Bilingual and Multilingual experiments with IR-n system

Elisa Noguera, Fernando Llopis, Rafael Muñoz and Rafael M. Terol
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

Miguel A. García-Cumbreras, Fernando Martínez-Santiago and Arturo Montejo-Raez
Department of Computer Science. University of Jaen, Jaen, Spain
magc,dofer,montejo@ujaen.es

Abstract

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

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Experimentation, Measurement, Performance

Keywords

Information Retrieval

1 Introduction

Information Retrieval (IR) systems [2] have to find the relevant documents to an user query from a document collection. We can find different kinds of IR systems at 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 like 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 like a bilingual (two different languages) or multilingual (more than two languages) IR system. Obviously, the document collection for multilingual systems is written in two different languages at least. IR-n system [3] is a monolingual, bilingual and multilingual IR system based on passages.

Language	Collections	TotalDocs	Size	SDAvg	WDAvg	WSAvg
French	Le Monde 94/95 SDA French 94/95	177452	487 MB	17	388	21
Portuguese	Público 94/95 Folha 94/95	210734	564 MB	18	433	23

Table 1: Data Collections

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.

This paper is organized as follows: next section describes bilingual task and training. Following, we describe the multilingual task. And finally, we present the archieved results and the conclusions.

2 Bilingual task

2.1 Description method

The participation of the system IR-n in the bilingual task this year has been focused on the following languages pairs:

- English-French
- English-Portuguese

For every language, the stemmers and the stopword lists used were provided by the clef organization (http://www.unine.ch/info/clef). Table 1 shows the characteristics collection which we have worked.

- SDAvg is the average of sentences in each document.
- WDAvg is the average of words in each document.
- WSAvg is the average of words in each sentence.

This year we have used two methods for bilingual task:

2.1.1 Method 1: Machine translation

We use different translator in order to obtain an automatic translation of queries. Three translators were used for all languages: FreeTranslation, Babel Fish and InterTran. It has been to carry out several test with CLEF-2004 collections for French and Portuguese.

Moreover, we have used a one more method merging all translations. It has been performed merging several translation built by an on-line translator. This strategy is based on the idea that the words which appears in different translations have more relevancy which those that only appear in one translation.

2.1.2 Method 2: Logic Forms

The last release of our IR-n system introduces a set of features that are based in the application of logic forms to topics and in the increment of the terms weight of the topics according to a set of syntactic rules. This reason produces that IR-n system includes a new module that increments the terms weights of the topics applying a set of rules based on the representation of the topics in the way of logic forms [7].

Task	Online tranlator	AvgP
English - Portuguese	Babelfish	38.68
	FreeTranslation	41.79
	InterTran	37.83
	Merging	42.18
English - French	Babelfish	36.06
	FreeTranslation	41.31
	InterTran	32.59
	Merging	40.88

Table 2: CLEF 2004. Training bilingual task

This process consist in that each one of the terms of the topic in the logic form can modify its weight term according to the type of assert of the term in the logic form and the relationships between these asserts of the topic in the logic form. The logic form of a topic (or sentence) is calculated through the analysis of dependency relationships between the words of the sentence.

2.2 Experimentation

This section describes the training process which has carried out this year in order to obtain optimum features to improve the performance of the system. The following subsections explain the specific experiments which we have carried out.

2.2.1 Method 1: Machine translation

This year, it has carried out several test with the objetive to establish the translator which obtain the best results for each task. In monolingual task was developed the combined passages method [4], for this reason it has also used in bilingual task. In the training test, it has been used the best input configuration for French and Portuguese.

Table 2 shows the scores achieved for each language in the CLEF-2004 collections. Best scores were achieved using the merge of translations in English-Portuguese and FreeTranslation in English-French.

2.2.2 Method 2: Logic Forms

Several tests were performed applying this method based on logic forms to the release of the document collection of the year 2004. These tests consisted in to increment the weights of several terms according to the rules defined in this method [7]. The increment of the terms weights were about 15% of these original scores.

3 Multilingual task: Mixed 2-step RSV merging algorithm and IR-n, a passage retrieval

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

IR-n has been used as Information Retrieval 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. In this way we have evaluated the performance of IR-n in several new ways:

- Some of the eight languages are unheard-of. Thus, we have evaluated IR-n with languages such as Swedish by first time.
- This is a very good opportunity to compare a passage IR system such as IR-n with IR systems based on document retrieved, such as ZPrise system with OKAPI weighting function.
- Finally, we are interested in the evaluation of IR-n in combination with 2-step RSV merging algorithm.

3.1 Experimentation

This section depicts the experiments briefly. A description in deep 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 to the German, Swedish, Finnish and Dutch languages. 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 the terms whose translation is known (aligned terms) in a different way that those words whose translation is unknown[1](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: $0.6* < RSV_aligned_doc > +0.4* < RSV_notaligned >$
2. By using Logistic Regression. The formula: $e^(alpha* < RSV_aligned_doc > +beta* < RSV_notaligned >)$
3. The last one also uses Logistic Regression but include a new component, the ranking of the doc. It applies the formula: $e^(alpha* < RSV_aligned_doc > +beta* < RSV_notaligned > +gamma* < ranking_doc >)$

Twenty first queries has been used as training and the other forty has been used for evaluation.

The table 3 shows the bilingual result obtained by using IR-n and ZPrise-OKAPI. The experimental method (preprocessing of the collections and translation of the queries) is exactly the same for IR-n a ZPrise. The only difference is just the IR software. The evaluation has been realized by using CLEF 160-200 queries.

4 Results at CLEF-2005

IR-n system used in order to participate in CLEF'2005 the best IR-n configuration obtained in the training process.

Three different runs have been submitted for each task. The first run IRn-xx-vexp uses combined passages method and query expansion. The second run IRn-xx-fexp only uses query expansion. The third run IRn-xx-vnexp uses combined passages method and it do not use query expansion. Furthermore, a fourth run IRn-xx-fexpfl has been submitted for English-Portuguese task, it uses the logic forms method.

Table 4 shows the scores achieved for each run. IR-n system has obtained better results than the average scores of CLEF 2005 for English-French and English-Portuguese.

[1]Note that unknowing the translation of a word is a different thing that an untranslated term. By example,

Language	ZPrise+OKAPI	IR-n
Dutch	30.94	34.03
English	52.06	54.96
Finnish	34.11	33.47
French	42.14	42.84
German	33.01	33.99
Italian	33.38	34.82
Spanish	37.35	39.68
Swedish	23.29	25.23

Table 3: Bilingual results (except English which is a monolingual experiment).

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	

Table 4: CLEF 2005 official results. Bilingual tasks

	Merging approach		
IR system	formula 1	formula 2	formula 3
ZPrise+OKAPI	28.78	29.01	29.12
IR-n	28.85	29.09	29.18

Table 5: Multilingual results by using three variants of mixed 2-step RSV.

Table 5 shows the official results for "Multi-8 Two-years on task. In spite of IR-n overcomes to ZPrise except of Finnish results (see bilingual results, table 3), the differences of average precision between both multilingual experiments is poor. The reason is that the merging algorithm is very independent of the initial selection of relevant documents. This feature is briefly discussed above and more profusely in [6].

5 Conclusions and Future Work

In bilingual task IR-n system has obtained better results merging translations than others translations. On the other hand, the combined passages method allows to improve the scores in the bilingual task on the fixed passages method. Like it happens in monolingual task.

Thus, we conclude that IR-n is a good information retrieval system for CLIR systems. It overcomes to 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. On the other hand, the improvement of IR-n respect of OKAPI-ZPrise is not fully exploited by 2-step RSV merging algorithm since this algorithm creates a dynamic index based on classic document retrieval models (more precisely the dynamic index created by 2-step RSV uses an OKAPI weighting schema). Possibly, if an IR-nlike system were implemented for the creation of such dynamic index the multilingual results would be improved in the same way that the monolingual results are.

6 Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia 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 International ACM Philadelphia SIGIR*, 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] Llopis F., Noguera E. Combining passages in monolingual experiments with ir-n system. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, In this volume, Vienna, Austria, 2005.

[5] F. Martínez-Santiago, Miguel García-Cumbreras, and L.A. Ureña. SINAI at CLEF 2004: Using Machine Translation Resources with Mixed 2-Step RSV Merging Algorithm. *Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer Verlag. In press.*, 2005.

[6] F. Martínez-Santiago, L.A. Ureña, and M. Martín. A merging strategy proposal: two step retrieval status value method. *Information Retrieval. In press*, 2005.

[7] Rafael M. Terol. The university of alicante at cl-sr track. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, In this volume, Vienna, Austria, 2005.

Machine Translation translates the whole of the phrase better than word by word. Thus, we don't know which word is translated for each word. An alignment algorithm at word level is required.