Named-Entity-Tagger for Persian and Slovak

1 Introduction

The goal of a named entity recognition (NER) system is to identify all textual mentions of the named entities. This can be broken down into two sub-tasks: identifying the boundaries of the NE, and identifying its type. While named entity recognition is frequently a prelude to identifying relations in Information Extraction, it can also contribute to other tasks.

The following lines show an example output of a tagger in Persian (written in English scripts):

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.

2 Steps

- Manually tagging a training, development, and test set for both languages so that we can test our taggers' performance. (quite laborious task!)
- Making gazetteers and regular expressions for some named-entities
- Implementing one or more statistical approaches for NER and evaluating them over Persian and Slovak Evaluating the taggers
- Developing a user interface that prints a list of the named entities, which texts they occurred in, and some context information like patterns of the context. (secondary goal)

3 Candidate Statistical Strategies

- We might build a trigram hidden Markov model to identify different named entities.
- We might train a (global) linear model for named-entity recognition using the perceptron algorithm.
- We might be able to train one of the available tools using our data.