Applying Logic Forms and Statistical Methods to CL-SR Performance

R. M. Terol, P. Martínez-Barco and M. Palomar
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera de San Vicente del Raspeig - Alicante - Spain
Tel. +34965903653
{rafamt, patricio, mpalomar}@dlsi.ua.es

Abstract

This paper describes in detail the combination of NLP methods applied to the treatment of logic forms in the topic processing and statistical methods applied to the search engine in the frame of the CL-SR performance. The method that infers the logic form of a topic is based on dependency analysis between the words of the topic. These dependencies between the words of the topic are calculated using the MINIPAR parser. Different combinations of the topic, description and narrative fields are used in the runs to perform the retrieval process. The based on logic forms method processes the description and narrative fields of the topics. This processing task consists on the removal of several terms according to the logic structure of the processed field in the logic form. On the other hand, the statistical processing applied to the search engine consists on using IR-n system. IR-n system is a passage retrieval system that manages overlapping of variable passages that are composed by a number of sentences. Different statistical similarity measures are managed by IR-n system to acquire the topic terms weight. The removal of several topic terms according to the logic structure of the topic origines that the rest of the topic terms acquire a better relevance.

Keywords

Speech Retrieval, Information Retrieval, Logic Forms

1 Introduction

Different combinations of the topic, description and narrative fields can be applied to perform the information retrieval process. This fact implies that all the terms of these fields are used by the search engine to accomplish its goal. The search engine usually removes many terms that can be considered as stop-words (prepositions, articles and so on). If we have a look to the structure of the description and narrative fields of a topic (see table 1), we can deduce that there exists many terms that would not be as relevant as other terms in the information retrieval process. Our system processes the topics according to an NLP based approach. The topic processing basically consists on removing several terms of the description and narrative fields of the topic. Obviously, these removed terms are consider as not relevant and then they will not be processed by the information retrieval engine.

In this new Cross-Language Speech Retrieval (CL-SR) Track, our research effort has been focused on combining the use of NLP and statistical methods in the CL-SR performance. Concretely,

Topic	Description	Narrative
Child survivors in **Sweden**	Describe survival mechanisms of **children born** in **1930-1933** who **spend** the **war** in **concentration camps** or in ...	The relevant material should describe the **circumstances** and **inner resources** of the **surviving children**

Table 1: The most relevant terms of the topic (in bold)

our main research goal has been centered on demonstrating that the use of NLP methods by way of processing the topics according to the logic structure of their associated logic form increases the results obtained by the statistical search engine. We applied the new version of IR-n system [1] as statistical search engine.

The following section shows the topic processing by way of applying NLP rules based on the logic structure of their associated logic forms. Finally, we describe the submitted runs, the obtained results in these submitted runs, and discuss the application of NLP methods to the statistical IR-n system.

2 System Description

This section presents the topic processing applying NLP rules based on logic forms [2]. The format of the applied logic forms is based on the format of the logic form defined by eXtended WordNet [3]. For example, the associated logic form of the topic "The liberation of Buchenwald and Dachau" is instantiated as "liberation:NN(x4) of:IN(x4, x2) buchenwald:NN(x3) and:CC(x2, x3, x1) dachau:NN(x1)".

The topic processing by way of logic forms consists on removing many terms of the topic according to the logic structure of its logic form. A combination of the text, description and narrative fields of the topic has been employed to perform the information retrieval process according to the submitted runs. The rules are only applied to the description and narrative fields of the topic because these fields contains a lot of information (see table 1) that would be previously filtered before to be processed by the information retrieval process. These rules are independently applied to the description and narrative fields of the topic when the number of words of these fields are upper to 10 words. In other case, there is not necessary to remove any word (term).

These rules consist on the removal of the firsts words until a preposition (predicate type IN), or a main verb (predicate type VB or VBE), or a compositional structure (predicate type CC), both included. If a preposition or a verb are the firsts words in the sentence, we removed them and then the processing continues until finding another preposition, main verb or compositional structure. If in this search process the system detects a noun (predicate type NN) coinciding with the nouns of the topic field then the search process is aborted until this noun. Table 2 shows how an example of the application of these rules.

Once these rules are applied, the next process consists on performing the search in the document collection according to combination of updated fields by the application of these rules. The statistical IR-n system [1] accomplishes this goal.

3 Submitted Runs

This section describes the submitted runs in which our system has participated in. The differences between these five submitted runs are basically based on the combination of the topic fields and

Field	Logic structure
Describe survival mechanisms of children born in 1930-1933 who spend the war in concentration camps or in ...	describe:VB(e2, x11, e1) survival:NN(x1) NNC(x8, x1, x9) mechanism:NN(x9) of:IN(x8, x2) child:NN(x2) bear:VB(e1, x8, x10) in:IN(e1, x4) ...
The relevant material should describe the circumstances and inner resources of the surviving children ...	relevant:JJ(x1) material:NN(x1) describe:VB(e1, x1, x5) circumstance:NN(x6) and:CC(x5, x6, x3) inner:NN(x2) NNC(x3, x2, x4) resource:NN(x4) ...

Table 2: Removed terms of the field (in bold)

on the indexation of a combination of different segment fields from 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:

- **UA_TDN_FL_ASR06BA1A2 Run.** In this run IR-n system indexes the combination of the **ASRTEXT2006B**, **AUTOKEYWORD2004A1** and **AUTOKEYWORD2004A2** segment fields of the document collection. The English title, description and narrative topic fields are used in the construction of the queries. This was the unique submitted run in which we apply the rules based on the topic processing by way of logic forms described in previous section.
- **UA_TDN_ASR06BA1A2 Run.** In this run, as previous submitted run, IR-n system indexes the combination of the **ASRTEXT2006B**, **AUTOKEYWORD2004A1** and **AUTOKEYWORD2004A2** segment fields of the document collection. The English title, description and narrative topic fields are used in the construction of the queries.
- **UA_TD_ASR06BA2 Run.** In this run IR-n system indexes the combination of the **ASRTEXT2006B** and **AUTOKEYWORD2004A2** segment fields of the document collection. Only the title and description topic fields are used in the construction of the queries.
- **UA_TDN_ASR06BA2 Run.** In this run, as in previous run, IR-n system indexes the combination of the **ASRTEXT2006B** and **AUTOKEYWORD2004A2** segment fields of the document collection. The English title, description and narrative topic fields are used in the construction of the queries.
- **UA_TD_ASR06B Run.** In this required run, IR-n system only indexes the **ASRTEXT2006B** segment field of the document collection. Only the title and description topic fields are used in the construction of the queries.

4 Results

Table 3 shows the results obtained by our system for each one of the submitted runs. These scores demonstrate that the application of the NLP rules based on the logic structure of the topic fields improve the the results of the statistical IR-n system.

5 Conclusion

In this research we have demonstrated that the previous preprocessing of the topics according to NLP methods produces an improvement in the statistical retrieval process. The NLP methods are based on the logic structure of the narrative and description topic fields.

run	map	R-prec	bpref	rr	p5	p20	p100	p1000
UA_TDN_ASR06BA1A2	0.0369	0.0757	0.0651	0.1800	0.0882	0.1029	0.0821	0.0262
UA_TDN_ASR06BA1A2	0.0365	0.0714	0.0640	0.2255	0.0882	0.0956	0.0785	0.0260
UA_TD_ASR06BA2	0.0328	0.0727	0.0660	0.1681	0.1000	0.0868	0.0700	0.0264
UA_TDN_ASR06BA2	0.0345	0.0756	0.0691	0.1671	0.0765	0.0926	0.0774	0.0278
UA_TD_ASR06B	0.0339	0.0736	0.0792	0.2010	0.1235	0.1088	0.0709	0.0297

Table 3: Evaluation Results

Acknowledgment
This research work has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01.

References

[1] 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.

[2] 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.

[3] S. Harabagiu, G.A. Miller, and D.I. Moldovan. WordNet 2 - A Morphologically and Semantically Enhanced Resource. In *Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources*, Maryland, June 1999, pp.1-8.