The SYSTRAN NLP Browser

An Application of Machine Translation Technology in Multilingual Information Retrieval

Denis A. Gachot, Elke Lange and Jin Yang

{denis | elke | yang}@systranmt.com
SYSTRAN Software, Inc.
7855 Fay Avenue, Suite 300
La Jolla, CA 92037

Abstract

The approach of using an existing machine translation system in multilingual information retrieval, as usually proposed, consists of automatically translating queries, or even the entire textual database, from one language to another. The information that machine translation technology can provide to multilingual information retrieval has been more extensively explored at SYSTRAN. An existing information retrieval tool, which is based on SYSTRAN parsing and machine translation technology, has been investigated for use in information retrieval. This paper is a description of the implementation of what is now called the SYSTRAN NLP Browser, a cross-linguistic multilingual information retrieval system. The first section discusses the utilization of machine translation technology in multilingual information retrieval in general. The second section describes the implementation of the NLP Browser. The third section discusses the present approach and the current development status, followed by the conclusion.

1. Introduction

Multilingual information retrieval, referred to as retrieval of text in one language, based on queries in another, has been explored mainly via three approaches: thesaurus-based, corpus-based and machine translation based [4]. The approach of using an existing machine translation system usually involves automatic translation of the queries, or even the entire textual database, from one language to another. Only few experiments which reflect this implementation have been reported. There are no reports to date, which indicate any other uses of machine translation techniques in information retrieval. In 1995, SYSTRAN Software, Inc., which is well-known for its machine translation systems, received funding from the US Government to develop a multilingual information retrieval system based on its natural language parsing and machine translation technology. The development of this system explores the potential of using machine translation for multilingual information retrieval in a way that goes beyond the mere translation of queries and/or documents.

Machine translation is the now traditional and standard name for computerized systems responsible for the production of translations from one natural language into another, with or without human assistance [3]. The possibility of using an existing machine translation system to translate queries and/or databases, although suggested repeatedly, has so far not received enthusiastic support [4]. The limitations mentioned usually center around the fact that machine translation systems make errors and only work on a limited domain [1]. In Davis and Dunning's study, various approaches for automatic translation of queries for creating multilingual queries has also been proposed, but using a fully automatic machine translation system was not considered [6]. Perhaps because "the disappointing past and present" of machine translation [2] is far from producing high quality translation, the feasibility of integrating it in multilingual information retrieval is too easily rejected. Also perhaps because machine translation and information retrieval fill different niches, the two disciplines have not seriously exploited each other. However, the quickly growing information in many different languages available on the Web makes multilinguality an important characteristic of a rapidly increasing tasks. This fact alone encourages another look at the potential usage of machine translation technology in multilingual information retrieval.

The major task of a machine translation system is to process natural language. There is some evidence that natural language processing techniques may be applied for information retrieval to be effective [5, 7]. Improvements in information retrieval are reported to have been made possible by a very fast syntactic parser in deriving certain types of phrases from the text [5]. Linguistic analysis was also used in the DR-LINK system to achieve a high level of intelligent information retrieval [7]. The START (SynTactic Analysis using Reversible Transformations) natural language system at Massachusetts Institute of Technology stresses the power of advanced techniques of language analysis, understanding, and generation in intelligent information retrieval.

Despite different approaches to machine translation, the basic tasks of a machine translation system can be generalized as source text analysis (including morphological, syntactic, semantic, and knowledge representation), source-target transfer (or mapping to a language independent representation), and target language generation, in conjunction with extensive bi- or multilingual dictionaries. A variety of information (morphological, syntactic and semantic) is accumulated and recorded throughout the entire process. We postulate that this accumulated linguistic information provides a powerful means for multilingual information retrieval: the source language parsers may be used to parse (not only translate) the database and/or the queries. The multilingual aspect is a given for machine translation, since it works with at least two languages at a time, and must be based on bi- or multilingual dictionaries.

2. SYSTRAN NLP Browser

2.1 Background

SYSTRAN has demonstrated success in the machine translation field with its long history spanning nearly 30 years. Constant development and modernization have resulted in today' s state-of-the-art machine translation systems. Language-pair capability was gradually expanded: 13 language-pair systems are currently available in production quality, other language pairs are constantly being added and are in various stages of development. Table 1 shows the current operational and pilot language pairs.

Table 1 SYSTRAN Machine Translation Systems

Operational Systems	English	↔	French, German, Spanish, Italian, Portuguese
	English	→	Arabic
	Japanese	→	English
	Russian	→	English

Pilot Systems	English	→	Dutch, Danish, Swedish, Norwegian, Finnish
	English	→	Russian
	English	→	Japanese
	English	→	Greek
	French	→	German
	German	→	French, Italian, Spanish
	Chinese	→	English
	Serbo-Croatian	→	English
	Korean	↔	English

SYSTRAN, a general-purpose fully automatic machine translation system, employs a transfer approach. A unified and highly modular architecture applies to all language-pair systems. This results in an internal representation of the source language structure and translation information that is formally identical no matter which languages are involved. In most machine translation systems, the crucial components are the dictionaries and the source language parsers. SYSTRAN dictionaries and parsers have evolved over a long time, have been tested on huge amounts of text, and contain extremely detailed linguistic rules and a large terminology database covering various domains. The dictionaries for all language-pair systems altogether contain over two million terms.

In addition to the machine translation systems per se, SYSTRAN has employed a development tool which lent itself to becoming the precursor of the new NLP Browser. For over 16 years, the linguists have been using a parser-based data retrieval tool, PDIAG ("Parsing DIAGnostic"), to collect data exhibiting particular grammatical or semantic features for use in development of machine translation systems. PDIAG utilizes SYSTRAN's source language dictionaries and the source language analysis modules to produce a parsed database. User queries are then processed, and sentences fulfilling query requirements are selected. In house, it has been an extremely useful means to gather data for the study of grammatical phenomena or lexical ambiguity. For example, given a large corpus of Spanish text, linguists would be able to extract all sentences containing any form of the word "*todo*" ("*todo*", "*todos*", "*toda*", "*todas*") to see its usage in live text, or they might restrict the search to subgroups (e.g. "*todo*" used as an adverb, or "*todo*" modifying abstract nouns, etc). Another example for the use of PDIAG would be to extract all sentences with reflexive verbs in order to construct generalized translation rules. Nowadays, the increasing accuracy and detail of the parsers and the richness and depth of semantic coding in the lexicons have allowed PDIAG to evolve into a tool for retrieving information for a wider user base.

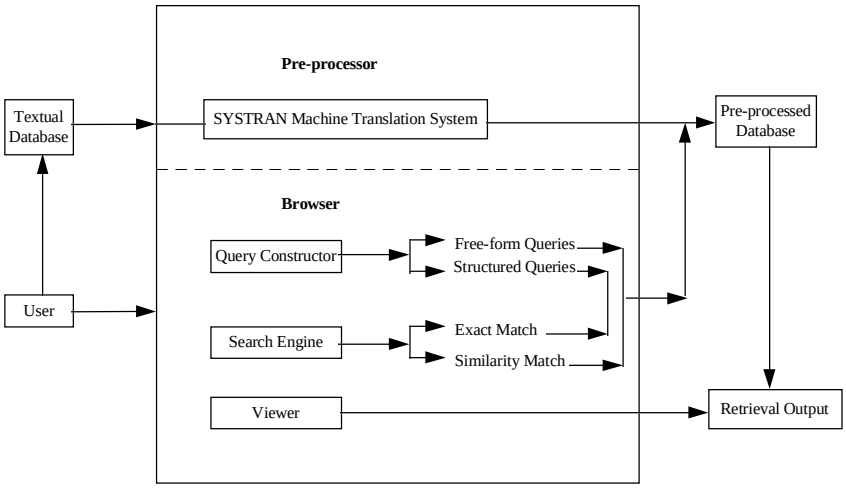
In response to increasing interest in multilingual information retrieval, the PDIAG tool has been expanded to what is now called the SYSTRAN NLP Browser, which is a multilingual information retrieval tool. Current work on the NLP Browser is sponsored by the US Government. The work, started in 1995, has produced a working prototype system, which is now being tested on a set of English and French texts taken from English and French editions of the journal "*Air and Cosmos*". This paper outlines the basic concept and discusses its potential and challenges.

2.2 System Description

2.2.3 The Components

The NLP Browser consists of two parts: document pre-processor and browser. The pre-processor works on documents separately from the query and retrieval process. Pre-processing of documents results in a searchable file which contains detailed information on the morphological, syntactic, and semantic level, as well as information on the output translation. Documents are automatically parsed and translated by SYSTRAN machine translation systems. The parse information is an intermediate product of the translation process which is ordinarily discarded. However, the pre-processor captures this information and stores it in a condensed format. The separation of pre-processing from the actual information retrieval allows high-speed query processing. From the pre-processed database, the browser performs retrieval with three components: a query constructor allowing the user to input various types of queries, a search engine to search and retrieve the relevant sentences, and a viewer to display the retrieved sentences (and their context) in the source language and/or as machine translated target language output. Figure 1 gives an overview of the system.

Figure 1 SYSTRAN NLP Browser - System Configuration



2.2.2 Pre-processing of Linguistic Information

The linguistic information contained in the pre-processed file represents the results of a multitude of decisions made, examined, and revised during the entire machine translation process. The output of the parser is an abstract representation, in which homographic words have been resolved for their appropriate part-of-speech, multiple-meaning words have been disambiguated, and syntactic and semantic relationships

between words or phrases have been established. In a similar fashion, all decisions made during the transfer and target language generation steps are also preserved.

The morphology of both source and target language words is transparent. Each text word is tagged for its inflectional value within its context as the result of the translation process. For example, the French word "*fait*" in the sentence "*il fait sa fortune*" is marked as a third person, singular, present tense, indicative verb form. However, the entire inflectional pattern of this verb is also available, so that a user may ask to retrieve sentences with any form of the verb "*faire*" and receive matching sentences with the forms "*faire*", "*font*", "*faisons*", "*fait*", etc. Note that the sentence "*ceci est un fait*" would not be matched in response to the query for the verb, since homograph resolution would have tagged "*fait*" as a noun in this context. The query does not necessarily have to include the specification "verb" in the above example. It may simply state that it intends to retrieve sentences in which the "object of *faire*" is a certain type of word. The system then knows to look for certain syntactic or semantic relationships between a verb (any form of "*faire*") and a noun or noun phrase.

The pre-processed file contains detailed indicators of the syntactic and semantic function of words or phrases and their relationship to each other. This syntactic and semo-syntactic information includes the standard surface syntactic designators, such as Verb - Direct / Indirect Object, Attributive Modifier, Apposition, etc. Identification of subject and predicate, and, in addition, deeper relationships which combine several surface structures into one designation like Action - Agent, Semantic Modifier - Modified, etc. are available. For example, in the English phrase "*repairing equipment*", the noun "*equipment*" is identified as the "semantic object" of the verb "*repair*". That same designation of the variants of these words is also given in the following sentences. Thus all of these sentences could be retrieved with a single query for "*repairing equipment*" :

(a) He repaired the equipment.
(b) Repairing the equipment is difficult.
(c) The equipment was repaired by him.
(d) The equipment repaired by him is still defective.
(e) The equipment which he repaired is new.
(f) The repaired equipment broke again.

Noun phrases, important for information retrieval, are marked as units, with identification of the head noun and complete specification of their internal structure. This means that a query can ask for the entire noun phrase or certain portions.

Stored information also includes semantic categories (e.g. Device, Combustible) and domain designators (e.g. Aviation, Medicine) which are encoded in the dictionaries. The translation is also included in the pre-processed database so that queries may specify target language words, again in any inflected form. Currently a hierarchical set of about 500 semantic categories are used.

2.2.3 Query Construction

The pre-processed database can be searched by two types of queries: structured and free-form queries. Structured queries allow great flexibility in expressing the desired information and at the same time make it possible to impose restrictions limiting the hits to relevant data only. These constructions are based on the PDIAG technology used by SYSTRAN linguists over many years and are fully functional at this time. Free-form queries allow the user to input queries as regular sentences in English or French, or one of the other languages for which SYSTRAN has a production-quality machine translation system. This is the area of the NLP Browser which is currently under development and for which all information

contained in the pre-processed database is of utmost importance. The aim of providing different query constructions is to present the user with several options to express what he/she is looking for. This may involve a simple key word or phrase search and may range to complex conceptual information retrieval.

2.2.3.1 Structured Queries

Structured queries consist of formal constructs with a pre-determined syntax in which a number of simple macros can be used. These queries are flexible and can take full advantage of all information discovered in the parsing process. Structured queries usually begin by identifying a key word or key feature. A key word may be a single word or a sequential string of words. A key feature may be any syntactic or semantic information on a single word whether the information is encoded in the dictionary or generated by the parser. Queries can be formed using AND, OR, NOT, and special macros to make various sophisticated linguistic inquiries. Macros are preceded by a period to distinguish them from natural language words. An illustrative example of a structured query containing only macros might be the following:

.HUMANS+.PROF+.SUBJ+.BPRED+.LOOK+.BSEMOBJ+.SPBDY

If a word is HUMAN and PROFessional, and if it is the SUBJect of the sentence, and if the PREDicate has the semantic category LOOK, and if its SEManticOBJect has the semantic category SPBDY (spacial body), then retrieve the sentence.

This will retrieve all sentences referring to any professional studying/observing/investigating/… any spatial body (e.g. planet, asteroid, Mars, …). The search can be restricted by using key words instead of category names or by adding further qualifiers. In its abstract form, the query may be used for retrieving information from documents in any language. When key words are added, they may be given in the foreign language or in English. Table 2 gives some examples of simpler structured queries which may be used alone or embedded in longer constructs.

Table 2 Samples of Structured Queries

Structured Queries	Explanations
.WD=nasa.+.AC	exact word match “NASA” if all-caps
.KD=fly+.VERB	all forms of VERB fly (fly, flies, flew, flown, flying)
.WD=ph*ll*s	word “Phyllis”, but unsure of spelling
.SCIOR	references to scientific organizations
.WD=aircraft/.MNG=avion	word in either the source text or its translation
.WD=rocket+.BR+.KD=society	contiguous phrase “rocket society”
.KD=fuel+.GGV/.MNG=gaz+.GGV	noun phrases whose head noun is “fuel” (English) or whose head noun has the meaning “gaz”(French)
.KD=rocket+.BMOD+.MACRO	“big” (has a semantic category MACRO) rocket(s)
.KD=russian+.MODNOUN+.colonel	colonel who is Russian
.ACRO+.AVIA	acronyms referring to aviation
.KD=repair+.BSEMOBJ+.EQWAR	references to the repairing of military equipment

WD=exact word, KD=root word, MNG=target language word, all other capitalized forms = macros

Since the structured queries are derived from an in-house development/diagnostic tool, the potential complexity is great and mastering every aspect cannot be expected of the average user of the NLP Browser. The most common and most useful query types are presented to the user in a pre-defined form. Selection menus will be provided with categories and macros that may be substituted in certain slots of the proposed query. A major effort in the development of the NLP Browser concentrates on identifying query patterns which can be presented to the user in a simple format that masks pre-programmed complexity.

Beyond providing help menus for query construction, our goal for the use of structured queries is to allow the user to state the query in plain English and have the Browser infer the intention and automatically construct the query. It is not unreasonable to ask the user to restrict the "plain English" query to a certain format, for instance to the type of simple sentences given in the Explanations column of the above examples. This is a transition to allowing free-form queries.

2.2.3.2 Free-form Queries

Free-form or natural language queries allow the user to input queries in unrestricted language as opposed to the restricted language mentioned above, or rigid Boolean syntax and sophisticated linguistic constructions. The queries consist of a simple word, a phrase, or a sentence. A single word free-form query (e.g. “fighter”), is similar to the traditional keyword match, a phrase free-form query (e.g. “big rocket”) is usually a noun phrase. A sentence as a free-form query is ideally formulated as a statement (e.g., “the giant rocket was launched”) for which the user expects to find similarity matches. Future work on free-form queries will include full sentences, such as “I am interested in launching dates of giant rockets”, “When were giant rockets launched?”, “Tell me when the last giant rocket was launched”, etc. These types of queries will not be reduced to structured queries, but will be analyzed as full sentences by the machine translation system.

2.2.4 Search Mechanisms

Corresponding to the two kinds of queries, two mechanisms perform search based on exact match or similarity match.

Structured queries imply an exact match in the sense that all conditions specified in the query must be fulfilled. The query’s conditions, however, may be written to allow several degrees of fuzziness if desired. For example, in a word match, the user may ask for an exact match of the word as specified or for inclusion of all inflected forms of the word, or for words with similar spelling (by the use of wild card characters). Similarly, structural requirements may be rigid (e.g. “find direct object”) or they may leave room for inclusion of several surface structures (e.g. “find semantic object”).

The other type of match, similarity match, is the matching scheme for free-form natural language queries. When a free-form query is submitted, the browser first runs it through the machine translation system, processing it in the same way as the text database was pre-processed. Hereafter, the browser performs a similarity match between the stored internal representation of the query and the internal representation of each sentence in the database. In this way, not only surface similarities but also structural similarities can be matched. Ranking criteria are based on the “similarity” of two sentences on various linguistic levels: word-level, phrase-level and sentence-level comparison and ranking. Ranking at each level includes morphological, syntactic and semantic comparison. For example in a search for information about “big rockets”, sentences containing “giant rocket” have a higher ranking than “military rocket”. The “similarity match” engine is in the initial stages of development

2.2.5 Output Display

After the input of queries, the search engine sequentially matches the queries with the sentences in the designated database. When a sentence which satisfies the query condition is matched, it is tagged and retrieved from the database. Since the locations of the matched sentences in the database are recorded, the context surrounding the matched sentence, the title of the document in which the sentence occurs, as well as other format information provided in the database can be easily shown upon the user's request. The retrieved output may be an isolated matched sentence, or the entire article containing the matched sentence. This is provided in the source language and/or in its machine translated target language.

3. Discussion

With increasing interest in multilingual information retrieval, the possibility has been examined that machine translation might be useful at the fringes of a multilingual information retrieval system (by machine translating either the documents or the queries). Since the fast growth of multilingual resources on the Web, it is time that a closer connection between machine translation and multilingual retrieval be examined and put into practice. With the NLP Browser, we are attempting to bridge the gap and need to ask some important questions: In which way are the goals of machine translation and multilingual information retrieval similar? Where do they diverge? What can machine translation contribute to multilingual information retrieval? What can be learned from four decades of machine translation? What about the still "disappointing present" [2] of machine translation? Not all of these questions can be fully answered at this time, but we will attempt to address some of them.

In which way are the goals of machine translation and multilingual information retrieval similar? First, both machine translation and information retrieval users are interested in information gathering and/or dissemination. SYSTRAN' s users at US Government agencies have long used the machine translation output for "information scanning". Operators, who do not necessarily know the language of the source documents but have an idea what they are looking for, have been reading raw machine translation output searching for information of interest to them. It is a small step from this activity to requesting an information retrieval search engine that can automatically select relevant information. Secondly, both machine translation and multilingual information retrieval face the common challenge: multilinguality. Currently, the NLP Browser is tested on English-French and French-English pairs. In principle, the NLP Browser works on any language-pair for which a SYSTRAN machine translation system exists. French and English have been chosen by the Government as the languages for initial development of the Browser mainly because these are mature machine translation systems and because English will be the language of choice for the user queries.

What can machine translation contribute to multilingual information retrieval? An information retrieval system that goes beyond simple Boolean word matching can profit from an engine with knowledge and understanding of the language to be searched and also of the language preferred by the user. There are various levels of knowledge, but machine translation certainly has to deal with a number of them, such as detailed lexical and grammatical information on words, ambiguities, the structure of sentences, as well as syntactic and semantic functioning of their components. This is useful information for any attempt to automate language analysis. Text understanding (of document and user query) is the area where machine translation can make the greatest contribution. SYSTRAN' s approach is to make use of machine translation components that have grown and that have been refined over many years, namely large

dictionaries and parsers, as well as an already existing search tool (PDIAG) and to integrate them into a new information retrieval tool, the NLP Browser.

What about the still "disappointing present" of machine translation? First, due to the complex nature of language, perfection in automatic text understanding and translation cannot be expected any time soon, whether it is developed for machine translation or specifically for information retrieval. SYSTRAN at least has the advantage of already having a highly developed body of data and rules, although no claims to perfection are made. Secondly, the NLP Browser as outlined in this paper does not heavily rely on the final translation output. Instead of searching the translated surface rendering of the document, many of its search mechanisms look at a deeper level by searching the internal representation of the entire machine translation process in the pre-processed files. It is quite possible that, for example, certain relationships between words may be correct at this level although the final translation may be awkward.

In which way can work on information retrieval benefit machine translation? By using the same parsers and dictionaries for both machine translation and information retrieval, we expect to find problem areas that will lead to the creation of new mechanisms, and whose implementation will benefit both the machine translation and the information retrieval tasks. There is the possibility that translation results can be quite acceptable despite some inadequacies in full understanding of the text translated. For example, in translating between related languages, it is often not so important that all words have complete semantic categorization. For information retrieval, at least in the current approach, rich semantic tagging in the dictionary is desirable. We are planning to review and enhance the large set of semantic categories and to increase semantic coding in the dictionaries.

The SYSTRAN NLP Browser is currently in its early development stage, therefore no extensive test results can be given at this time. However, preliminary testing on a multi-lingual database clearly shows its promising potential. Current work concentrates on refining the user interface, providing easy-to-use instructions for structured query construction, and further exploring the similarity matching scheme for free-form natural language queries. Future work will entail:

Further refinements of the machine translation components. This includes 1) the further refinement of dictionaries needed, especially in terms of semantic categorization. 2) continued development of size and quality of parsers and dictionaries for newer language pairs. Currently, English and French are the most mature SYSTRAN systems, with Russian and German following closely. All other production systems may be good enough for producing reasonable translations, but could benefit from more development before becoming a reliable base for information retrieval.

Simplification of the query construction process. This includes 1) constructing user-friendly structured queries. The current format of structured queries, as used in-house, would require detailed user training which is not always convenient. 2) improving the handling of free-form queries and the corresponding similarity match engine.

Evaluation of the information retrieval performance on a large-scale, and comparing its performance with results from other systems. It is a non-trivial step to go from machine translation development to creating a user-friendly and effective multilingual information retrieval system. The possibility of integrating the NLP Browser with other information retrieval approaches will also be examined.

4. Conclusion

In summary, the SYSTRAN NLP Browser is unique in that it integrates machine translation technology at all levels of information retrieval. The database is pre-processed, and a wealth of intermediate information on each sentence as well as the translation is stored in compressed format. Queries can be structured to describe sophisticated linguistic information at various levels, or free-form queries can be parsed before searching. This integration spans the use of SYSTRAN's large multilingual dictionaries, its fine-grained parsing capability, its semantic disambiguation tools, and actual translation. Multilingual retrieval capability is realized by allowing a foreign language text to be queried in its original language and/or in English or in an abstract language-independent form, and by supplying the user with the results of the search in both the language of the document and in the user's language. The languages for which the browser will be available are determined by the language pair translation systems in production. Their number is constantly growing.

The state-of-the-art of machine translation technology is far from producing "high quality" translation. Drawbacks can be easily found if the employment of machine translation is limited to automatically translating queries or even the entire textual database from one language to another, since current machine translation systems certainly make errors. However, using machine translation technology in multilingual information retrieval can go far beyond the application mentioned in the literature so far, which was limited to translating the document and/or the query. The machine translation dictionaries, linguistic parser and its intermediate results can be employed in various applications, including multi-lingual information retrieval. The SYSTRAN machine translation parser makes mistakes, but the multi-level approach outlined here anticipates a high success rate.

Acknowledgments

This project is supported under a contract with National Air Intelligence Center (NAIC). We would also like to thank Dale Bostad from NAIC for his continuous interest and long-time support.

References

[1] Christian Fluhr (1995) Multilingual information retrieval. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology*, pages 291-305. http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html

[2] Martin Kay (1995) Machine Translation: The Disappointing Past and Present. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology. http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node4.html#SECTION82*

[3] John Hutchins & Harold Somers (1992). *An Introduction to Machine Translation.* Academic Press.

[4] Douglas W. Orad & Bonnie J. Dorr (1996). *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19. Institute for Advanced Computer Studies. University of Maryland http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

[5] Tomek Strzalkowski (1992). *Information Retrieval Using Robust Natural Language Processing*. ACL-Proceedings 30th'92, 1992.

[6] Mark Davis & Ted Dunning (1995). A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval. In D. K. Harman, editor, *The Fourth Text Retrieval Conference (TREC-4)*. NIST. November 1995. http://crl.nmsu.edu/ANG/MWD/Book2/trec4.ps

[7] Liddy, Elizabeth D., Woojin Paik, Edmund S. Yu, and Mary McKenna (1994) *Document Retrieval Using Linguistic Knowledge*. Proceedings of RIAO '94. New York, 1994.

Note: The Uniform Resource Locators (URL) included in the references were correct at the time of writing the present paper.