NLP for Shallow Question Answering of Legal Documents Using Graphs*

Alfredo Monroy1, Hiram Calvo1,2, and Alexander Gelbukh1

1 Center for Computing Research, National Polytechnic Institute
Mexico City, 07738, Mexico
alopezm301@ipn.mx, hcalvo@cic.ipn.mx, gelbukh@gelbukh.com
2 Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0192, Japan
calvo@is.naist.jp

Abstract. Previous work has shown that modeling relationships between articles of a regulation as vertices of a graph network works twice as better than traditional information retrieval systems for returning articles relevant to the question. In this work we experiment by using natural language techniques such as lemmatizing and using manual and automatic thesauri for improving question based document retrieval. For the construction of the graph, we follow the approach of representing the set of all the articles as a graph; the question is split in two parts, and each of them is added as part of the graph. Then several paths are constructed from part A of the question to part B, so that the shortest path contains the relevant articles to the question. We evaluate our method comparing the answers given by a traditional information retrieval system—vector space model adjusted for article retrieval, instead of document retrieval—and the answers to 21 questions given manually by the general lawyer of the National Polytechnic Institute, based on 25 different regulations (academy regulation, scholarships regulation, postgraduate studies regulation, etc.); with the answer of our system based on the same set of regulations. We found that lemmatizing increases performance in around 10%, while the use of thesaurus has a low impact.

1 Introduction

Previous work [20] has shown that modelling relationships between articles of a regulation as vertices of a graph network works twice as better than traditional information retrieval systems for returning articles relevant to the question. Despite being that approach language independent, in this work we experiment by using natural language techniques such as lemmatizing and using manual and automatic thesauri for improving question based document retrieval. We focus in Spanish language. For automatic thesaurus we used a distributional thesaurus [19], and for the manual thesaurus we used a human oriented dictionary (Anaya) [21]. The advantage of using a distributional thesaurus is that the approach remains language independent—not being the

* We thank the support of Mexican Government (SNI, SIP-IPN, COFAA-IPN, and PIFI-IPN) and Japanese Government. Second author is a JSPS fellow.
case with the human oriented dictionary. On the other hand, using a lemmatizer would make this approach language dependent, as for a particular a lemmatizer is needed. In particular, we want to measure the advantage of using these resources and we want to measure the possible benefit from adding this kind of information. Our system gives answers which consist of set of articles related to the question and also the relevant articles related with them to complement the answer. This is called shallow QA, because its operation lies in the middle of snippet retrieval and giving the exact answer.

We test our system with regard to a traditional vector space model information retrieval system to answer questions particularly for the Spanish language given a set of 25 regulation documents from the National Polytechnic Institute. For details of the rest of the System, see sections 2 and 3. The addition of NLP techniques is explained with further detail in Section 4. Evaluation and experiments are shown in Sections 5 and 6.

2 Related Work

There are not many works particularly devoted to the legal domain, despite of its wide use and application. Particularly for the legal domain, the workshop Question Answering for interrogating legal documents took place in 2003, in the framework of the JURIX Forum (The foundation for Legal Knowledge Based Systems). Several works showed that a common problem is that traditional Information Retrieval Methods are not adequate to find the relevant fragments which answer legal questions because they do not consider the logic relationships between articles. In addition, many questions require an answer which cannot be found explicitly in a single article, or fragments of them, but intrinsically in the relationship between articles [9,10]. Some works use logic inference mechanisms such as COGEX System [14] and the system by Quaresma et al. [7]. However, these systems need expensive resources such as ontologies, axioms, and are language dependent. To avoid such requirements, we use a graph for capturing the relationships between articles in regulations as proposed in [20].

3 System Design

The architecture of this SQAS is based on the work shown in [20]. It was designed considering common characteristics posed by regulation documents, as well as the kind of questions and answers expected by the user. It is important to mention that regulation texts have a defined structure, they are composed of chapters, and these, in turn, are subdivided in articles. This makes possible to use different techniques which with other kind of texts would not be possible. Articles from a single regulation text are related between them, and also there are links between different regulations.

We focus in questions where the answer can be given as a set of articles from a regulation. For example, for the question: Is it possible to award a honourable mention to a bachelor if he chose to graduate using the qualification option? the answer can be given as a set of articles: See Chapter II of “On Graduating Options” and article 13, Chapter VII, “On the Professional exam”, article 43. When one looks to such articles, they say: