Named-Entity-Tagger for Persian and Slovak

1 Introduction

Information extraction is one of the main tasks in Natural Language Processing (NLP) which mainly focuses on finding relations between named entities. The extracted relations are then used to build knowledge bases. Various NLP applications can benefit from such knowledge bases, such as question answering systems in such a way that they can answer complex questions which require information from different sources. But for now we are just going to investigate the first step for Persian and Slovak, that is "identifying the Named-Entities". :)

2 Data

Availability of a corpus of raw text about people biographies, is necessary so that we can test and evaluate our approach. This section describes the text collection used in our experiments.

2.1 Corpus

The corpus is supposed to be a collection of biographies in Persian and Slovak. It should contain biographies of scientists, poets and other prominent people around the world including both contemporary and ancient ones. It is very important for us that the corpus contains typical texts on biography and exhibits various writing styles. We are going to prepare one by merging several sources available on the web. We will develop a specific crawler to accumulate these texts. We will then normalize the text collection by removing HTML tags. It is worthwhile to mention that we already have a corpus but a thin one, consisting of just 1147 text files. We hope we will be able to enrich that much more. :)

According to what already exists, for some people, more than one biography is available in the corpus, but it is assumed that the biographical data, such as the `date-of-birth`, is the same (not necessarily in the form of representation) in all of them. The text of each biography in our current corpus may range from 12 to 6412 tokens. The distribution of the documents according to their length is depicted in Figure 1.

2.2 Pre-processing

One of the important factors in biographical text is co-reference resolution, since only in few sentences the person name appears in the text and in the rest of the cases only a pronoun is used to refer to the person. In such cases the pronouns should be replaced by the person name which the text is about.

2.3 Named Entity Tagging

Our current entity-tagging strategy consists of both gazetteers and regular expressions which one or both may be used to tag a specific entity type in the text. Entities in our ontology are listed in Table 1.

Figure 1: Corpus distribution according to the size of documents

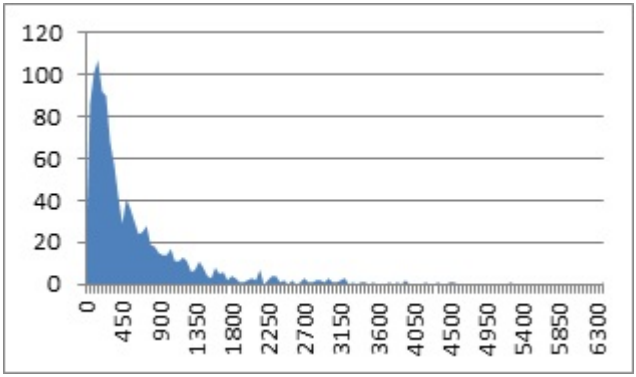

Table 1: Resources for named entity recognition

Entity	Method	Format
`place`	gazetteer + regular expression	<[PLACE:xxxxx]>
`country`	gazetteer	<[COUNTRY:xxxxx]>
`nationality`	gazetteer	<[NATIONALITY:xxxxx]>
`religion`	gazetteer	<[RELIGION:xxxxx]>
`literary-style`	gazetteer	<[STYLE:xxxxx]>
`literaty works`	regular expression	<[BOOKS:xxxxx]>
`date`	regular expression	<[DATE:xxxxx]>

The following lines show an example output of the tagger:

doctor <[PERSON:eric bern]> dar <[DATE:10 mei 1910]> dar šahre <[PLACE:montreāl]> be doniā āmad.
doctor <[PERSON:Eric Bern]> on <[DATE:10 May 1910]> in city <[PLACE:Montreal]> to world came

Doctor <[PERSON:Eric Bern]> was born on <[DATE:May 10, 1910]> in <[PLACE:Montreal]> city.