IR-n system at CLEF-2002

Fernando Llopis, José L. Vicedo and Antonio Ferrández
Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
{llopis,vicedo,antonio}@dlsi.ua.es

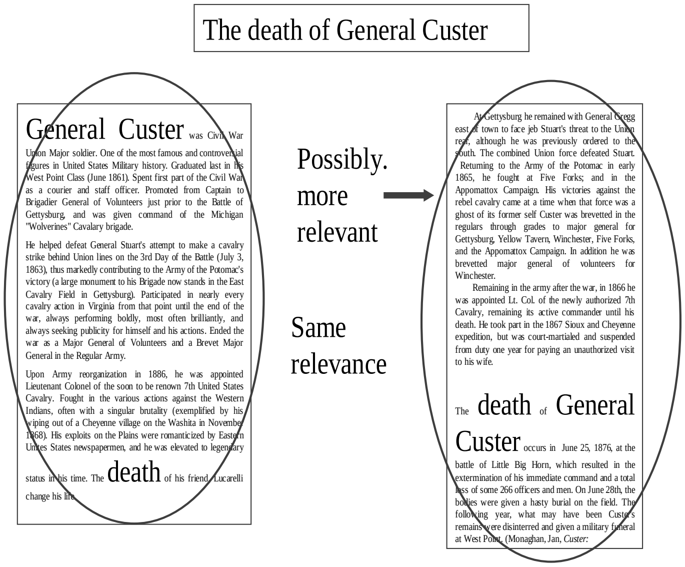
Abstract

Passage Retrieval is an alternative to traditional document-oriented Information Retrieval. These systems use contiguous text fragments (or passages), instead of full documents, as basic unit of information. IR-n system is a passage retrieval system that use groups of contiguous sentences as unit of information. This paper reports on experiments with IR-n system at Clef-2002 where it has obtained considerable better results than last participation.

1 Introduction

Information Retrieval (IR) systems receive as input a user's query and as result, they return a set of documents ranked by their relevance to the query. There are different techniques for measuring the relevance of a document to a query, but most of them take into account the number of times that query terms appear in documents, the importance or discrimination value of these terms in the document collection, as well as the size of each document.

One of the main problems related to document-oriented retrieval systems is that they not consider the proximity of appearance of query terms into the documents [6](see Figure 1).



General Custer was Civil War Union Major soldier. One of the most famous and controversial figures in United States Military history. Graduated last in his West Point Class (June 1861). Spent first part of the Civil War as a courier and staff officer. Promoted from Captain to Brigadier General of Volunteers just prior to the Battle of Gettysburg, and was given command of the Michigan "Wolverines" Cavalary brigade.

He helped defeat General Stuart's attempt to make a cavalry strike behind Union lines on the 3rd Day of the Battle (July 3, 1863), thus markedly contributing to the Army of the Potomac's victory (a large monument to his Brigade now stands in the East Cavalry Field in Gettysburg). Participated in nearly every cavalry action in Virginia from that point until the end of the war, always performing boldly, most often brilliantly, and always seeking publicity for himself and his actions. Ended the war as a Major General of Volunteers and a Brevet Major General in the Regular Army.

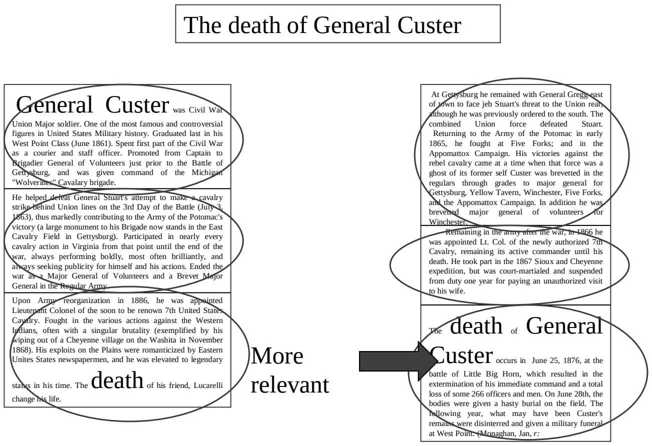
Upon Army reorganization in 1886, he was appointed Lieutenant Colonel of the soon to be renown 7th United States Cavalry. Fought in the various actions against the Western Indians, often with a singular brutality (exemplified by his wiping out of a Cheyenne village on the Washita in November 1868). His exploits on the Plains were romanticized by Eastern Unites States newspapermen, and he was elevated to legendary status in his time. The death of his friend, Lucarelli change his life

At Gettysburg he remained with General Gregg east of town to face jeb Stuart's threat to the Union rear, although he was previously ordered to the south. The combined Union force defeated Stuart. Returning to the Army of the Potomac in early 1865, he fought at Five Forks; and in the Appomattox Campaign. His victories against the rebel cavalry came at a time when that force was a ghost of its former self Custer was brevetted in the regulars through grades to major general for Gettysburg, Yellow Tavern, Winchester, Five Forks, and the Appomattox Campaign. In addition he was brevetted major general of volunteers for Winchester.

Remaining in the army after the war, in 1866 he was appointed Lt. Col. of the newly authorized 7th Cavalry, remaining its active commander until his death. He took part in the 1867 Sioux and Cheyenne expedition, but was court-martialed and suspended from duty one year for paying an unauthorized visit to his wife.

The death of General Custer occurs in June 25, 1876, at the battle of Little Big Horn, which resulted in the extermination of his immediate command and a total loss of some 266 officers and men. On June 28th, the bodies were given a hasty burial on the field. The following year, what may have been Custer's remains were disinterred and given a military funeral at West Point. (Monaghan, Jan, *Custer:*

Figure 1: Document-oriented retrieval



The death of General Custer

General Custer was Civil War Union Major soldier. One of the most famous and controversial figures in United States Military history. Graduated last in his West Point Class (June 1861). Spent first part of the Civil War as a courier and staff officer. Promoted from Captain to Brigadier General of Volunteers just prior to the Battle of Gettysburg, and was given command of the Michigan "Wolverines" Cavalary brigade.

He helped defeat General Stuart's attempt to make a cavalry strike behind Union lines on the 3rd Day of the Battle (July 3, 1863), thus markedly contributing to the Army of the Potomac's victory (a large monument to his Brigade now stands in the East Cavalry Field in Gettysburg). Participated in nearly every cavalry action in Virginia from that point until the end of the war, always performing boldly, most often brilliantly, and always seeking publicity for himself and his actions. Ended the war as a Major General of Volunteers and a Brevet Major General in the Regular Army.

Upon Army reorganization in 1886, he was appointed Lieutenant Colonel of the soon to be renown 7th United States Cavalry. Fought in the various actions against the Western Indians, often with a singular brutality (exemplified by his wiping out of a Cheyenne village on the Washita in November 1868). His exploits on the Plains were romanticized by Eastern Unites States newspapermen, and he was elevated to legendary status in his time. The death of his friend, Lucarelli change his life.

At Gettysburg he remained with General Gregg east of town to face jeb Stuart's threat to the Union rear, although he was previously ordered to the south. The combined Union force defeated Stuart. Returning to the Army of the Potomac in early 1865, he fought at Five Forks; and in the Appomattox Campaign. His victories against the rebel cavalry came at a time when that force was a ghost of its former self Custer was brevetted in the regulars through grades to major general for Gettysburg, Yellow Tavern, Winchester, Five Forks, and the Appomattox Campaign. In addition he was brevetted major general of volunteers for Winchester.

Remaining in the army after the war, in 1866 he was appointed Lt. Col. of the newly authorized 7th Cavalry, remaining its active commander until his death. He took part in the 1867 Sioux and Cheyenne expedition, but was court-martialed and suspended from duty one year for paying an unauthorized visit to his wife.

The death of General Custer occurs in June 25, 1876, at the battle of Little Big Horn, which resulted in the extermination of his immediate command and a total loss of some 266 officers and men. On June 28th, the bodies were given a hasty burial on the field. The following year, what may have been Custer's remains were disinterred and given a military funeral at West Point. (Monaghan, Jan, *r:*

Figure 2: Passage retrieval

A possible alternative to these models consists on computing the similarity between a document and a query in accordance with the relevance of the passages each document is divided (see Figure 2). This approach, called Passage Retrieval (PR), is not so affected by the length of the documents and besides, they add the concept of proximity to the similarity measure by analysing small pieces of text instead of whole documents. Figures 1 and 2 show the main differences between both approaches.

PR systems can be classified in accordance with the way of dividing documents into passages. PR community generally agrees with the classification proposed in [1], where the author distinguishes between discourse models, semantic models, and window models. The first one uses the structural properties of the documents, such as sentences or paragraphs [2] in order to define the passages. The second one divides each document into semantic pieces according to the different topics in the document [3]. The last one uses windows of a fixed size (usually a number of terms) to determine passage boundaries [5].

At first glance, we could think that discourse-based models would be the most effective, in retrieval terms, since they use the structure of the document itself. However, this model greatest problem relies on detecting passage boundaries since it depends on the writing style of the author of each document. On the other hand, window models have as main advantage that they are simpler to accomplish, since the passages have a previously known size, whereas the remaining models have to bear in mind the variable size of each passage. Nevertheless, discourse-based and semantic models have the main advantage that they return full information units of the document, which is quite important if these units are used as input by other applications.

The passage extraction model that we propose (IR-n) allows us to benefit from the advantages of discourse-based models since self-contained information units of text, such as sentences, are used for building passages. Moreover, the relevance measure which, unlike other discourse-based models, is not based on the number of passage terms, but on a fixed number of passage sentences. This fact allows a simpler calculation of this measure unlike other discourse-based or semantic models. Although each passage is made up by a fixed number of sentences, we consider that our proposal differs from the window models since our passages do not have a fixed size (i.e. a fixed number of words) since we use sentences with a variable size.

This paper is structured as follows. The following section presents the basic features of IR-n system. Third section describe the main improvements introduced for Clef-2002 Conference. Fourth section describes the different runs performed for this campaign and discusses the results obtained. Finally, last section extracts initial conclusions and opens directions for future work.

2 IR-n system

The system proposed has the main following features:

1. A document is divided into passages that are made up by a number N of sentences.

2. Passages overlap. First passage contains from sentence *1* to N, second passage contains from sentence *2* to $N+1$, etc.

3. The similarity between a passage p and a query q is computed as follows:

$$Passage\_similarity = \sum_{t \in p \wedge q} W_{p,t} * W_{q,t} \tag{1}$$

   Where
   $W_{p,t} = log_e(f_{p,t} + 1)$,
   $f_{p,t}$ is the number of appearances of term t in passage p,
   $W_{q,t} = log_e(f_{q,t} + 1) * idf$,
   $f_{q,t}$ represents the number of appearances of term t in question q,
   $idf = log_e(n/f_t + 1)$,
   n is the number of documents of the collection and
   f_t is the number of documents term t appears in.

   As it can be observed, this formulation is similar to the cosine measure defined in [9]. The main difference is that length normalisation is omitted. Instead, our proposal accomplishes length normalisation by defining passage size as a fixed number of textual discourse units. In this case, the discourse unit selected is the sentence and a passage is defined as a fixed number N of sentences. This way, although the number of terms of each passage may vary, the number of sentences is constant.

IR-n system has been developed in C++ and runs in a Linux cheap computer, without additional software requirements.

3 IR-n system from Clef-2001 to Clef-2002

In last Clef edition IR-n system [8] was used in two retrieval tasks: monolingual (Spanish) and bilingual (Spanish-English). Bilingual task results were satisfactory however, monoligual results were very poor ranging below the average of the results obtained by all the participant systems.

After analysing those results we arrived at a series of conclusions that are summed up in the following points:

- We had several problems on processing SGML original files. Consequently, some documents were not indexed correctly.

- The Spanish lemmatizer that we selected (conexor) produced a high number of errors.

- The type of document collection used, press reports of small size, did not allow big differences between passage retrieval and document retrieval approaches. This fact was confirmed when verifying that the results obtained by our system were similar to the baseline system (cosine model) whereas when retrieving from Los Angeles Times collection the improvement achieved by the passage approach was considerable.

		Precision at N documents						
	Recall	5	10	20	30	200	AvgP	Inc
Baseline	94.02	0.6000	0.5408	0.4582	0.4054	0.1826	0.4699	0.00
IR-n 7 sentences	94.54	0.6612	0.5796	0.4939	0.4490	0.1917	**0.5039**	7.23%
IR-n 8 sentences	94.95	0.6735	0.6061	0.4929	0.4537	0.1924	0.5017	6.76%

Table 1: Results for short questions

		Precision at N documents						
	Recall	5	10	20	30	200	AvgP	Inc
Baseline	95.62	0.6163	0.5612	0.4857	0.4367	0.1943	0.5010	0.00
IR-n 6 sentences	96.18	0.6653	0.5918	0.5020	0.4469	0.1995	**0.5156**	2.92%
IR-n 7 sentences	95.99	0.6816	0.5939	0.4990	0.4490	0.1983	0.5150	2.79%

Table 2: Results for long questions

- We could not make any previous experiment for determining the optimum size of the passage since it was the first time this approach was applied.

The main changes proposed for Clef-2002 were designed to solve these problems. Therefore, the following changes were introduced:

- Documents and questions preprocess was improved.
- The Spanish lemmatizer was replaced by a simple stemmer.
- A serie of experiments was performed to determine the suitable size of the passages (the number N of sentences).
- The relevance measure was modified in order to increase the score of the passages when a sentence contained more than a consecutive word of the question.
- Long questions treatment was changed.
- We added a question expansion module that could be applied optionally.

3.1 Training process

We developed a serie of experiments in order to optimize system performance. These experiments were carried out on the same document collection (EFE agency), but using the 49 test questions proposed in Clef-2001.

As baseline system we selected the well-known document retrieval model based on the cosine similarity measure [9]. The experiments were designed for detecting the best value for N (the number of sentences that make up a passage). Initially, we detected the interval where the best results were obtained and then, we proceeded to determine the optimum value for N. System performance was measured using the standard average interpolated precision (AvgP).

For short questions, best results were obtained when passages were 7 or 8 sentences length. For long questions, best results were achieved for passages of 6 or 7 sentences. Tables 1 and 2 show these results for short and long questions respectively.

In both cases better results are obtained although, the difference with baseline is more considerable when using long queries. After analysing these results, we determined to fix the size of passages to 7 sentences since this length achieved the best results for short questions and they also were nearly the best for long queries.

Once we had determined the optimum length for passages, we designed a second experiment for adapting the similarity measure described before in such a way that allowed increasing this measure when more than one question term was found into a sentence and they presented the

		Precision at N documents						
	Recall	5	10	20	30	200	AvgP	Inc
IR-n base	94.54	0.6612	0.5796	0.4939	0.4490	0.1917	0.5039	0.00
IR-n factor 1.1	94.95	0.6653	0.5918	0.5041	0.4497	0.1935	0.5102	1.25%
IR-n factor 1.2	94.84	0.6694	0.5878	0.5010	0.4510	0.1933	0.5127	1.74%
IR-n factor 1.3	94.47	0.6735	0.5857	0.4990	0.4537	0.1930	0.5100	1.21%
IR-n factor 1.4	94.28	0.6653	0.5878	0.5041	0.4531	0.1914	0.5081	0.83%

Table 3: Results for short questions.

		Precision at N documents						
	Recall	5	10	20	30	200	AvgP	Inc
IR-n base	95.99	0.6816	0.5939	0.4990	0.4490	0.1983	0.5150	0.00
IR-n factor 1.1	95.88	0.6735	0.5898	0.5010	0.4510	0.1969	0.5098	-1.00%
IR-n factor 1.2	95.47	0.6694	0.5898	0.5082	0.4510	0.1959	0.5047	-2.00%
IR-n factor 1.3	95.40	0.6490	0.5959	0.5031	0.4524	0.1945	0.4975	-3.39%
IR-n factor 1.4	94.95	0.6449	0.6000	0.5061	0.4517	0.1919	0.4930	-4.27%

Table 4: Results for long questions.

same order in both question and sentence. This experiment consisted on optimizing the value α that increases the score of a question term when this circumstances happen. Thus, the passage similarity formula previously mentioned changed as follows:

$$Passage\_similarity = \sum_{t \in p \wedge q} W_{p,t} * W_{q,t} * \alpha \qquad (2)$$

The factor α takes value 1 for a term that appears into a sentence whose terms previous and later in the question are not in the same phrase, and another value in the opposite case. This experiment has applied several coefficients in order to obtain the optimum value for α. Tables 3 and 4 shows the results obtained for short and long questions respectively.

In these tables it is possible to observe that, for short questions, results improve for α values of 1.1 and 1.2 whereas results slightly get worse for long questions.

4 Clef-2002: Experiments and Results

As the results obtained in Clef-2001 for monolingual task were not the expected, this year our participation was focused to improve the Spanish monolingual task.

4.1 Runs Description

We carried out four runs for monolingual task. Two with *title + description* and two with *title + description + narrative*. For all the runs passage length was set to 7 sentences and the value 1.1 was assigned to the α proximity coefficient. These runs are described below.

To clarify the differences between the four runs we will consider the following example question:

<top>
<num> C103 </num>
<ES-title> *Conflicto de intereses en Italia* </ES-title>
<ES-desc> *Encontrar documentos que discutan el problema del conflicto de interes es del primer ministro italiano, Silvio Berlusconi.* </ES-desc>
<ES-narr> *Los documentos relevantes se referirán de forma explícita al conflicto de intereses entre el Berlusconi político y cabeza del gobierno italiano, y el Berlusconi hombre de negocios.*

También pueden incluir información sobre propuestas o soluciones adoptadas para resolver este conflicto. </ES-narr>
</top>

4.1.1 IR-n1.

This run takes only short questions (title + description). The example question was processed as follows:

Conflicto de intereses en Italia. Encontrar documentos que discutan el problema del conflicto de intereses del primer ministro italiano, Silvio Berlusconi.

4.1.2 IR-n2.

This run is a little more complex. The question is divided into several queries. Each query contains an isolated idea appearing into the whole question. Then each query is posed for retrieval, evaluating this way, how passages respond to each of them. This approach is fully described in [7] and basic steps are summed up as follows:

1. Question narrative is divided according to the sentences it contains.
2. The system generates as many queries as sentences are detected. Each query contains title ,description and a sentence of the narrative.
3. Each generated query is processed separately recovering best 5,000 documents.
4. Relevant documents are punctuated with the maximum similarity value obtained for all the generated queries processed.
5. Best 1,000 relevant documents are finally retrieved.

In this case, from the example question described before the system generates the following two queries:

Query 1. *Conflicto de intereses en Italia. Encontrar documentos que discutan el problema del conflicto de interes es del primer ministro italiano, Silvio Berlusconi. Los documentos relevantes se referirán de forma explícita al conflicto de intereses entre el Berlusconi político y cabeza del gobierno italiano, y el Berlusconi hombre de negocios.*

Query 2. *Conflicto de intereses en Italia. Encontrar documentos que discutan el problema del conflicto de interes es del primer ministro italiano, Silvio Berlusconi.También pueden incluir información sobre propuestas o soluciones adoptadas para resolver este conflicto.*

4.1.3 IR-n3.

This run is similar to IR-n1 but applies query expansion according to the model defined in [4]. This expansion consists on detecting the 10 more excellent terms of first 5 recovered documents, and adding them to the original question.

4.1.4 IR-n4.

This run uses long questions formed by title, description and narrative. The example questions was posed for retrieval as follows.

Conflicto de intereses en Italia. Encontrar documentos que discutan el problema del conflicto de interés es del primer ministro italiano, Silvio Berlusconi. Los documentos relevantes se referirán de forma explícita al conflicto de intereses entre el Berlusconi político y cabeza del gobierno italiano, y el Berlusconi hombre de negocios. También pueden incluir información sobre propuestas o soluciones adoptadas para resolver este conflicto.

		Precision at N documents						
	Recall	5	10	20	30	200	AvgP	Inc
Median clef2002 systems							0.4490	0.00
IR-n1	90.08	0.6800	0.5820	0.5140	0.4620	0.1837	0.4684	+4.32%
IR-n2	92.64	0.7200	0.6380	0.5600	0.4813	0.1898	0.5067	+12.85%
IR-n3	93.51	0.6920	0.5920	0.5190	0.4667	0.2018	0.4980	+10.91%
IR-n4	91.83	0.7120	0.6120	0.5380	0.4867	0.1936	0.4976	+10.82%

Table 5: Results comparison.

4.2 Results

In this section the results achieved by our four runs are compared with the obtained by all the systems that participated at this conference. Table 5 shows the average precision for monolingual runs and computes the increment of precision achieved. This increment (or decrement) was calculated by taking as base the median average precision of all participant systems.

As it can be observed, our four runs performed better than median results. Our baseline (IR-n1) improved around a 4% and the remaining runs performed better between 11 and 13%.

5 Conclusions and Future Work

General conclusions are positive. We have obtained considerably better results than in previous edition. This fact has been caused mainly by three aspects. First, the better preprocessing of documents carried out. Second, the system has been correctly trained to obtain the optimum size of passage. Third, the errors introduced by the Spanish lemmatizer have been avoided by using a simple stemmer.

After this new experience, we are examining several lines of future work. We want to analyse the possible improvements that could be obtained using another type of lemmatizer instead of the simple stemmer that we have used this year. On the other hand we are going to continue studying modifications for the relevance formula in order to improve the application of vicinity factors.

6 Acknowledgements

This work has been supported by the Spanish Government (CICYT) with grant TIC2000-0664-C02-02.

References

[1] James P. Callan. Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, pages 302–310, London, UK, July 1994. Springer Verlag.

[2] J. Allan G. Salton and C. Buckley. Approaches to passage retrieval in full text information systems. In *Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, Pittsburgh, PA, jun 1993.

[3] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proc. 16th ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 59–68, 1993.

[4] P. Jourlin, S.E. Johnson, K. Spärck Jones, and P.C. Woodland. General query expansion techniques for spoken document retrieval. In *Proc. ESCA Workshop on Extracting Information from Spoken Audio*, pages 8–13, Cambridge, UK, 1999.

[5] Marcin Kaszkiel and Justin Zobel. Passage Retrieval Revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Text Structures, pages 178–185, Philadelphia, PA, USA, 1997.

[6] Marcin Kaszkiel and Justin Zobel. Effective Ranking with Arbitrary Passages. *Journal of the American Society for Information Science (JASIS)*, 52(4):344–364, February 2001.

[7] Fernando Llopis, Antonio Ferrández, and José L. Vicedo. Using Long Queries in a Passage Retrieval System. In O. Cairo, E. L. Sucar, and F. J. Cantu, editors, *Proceeding of Mexican International Conference on Artificial Intelligence*, volume 2313 of *Lectures Notes in Artificial Intelligence*, Mérida, Mexico, 2002. Springer-Verlag.

[8] Fernando Llopis and José L. Vicedo. IR-n system, a passage retrieval systema at CLEF 2001. In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Darmstadt, Germany, 2001. Springer-Verlag.

[9] Gerard A. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, New York, 1989.