Combining passages in monolingual task with IR-n system

Fernando Llopis and Elisa Noguera
Grupo de investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
llopis,elisa@dlsi.ua.es

Abstract

This paper describes our participation in monolingual tasks at CLEF-2005. In this research we have worked in the following languages: English, French, Portuguese, Bulgarian and Hungarian. Our task has been focused on using combined different size passages to improve the Information Retrieval process. Once we have studied the experiments which have been carried out and the official results at CLEF, we have realized that this combining model gets better the achieved scores considerably.

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 systems based on passages (PR) [2] determine the relevance of a document regarding to a question. This relevance is obtained from the similarity of different fragments in this document regarding to the same question. This models not only let to improve the location of relevant documents, but also let us to find the most relevant part of the document accurately. This last advantage allows us that these systems which are used in other tasks as Question Answering (QA).

PR systems are classified according to how the passages are determined in each document. IR-n system is a PR system which defines the passages based on a fixed number of sentences. This provides the passages with some syntactical content. Last years our researches with IR-n system are based on detecting the suitable size for each collection (to experiment with test colletion), but determining the similarity of a document based on the passage with more similarity. This year the score which is given to each document is based on the similarity of several size passages.

This paper is organized as follows: next section describes IR-n system and its new changes. Following, we describe the task developed at CLEF 2005 by our system and the training. And finally, we present the archieved results and the conclusions.

2 IR-n system

IR-n system [3] was developed in 2001. It was written in C++ entirely, although it has been used external tools (stemmers) in few occasions. Last year [4] the system was designed and developed again with the aim of getting more information (size of passages in words) and to improve the process speed.

The system works in Linux system without excesive requeriments, nevertheless, in view of the fact that the search process is carried out with structures load in memory, it is convenient that the computer has the enough memory.

In this section is presented the main charateristics of IR-n system and it is detailed the combined passages method used in this edition.

2.1 Resources: stemmers and stopword lists

This was the first year that we have worked with Bulgarian and Hungarian languages. It has been used the stemmers and stopwords lists available on the web http://www.unine.ch/info/clef. We can highlight that Hungarian and Bulgarian collections are encoded in UTF-8. In addition Bulgarian stemmer is developed in perl to support UTF-8. The rest of the stemmers are developed in C.

2.2 Similarity measures

IR-n system is ready for using several similarity measures: cosine [5], pivoted cosine [7] and okapi [6]. To the last one the values of parameters (k1,b,avg) could be update in a easy way, in order to get the best results.

On the whole, experiments carried out by okapy measure show us that we could obtain the best results. Furthermore, we have contrasted normalization concept with size passage again. Previous versions of IR-n system does not use size passage in the similarity measures because all the passages had the same size. Last edition we could check that results improved if it was considered size passages.

2.3 Query expansion

Most of IR systems use query expansion techniques [1] based on adding the most frequent terms contain in the more relevant documents to the original query. Architecture IR-n allows us to use query expansion based on more relevant passages or documents. In fact, last edition we got better results using the more relevant passages.

2.4 Combined passages

The present year, technique called 'combined passages' has been developed. The model consists of applying similar techniques for merging relevant document lists in multilingual task but using relevant passage lists of different size.

This model consists of using different size passages in order to get relevant document lists. The list which have been obtained are combined sequently. Table 1 shows different methods used to obtain the ranking of scores.

We have used four methods: MAX merges the n list and if a document is in several lists it will provide the highest score. SUM carries out the average of the scores. The methods 3 and 4 are as the previous ones but using normalization. This normalization is carried out subtracting the score of each document RSV_k from the minimum score of the list and dividing by $max(RSV_K) - min(RSV_k)$.

Obviously, this model improves and involves speed process. However in IR-n system architecture, this trouble does not increase the speed process of the system. This happens because

Number	Method	Formula
1	MAX	$max(RSV_k)$
2	SUM	$sum(RSV_k)$
3	MAX RSVnorm	$max((RSV_k - min(RSV_k))/(max(RSV_k) - min(RSV_k))$
4	SUM RSVnorm	$sum((RSV_k - min(RSV_k))/(max(RSV_k) - min(RSV_k))$

Table 1: Data fusion method

Language	Collections	TotalDocs	Size	SDAvg	WDAvg	WSAvg
English	The Angeles Times 94 Glasgow Herald 95	169477	579 MB	25	529	20
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
Hungarian	Magyar Hirlap 02	49530	105 MB	11	245	20
Bulgarian	Standart 02 - Sega 02	69195	213 MB	246	157	18

Table 2: Data Collections

IR-n system produces a segmentacion of the documents in passages in the search time and the calculation of similarity is carried out on structures load in memory.

3 Training

This section describes the training process which has been carried out in order to obtain the best features to improve the performance of the system. Firstly, the collections and resources are described. The following section explains the specific experiments which we have carried out.

3.1 Data Collections

This year our system has participated in the following tasks: Monolingual: English, French, Portuguese, Bulgarian y Hungarian. Table 2 shows the characteristics of the 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.

3.2 Experiments

This year several tests have been carried out in order to establish the best similarity measure for each language and to provide the value of input parameters in the system. We have evaluated the following languages: English, French and Portuguese. We could not evaluated Bulgarian and Hungarian languages because we did not have data of last years, as result of this, we have choosed the similarity measures and the parameters for these languages comparing them with the rest ones.

The aim of the experiments phase is set up the optimum value of the input parameters for each collection. For training has been used the collections CLEF-2003 (English and French) and CLEF-2004 (Portuguese). Query expansion techniques have also been used in all languages.

Language	size P	k1	b	avg	avgP
English	8	2	0.5	300	0.5083
French	9	1.2	0.3	300	0.5240
Portuguese (2004)	8	2	0.5	500	0.4741

Table 3: Training results in fixed passages system

Language	size P	k1	b	avg	exp	expd	expq	avgP
English	8	1	0.5	500	4	5	10	0.5267
French	8	2	0.7	400	3	5	10	0.5488
Portuguese (2004)	10	2	0.5	500	3	10	10	0.5084

Table 4: Training results in fixed passages system with query expansion

3.2.1 Fixed size passages

It has been performed experiments for setting the size passages and the values of parameters in okapi system which allow us to obtain the best results. As the table shows 3 the size passage is the same for all languages (8 sentences), however in French is 9 sentences.

3.2.2 Fixed size passages with query expansion

On the one hand, experiments which are carried out with query expansion tried to fix the number terms to add in the original query and the number of documents (passages) to take into account. Furthermore, we have evaluated the use of different size passages. We have got the best results with 10 terms in every test and it has been used the 5 or 10 passages more relevants depending on the specific language. On the other hand, it can be appreciated that the size passage is 8 in this case also.

As we check in the table 5 query expansion in fixed system allows us to improve scores between 3.6% and 7.2% according to the different languages.

3.2.3 Combined passages

Combined passages method consist on using the similarity values which are provided from different size passages of the same document to obtain the document similarity. Because of that, it has been defined three types of passages: small, medium and big passage. The number of sentences which composes each passage is the following:

$P1 = (3, 4, 5, 6)$

$P2 = (7, 8, 9, 10)$

$P3 = (11, 12, 13, 14)$

Experiments have been carried out by means of using one of each type. In this way is obtained the similarity of each passage. Document similarity is got using one of the four method described previously (see table 1).

Language	avgP	avgP with exp	Dif
English	0.5083	0.5267	+3.6%
French	0.5240	0.5488	+4.7%
Portuguese (2004)	0.4741	0.5084	+7.2%

Table 5: Comparative results in fixed system

Language	Model	P1	P2	P3	avgP
English	1	3	8	11	0.5034
English	**2**	**3**	**7**	**13**	**0.5182**
English	3	5	10	12	0.4987
English	4	3	7	14	0.5139
French	1	5	8	12	0.5129
French	**2**	**3**	**9**	**12**	**0.5208**
French	3	3	8	12	0.5106
French	4	3	7	12	0.5163
Portuguese (2004)	1	6	8	13	0.4625
Portuguese (2004)	**2**	**4**	**8**	**14**	**0.4792**
Portuguese (2004)	3	4	8	14	0.4724
Portuguese (2004)	4	4	8	14	0.4788

Table 6: Training results in combined passages method

Language	Model	P1	P2	P3	avgP
English	1	5	9	11	0.5074
English	**2**	**4**	**9**	**11**	**0.5344**
English	3	6	7	11	0.5259
English	4	4	9	11	0.5317
French	**1**	**5**	**9**	**13**	**0.5614**
French	2	3	8	11	0.5599
French	3	3	9	11	0.5486
French	4	6	10	14	0.5576
Portuguese (2004)	1	4	8	11	0.4892
Portuguese (2004)	**2**	**4**	**8**	**12**	**0.5089**
Portuguese (2004)	3	4	8	12	0.5046
Portuguese (2004)	4	4	8	12	0.5085

Table 7: Training results in combined passages method with query expansion

The tests which get better results are showed in table 6. 'Model' belongs to the combined method used and the columns P1, P2 and P3 are size passages which are provided by the best combined method. The columns P1, P2 and P3 are a small, medium and big passages respectively.

As we show the combined method, which provides the best results, is the method 2 (SUM without normatization) for all languages.

We have proved that the results increases comparing with fixed system in all languages except in French.

3.2.4 Combined passages with query expansion

We have carried out the same tests with query expansion and the results improves in all languages, although the increase is not meaningful in Portuguese.

The best combined method for English and Portuguese carries on being the method 2 (SUM), but in French is 1 (MAX) (see table 7).

As we check in table 5 combined passages system improves the results between 3.1% and 7.7% according to each language.

Tables 9 and 10 compare both methods showing the best results obtained in each test.

Language	avgP	avgP with exp	Dif
English	0.5182	0.5344	+3.1%
French	0.5208	0.5614	+7.7%
Portuguese (2004)	0.4792	0.5089	+6.1%

Table 8: Comparative results in combined system

Language	avgP fixed	avgP comb	Dif
English	0.5083	0.5182	+1.9%
French	0.5240	0.5208	-0.6%
Portuguese (2004)	0.4741	0.4792	+1%

Table 9: Comparative: fixed system vs combined system without query expansion

4 Results at CLEF-2005

We have submitted three runs for each language in our paticipation at the CLEF-2005. The best parameters, which provided the best results in the training, have been used in all cases. We did not have any training data about Bulgarian and Hungarian languages, therefore we have used the parameters of English language.

- IRn-xx-fexp is based on using the fixed size passages system which obtains the best results in training. Query expansion techniques have been used in these runs.
- IRn-xx-vnexp Combined passages system has been used without applying query expansion techniques.
- IRn-xx-vexp Combined passages system has been used with query expansion techniques.

Officials results for each run are showed in table 11. The model IRn-xx-vnexp is taken as a reference. As other models which use query expansion techniques, our model also increases the performance on the base system.

In this table 11 the two models with query expansion are compared. This one presents that the percentage of improvement in the combined model is around 4% of increase avgP in every language (except for Bulgarian).

As shown table 11, our results are above average in all languages appreciably, except for Bulgarian that the results are below average.

5 Conclusions and Future Work

This paper has described PR model which uses similarity values of three different size passages for each document in order to obtain the similarity of document regarding the question. This model has allowed us to improve the results around 4% according to models which used only a fixed size passage.

Language	avgP fixed	avgP comb	Dif
English	0.5267	0.5344	+1.4%
French	0.5488	0.5614	+2.2%
Portuguese (2004)	0.5084	0.5089	+0.09%

Table 10: Comparative: fixed system vs combined system with query expansion

Language	Run	AvgP	Dif
English	IRn-en-vexp	41.88	
	IRn-en-fexp	41.08	
	IRn-en-vnexp	40.24	
French	CLEF Average	35.30	
	IRn-fr-vexp	35.90	+1.7%
	IRn-fr-fexp	34.85	
	IRn-fr-vnexp	30.70	
Portuguese	CLEF Average	33.29	
	IRn-pt-vexp	36.03	+8.2%
	IRn-pt-fexp	34.46	
	IRn-pt-vnexp	33.15	
Hungarian	CLEF Average	29.00	
	IRn-hu-vexp	31.74	+9.4%
	IRn-hu-fexp	30.55	
	IRn-hu-vnexp	30.36	
Bulgarian	CLEF Average	22.00	
	IRn-bu-vexp	17.46	
	IRn-bu-fexp	17.58	
	IRn-bu-vnexp	17.87	-18.0%

Table 11: CLEF 2005 official results. Monolingual tasks

IR-n architecture allows us to realize this increase of steps in combined model without improving speed process considerably.

Lastly, we outline the future directions that we plan to undertake not only to improve this model, but also to be applied it in Question Answering task.

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] Aitao Chen and Fredric C. Gey. Combining query translation and document translation in cross-language retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, pages 108–121, Trondheim, Norway, 2003. Springer-Verlag.

[2] M. Kaskziel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185, 1997.

[3] F. Llopis. *IR-n: Un Sistema de Recuperación de Información Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[4] Fernando Llopis, Rafael Muñoz, Rafael M. Terol, and Elisa Noguera. Ir-n r2 : Using normalized passages. In Carol Peters and Francesca Borri, editors, *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop*, pages 65–72, Pisa, Italy, 2004. IST-CNR.

[5] G. Salton. Automatic text processing: The transformation, analysis, and retrieval of information by computer. 1989.

[6] Savoy J. Fusion of probabilistic models for effective monolingual retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[7] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, Experimental Studies*, pages 21–29, 1996.