross-lingual (cf. Harmann, 1998; Hudon, 1997).

Other techniques are now being tested. They have only given limited satisfaction so far. Among those we can mention Lexical Conceptual Structures (LCS) (for a detailed description of this method cf. Dorr et al. 1998).

Notes

1. This term is defined as "the retrieval of documents based on explicit queries formulated by a human using natural language when the documents are expressed is not the same as the language in which the queries are expressed. It is the ability to issue a query in one language and receive a document in another that distinguishes cross-language information retrieval from monolingual information retrieval. Although monolingual information retrieval is outside the scope of this definition, cross-language and monolingual retrieval functionality can certainly both be provided by a single system. (...) The term "cross-lingual" still appears fairly often, and it is probably slightly better from a grammar perspective. But "cross-language" is now used by enough people that it is clearly the preferred term between those two", Douglas W. Oard, http://www.clis.umd.edu/faculty/oard/oard.html.

2. It is a research project conducted within the framework of a program sponsored by the "Association des Universités Francophones " (AUF) - an international association whose mission is to promote the dissemination of French as a scientific medium (cf. see Mustafa el Hadi and Jouis, 1996; Béguin, Jouis and Mustafa el Hadi, 1997). The main characteristics and purpose of these systems i.e. their description as far as linguistic approaches are concerned, are documented in
previous works.

3. With the growth of practical systems such as machine translation or natural language interface in the field of computational linguistics, the need for managing lexical units specific to an application domain or a certain sublanguage is increasing. One of the major problems is the treatment of terminology representing a domain-specific knowledge or concepts. IR work and computational linguistics are both used for extracting a content-bearing lexical unit for a certain topic or a domain. They also both use the same methodology, i.e., linguistic and/or statistical methods.

4. SPIRIT means "Syntactic and Probabilistic Indexing and Retrieval of Information in Texts."

5. To sum up the project aims at dealing with the following situations:

- The user queries in his mother tongue textual databases containing documents written in a foreign language. The user can read this language but does not master its vocabulary.

- The user queries in his mother tongue textual databases containing documents written in a foreign language for which he has no understanding at all. In this situation, the bilingual query system helps the user decide whether or not fully-fledged translation is necessary by assessing the degree of relevance of the document in relation to the query. In this scenario, it might be interesting to combine the multilingual query system with an even rudimentary automatic translation systems to further enlighten the user and help him decide whether or not the text deserves being translated.

The user queries with a single question in his mother tongue databases containing texts written in a foreign language to get a global view of the problem. This last scenario is of help to many users, even if they master the various languages encountered in the database, because the system enables the user to speed up and increase the efficiency of his research (cf. Fluhr, 1994).

References


