Integrating Different Strategies for Cross-Language Information Retrieval in the MIETTA Project

Paul Buitelaar, Klaus Netter, Feiyu Xu
DFKI Language Technology Lab
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
{paulb,netter,feiyu}@dfki.de

ABSTRACT

In this paper we describe an integrated approach to cross-language retrieval within the MIETTA project, whose objective is to build a special purpose search engine in the tourism domain that covers information from a number of geographical regions. MIETTA is designed to enable users to search and retrieve information on the regions covered in their own language preferably. In order to facilitate the user with such functionality, the system includes document translation, cross-language query translation, multilingual generation from information extraction templates and document classification. In addition, query expansion is offered to identify proper query translation and enable template matching for information extraction purposes.

Keywords: Cross Language Information Retrieval, Information Extraction, Natural Language Generation, Machine Translation, Query Translation, Query Expansion

1 INTRODUCTION

MIETTA (Multilingual Information Extraction for Tourism and Travel Assistance) is a project in the Language Engineering Sector of the Telematics Application Program of the European Union, that combines amongst others technologies from information retrieval with the areas of shallow natural language processing and information extraction (see e.g. [1]). The main objective of the project is to facilitate cross-language retrieval of tourist information in several languages (English, Finnish, French, German, Italian) and on a number of different geographical regions (the German federal state of Saarland, the Southwestern Finnish region centered around Turku and the Italian city of Rome).

Tourism is an application domain, which is almost by its very nature multilingual and which is highly dependent on providing easy access to information on target regions. The world wide web is a growing resource for such information next to other data bases, and one of the main objectives of MIETTA is to collect and disclose such information through a specialized server, helping the user to find and navigate through this information preferably in his own language.

Approaches to cross-language information retrieval, which become relevant in this context, typically include the translation of the user query, or alternatively a full translation of the document base (see e.g. [2] [3] [4]). In the MIETTA project, a third type of multilingual information access is added, where the system can produce the relevant information in different languages by generating from a set of filled-in language independent templates which are obtained through information extraction technology. Information extraction can be used as a restricted, but goal directed search strategy that supplies the user with a fixed set of query options (templates) from which the system can generate natural language representations in preferred languages. Such a strategy presupposes domain specific natural language processing, term extraction and term translation for all languages involved.

Also, in connection to this template-based search the user will be offered the possibility to navigate through a classification hierarchy, which on the surface is language specific, but in its underlying form is language independent, thereby enabling access to documents in multiple languages. Classification-based navigation is meant to allow the user to browse through a refined conceptual classification tree or graph with categories (some of them corresponding to templates) in his own language that will take him to documents in any language. This, like query translation, requires some passive knowledge of the foreign languages.

2 THREE STRATEGIES FOR CROSS-LINGUAL INFORMATION ACCESS

In the following we discuss the strategies on cross-lingual information retrieval in more detail as we briefly presented them above, concentrating above all on possible strengths and weaknesses. The first two strategies are geared more towards standard information retrieval approaches, the third one is based on information extraction.

2.1 DOCUMENT TRANSLATION

Full document translation can be applied offline to produce translations of an entire document. The function of this translation is twofold, viz., to provide the basis for constructing an index for information retrieval or to offer the user the possibility to browse through a translated version of an original translated in his own language or in a language which he understands. The ideal scenario is that the user enters a query term in his own language, this term is matched against an index constructed from the translation of a document, which is then presented to the user. If he is satisfied with the relevance of the document he can then also access the original document and verify the content in this version.

In a genuine multilingual document base, offline document translation can result in a multiplication of the entire sets of documents in all languages covered. In such a scenario, every monolingual index is constructed from such a set of translated and original documents and covers the content of the entire multilingual original document set. Such an approach is realistic if storage space to account for the multiplication of documents and indices is not a relevant factor, i.e., above all in those cases when a specialised and limited subject domain is addressed.

Document Translation can be the preferred strategy in cross language retrieval, if the purpose is to allow the users to search for foreign documents in their own language and receive results back in that language. In this sense it is clearly a superior option which does not even require passive knowledge of the foreign language from the user.

Machine or (large scale) human translation, however, is not always available as a realistic option for every language pair. Typically machine translation systems only translate between language pairs which involve one of the major languages, such as English, German or Spanish, and ever so often only English will be the common language paired with all other languages. Also, without careful adaptation to a particular domain, machine translation may not always fulfill the necessary quality standards, and will not be sufficiently satisfying for a user even as a purely informative translation which is meant to give him only a rough indication of the relevance of the foreign language document.

Still, it must not be forgotten that even in those cases offline document translation can be useful to construct a 'translated' index, which can be matched against the user's query. Even if it is not foreseen that a user inspects the translations, the 'translated' index can point directly to the corresponding original documents. Since indexing is typically on the basis of words, terms or maximally phrases and since machine translation systems normally offer quite satisfactory quality at this level, such an approach can be a viable or even superior alternative approach to online query translation. One of the main differences between the two alternatives is of course that the user can still select and determine the correct translations in an additional interactive step (see below)

Document translation in the MIETTA project can be implemented on the basis of the LOGOS machine translation server. The translation directions which can be provided are:

- German $\Rightarrow$ English, French, Italian

- English $\Rightarrow$ French, German, Italian, Spanish

Thus, a translation from Italian or Finnish documents into other languages will be impossible for the moment, unless this document is already manually translated into other languages, preferably English.

2.2 QUERY TRANSLATION + EXPANSION

Online translation can be applied to the query terms entered by the user. Online query translation will help the user to formulate his query in another language than his own by offering possible translations as options for him to choose from. It makes sense provided that the user either has at least some passive knowledge of the foreign language or that he can have the retrieved document(s) translated automatically or by hand. Thus, the degree of passive knowledge required will very much depend on the additional facilities available.

If a user has no understanding of the foreign language, a retrieved document will be of no use without translation facilities. If he is not able to narrow down the scope by means of disambiguation or determine the relevance of retrieved foreign language documents, even online automatic translation could be of no avail for sheer reasons of volume. The possibility to disambiguate or expand a query in the process or for the purpose of query translation is therefore quite essential.

In MIETTA, we therefore provide a close interaction with query expansion as a support for query translation, offering to the user only semantically adequate translations. Such an approach originates from the believe that the user can be more easily asked about the *meaning* of a search term in his own language, than about the proper *translation*.

Query expansion can add semantic knowledge to the original query by narrowing down the meaning through introducing related terms, synonyms or hyponyms. This extension will exclude those documents in which the related terms do not occur or rather where terms related to alternative readings are prominent and thereby raise precision. Conversely, query expansion can also raise recall by broadening the search space with semantically related terms, synonyms and hyponyms.

For instance, the query “rabbit” can be expanded in a number of ways. By adding the related terms “pork” and “meat”, the search will go in the direction of “cooking and recipes”, whereas adding ”deer” and “wildlife” will look into things like “nature and preservation”.

2.3 INFORMATION EXTRACTION

A third method for overcoming the language barrier is to use an interlingual representation as provided by information extraction technology. In the scenarios above the content of the documents is disclosed through an index and a corresponding query and retrieval process, which takes the user to the appropriate document(s). Information extraction on the other hand is based on the assumption that a targeted partial analysis of the texts will provide the possibility to extract selected relevant information and to store this information in the form of templates (frames with predefined slots). Additionally, information extraction combined with natural language generation offers the potential to present selected information in other languages.

One problem with this approach is of course the limited and predefined amount of information that can be disclosed, as it presupposes a partial understanding of the original text. A closely related issue is whether the extracted information can be represented by an interlingua or language neutral format. Clearly, this is not a problem for the kind of data that will be more or less independent from human languages anyway, e.g., numeric information or most kinds of named entities. However, terms in general, corresponding to all kinds of nominal phrases, other than dates and names, will be difficult to represent in a truly interlingual format unless extensive multilingual domain modelling is achieved before hand, identifying important terms and their possible translations.

Thus, in some sense, the employment of information extraction for multilingual information handling can be interpreted as an attempt to build a highly intelligent machine translation system, which not only translates (relatively close to the surface of the language pairs), but which attempts to understand and isolate some important relations and to present them in different target languages.

As such, it is quite clear that the approach will be less robust and also less flexible than the considerably more shallow text retrieval methods. However, it also offers a substantially more structured access to the content of a document base than standard text retrieval methods do.

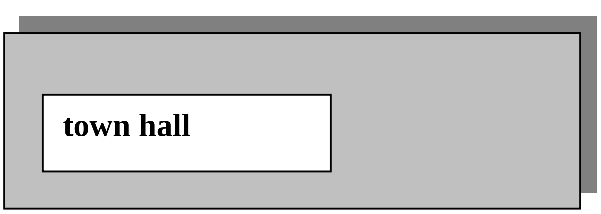
It should be pointed out that the assumption of a full fledged natural language generation component is of course not mandatory, but that the information can be equally well presented to the user in a tabular format employing the slot labels of the template. The generation and translation part in this case would then be reduced only to processing those terms that fill the slots.

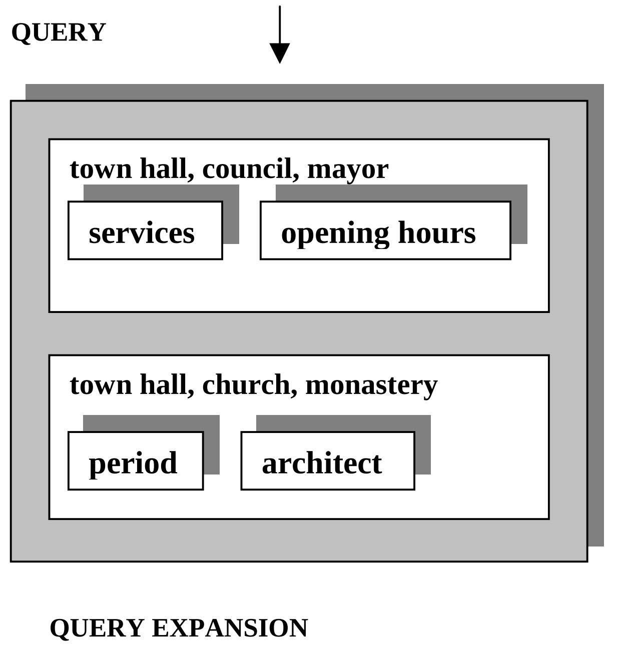
2.4 A SIMPLE EXAMPLE

The following example serves to illustrate the combination of query expansion and information extraction. It shows how the user can navigate along different meaning dimensions of a search term and how he can exploit information extraction technology to obtain access to deeper and more structured information related to such meaning dimensions.

The search term “town hall” in English is ambiguous between a BUILDING and GOVERNMENT interpretation. Both for clarification within the query language and for query translation into other languages, we need to disambiguate between these. By giving the user a set of related terms, corresponding to one of the interpretations (semantic classes in the classification hierarchy) matching templates can be found.

So, for instance, the related terms “council” and “mayor” corresponding to the GOVERNMENT class will lead to a matching template that includes slots for “services offered” and “opening hours”. On the other hand, related terms like “church” and “monastery”, which correspond to BUILDING, will lead to a template that includes slots describing “building period” and “architect”.





3 LANGUAGE TECHNOLOGY

In the MIETTA project a broad range of language technology methods will be employed to support and enable the above mentioned approaches to information retrieval and information extraction. Most of these are relatively standard technologies that have proven useful in many projects targeting cross-language information retrieval, others are more specific to information extraction and its use in cross-lingual information access.

3.1 NATURAL LANGUAGE ANALYSIS

Information extraction in particular, but also the construction of full phrasal indices for information retrieval presumes robust natural language analysis as a pre-processing step. This will enable filling out templates with appropriate data from natural language texts and identifying suitable phrases as index terms.

Natural language analysis will take place in four stages:

- Tokenisation: recognition of sentence boundaries, proper names (named entities should be recognised before morphological processing) and abbreviations

- Morphological Analysis: lemmatisation, part of speech (POS) and other morphological information

- Part of Speech Tagging: disambiguation of POS by using contextual information

- Shallow Parsing: phrase recognition using POS tagged word stems and the corresponding morpho-syntactic information

For the morphological and grammatical analyses, language specific data as well as additional language independent techniques and tools are needed:

- Language specific monolingual dictionaries

  This includes the use of gazetteers on company names, countries, person names, etc. At least partly, these will be language independent.

- Language specific morphology tools for lemmatisation and decomposition

  In addition to language specific morphological processing, the MIETTA system will also use socalled "fuzzy matching" for IR indexing purposes. Fuzzy matching simply tries, through string manipulation, to match the longest substring or combinations thereof. In some sense this could be seen as language independent morphological processing.

- Language specific NP grammars

  NP grammars in MIETTA need to be dependency grammars, because we consider head-modifier combinations as a basic term unit, instead of for instance a string, keyword or phrase. See below for more details on this.

3.2 TERM EXTRACTION + TRANSLATION

One option to facilitate a precise matching of query terms on to the requested documents is to analyze the document set and identify relevant terms. This is the profiling part of information retrieval, in which documents are indexed by the terms that occur in them. The big question then is what should constitute a term.

Traditionally, any string (or parts thereof) in a document will be considered a term. This secures a high recall, but lowers precision considerably. Therefore a more extensive definition of what constitutes a term is needed. For instance, experiments with terms based on (nominal) phrases have shown promising results (e.g. [5]).

Head-Modifier Terms

In MIETTA we take this idea even a step further by normalizing phrases into head-modifier constructions that are semantic abstractions over a number of variations of phrases (using [6]). An example will clarify this more clearly. The following nominal phrases:

Bill Clinton's computer terminal
the terminal of Bill Clinton's computer
the terminal of my 20 year old computer
the terminal of my 20 year old son's computer

can be all reduced to the same head-modifier construction:

HEAD: terminal MODIFIER: computer

By taking head-modifier constructions as the basic unit, terms become much more closer to the abstract idea of a "concept" than is possible with just phrase recognition, or simple strings for that matter.

Independent from what a term looks like, is the question what constitutes a relevant term, where relevance is measured relative to the document it occurs in. For this purpose we use standard TF/IDF techniques, combined with a measure of mutual information between heads and modifiers. Mutual information is well known in statistical approaches to natural language processing and is used in a range of methods for determining the strength of connection between words or terms. For more details on mutual information see for instance [7].

Multilingual Thesaurus

Identifying relevant terms in a document set can be used to construct a thesaurus of such terms, that can then be used for efficient query expansion. Additionally, by performing term extraction in parallel for each language involved in MIETTA, we can construct a multilingual thesaurus that can also assist in query translation.

Importantly, by synchronizing thesaurus construction with class-based navigation, template definition and query expansion, we arrive at a completely homogeneous approach to concept-based search that allows the user to browse freely and consistently through a network of concepts that take shape in either classes, templates or related terms.

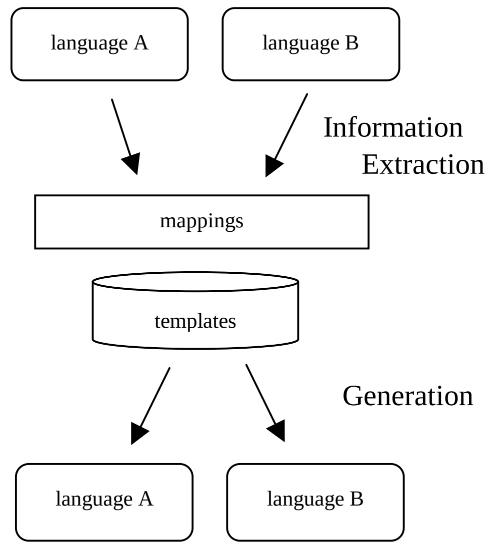
A multilingual thesaurus can be constructed in a number of different ways. First, by a direct transfer from terms in one language onto those in other languages. Secondly, through the use of abstract concept labels that are language independent and map to all languages as an interlingua or, thirdly, by choosing one language (e.g. English) as an interlingua and mapping all other languages onto this one. In MIETTA we take the latter approach, because it is the most straightforward and practical one.

3.3 NATURAL LANGUAGE GENERATION

Templates and sets of templates represent an ideal input to natural language generation. In the simplest case they can feed a system of canned text generation. In more complex cases some variations on the syntactic structure might be required. Finally, the most interesting situation is raised by cases when the system has to deal with *sets* of templates that jointly provide an answer to the query posted by the user. Then there is room for a full-fledged system, exploiting sophisticated text planning strategies (see e.g. [8], describing the natural language generation system we use in MIETTA).

Multilingual Generation

The figure below (due to [9]) shows how language independent templates interact with generation into individual languages on the one side and mappings between information extracted in these languages on the other.

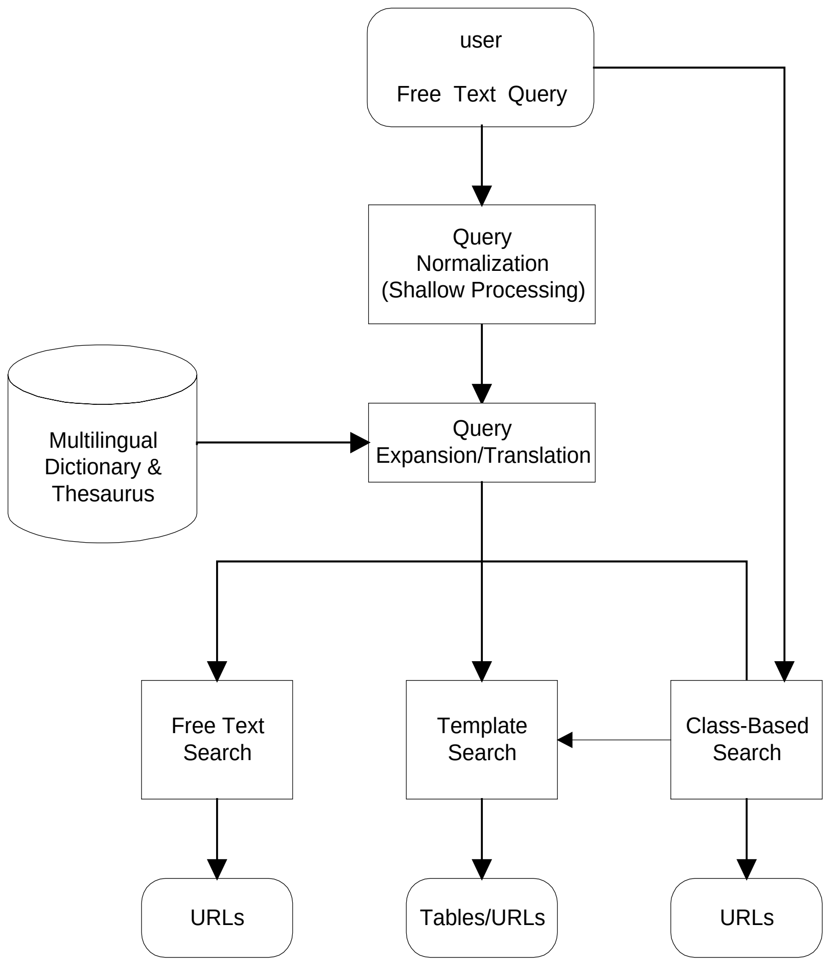


These mappings are crucial to the multilinguality of the information extraction and natural language generation systems and will be, as mentioned before, organized in a multilingual thesaurus that maps terms in one language onto those of another.

4. INTEGRATION

To the user, the different strategies (Document Translation; Query Translation combined with Query Expansion; Information Extraction combined with Multilingual Natural Language Generation and Classification Navigation) should be completely transparent, except for the input and output languages, which can be set by options.

Such transparency, however, requires an integrated approach to cross-language retrieval that combines these strategies, depending on availability and need. The following figure gives an overview of how query expansion is envisioned to take a central role in this.



We assume that the user has basically two choices for searching; one is on the basis of free text queries, the other on the basis of concept-oriented classifications. Free text queries can take the user either to different kinds of full text indices and/or it can take the user to a classification hierarchy, if the search term happens to map onto on of the conceptual classes. In the class-based navigation strategy the user can then browse through the conceptual tree, whose nodes are either directly associated with sets of URLs or document titles, or alternatively with templates related query forms (also hierarchically ordered) that can lead to URLs or document titles indirectly.

The query forms are meant to support the user by guiding his search, i.e., making it clear to him on what aspects of a certain class or category he can pose more specific queries. To further support a precise formulation of the (free text) query, the user can submit it to the query expansion component. Here, the query is first processed and normalized, where appropriate, and can then be enriched or made more precise on the basis of a thesaurus.

While these scenarios so far can be implemented also in a monolingual environment, multilingual search adds another layer of complexity. Thus, depending on the input and output languages, that is, query and document language, we can envision a number of different scenario's. The most simple one is where the query and document language are the same. This is for instance the case when a German user asks for information about the Saarland. This region only has information available in German, so for this scenario no translation is necessary at all. The German user will simply present his query to the system that will then normalize it into a head-modifier construction as far as possible and allow the user to expand his query with additional German terms from the thesaurus. Finally, given the user's preference, the simple or expanded query will be transmitted to one out of the sub-processes of query processing: free text query or template query. In this scenario both of these are possible.

A somewhat more complicated case is the situation where the query language is different from the document language, although translations of these documents exist in the query language. This is for instance the case when a German user requires information about Rome. Documents in Rome are available in five languages (English, French, German, Italian and Spanish), in parallel human translation. This allows the system to access term indexes in the query language (German) and the whole process will be equal to the previous one. Where no human translation exists, but an automatic translation of the document into the query language has been carried out, the user should be given the choice to see the translated document as a response to his query or to be taken to the original because of inadequate translation quality.

Finally, the really complicated cases are where the query language is different from the document language, and no translations of these documents exist in the query language. This is the case when a German user requires information about Turku, or when a Finnish user asks for information about the Saarland. In both cases, no translation from or into Finnish is available and the system has to perform a combination of query translation and query expansion in order to facilitate template query: information extraction in the document language and natural language generation in the query language.

Here, for instance, a German user submits a query to the system that will be normalized to a German head-modifier construction and then translated into Finnish by the query expansion and translation component. This translation, however, will be only available in the background, because it is not assumed that the German user will be able to judge the translation on correctness anyway. Also, the expansion component, given the information available in the multilingual thesaurus, already expanded this translation into a template that can be matched on Finnish documents. The results thereof are then synchronized with their German translations that again are available through the multilingual thesaurus, and these are used in generating a German text or table.

All of this, of course, depends on the coverage of the terms in the thesaurus, which should therefore be quite extensive and to the point, that is, covering many domain specific (simple or complex) terms. Nevertheless, if a query term is not matched by the thesaurus, the fall back option is a free text query with the translated query. This, however, has two big disadvantages. First, the translation is not semantically guided and will therefore be less accurate. Secondly, even if good results are found these can be only presented in Finnish, which will largely be useless to the German user.

5. CONCLUSION

We presented a cross-language retrieval strategy that combines document translation, query translation, query expansion, information extraction, natural language generation and class-based navigation. By integrating all of these into a coherent approach, the different strategies for cross-language retrieval that are needed can be left transparent to the user.

6. REFERENCES

[1] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In: E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*, pages 383-406, MIT Press, 1997.

[2] G. Grefenstette. (ed.) Proceedings of the SIGIR96 Workshop on Cross-Linguistic Information Retrieval, 1996.

[3] W.G. ter Stal, J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter,

G. Smart (1998) Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development. In: Computer Networks and ISDN Systems 30(13), pages 1237-1248, Elsevier Science BV

[4] MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. Proceedings AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Menlo Park CA, 1997.

[5] D. Evans. Lessons from the CLARIT Project In: Proceedings of SIGIR93, Pittsburgh, PA, USA, 1993.

[6] R. Backofen, J. Baur, M. Becker, C. Braun and G. Neumann. An Information Extraction Core System for Real World German Text Processing. Proceedings of the 5th ANLP, Washington DC, 1997.

[7] K. W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, pages 22-29, 16, 1990.

[8] S. Busemann and H. Horacek. A Flexible Shallow Approach to text Generation. In: E. Hovy (ed.) Proceedings of the Nineth International Natural Language Generation Workshop (INLG98), Niagara-on-the-Lake, August 1998.

[9] L. Dini. Parallel Information Extraction Systems for Multilingual Information Access. Paper presented at Euriscon, 1998.