Lemmatization and Morphosyntactic Tagging of Croatian and Serbian

Željko Agić* Nikola Ljubešić* Danijela Merkler†

*Department of Information and Communication Sciences
†Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb

BSNLP 2013, Sofia, 8 August 2013

Motivation

- Croatian a highly flective language
- no freely available morphosyntactic tagger and lemmatizer
- starting NLP research almost impossible
- Serbian very similar, same situation regarding basic language technologies
- natural idea – tag data and train stochastic models
- dataset – SETimes corpus, *news and views of Southeast Europe* in ten languages, contains both Croatian and Serbian – parallel data with possibility of annotation projection

Corpus construction and annotation

- SETimes corpus
  `http://www.nljubesic.net/resources/corpora/setimes/`
- pre-annotated with the Croatian Lemmatization Server (HML)
- disambiguated and additionally annotated by experts
- HML tagset adopted to MTEv4
- draft of a new tagset developed – MTEv5
  `http://nl.ijs.si/ME/V5/msd/html/`
- homonymy numbering left out from lemmatization (*biti1*, *biti2*)
- corpus published under CC-BY-SA license on
  `http://nlp.ffzg.hr/resources/corpora`

Stats for SETIMES.HR corpus and test sets

Corpus	Sent's	Tokens	Types	Lemmas
SETIMES.HR	4 016	89 785	18 089	8 930
set.test.hr	100	2 297	1 270	991
set.test.sr	100	2 320	1 251	981
wiki.test.hr	100	1 887	1 027	802
wiki.test.sr	100	1 953	1 055	795

Tagset variation in tag counts

		set.test		wiki.test	
Tagset	SETIMES.HR	hr	sr	hr	sr
MTE v4	660	235	236	188	192
MTE v5	663	233	234	192	195
MTE v5r1	618	213	216	176	180
MTE v5r2	634	216	217	178	181
MTE v5r3	589	196	199	162	166

Experiment setup

1. tagger and lemmatizer selection experiments
    - use freely available tools for building and applying statistical models
    - tool selection on set.test.hr
    - BTagger, CST, HunPos, PurePos, SVMTool, TreeTagger
2. tagset selection experiments
    - use only the best performing tool(s)
    - tagset – v4 vs. v5, three reductions
    - language – Croatian, Serbian
    - domain – in-domain, out-of-domain

Tagger and lemmatizer selection experiment

Tool	Lem.	MSD	Train (sec)	Test (sec)
BTagger	96.22	86.63	24 864.47	87.01
CST	97.78	/	1.80	0.03
+ lex	97.04	/	1.87	0.12
HunPos	/	87.11	1.10	0.11
+ lex	/	84.81	10.79	0.45
PurePos	74.40	86.63	5.49	4.42
SVMTool	/	84.99	1 897.08	3.28
TreeTagger	90.51	85.07	7.49	0.19
+ lex	94.12	87.01	17.48	0.31

Tagging accuracy

	set.test		wiki.test	
POS	hr	sr	hr	sr
HunPos	97.04	95.47	94.25	96.46
+ lex	96.60	95.09	94.62	95.58
MSD				
HunPos	87.11	85.00	80.83	82.74
+ lex	84.81	81.59	78.49	79.20

Lemmatization accuracy

	set.test		wiki.test	
Model	hr	sr	hr	sr
CST	97.78	95.95	96.59	96.30
+ lex	97.04	95.52	96.38	96.61

Tagset selection experiment

Tagset	set.test		wiki.test	
POS	hr	sr	hr	sr
MTE v4	96.08	94.61	93.96	95.85
MTE v5	97.04	95.52	94.30	96.40
MTE v5r1	97.04	95.47	94.25	96.46
MTE v5r2	97.00	95.60	94.20	96.30
MTE v5r3	97.13	95.56	94.09	96.15
MSD				
MTE v4	86.24	83.45	80.45	81.98
MTE v5	86.77	84.48	80.46	82.43
MTE v5r1	87.11	85.00	80.83	82.74
MTE v5r2	87.11	84.96	81.20	82.38
MRE v5r3	87.72	85.56	81.52	82.79

Lemmatization accuracy on different tagsets

	set.test		wiki.test	
Tagset	hr	sr	hr	sr
MTE v4	97.78	95.82	96.66	96.11
MTE v5	97.82	95.86	96.81	96.30
MTE v5r1	97.78	95.95	96.59	96.30
MTE v5r2	97.87	95.99	96.75	96.20
MTE v5r3	97.74	95.99	96.54	96.20

Statistical significance of differences in full MSD tagging

- approximate randomization with 1000 iterations

Tagsets	v5	v5r1	v5r2	v5r3
v4	0.268	<0.05	<0.05	<0.01
v5	/	<0.01	<0.05	<0.01
v5r1	/	/	0.877	<0.05
v5r2	/	/	/	<0.01

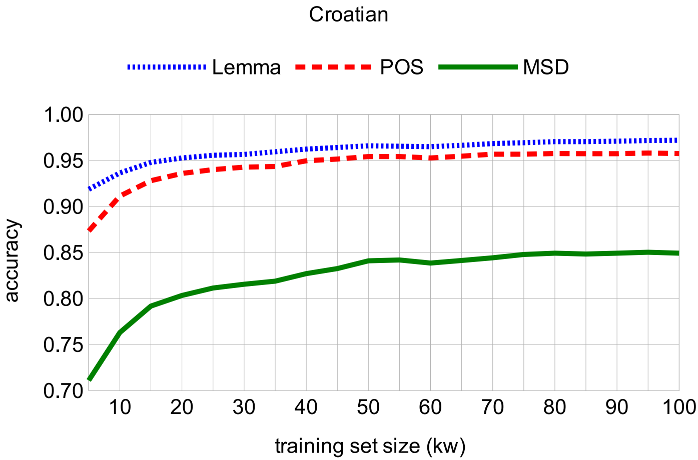
Precision, recall and F_1

	Croatian			Serbian		
POS	P	R	F_1	P	R	F_1
Adj	66.80	63.83	65.28	66.79	66.54	66.66
Adv	84.56	82.73	83.63	82.57	73.77	77.92
Conj	94.12	92.66	93.38	96.89	94.28	95.57
Noun	76.78	77.30	77.04	75.38	76.30	75.84
Num	91.30	94.38	92.81	94.19	91.01	92.57
Prep	95.93	97.52	96.72	94.30	94.55	94.42
Pron	81.85	83.20	82.52	81.43	82.83	82.12
Verb	93.81	95.96	94.87	93.36	93.84	93.60

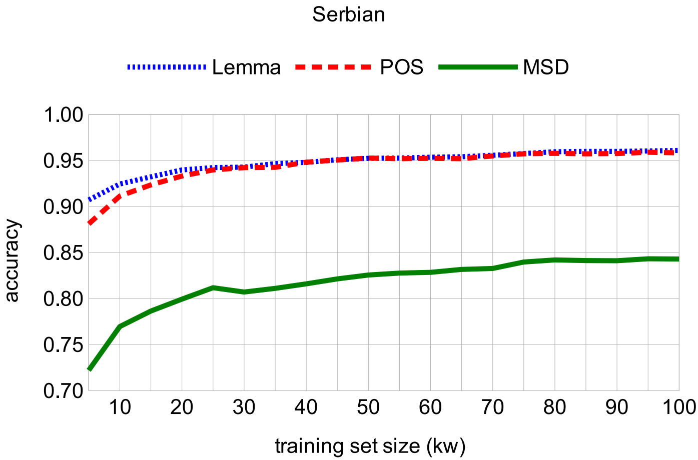
POS confusion matrix

POS	Abbr	Adj	Adv	Conj	Noun	Num	Part	Prep	Pron	Res	Verb
Abbr		0	0	0	1	3	0	0	0	0	0
Adj	0		20	0	50	0	1	0	3	1	4
Adv	0	10		9	12	0	0	2	0	0	2
Conj	0	0	5		2	0	5	5	7	0	0
Noun	0	37	28	0		4	0	1	5	7	25
Num	2	4	0	0	2		0	0	0	0	0
Part	0	0	0	3	0	0		0	0	0	3
Prep	0	0	2	3	2	0	1		0	0	0
Pron	0	2	1	9	3	0	1	0		0	1
Res	0	0	1	0	4	0	0	2	0		0
Verb	0	9	4	0	35	1	2	1	0	1	

Learning curves



Learning curves



Lemmatization and Morphosyntactic Tagging of Croatian and Serbian

Željko Agić* Nikola Ljubešić* Danijela Merkler†

*Department of Information and Communication Sciences
†Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb

BSNLP 2013, Sofia, 8 August 2013