The University of Alicante at CL-SR track

Rafael M. Terol and Manuel Palomar and Patricio Martinez-Barco
and Fernando Llopis and Rafael Muñoz and Elisa Noguera
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera de San Vicente del Raspeig - Alicante - Spain
Tel. +34965903653
{rafamt, mpalomar, patricio, llopis, rafael, elisa}@dlsi.ua.es

Abstract

This paper describes the participation of the University of Alicante in the new CL-SR Track at CLEF conference. In this track we introduce a set of features in the topics processing applied by our IR-n system. This set of features 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.

Keywords

Speech Retrieval, Information Retrieval, Logic Forms

1 Introduction

In the same line of active participation of University of Alicante in previous CLEF conferences, IR-n system takes part in this new Cross-Language Speech Retrieval (CL-SR) Track at the present CLEF 2005 conference. As novelty, 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].

Following section shows the main features of this new release of IR-n system. The logic form derivation module and the rules applied to these logic forms are also presented in following sections. Finally, we describe each one of the submitted runs, the scores obtained by the IR-n system in these submitted runs, the conclusions and the future works in our research activity.

2 IR-n as Passage Retrieval System

IR-n is a passage retrieval system (RP). RP systems [2] studies the appearance of query terms in contiguous fragments of the documents (also called passages). One of the main advantages of these systems is that these allow us to determine not only if a document is relevant or not, but also that these systems detect the relevant part of the document.

These passages are usually composed for a fixed number of sentences, but the but the format of the document collection of this CL-SR track does not allow this feature. These documents are composed by a contiguous set of words without punctuation marks. Moreover, we can't know the limit between each sentence. As a result, we have chosen a fixed number of words to compose the passages. Furthermore, IR-n system uses overlapping passages in order to avoid that some

documents could be considered not relevant if it appears words of the question in adjacent passages.

IR-n system allows the use of distinct similarity measures (Ex. Okapi [6]) to calculate the weights of the words of the topic according to the document collection.

Once the weights of the words have been calculated and with the aim of increment the weights of several words, IR-n system incorporates a new module that apply a set of heuristics to the representation of the topics in the way of logic forms.

According to others IR systems, IR-n system uses different techniques of the query expansion. Previous researches [1] have shown that the approaches get better results where they are based on passage retrieval in opposition to full document retrieval.

On the other hand, in present conference and for the ad-hoc track, a new technique called variable passages [3] has been implemented. It applies fusion methods which are used in multilingual tracks to combine results with different size of passages.

Following section shows in detail the main features of the treatment of topics in the way of logic forms performed by IR-n system. The process that automatically derives the logic form applying a set of inference rules to the analysis of dependencies between the words of the topic is also described.

3 Logic Form Derivation

To enhance the performance of our IR-n system we use the logic form of the topics. 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. MINIPAR [4] is the toolkit that obtains this analysis of dependency relationships between the words of the sentence. Following subsections describe the process of Logic Form Derivation applying this process to a topic as example.

3.1 Analysis of dependency relationships between words

This task obtains the different relationships between the words of the sentence. These dependency relationships between words are calculated by MINIPAR [4]. Figure 1 shows the dependence relationships between the words of the topic "*The story of Variant Fly and the Emergency Rescue Committee who saved thousands in Marseille*".

3.2 Logic Form Inference

The logic form of the sentence is calculated via this analysis of dependency relationships between the words of the sentence. Our approach employs a set of rules that infer several aspects such as the assert, the type of assert, the identifier of the assert and the relationships between the different asserts in the logic form. This technique improves the Moldovan technique [5] that constructs the logic form through the syntactic tree obtained from the output of the syntactic parser. Our logic form, as Moldovan logic form, is based in the format of logic form defined by eXtended WordNet [8]. The logic form "*story:NN(x14) of:IN(x14, x13) varian:NN(x10) NNC(x11, x10, x12) fry:NN(x12) and:CC(x13, x11, x6) emergency:NN(x5) NNC(x6, x5, x7) rescue:NN(x8) NNC(x7, x8, x9) committee:NN(x9) who:NN(x1) save:VB(e1, x1, x2) thousand:NN(x2) in:IN(e1, x3) marseille:NN(x3)*" is inferred by the application of our system based rules to the analysis of dependency relationships between the words of the topic "*The story of Variant Fly and the Emergency Rescue Committee who saved thousands in Marseille*". In this format of logic form

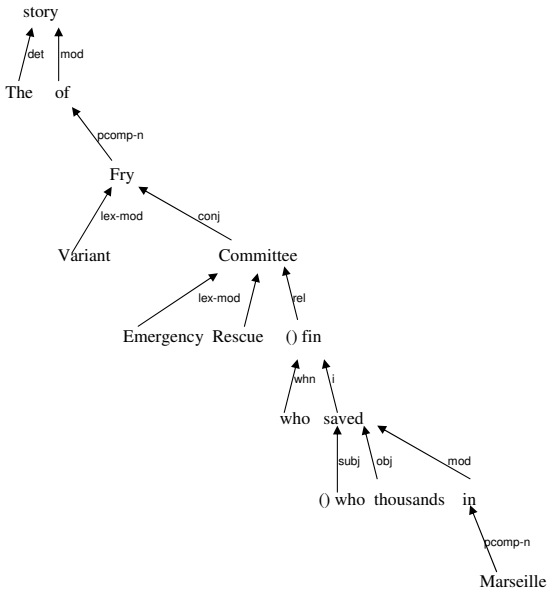

Figure 1: Analysis of dependency relationships between words of the topic

each assert has at least one argument. The first argument is usually instantiated with the identifier of the assert and the rest of the arguments are corresponded to the identifiers of other asserts that are related with this assert. As instance, in the assert "*story:NN(x14)*", its type is corresponded to noun *(NN)* and its identifier is instantiated to *x14*; in the assert "*NNC(x11, x10, x12)*", its type is corresponded to composed entity *(NNC)*, its identifier is instantiated to *x11*, and the other two arguments indicate the relationships with other asserts: *x10* and *x12*; and so on.

4 Applying rules to logic form to increment the topic terms weights

When the type of assert is a preposition *(IN)* which second argument instantiates an assert which type is noun *(NN)* or derives in a assert which type is noun, then the weight term associated to this last assert is modified. This rule generally describes those grammatically utterances that have a circumstantial behaviour in the sentence (ej. in Marseille, in concentration camps, in Sweden, of Holocaust experience and so on) and then we consider their words which POS are nouns (type of predicate NN) as very relevant words in the topic. This reason produces that we increment the weight terms of these words (terms) in 15%. Table 1 shows the terms weights that IR-n system associates to the topic the topic "*The story of Variant Fly and the Emergency Rescue Committee who saved thousands in Marseille*". These terms are expressed by their stem.

According to the rule described in this section, the logic form inferred for this topic ("*story:NN(x14) of:IN(x14, x13) varian:NN(x10) NNC(x11, x10, x12) fry:NN(x12) and:CC(x13, x11, x6) emergency:NN(x5) NNC(x6, x5, x7) rescue:NN(x8) NNC(x7, x8, x9) committee:NN(x9) who:NN(x1) save:VB(e1, x1, x2) thousand:NN(x2) in:IN(e1, x3) marseille:NN(x3)*") have two asserts which types are *IN*. The second argument of these asserts is instantiated to *x13* and *x3* respectively. *x13* derives in the asserts *x10*, *x12*, *x5*, *x8* and *x9* which types are *NN*, while the type of *x3* is directly *NN*. According to this rule, these fact produces that the terms weight associated to all these asserts increment their value in 15%. Table 2 shows the weight terms once this rule has been applied.

Term (stem)	Weight
stori	1.84449
fry	6.19484
emerg	6.47296
rescu	6.19484
committe	4.08194
save	3.06725
thousand	2.33944
marseil	5.13363

Table 1: Terms weights assigned by IR-n system

Term (stem)	Weight
stori	1.84449
fry	7.124066
emerg	7.443904
rescu	7.124066
committe	4.694231
save	3.06725
thousand	2.33944
marseil	5.9036745

Table 2: Terms weights according to logic form rules

5 Submitted Runs

This section describes the different submitted runs of our IR-n system. The differences between these five submitted runs are basically based in the treatment of the topics and in the indexation of a combination of different field of the segments in the document collection. In all submitted runs we use the indexing and searching processes developed by our IR-n system using the English as query language. There is not used any kind of thesaurus terms as keywords in the indexing and in the searching processes. Following subsections show the features of these five submitted runs according to the judgment pool priority order.

5.1 UATDASR04FL Run

In this run IR-n system indexes the automatically created transcript using the best presently available ASR system (**ASRTEXT2004A** field of the segments in the document collection). The English title and description fields of the topics are used in the construction of the queries. This is the unique submitted run in which we apply the rules based on the processing of queries in the way of logic forms described in previous section.

5.2 UATDASR04 Run

In this run, as previous submitted run, our IR-n system indexes the **ASRTEXT2004A** field of the segments in the document collection. The English description field of the topics is used in the construction of the queries.

5.3 UATDASR04AUTOA1 Run

In this run we index the **ASRTEXT2004A** field and a set of thesaurus keywords that were assigned automatically using a k-Nearest Neighbor (kNN) classifier using only words from the

ASRTEXT2004A field of the segment (**AUTOKEYWORD2004A1** field of the segments in the document collection). The English description field of the topic is used in the construction of the queries.

5.4 UATDASR04AUTOA2 Run

In this run IR-n system indexes the **ASRTEXT2004A** field and a a set of thesaurus keywords that were assigned using a different kNN classifier that was trained (fairly) on different data (**AUTOKEYWORD2004A2** field of the segments in the document collection). The English description field of the topic is used in the construction of the queries.

5.5 UATDASR04AUTOS Run

In this run our IR-n system indexes the **ASRTEXT2004A**, **AUTOKEYWORD2004A1** and **AUTOKEYWORD2004A2** fields of the segments in the document collection. The English description field of the topics is used in the construction of the queries.

6 Results

Table 3 shows the results obtained by our system for each one of the submitted runs. UATDASR04AUTOA2 is the submitted run that better scores has obtained in comparison with the rest of our submitted runs. In this run IR-n system indexes the **ASRTEXT2004A** and **AUTOKEYWORD2004A2** fields of the segments in the document collection.

run	map	rprec	bpref	rr	p5	p20	p100	p1000
UATDASR04LF	0,0768	0,1230	0,0949	0,4622	0,2160	0,1740	0,1088	0,0324
UATDASR04	0,0724	0,1246	0,0899	0,4377	0,1840	0,1660	0,1036	0,0313
UATDASR04AUTOA1	0,0727	0,1206	0,1018	0,4509	0,2800	0,1740	0,0916	0,0277
UATDASR04AUTOA2	0,0769	0,1181	0,0980	0,4744	0,2640	0,1920	0,0928	0,0290
UATDASR04AUTOS	0,0739	0,1274	0,1056	0,4354	0,2640	0,1880	0,0920	0,0260

Table 3: Evaluation Results

7 Conclusions

In this new release of the CL-SR track at the CLEF 2005 conference we have participated applying our IR-n system to the English language. Our main aim is to evaluate the goodness of the new Logic Form Module of IR-n system. According to our foresight, the obtained scores applying this module (UATDASR04LF) are upper than the obtained scores without the use of this new module (UATDASR04).

According to the format of the document collection, each document is considered as a sentence (continuous set of words). Then, this fact produces that IR-n system runs as a document retrieval system and not as a passage retrieval system. This feature resumes that the powerful of the use of the new logic form module must be combined with the passage overlapping technique in document collections where documents are composed by many passages (see our paper in the bilingual IR track at present conference). The combination of these two techniques would obtain better scores.

Acknowledgment
This research work has been partially funded by the Spanish Government under project CICyT

number TIC2000-0664-C02-02 and PROFIT number FIT-340100- 2004-14 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 *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, pages 108–121, Trondheim, Norway.

[2] Marcin Kaszkiel and Justin Zobel. Passage Retrieval Revisited. In *Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185.

[3] Fernando Llopis and Elisa Noguera. Combining Passages in Monolingual Experiments with IR-n system. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, in this volume, Vienna, Austria.

[4] MINIPAR parser. In *http://www.cs.ualberta.ca/ lindek/minipar.htm*

[5] Dan Moldovan and Vasile Rus. Logic Form Transformation of WordNet and its Applicability to Question-Answering. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July 2001.

[6] Jacques Savoy. Fusion of Probabilistic Models for Effective Monolingual Retrieval. In *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Trondheim, Norway.

[7] Rafael M. Terol, Patricio Martínez-Barco and Manuel Palomar. Applying Logic Forms to Biomedical Q-A. In *International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2005)*, pages 29–32, Istambul, Turkey, Juny 2004.

[8] eXtended WordNet. In *http://xwn.hlt.utdallas.edu/*