# Supporting named entity recognition and syntactic analysis with full-text queries

Luísa Coheur, Ana Guimarães, Nuno Mamede

$L^2F$/INESC-ID Lisboa
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
{lcoheur,arog,nuno.mamede}@l2f.inesc-id.pt

**Abstract.** JaTeDigo is a natural language interface (in Portuguese) to a cinema database that has to deal with a vocabulary of more than 2500000 movies, actors and staff names. As our tools were not able to deal with such a huge amount of information, we decided to profit from full-text queries to the database to support named entity recognition (NER) and syntactic analysis of questions. This paper describes this methodology and evaluates it within JaTeDigo.

## 1 Introduction

JaTaDigo is a Natural Language Interface (in Portuguese) to a cinema database. Its database has information from IMDB, OSCAR.com and PTGate[1].

Regarding question interpretation, JaTeDigo runs over a natural language processing chain, responsible for a morpho-syntactic analysis and for a semantic interpretation based on domain terms. Due to the huge quantity of names that can be asked about, two problems arose: first, it was impossible to load that information in the dictionaries and run the system in a reasonable time; second, as the part-of-speech tagger suggests labels for unknown words – and names can be entered both in Portuguese and English –, the morpho-syntactic analysis became erratic, and it was impossible to build syntactic rules to capture those sequences.

Considering other NLIDB, the last decade brought some interesting and promising NLIDB/QA systems such as BusTUC [1] or Geo Query [2]. However, typically these NLIDB run over small databases, and they do not accept named entities in other languages than the query language. As so, after several experiments, and before going into sophisticated QA techniques, we decided to use full-text queries and see the results. In this paper we explore this hypothesis. Further details about JaTeDigo can be found in [3].

## 2 Overview of the methodology

1. The question – possibly without some words, such as interrogative pronouns – is submitted to the database in a full-text query;

[1] http://www.imdb.com/, http://www.oscars.org/awardsdatabase/ and http://www.cinema.ptgate.pt, respectively.

2. The resulting first 100 rows – value obtained empirically – are compared with the question sequences and named entities are captured;
3. A disambiguation step is performed, if there is more that one entity with the same name;
4. A local grammar is created and added to the main grammar;
5. A morpho-syntactic and a semantic analysis are performed over the question and if during morpho-syntactic analysis the part-of-speech tagger suggests wrong tags, the local grammar overrides them;
6. Syntactic and semantic analysis are performed, an SQL query is built and the answer is retrieved from the database.

## 3 Evaluation

For evaluation proposes, 20 users performed 198 questions. From these, 41 got no response and 157 were answered. From the 157, 7 got wrong answers. Although there were many causes for these errors, if names were correct and complete, NER based in full-text queries was 100% accurate. It should be noticed however that if names are misspelled or incomplete, results can be unanswered questions or – worse – wrong answers.

## 4 Conclusions and Future Work

In this paper, we have presented a methodology to support NER and a syntactic analysis, based on full-text queries. The questions is submitted to the database in a full-text query, named entities are identified from the returned results and a local grammar is created, overriding possible errors from the part-of-speech tagger. This method provides recognition of named entities with little effort and although a simple solution, it can be useful to NLIDB developers that are using tools that cannot deal with large vocabularies.

## Acknowledgment

This work was partially funded by FCT project CMU-Pt/0005/2007.

## References

1. Amble, T.: BusTUC - a natural language bus route oracle. In: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, Washington, USA, Association for Computational Linguistics (2000) 1–6
2. Kate, R.J., Wong, Y.W., Mooney, R.J.: Learning to transform natural to formal languages. In: AAAI. (2005) 1062–1068
3. Guimarães, R.: Játedigo – uma interface em língua natural para uma base de dados de cinema. Master's thesis, Instituto Superior Técnico (2007)