Named-Entity Tagger for Persian and Slovak

Omid Moradiannasab and Michal Petko
NLP Tools for Low-Resource Languages, Saarland University, Germany
omidm@coli.uni-saarland.de, miso231@gmail.com

1 Introduction

What is NE tagging?
Here is a sample input and output for a named entity tagging tool:

INPUT: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT: Profits soared at Boeing Co. , easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

What are the primary goals of this project?

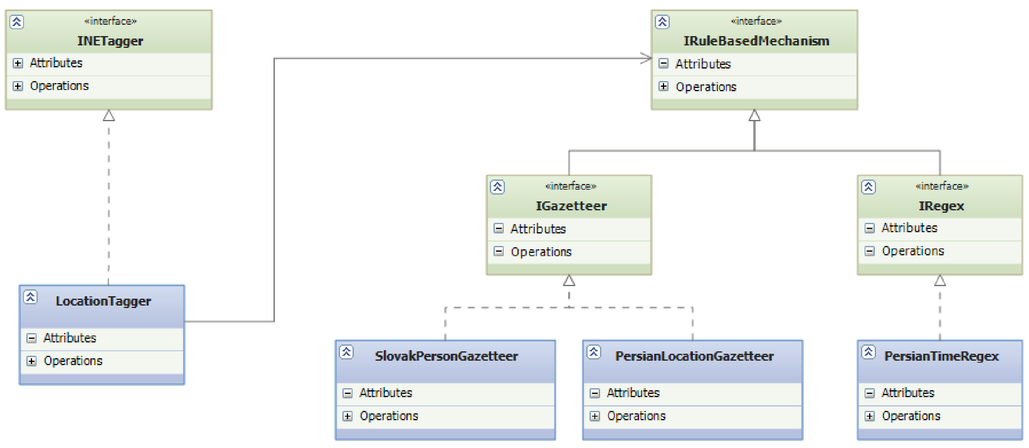
- Making named-entity taggers for Persian and Slovak using rule-based approaches (gazetteers and regular expressions)
- Tagging a corpus for each of the two languages using the rule-based taggers
- Evaluating the results

2 Datasets

- The Bijankhan POS tagged Corpus for Persian
- HC corpora for Slovak

3 Structure

Class diagram of the application



4 Solved problems

- Sufix problem in Slovak
- Excluding adjectives

  v slovenskom meste, Česko spolu so Slovenskom

- Ambiguity

  Martin (male name and also city)
  Omid (male name and also "hope")

- Boundary errors

  Národná rada Slovenskej republiky

- Nicknames / name's abbreviation

  → Positive meaning

  Michal → Michalko, Miško
  Katarína → Katka

  → Negative meaning

  Michal → Mišo
  Katarína → Kata

  → Crazy change

  Alexander → Saška, Šaňko / Saša, Šaňo

- No capital letter for proper nouns in Persian

5 Results

- Manual evaluation
- Test-set statistics:

  → **Slovak**

  35975 words
  data from Twitter, blogs and online news

  → **Persian**

  around 600 paragraphs for Persian including: articles, interviews, scientific papers, stories, ...

- results:

	Slovak			Persian		
	Person	**Location**	**Temporal Exp.**	**Person**	**Location**	**Temporal Exp.**
boundary errors	7	8	6	53	12	23
tags	216	218	111	103	96	107
precision	0.84	0.84	0.96	0.93	0.92	0.97
recall	0.51	0.73	0.95	0.82	0.92	0.83
F1 score	0.63	0.78	0.96	0.87	0.92	0.89

boundary errors are counted as true positives

6 Still-to-be-solved problems

- Problems (Person)

  → **Slovak**

  No surnames in gazetteer
  Name initials (e.g. M.R. Štefánik)
  Misspelling
  Names of writers, artists etc. (e.g. Shakespeare, Picasso)

  → **Persian**

  Enormous number of surnames, not easy to handle with just a gazetteer
  Surnames could be consisted of more than one part
  Efficient tokenization is still a problem in Persian which propagates error to NER as well
  Proper nouns do not start with capital letters and this makes it defficult to detect them
  Most of the names have a general meaning as well and could be used in their general sense in the text. e.g. Omid means hope in English

- Problems (Location)

  States abbreviations (e.g. UK, SVK, RUS)
  Insufficient words in gazetteer
  Names of the streets, boulevards, squares etc. are too many to be just dealt with gazetteers
  Misspelling

- Problems (Temporal Expression)

  range: $\langle 1900, 2099 \rangle$

7 Future Work

- Covering more entity types and including a hierarchical entity set
- Extracting some contextual information for every entity type
- Using such contextual information as a means to iteratively enrich gazetteers
- Training some statistical approaches using our tagged corpora

8 References

- "Tagging Problems, and Hidden Markov Models", Michael Collins, Columbia University
- "Information Extraction and Named Entity Recognition", Christopher Manning, Stanford University
- "Investigation on a Feasible Corpus for Persian POS Tagging", Hadi Amiri, Hosein Hojjat, Farhad Oroumchian. 12th international CSI computer conference, Iran, 2007
- HC Corpora, Hans Christensen, URL: <http://www.corpora.heliohost.org/>