Determining the Parts of Speech in Slovak Language

Dalibor Mészáros, supervisor: Márius Šajgalík

STU FIIT

PeWe@FIIT
personalized web group

Overview

- We focus on Part-of-Speech tagging in Slovak language
- We try to adapt the current State-of-the-Art approaches in English language (97,3% accuracy)
- Solution based on Conditional Random Fields
- We evaluate our approach on annotated dataset obtained from Slovak Corpus of Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Conditional Random Fields

- Belongs to statistical graph models - discriminative models
- Can be used to categorize words into parts of speech
- Learns relationships between tags, words and features
- Can predict labels with learned weights
- Always labels an unknown word with “something”
- Model represents sequences, rather than individual words
- Seems superior to a dictionary solution in many ways

Results

- The CRF model has been trained on 515,624 words
- The accuracy has been evaluated on 141,939 words

Word Case	Uppercase	Digits or Punct.	Word Length	Sentence Position	Prefix & Suffix Length	Accuracy	Unknown Words Acc.
Lowercase	Yes	No	No	No	3	95,60%	84,34%
Lowercase	Yes	No	No	No	4	94,97%	81,04%
Lowercase	Yes	No	No	No	1,2,3	96,04%	86,44%
Lowercase	Yes	No	No	No	1,2,3,4	96,41%	87,16%
Lowercase	Yes	Yes	Yes	Yes	1,2,3	96,25%	86,54%
Lowercase	Yes	Yes	Yes	Yes	1,2,3,4	96,63%	87,34%
Unchanged	No	No	No	No	3	94,94%	81,53%
Unchanged	No	No	No	No	4	94,31%	78,34%
Unchanged	No	No	No	No	1,2,3	95,84%	85,51%
Unchanged	No	No	No	No	1,2,3,4	96,27%	86,49%
Unchanged	Yes	Yes	Yes	Yes	1,2,3	96,11%	85,95%
Unchanged	Yes	Yes	Yes	Yes	1,2,3,4	96,51%	87,04%

Features

- CRF views words and tags as mere strings
- Therefore we need additional information to support CRF model
- Features have direct impact on achieved accuracy and performance

Word transformation

- Word converted into lowercase

Binary Features

- Which characters are uppercase
- Word contains digits or punctuation marks

Simple Features

- Word’s length and sentence position
- Word’s prefix and suffix of set length

Vector Representation of Words

- We use feature vectors, to create additional relationships

Label	Word	1st Up.	Full Up.	Semi Up.	Punct.	Digit	Length	Position	Prefix 1	Prefix 2	Prefix 3	Prefix 4	Suffix 1	Suffix 2	Suffix 3	Suffix 4
0	1685	0	0	0	0	1	4	4	1	16	168	1685	5	85	685	1685
R	si	0	0	0	0	0	2	5	s	si	si	si	i	si	si	si
S	Turci	1	0	1	0	0	5	6	T	Tu	Tur	Turc	i	ci	rci	urci
V	podrobili	0	0	0	0	0	9	7	p	po	pod	podr	i	li	ili	bili
T	aj	0	0	0	0	0	2	8	a	aj	aj	aj	j	aj	aj	aj
A	Nové	1	0	1	0	0	4	9	N	No	Nov	Nové	é	vé	ové	Nové
S	Zámky	1	0	1	0	0	5	10	Z	Zá	Zám	Zámk	y	ky	mky	ámky
Z	.	0	0	0	1	0	1	11	.	.	.	.	.	.	.	.