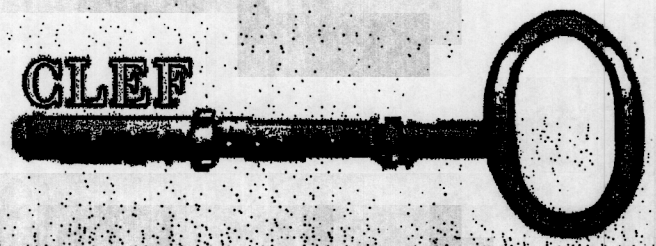
Cross Language Evaluation Forum



Results of the CLEF 2005 Cross-Language System Evaluation Campaign

EXTENDED ABSTRACTS

CLEF 2005 Workshop, 21-23 September, Vienna, Austria

CONTENTS

Bilingual and Multilingual Experiments with the IR-n System

Elisa Noguera[1], Fernando Llopis[1], Rafael Muñoz[1],
Rafael M. Terol[1], Miguel A. García-Cumbreras[2], Fernando Martínez-Santiago[2],
and Arturo Montejo-Raez[2]

[1] Grupo de investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{elisa, llopis, rafael, rafaelmt}@dlsi.ua.es
[2] Department of Computer Science. University of Jaen, Jaen, Spain
{magc, dofer, montejo}@ujaen.es

Abstract. Our paper describes the participation of the IR-n system at CLEF-2005. This year, we participated in the bilingual task (English-French and English-Portuguese) and the multilingual task (English, French, Italian, German, Dutch, Finish and Swedish). We introduced the method of combined passages for the bilingual task. Futhermore we have applied the method of logic forms in the same task. For the multilingual task we had a joint participation with the University of Alicante and University of Jaén. We want to emphasize the good score achieved in the bilingual task improving around 45% in terms of average precision.

1 Introduction

Information Retrieval (IR) systems [2] try to find the relevant documents given a user query from a document collection. We can find different types of IR systems in the literature. On the one hand, if the document collection and the user query are written in the same language then the IR system can be defined as a monolingual IR system. On the other hand, if the document collection and the user query are written in different languages then the IR system can be defined as a bilingual (two different languages) or multilingual (more than two languages) IR system. Obviously, the document collection for multilingual systems is written in at least two different languages. The IR-n system [3] can work with collections and queries in any language.

Passage Retrieval (PR) systems [1] are information retrieval systems that determine the similarity of a document with regard to a user query according to the similarity of fragments of the document (passages) with regard to the same query.

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 79–82, 2006.

2 Bilingual Task

2.1 Method 1: Machine Translation

We use different translators in order to obtain an automatic translation of queries. Three of them were used for all languages: FreeTranslation, Babel Fish and InterTran. Moreover, we have used one more method merging all translations. This is performed by merging several translations from the on-line translators. This strategy is based on the idea than the words which appear in multiple translations have more relevancy than those that only appear in one translation.

The method of combined passages was developed in the monolingual task [4], for this reason it has been also used in the bilingual task. In the training test, the best input configuration has been used for French and Portuguese. Best scores were achieved using the merge of translations in English-Portuguese and FreeTranslation in English-French.

2.2 Method 2: Logic Forms

The last release of our IR-n system introduced a set of features that are based on the application of logic forms to topics and in the incrementing of the weight of terms according to a set of syntactic rules. The reason for this is that IR-n system includes a new module that increments the weight of terms, applying a set of rules based on the representation of the topics in the way of logic forms [7].

3 Multilingual Task

This year we made a combination between the fusion algorithm 2-step RSV [6], developed by the University of Jaén, and the passage retrieval system IR-n, developed by the University of Alicante. A full detailed description of the experiments is available in this volume.

IR-n has been used as IR system in order to make some experiments in Multi-8 Two-years-on task. Thus, it has been applied over eight languages: English, Spanish, French, Italian, German, Dutch, Finnish and Swedish.

An in depth description of the training test is available in [6]. Firstly, each monolingual collection is preprocessed as usual (token extraction, stopwords are eliminated and stemming is applied to the rest of words). In addition, compound words are decompounded as possible for German, Swedish, Finnish and Dutch. We use the decompounding algorithm depicted in [5]. The preprocessed collections have been indexed using the passage retrieval system IR-n and the document retrieval system ZPrise. The IR-n system has been modified in order to return a list of the retrieved and relevant documents, the documents that contain the relevant passages. Finally, given a query and its translations into the other languages, each query is searched in the corresponding monolingual collection.

When the monolingual lists of relevant documents are returned, we apply the 2-step RSV fusion algorithm. This algorithm deals with the translations

whose terms are known (aligned terms) in a different way that those words whose translation is unknown (non-aligned words) by giving two scores for each document. The first one is calculated taking into account aligned words, and the second one only uses non-aligned terms. Thus, both scores are combined into a only RSV per document and query by using some formulae:

1. Combining the RSV value of the aligned words and not aligned words with the formula:

$$0.6 < RSVAlignedDoc > +0.4 < RSVNotAligned >$$

2. By using Logistic Regression. The formula:

$$e^{\alpha \cdot < RSVAlignedDoc > + \beta \cdot < RSVNotAligned >}$$

3. The last one also uses Logistic Regression but include a new component, the ranking of the doc. It applies the formula:

$$e^{\alpha \cdot < RSVAlignedDoc > + \beta \cdot < RSVNotAligned > + \gamma \cdot < RankingDoc >}$$

4 Results at CLEF-2005

The IR-n system used the best configuration obtained in the training process. Three different runs have been submitted for each task. The first run IRn-xx-vexp uses the method of combined passages with query expansion. The second run IRn-xx-fexp only uses query expansion. The third run IRn-xx-vnexp uses the method of combined passages without query expansion. Furthermore, a fourth run IRn-xx-fexpfl has been submitted for English-Portuguese task. It uses the method of logic forms. Table 1 shows the scores achieved for each run.

Table 1 shows the official results for "Multi-8 Two-years on task. IR-n performs better than ZPrise except for Finnish results, the differences of average precision between both multilingual experiments is very small. The reason is that the merging algorithm is independent of the initial selection of relevant documents. This feature has been briefly discussed above and in more detail in [6].

5 Conclusions and Future Work

In the bilingual task the IR-n system has obtained better results merging translations than using single translations. On the other hand, the method of combined passages improves the scores in the bilingual task compared to the method of fixed passages, as it happens in the monolingual task.

Thus, in the multilingual task we conclude that IR-n is a good information retrieval system for CLIR systems. It improves on document-based systems such as OKAPI-ZPrise in bilingual experiments. In addition, the integration of this system with complex merging algorithms such as 2-step RSV is straightforward.

Possibly, if an IR-n like system were implemented for the creation of a dynamic index the multilingual results would be improved in the same way that the monolingual results are.

Table 1. CLEF 2005 official results. Bilingual and Multilingual tasks.

BILINGUAL TASK			
Language	**Run**	**AvgP**	**Dif**
English - Portuguese	CLEF Average	21.71	
	IRn-enpt-vexp	29.18	+34.4%
	IRn-enpt-fexp	28.94	
	IRn-enpt-vnexp	25.22	
	IRn-enpt-fexpfl	27.27	
English - French	CLEF Average	24.76	
	IRn-fr-vexp	35.90	+45.3%
	IRn-fr-fexp	29.12	
	IRn-fr-vnexp	29.13	

MULTILINGUAL TASK			
IR system	**Formula 1**	**Formula 2**	**Formula 3**
ZPrise+OKAPI	28.78	29.01	29.12
IR-n	28.85	29.09	29.18

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencian Government under project numbers GV04B-276 and GV04B-268.

References

1. M. Kaskziel and J. Zobel: Passage Retrieval Revisited. In *Proceedings of the 20th annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97)*, Philadelphia, pages 178–185, 1997.
2. F. W. Lancaster: *Information Retrieval Systems: Characteristics, Testing and Eval.*, John Wiley and Sons, New York, 1979.
3. F. Llopis: *IR-n: Un Sistema de Recuperación de Información Basado en pasajes*. PhD thesis. University of Alicante, 2003.
4. F. Llopis and E. Noguera: Combining Passages in Monolingual Experiments with IR-n System. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, Vienna, In this volume.
5. F. Martínez-Santiago, M. García-Cumbreras and L. A. Ureña: SINAI at CLEF 2004: Using Machine Translation Resources with Mixed 2-Step RSV Merging Algorithm. *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum*, Bath, page 156-164, 2005.
6. F. Martínez-Santiago, L. A. Ureña, and M. Martín: A Merging Strategy Proposal: Two Step Retrieval Status Value Method. *Information Retrieval*, 9(1):95–109, 2006.
7. R. M. Terol: The University of Alicante at CL-SR Track. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, Vienna, In this volume.