Spoken Document Retrieval experiments with IR-n system

Fernando Llopis and Patricio Martínez-Barco
Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{*llopis,patricio*}*@dlsi.ua.es*

19/07/2003

Abstract

This paper describes the first participation of IR-n system at Spoken Document Retrieval, focusing on the experiments we made before participation and showing the results we obtained. IR-n system is an Information Retrieval system based on passages and the recognition of sentences to define them. So, the main goal of this experiment is to adapt IR-n system to the spoken document structure by means of the utterance splitter and the overlapping passage technique allowing to match utterances and sentences

1 Introduction

Usually, research work on natural language processing has started from written documents instead of spoken documents due to spoken document processing has a lot of disadvantages induced by its informal disposition among other reasons.

As appointed by Dahlbäck [2]:

"... spoken input is often incomplete, incorrect and contains interruptions and repairs; full sentences occur only very occasionally. Therefore new basic units for the development of dialogue models have to be proposed ..."

Thus, some of the most important problems to solve in spoken document processing are [3]:

- The lack of punctuation marks, that impedes the well understanding of sentences because boundaries are unknown. This understanding must be induced by pause detection. This is the reason why the "sentence" concept is replaced by the "utterance" concept. Utterance is defined, from a pragmatic point of view, as a sequence of words chained by a speaker between two pauses. In the same way, the "paragraph" is replaced by the "turn" that is defined from a pragmatic point of view as the set of utterances that a speaker can express between two speaker changes (when several speakers participate in the dialogue), or the set of utterances that a speaker expresses about the same subject (in monologues or newsreels).

- Moreover, turns may be considered like null or empty when they do not contribute to the discourse, that is, turns having the function of pointing out the speaker is *on* the conversation: "ejem...", "yes...", "I know..."; as well as other turns without semantic content such as "good morning", "have a good weekend", and so on.

- Furthermore, turns can be interrupted due to overlaps, or speaker mistakes, causing repetitions and modifications of previous information.

This sort of problems is increased with problems derived from the automatic transcription process which incorporates noise, spelling mistakes, and unrecognizable words due to deficiencies in the original recording or speak recognition fails.

Due to this, the use of spoken documents in information retrieval tasks allows to test the system robustness against document mistakes. Then, the main goal of this paper is to test the robustness of IR-n System and to study some text processing techniques that could improve this robustness in spoken documents.

IR-n is an Information Retrieval system based on passages [4] [5]. Passages are defined using a fixed number of sentences from the original document. It seems obvious that IR-n has been developed to work on written documents with a clear structure based on known sentence boundaries. However, in order to test its robustness, IR-n has been submitted to the CLEF SDR Track.

SDR task is based on processing non-structured documents that proceed from an automatic transcription of radio news. Our main objective is to test if IR-n system can be applied to document collections where sentence boundaries are unknown. This experiment is focused on the estimation of sentence boundaries by means of the pauses recognized along the transcription process. So, the main hypothesis is based on the following ideas:

- longest pauses mean the end of utterances
- IR-n System can accept utterances instead of sentences to define passages.

So, the experiments will focus on determining what is the average length of a pause between utterances to build an utterance splitter that will feed the IR-n system.

However, using this model, passage definitions may be faulty. The terms of a query may be dispersed among several passages, and some relevant documents may be discarded. This problem can be avoid by using passage overlapping, since this technique allows more than one passage sharing the same fragment of document.

2 CL-SDR Track description

Cross-Language Spoken Document Retrieval (CL-SDR) is a new track proposed for CLEF 2003. The track is mostly based on existing resources, available by NIST, which were used at TREC-8 [6] and TREC-9 [7].

The benchmark track is an extension of evaluation data prepared by NIST for TREC 8-9 SDR tracks. It has a collection of automatic transcripts (557 hours) of American-English news recordings broadcasted by ABC, CNN, Public Radio International, and Voice of America between February and June 1998. Transcripts are provided with known story boundaries (21,754 stories); and a collection of 100 English topics, either in terse or short format. The TREC collection has been extended with translations of the short topics into five European languages: Dutch, Italian, French, German, and Spanish.

Technical specifications of the task are shown in table 1.

3 Passage definition at IR-n system

Taking advantage of using sentences in IR-n as a basic unit to the passage definition task, the sentence will be used to define the passage overlapping too.

The overlapping degree (G_{sol}) in IR-n system shows the sentence number from which the definition of the next passage starts. The main features of this value are the following:

1. G_{sol} must be lower than the passage size. Having the same value means that no overlapping is used.
2. The lower the value G_{sol} is, the higher the amount of text shared by two consecutive passages will be.

- Objective: the track aims at evaluating CLIR systems on noisy automatic transcripts of spoken documents with known story boundaries.
- Development data (from TREC 8 SDR):
    1. Document collection: B1SK Baseline Transcripts, known bounds download from NIST.
    2. Topics: Short topics in English, Dutch, French, German, Italian, and Spanish.
    3. Relevance assessments: Topics-074-123.
    4. Parallel document collections (optional and only available through LDC): Textual resources.
- Evaluation data (from TREC 9 SDR):
    1. Document collection: B1SK Baseline Transcripts, known bounds download from NIST.
    2. Topics: Short topics in English , Dutch, French, German, Italian, and Spanish.
    3. Relevance assessments: Topics-124-173.
    4. Parallel document collections (optional and only available through LDC): Textual resources.
- Primary Conditions (mandatory for all participants):
    1. Monolingual IR without using any parallel collection (contrastive condition).
    2. Bilingual IR from French or German.
- Secondary Condition (optional):
    1. Monolingual IR using any available parallel collections.
    2. Bilingual IR from other languages.
- Submission of runs:
    1. Maximum 12 runs per participant, with the limit of 3 runs for each considered source language.

Table 1: Technical specifications of the CLEF'2003 CL-SDR Track

3. As a result, the lower the value G_{sol} is, the more number of passages will be defined in the document.

The use of passage overlapping means to redefine the passage concept to IR-n in the following way:

- Given a document D consisting of N sentences.

$$D = f_1..f_N \tag{1}$$

- Taken into account that n is the number of sentences integrating a passage.
- Given an overlapping degree G_{sol}
- The following passages will be defined from the document D

$$P_i = f_{G_{sol}*(i-1)+1}, ..., f_{\min(G_{sol}*(i-1)+n,N)}, i \in [1..N/G_{sol} - 1] \tag{2}$$

Given that definition, and supposing a passage size of 15 sentences, an overlapping degree of 10, and a document size of 35 sentences, the passage generation will be performed in the following way:

1. $P_1 = f_1..f_{15}$
2. $P_2 = f_{11}..f_{25}$
3. $P_3 = f_{21}..f_{35}$

The increase of the efficiency in document retrieval is an immediately advantage of passage overlapping. However, the response time increases (to a large extent when the overlapping degree is lower) because the number of passages to be evaluated is greater.

Nevertheless, the use of lower overlapping degrees improves the system results noticeably, and it has not excessive influence on the searching time.

Overlapping does not increase the searching cost so much due to two main reasons:

1. IR-n does not evaluate each one of the document passages, since the similarity measure [1] in some cases may be avoided. The first passage to be evaluated is the one starting in the first sentence of the document in which a query term appears. That is due to passages starting in a previous sentence can not obtain a similarity measure higher than this first passage, by the way in which the similarity measure has been defined in IR-n.

   For this same reason, the last passage to be evaluated is the one finishing in the last sentence of the document in which a query term appears.

   These same conclusions may be extended to passages not located at the limits of the document, that is, internal passages. Given an overlapping degree G_{sol}, if a passage does not contain query terms during its first sentences then its evaluation can be omitted. For example, if G_{sol} is equal to 1, the evaluation of those passages which first sentence does not contain any query term is not needed.

   Because of this, the number of passages to be evaluated is reduced, and, consequently, to use of small overlapping degrees has not the same influence as if each passage of the document is evaluated.

2. Another important aspect is related to the system implementation. IR-n implementation is based on storing all the information about word occurrences in main memory. Thus, the segmentation process is performed during the execution over data structures located at main memory.

   Considering that the most influencing factors to time processing are related to disc access times, this minor increase of time when a greater number of passages is processed, it is not significant to the final time.

For this reason, IR-n uses an overlapping degree (G_{sol}=1) being the value that obtains the best performance.

4 Experimental work

According to the track specification, the test collection used in this experiment was TREC-8. During these experiments several passages sizes (from 1 to 9 sentences) and several pause recognition sizes (0.1, 0.2, and 0.3 seconds) have been valuated. Moreover, the IR-n system with and without query expansion has been tested.

Tables 2, 3 and 4 show the results without query expansion.

Tables 5, 6 and 7 show the results with query expansion.

These tables show that the best result is obtained using the model with query expansion, a passage size of 5 sentences and 0.2 seconds to recognize a pause between two utterances at the utterance splitter.

5 System evaluation

This system was evaluated with the TREC SDR-9 collection according to the track specification. Moreover, a bilingual test was performed using French queries that were translated into English by Power Translator, Free-translator and Babel Fish.

Both monolingual and bilingual tests were performed with and without query expansion. The best results for monolingual and bilingual queries are shown in tables 8 and 9 respectively.

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F K3	78.49	0.4980	0.4490	0.3398	0.2837	0.1041	0.3301
IR-n 2 F K3	79.26	0.5347	0.4633	0.3612	0.3102	0.1067	0.3540
IR-n 3 F K3	79.43	0.5429	0.4735	0.3602	0.3095	0.1099	0.3695
IR-n 4 F K3	79.81	0.5592	0.4735	0.3786	0.3204	0.1106	0.3774
IR-n 5 F K3	79.65	0.5469	0.4878	0.3786	0.3224	0.1107	0.3812
IR-n 6 F K3	80.09	0.5633	0.5102	0.3888	0.3293	0.1120	0.3845
IR-n 7 F K3	80.14	0.5796	0.4980	0.3878	0.3279	0.1123	**0.3852**
IR-n 8 F K3	80.20	0.5796	0.4980	0.3888	0.3265	0.1135	0.3850
IR-n 9 F K3	80.31	0.5755	0.5000	0.3888	0.3197	0.1141	0.3817

Table 2: Training results without query expansion using 0.1 seconds to discover pauses

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F K3	78.71	0.4898	0.4429	0.3490	0.2884	0.1052	0.3343
IR-n 2 F K3	79.26	0.5306	0.4694	0.3602	0.3095	0.1073	0.3600
IR-n 3 F K3	79.92	0.5633	0.4796	0.3786	0.3122	0.1101	0.3756
IR-n 4 F K3	79.70	0.5755	0.4959	0.3806	0.3204	0.1112	0.3825
IR-n 5 F K3	80.09	0.5755	0.5000	0.3888	0.3231	0.1121	0.3834
IR-n 6 F K3	80.14	0.5714	0.4878	0.3816	0.3245	0.1132	0.3823
IR-n 7 F K3	80.42	0.5714	0.4837	0.3857	0.3136	0.1145	0.3801
IR-n 8 F K3	80.64	0.5837	0.5000	0.3898	0.3177	0.1150	0.3842
IR-n 9 F K3	80.42	0.5796	0.5000	0.3949	0.3184	0.1142	**0.3856**

Table 3: Training results without query expansion using 0.2 seconds to discover pauses

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F K3	78.88	0.4653	0.4347	0.3429	0.3007	0.1063	0.3341
IR-n 2 F K3	79.48	0.5469	0.4796	0.3745	0.3068	0.1088	0.3687
IR-n 3 F K3	79.98	0.5469	0.4776	0.3714	0.3238	0.1110	0.3784
IR-n 4 F K3	80.36	0.5837	0.4816	0.3796	0.3211	0.1128	**0.3805**
IR-n 5 F K3	80.20	0.5837	0.4796	0.3837	0.3136	0.1138	0.3794
IR-n 6 F K3	80.25	0.5878	0.4755	0.3816	0.3082	0.1145	0.3729
IR-n 7 F K3	80.25	0.5837	0.4837	0.3847	0.3095	0.1140	0.3751
IR-n 8 F K3	80.58	0.5714	0.4735	0.3796	0.3156	0.1135	0.3712
IR-n 9 F K3	80.64	0.5796	0.4633	0.3755	0.3156	0.1131	0.3672

Table 4: Training results without query expansion using 0.3 seconds to discover pauses

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F1	83.44	0.5347	0.5082	0.3959	0.3320	0.1119	0.4029
IR-n 2 F1	83.94	0.6000	0.5143	0.4133	0.3422	0.1161	0.4307
IR-n 3 F1	84.98	0.5959	0.5204	0.4112	0.3544	0.1170	0.4373
IR-n 4 F1	85.42	0.6041	0.5388	0.4143	0.3517	0.1176	0.4392
IR-n 5 F1	85.15	0.5959	0.5408	0.4255	0.3612	0.1192	0.4494
IR-n 6 F1	85.20	0.6204	0.5306	0.4378	0.3653	0.1208	**0.4530**
IR-n 7 F1	85.42	0.6000	0.5327	0.4327	0.3680	0.1212	0.4503
IR-n 8 F1	85.31	0.6082	0.5327	0.4367	0.3653	0.1215	0.4528
IR-n 9 F1	85.37	0.6041	0.5388	0.4347	0.3639	0.1218	0.4489

Table 5: Training results with query expansion using 0.1 seconds to discover pauses

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F1	82.51	0.5429	0.5102	0.4020	0.3361	0.1144	0.4160
IR-n 2 F1	84.43	0.5837	0.5469	0.4194	0.3476	0.1167	0.4421
IR-n 3 F1	85.09	0.5837	0.5449	0.4265	0.3497	0.1172	0.4540
IR-n 4 F1	85.37	0.6163	0.5429	0.4306	0.3578	0.1194	0.4606
IR-n 5 F1	85.53	0.6041	0.5469	0.4408	0.3680	0.1206	**0.4620**
IR-n 6 F1	85.64	0.6041	0.5490	0.4398	0.3653	0.1212	0.4619
IR-n 7 F1	85.92	0.6041	0.5367	0.4337	0.3639	0.1219	0.4584
IR-n 8 F1	85.97	0.6000	0.5408	0.4378	0.3633	0.1220	0.4596
IR-n 9 F1	85.86	0.6041	0.5347	0.4398	0.3612	0.1226	0.4594

Table 6: Training results with query expansion using 0.2 seconds to discover pauses

		Precision at N documents					
	Recall	5	10	20	30	200	AvgP
IR-n 1 F1	83.44	0.5551	0.4959	0.4102	0.3435	0.1177	0.4154
IR-n 2 F1	84.76	0.6000	0.5245	0.4224	0.3558	0.1211	0.4385
IR-n 3 F1	85.26	0.6531	0.5286	0.4337	0.3605	0.1238	0.4527
IR-n 4 F1	85.15	0.6286	0.5367	0.4235	0.3680	0.1248	0.4520
IR-n 5 F1	85.15	0.6327	0.5388	0.4265	0.3653	0.1254	0.4540
IR-n 6 F1	85.04	0.6367	0.5306	0.4276	0.3694	0.1257	0.4544
IR-n 7 F1	85.26	0.6367	0.5347	0.4255	0.3639	0.1254	**0.4564**
IR-n 8 F1	85.53	0.6367	0.5327	0.4316	0.3646	0.1251	0.4554
IR-n 9 F1	85.48	0.6367	0.5367	0.4265	0.3565	0.1252	0.4518

Table 7: Training results with query expansion using 0.3 seconds to discover pauses

System	AvgP
ITC-irst	0.3944
Exeter	0,3843
Alicante	0,3637
JHU/APL	0,3192

Table 8: Monolingual results with query expansion using 0.3 seconds to discover pauses

System	AvgP
ITC-irst	0.3064
Alicante	0,3032
Exeter	0,2876
JHU/APL	0,1941

Table 9: Bilingual results with query expansion using 0.3 seconds to discover pauses

6 Conclusions and future work

Although we expected to know more information about other systems at the conference, we are pleased to see these results being above average for SDR track, taking into account that IR-n system was not designed to work on spoken documents.

Nevertheless, more experiments are expected to be done to increase the system performance.

7 Acknowledgements

This work has been partially supported by the Spanish Government (CICYT) with grant TIC2000-0664-C02-02 and (PROFIT) with grant FIT-150500-2002-416.

References

[1] J. Chen, A. Diekema, M. Taffet, N. McCracken, N. Ozgencil, O. Yilmazel, and E. Liddy. Question Answering: CNLP at the TREC-10 Question Answering Track. In *Tenth Text REtrieval Conference (Notebook)*, volume 500-250 of *NIST Special Publication*, Gaithersburg, USA, nov 2001. National Institute of Standards and Technology.

[2] N. Dahlbäck. Towards a dialogue taxonomy. In Elisabeth Maier, Marion Mast, and Susann LuperFoy, editors, *Dialogue Processing in Spoken Language Systems*, volume 1236 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, 1997.

[3] M.G. Fernández. *Un modelo para la especificación lingüística y la gestión computacional en diálogos hombre-máquina mediante instrucciones expresadas en lenguaje natural.* PhD thesis, Universidad de Sevilla, Departamento de Filología Inglesa. Facultad de Filología, Sevilla, 2000.

[4] Fernando Llopis and José L. Vicedo. IR-n system, a passage retrieval system at CLEF 2001. In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, pages 244–252, Darmstadt, Germany, 2001. Springer-Verlag.

[5] Fernando Llopis, José Luis Vicedo, and Antonio Ferrández. IR-n system at Clef-2002. In *Workshop of Cross-Language Evaluation Forum (CLEF 2002)*, Lecture notes in Computer Science, pages 177–184, Roma, Italy, 2002. Springer-Verlag.

[6] *Eighth Text REtrieval Conference*, volume 500-246 of *NIST Special Publication*, Gaithersburg, USA, nov 1999. National Institute of Standards and Technology.

[7] *Ninth Text REtrieval Conference*, volume 500-249 of *NIST Special Publication*, Gaithersburg, USA, nov 2000. National Institute of Standards and Technology.