CINDOR Conceptual Interlingua Document Retrieval: TREC-8 Evaluation.

Miguel Ruiz, Anne Diekema, Páraic Sheridan

MNIS-TextWise Labs
Dey Centennial Plaza
401 South Salina Street
Syracuse, NY 13202

Abstract:
The TREC-8 evaluation of the CINDOR system was based on English and French data from the cross-language retrieval track. Our objective was to continue our investigation of our conceptual interlingua approach to cross-language retrieval, specifically by measuring the contribution of conceptual retrieval over and above a baseline cross-language retrieval approach based on machine translation of queries. In both of the cross-language runs that were submitted for evaluation, corresponding to English-French and French-English retrieval, performance was measured at 75% of the equivalent monolingual searches. We noted however that absolute average precision values achieved were somewhat lower than many other systems in the cross-language track. Our hypothesis, that the underlying retrieval engine used in CINDOR was employing a simple retrieval function that was impacting performance, was confirmed through experiments with the SMART system configured with several different retrieval settings. Taken together, our TREC-8 experiments point to the value of our conceptual interlingua approach to retrieval, but indicate that our retrieval algorithm must be brought up to date so that valid comparisons may be made to other approaches used in other cross-language systems.

1. Introduction

The CINDOR project at MNIS-TextWise Labs is pursuing a 'conceptual interlingua' approach to cross-language information retrieval, based on a conceptual lexical resource modeled around WordNet [Miller 1990]. WordNet synonym groups, 'synsets', are taken to represent concepts which we assume are essentially language neutral. We have constructed the conceptual interlingua resource around the WordNet hierarchy by linking equivalent synonymous terms in several languages into the synsets representing a given concept. To date we have extended the conceptual interlingua to French, Spanish and Japanese, achieving approximately 20% coverage of WordNet synsets in each language. Our goal however is not to achieve complete coverage of WordNet content in each language. One of the objectives of our evaluation efforts has therefore been to investigate the extent to which conceptual interlingua coverage translates to vocabulary coverage in a typical document collection. For example, 55% of term occurrences in the TREC French collection match into our conceptual interlingua, which has 18% synset coverage in French.

Much of our research in the CINDOR project has been directed toward understanding the performance of a conceptual resource such as ours for cross-language retrieval. In particular, our participation in the TREC-8 evaluation was directed at verifying performance improvements in the current version of the CINDOR system over that which was used in our TREC-7 experiments, given significant re-development of the system over the period in between. The CINDOR system evaluated here sits on top of an Oracle database and interacts with the ConText text management system available with the Oracle relational database management system. While Oracle provides the data management capabilities, ConText supports text indexing and retrieval over Oracle data. On top of this, CINDOR provides text processing to extract indexing terms and map them into our conceptual interlingua vocabulary for indexing. An overview of this architecture is provided in Section 2.

For TREC-8 we submitted two official runs, restricting our attention to English and French – that subset of the cross-language track data which overlaps with our current research focus. Independently of TREC, we are conducting benchmark evaluation experiments of CINDOR in Spanish, using previous TREC test collections, and Japanese, using the new NACSIS test collection [Kando 1999]. A new feature of the CINDOR system used in TREC-8 experiments is the inclusion of the Systran machine translation system to provide automated translations of input queries. The machine translation output serves as an additional source of evidence for target-language query terms and is used to complement the conceptual translation provided through the conceptual interlingua. As part of our TREC-8 experiments, we have therefore had to opportunity to investigate and identify the usefulness of conceptual interlingua translations over and above those provided by Systran. A report and analysis of our TREC-8 experiments and performance across French and English is presented in Section 3.

The initial review of our results, together with some investigative experiments using TREC-7 cross-language track data, suggests that CINDOR retrieval performance is being negatively impacted by reliance on the standard Oracle ConText '*tf**idf' retrieval weighting algorithm. This is consistent with the well-established observation over past TREC evaluation experiments in general; that the weighting scheme is a crucial component in overall system performance. We therefore establish, in Section 4, the extent to which our runs using CINDOR with ConText may be impacted by this retrieval problem in order to suggest the extent to which performance can be further improved.

2. The CINDOR System.

The CINDOR (Conceptual Interlingua Document Retrieval) system is cross-language text retrieval system capable of accepting a user's query stated in their native language and then seamlessly searching, retrieving, relevance ranking and displaying documents written in a variety of foreign languages. CINDOR allows users of the system to state queries in any of the supported languages (currently English, French, Spanish, and Japanese) and search and retrieve documents from any of the supported languages.

The CINDOR system adopts a unique approach to cross-language information management based on a language-independent conceptual representation known as a 'Conceptual Interlingua'. This facilitates direct mapping between the interlingual representations of documents and user queries in multiple languages, a substantial advantage over systems, which rely on pairwise translations between languages. The CINDOR approach also ensures that documents and queries are matched at the underlying concept level, rather than relying on exact word matches. Queries are specified as natural language expressions rather than as keyword lists, as are commonly used for example in Internet search engines. This conceptual matching of natural language queries is designed to enhance retrieval effectiveness over keyword-based systems, which rely on exact, matching of words or word stems.

Conceptual Interlingua

We use the term conceptual interlingua to refer to a knowledge base of language-independent concept representations. Our current conceptual interlingua is a hierarchically organized concept lexicon in which concepts are related through various lexical relations. Concepts in the hierarchy are considered to be essentially language neutral and are then linked to their relevant terminological instantiations in various languages, currently English, French, Spanish and Japanese. Our Conceptual Interlingua therefore consists of two separate resources, which we refer to as the conceptual resource and the (multilingual) terminological resources.

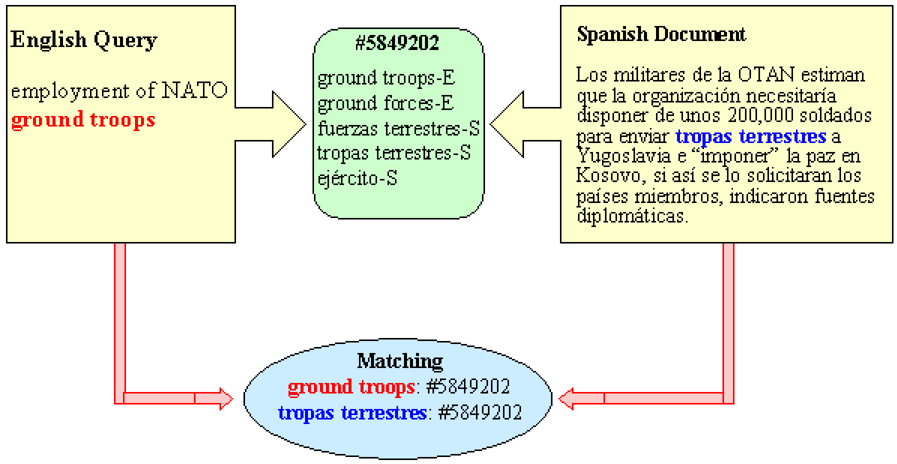
The CINDOR conceptual interlingua is built around WordNet [Miller 1990], a lexical resource, which contains approximately 165,000 different wordforms, organized into some 70,100 different concepts denoted by a group of synonyms, or *'synsets'*. Starting from the English Princeton WordNet, a large portion of the synsets has been translated into French, Spanish and Japanese. The conceptual interlingua consists of synset numbers; i.e. for document indexing words are "translated" by their synset IDs. Our Interlingua is therefore set up so that equivalent words in English, French, Spanish and Japanese are indexed by identical synset IDs. We "cross the language barrier" by mapping everything to synset IDs.

Following the distinction between 'conceptual' and 'terminological' resources outlined above, the 'conceptual' resource of our conceptual interlingua consists of the WordNet hierarchy of synset labels. Each synset label (concept) is linked then to a set of words or phrases which instantiate that concept in each of the languages supported – the 'terminological' resource. For example, the concept of *"elasticity:* the tendency of a body to return to its original shape after it has been stretched or compressed", which has the label 131186, is instantiated in English and French as follows:

131186 spring, give, springiness
131186 élasticité, flexibilité, moëlleux

We consider the label *131186* to represent the language independent concept of *elasticity* so this number is part of our conceptual hierarchy. The terminology related to this concept in each language is then linked to this concept label from each of our language resources. In terms of CINDOR document processing, this means that any document or query term, which is identified as an instantiation of the concept of 'elasticity', is indexed to the concept label *131186*. Whether the term occurs in an English, French, Spanish or Japanese document or query, the label will be the same and retrieval will be enabled, as illustrated in the Figure below using the term, *"ground troops":*

The architecture of the CINDOR system in its current form involves the use of the Oracle relational database management system (v8.0.5) with the ConText option for storage of source documents and management of inverted index tables. The conceptual interlingua resource is also transformed to a flat table and stored in an Oracle database. The Oracle system was chosen for its scalability and robustness, allowing CINDOR to be deployed over very large document collections and allowing for the full range of database management functions to be applied over stored text.



Example Cross-Language matching through Conceptual Interlingua

Integration of CINDOR functionality for cross-language retrieval is achieved in places where ConText provides for various "filters" to be applied to document content before being indexed and stored in inverted access tables. Although ConText filters were originally intended for simple format transformation, CINDOR subjects document content to full morphological analysis, part-of-speech tagging, and conceptual analysis against the conceptual interlingua. Word stems plus conceptual codes are returned to ConText from the CINDOR language analysis module and are then indexed into Oracle tables.

At retrieval time, natural language queries are analyzed in the same way as documents and are then transformed into SQL statements and submitted to ConText for evaluation. Through experimentation over time, we have found that the optimum query format consists of a number of query segments, each evaluated in turn and then combined into a final ranked result. To the extent that term and collection frequencies are computed and stored internally by ConText, the retrieval ranking algorithm is outside the control of the CINDOR system. We have augmented ConText retrieval through a standard document length normalization adjustemt, but we have suspected for some time that internally ConText was using a rather simple weighting mechanism that could be substantially improved upon. This is a topic to which we have given some attention in our TREC-8 experiments.

3. TREC-8 Experiments.

TREC-7 represented the first evaluation of the CINDOR system, development of which was completed with little time to spare before submission of official runs [Diekema *et al* 1999]. The intervening year has seen a substantial re-development effort of many components of the CINDOR system, particularly with a view to addressing shortcomings identified in TREC-7 experiments. An important component of our objectives in TREC-8 therefore involves establishing the extent to which this re-development effort has lead to improvements in system performance.

A new component of CINDOR processing which has been introduced over the past year is a machine translation system, which is used both for translating queries into the language of documents, and also translating foreign language documents back to the language of the user on demand. The use of a machine translation system at query time is intended to contribute a further source of target language terms for queries and complement the conceptual mapping provided through the conceptual interlingua, especially in cases where query terms are not present in the interlingua resource.

Although the cross-language track again set as the main task the retrieval of documents from a multi-lingual set of English, French, German and Italian documents, we focused on the sub-task which involved the language pair of English and French since these are the two of the four which are covered by our resources. Two official runs were submitted to NIST for evaluation; English queries against French documents (TW8E2F) and French queries against English documents (TW8F2E). These official runs are complemented here by a series of other unofficial runs which were undertaken to allow us to examine a range of evaluation questions which were of interest.

A primary question of interest in evaluating the CINDOR system relates to the contribution of our conceptual interlingua approach to retrieval. Although designed primarily to facilitate cross-language retrieval, we anticipate that the benefits of synonym expansion may be observed also in monolingual retrieval settings. We have therefore completed experiments in which the use of conceptual interlingua indexing was de-activated for retrieval and compared performance to that of the standard CINDOR system with the conceptual interlingua enabled. A comparison between results for monolingual French retrieval is included in Figure 1, while English-French cross-language retrieval is illustrated in Figure 2.

The difference between the baseline system and the CINDOR system in Figures 1 and 2 is that the baseline system completes morphological and part-of-speech analysis but does no further processing, while CINDOR takes the further step of assigning conceptual codes to index terms. Further, these experiments included the current CINDOR proper name recognition module, which is still under development. This module attempts to recognize and tag proper names such as people, places, organizations etc. and to categorize them into appropriate classes. The advantage of this module in these experiments is likely to come from the ability to recognize multi-word proper names and to treat them as a single unit. The baseline system for cross-language retrieval consists of the baseline monolingual system augmented by Systran machine translation of queries for matching English queries against French documents.

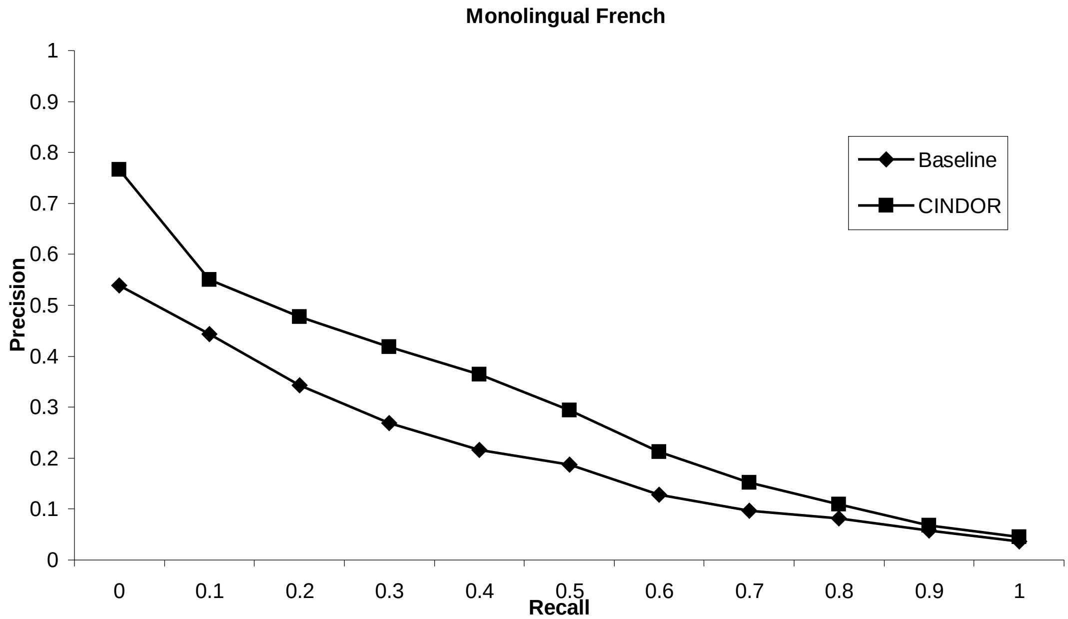


Figure 1: Conceptual Interlingua retrieval; French-French

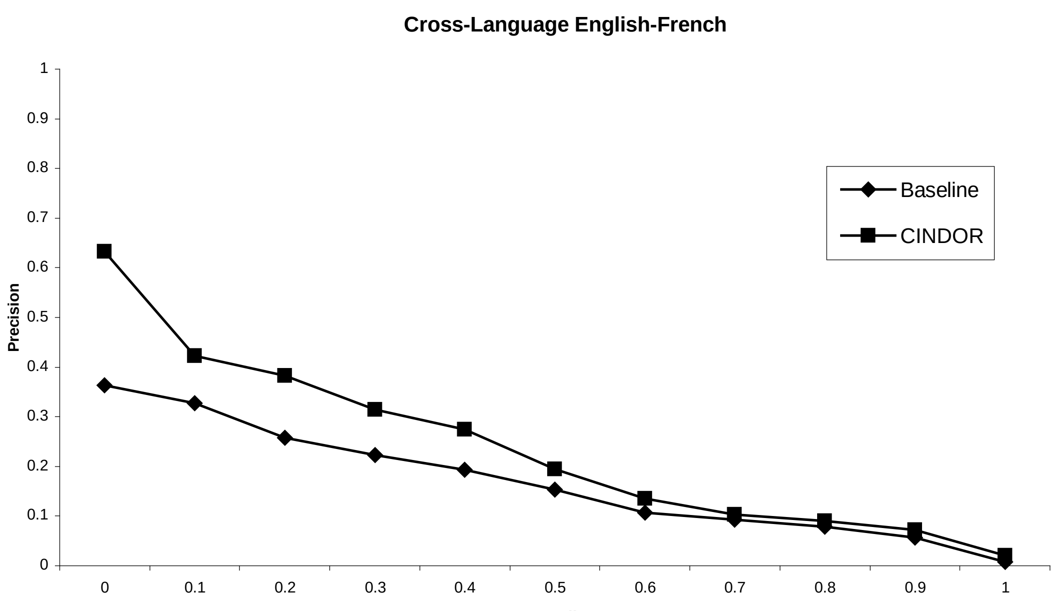


Figure 2: Conceptual Interlingua retrieval; English-French

In monolingual French retrieval, the baseline average precision is 0.2016, while CINDOR processing raises the value to 0.2921, a 45% increase. Similarly in our cross-language experiments, the baseline average precision is 0.1590, while CINDOR processing raises this to 0.2209, a 39% increase. This later result is particularly interesting because it represents the increase in cross-language retrieval performance to be achieved through the use of the conceptual interlingua above that achieved through straightforward translation of queries through automated machine translation. It is unclear from these experiments however, the extent to which performance gains above the baseline system may be attributed primarily to the conceptual matching enabled through the conceptual interlingua, or to the increased precision achieved through matching as single units the proper names which are frequent in TREC queries. In the ongoing evaluation of the CINDOR system, we will undertake in the near future a detailed analysis of the contribution of our proper name recognition module, which will enable us to determine exactly the contribution of each component.

One of the advantages of our conceptual interlingua approach to cross-language retrieval, at least in theory, is that by matching at the conceptual level, we can expect minimal loss in retrieval precision when matching across different languages compared to retrieval in a monolingual environment. The extent to which cross-language results mirror those of equivalent monolingual searches is easily facilitated using the TREC data, since topics are made available in each of the document languages. Our official submissions were intended to compare the performance of English-French retrieval compared to French-French monolingual, and French-English compared to monolingual English-English retrieval. The results of these runs are presented in Figures 3 and 4 below.

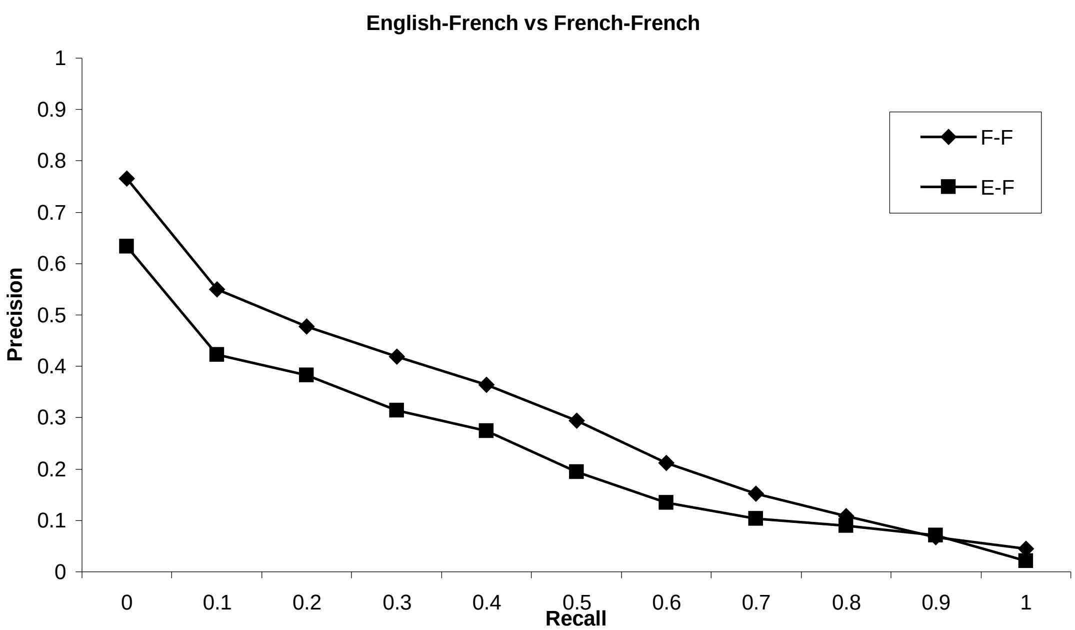


Figure 3: Cross-language versus Monolingual: French documents.

Comparing cross-language retrieval of French documents in response to English queries against the monolingual case where equivalent French queries are used, illustrated in Figure 3, indicates that our cross-language retrieval performance is at 75% of monolingual. The English-French run has an average precision of 0.2209, compared to an average precision of 0.2921 for the monolingual French-French run.

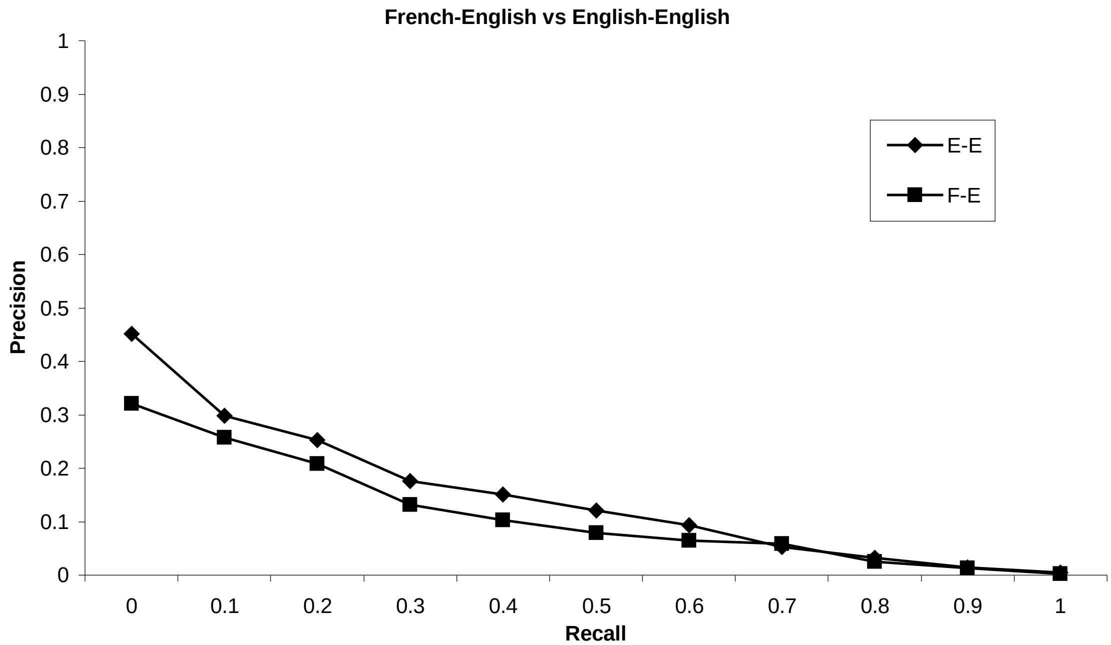


Figure 4: Cross-language versus Monolingual: English documents.

A similar comparison of English document retrieval in response to French queries versus the monolingual case where equivalent English queries are used, illustrated in Figure 4, indicates that this cross-language retrieval performance is also at 75% of monolingual performance. The French-English run has an average precision of 0.1010 versus an average precision of 0.1331 for the monolingual English-English run. These results show nice consistency, though not quite at the level of performance we would have hoped for.

4. Retrieval Performance.

Although the CINDOR system, as evaluated so far, has demonstrated performance improvements over a baseline system without the conceptual interlingua and proper name recognition modules, and has performed consistently at 75% of monolingual precision in both directions in the English and French language pair, the system consistently under performs against comparable systems in terms of average precision. This is obvious from examination of both TREC-7 and TREC-8 results, as well as other experiments we have undertaken, including experiments in other language combinations.

Since the low average precision scores have been in evidence in all experiments we have conducted, using many different configurations both in terms of system set-up and test environment (document collection, queries, language pairs, etc), it seems likely that the problem is inherent in the retrieval algorithm being used and is not due to mis-translations of terms from particular queries. Unfortunately, given the current architecture of the system, we are reliant upon the retrieval mechanism of the Oracle ConText system and are unable to determine exactly the parameters of the retrieval function being used.

We have however undertaken a standard benchmark comparison against the CINDOR system and the SMART retrieval system, which has been used in extensive experimentation on term weighting algorithms [Salton & Buckley 1997]. The objective of this experiment is to test our hypothesis that the ranking algorithm is impacting our performance. A particularly useful feature of the SMART system in this case is its ability to use a range of different ranking functions which can be specified quite simply using a standard notation. Our initial hypothesis is that the ConText system provides a straightforward *tf*idf* ranking of documents, denoted as *ntn.ntn* in SMART notation. The final CINDOR ranking, which applies pivoted length normalization in order to re-rank the initial results, therefore equates to *ntu.ntn* in SMART notation.

We tested our hypothesis using the collection of 243,000 Associated Press documents used in the TREC cross-language track and the 28 TREC-7 cross-language track queries in English, for which we already had results using CINDOR. This is a simple monolingual English experiment for the purpose only of establishing the performance of our hypothesized CINDOR retrieval function versus one which has been shown to perform well over TREC data in the SMART system (*Lnu.ltu*). The results of our experiment are presented in Table 1 below.

These results confirm our hypothesis, that CINDOR's final ranking is equivalent to a simple *ntu.ntu* function, while the SMART *Lnu.ltu* formula (Singhal *et al* 1996) provides substantially better performance – 80% better in this case. This is in fact only a subset of a range of detailed ranking experiments we have conducted, which confirm the importance of the retrieval weighting function in our overall results, cross-language or otherwise.

		Average Precision
CINDOR		0.2515
SMART		
	ntu.ntu	0.2426
	Lnu.ltu	0.4531

Table 1: CINDOR ranking versus SMART variants.
(*Associated Press collection – TREC-7 CLIR English queries)*

This conclusion is of course not news, especially in the context of the *eighth* text retrieval conference, but this straightforward investigation has served to succinctly pinpoint the problem which has resulted in the CINDOR system comparing poorly to published results in a range of experiments in various language combinations. More importantly, it points immediately to the solution of this problem and indicates the direction in which our work should proceed. It is critical that we put the CINDOR system on equal footing with other systems in terms of the retrieval function used so that we can then more clearly establish the advantages and disadvantages of our conceptual interlingua approach to cross-language retrieval.

5. Conclusion.

Our TREC-8 experiments reported here are part of a wider and ongoing series of evaluation experiments designed to establish the performance of the CINDOR retrieval system over a range of language combinations and text types, and more broadly to evaluate the usefulness of our conceptual interlingua approach to cross-language information retrieval.

The results presented here for English and French suggest that there are benefits to be had from the use of our conceptual interlingua resource. In comparing the CINDOR system against a simple baseline for monolingual retrieval, and against that baseline system using Systran machine translation of queries for cross-language English-French retrieval, CINDOR provided 40-45% gains in average precision over the baseline system. We have also established a consistent level of cross-language performance using the CINDOR system, when compared to equivalent searches in a monolingual environment using same-language queries and documents. In both English-French and French-English, average precision in cross-language searches was measured at 75% of the level achieved in equivalent monolingual experiments.

We have noted however, that although these comparative results between different experiments with the CINDOR system are informative, the low absolute level of precision achieved using CINDOR across a range of experiments is an impediment to useful comparisons between our conceptual interlingua approach to cross-language retrieval and other approaches which have been tried and evaluated in the TREC cross-language track and elsewhere. A straightforward investigation using the SMART retrieval system was enough to verify our hypothesis that the low level of performance was attributable to an overly simplistic retrieval function, and that replacement of this algorithm with a state-of-the-art weighting scheme could deliver on the order of 80% improvement in average precision. Addressing the retrieval weighting function problem is therefore an important component of our future work.

References:

[Diekema *et al* 1999]

Diekema A, Oroumchian F, Sheridan P, and Liddy E, "TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French", *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, NIST Special Publication, 1999.

[Miller 1990]

Miller G., "WordNet: An On-line Lexical Database", *International Journal of Lexicography*, Vol. 2, No. 4, Special Issue, 1990.

[Salton & Buckley 1997]

Salton G, and Buckley C, "Term-weighting Approaches in Automatic Text Retrieval", In *Readings in Information Retrieval*, K. Spark-Jones & P. Willet (Eds), Morgan Kaufmann, San Francisco, CA, 1997, pages 323-238.

[Singhal *et al* 1996]

Singhal A, Buckley C, and Mitra M, "Pivoted Document Length Normalization", In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, August 1996, ACM Press, pages 21-290.