Parsing Croatian and Serbian by Using Croatian Dependency Treebanks

Željko Agić, Danijela Merkler, Daša Berović

University of Zagreb
Faculty of Humanities and Social Sciences

2013-10-18
SPMRL 2013, Seattle, WA, USA

Talk overview

- Motivation
- Workflow
- Resources and tools
- Experiment setup
- Results
- Conclusions
- Future work

Motivation

Croatian and Serbian language

- South Slavic languages
- highly inflectional
  - morphosyntactic tagsets with 1500+ tags
- relaxed word order
  - mainly SVO
    Ivo pije vodu. Ivo vodu pije. Vodu Ivo pije.
    Vodu pije Ivo. Pije vodu Ivo. Pije Ivo vodu.
- BCS macro-language
  - Bosnian, Croatian, Montenegrin, Serbian
  - 20+ M native speakers
  - mutually intelligible
  - real and policy-induced differences

Motivation

State of the art in Croatian and Serbian parsing

- Croatian
  - preprocessing resources exist
  - prototype chunker
  - Croatian Dependency Treebank (HOBS)
  - parsing using standard parsers
    - MST beats Malt due to non-projectivity
    - accuracy at 73% LAS
    - k-best parsing and valency lexicon reranking adds 3 LAS points
- Serbian
  - preprocessing resources exist
  - rule-based NP-chunking
  - no experiments in parsing, dependency or otherwise

Motivation

General observations

- both languages under-resourced
- very limited availability
  - web services
  - restrictive licensing
  - HOBS is BY-NC-SA 3.0, but without syntactic tags
- perspective
  - under-resourced + unavailable = ?
  - regional monopoly, general invisibility

Workflow

- build on language similarity
  - develop free resources and test them on both languages
- our recent development
  - created SETimes.HR corpus
  - manual preprocessing
  - manual syntactic annotation using novel formalism
  - state of the art in lemmatization and tagging
  - freely available (CC-BY-SA 3.0)
- in this contribution
  - enlarged the treebank
  - created dependency parsing test sets for both languages
  - compared with HOBS
  - observed influence of preprocessing and features

Treebanks

- Croatian Dependency Treebank (HOBS)
  - newspaper text
  - Multext East v4 morphosyntactic tagset (MTE v4)
  - syntactic annotation as in Prague Dependency Treebank
    - two versions: HOBS, HOBS + Sub
    - HOBS + Sub has additional tags for subordinate clauses
  - tenfold cross-validation: 73% LAS
  - available to us as MSTParser models
- SETimes.HR treebank of Croatian
  - newspaper text from SETimes
  - MTE v4 and MTE v5 morphosyntactic annotation
  - 15-tag HOBS-based syntactic formalism
  - tenfold cross-validation: 80% LAS

Treebanks

Statistics for Croatian treebanks

- note throughout the experiment that HOBS is more than 25% larger than SETimes.HR

Features	HOBS	HOBS + Sub	SETimes.HR
Sentences	4 626	4 626	3 853
Tokens	117 369	117 369	86 991
Types	25 038	25 038	17 723
Lemmas	12 388	12 388	8 773
MSD tags	914	911	662
Syn. tags	27 (70)	28 (81)	15

Test sets

- four text samples
  - {newspaper, Wikipedia} x {Croatian, Serbian}
  - 100 sentences each
  - parallel sentences where applicable
- manually annotated
- MTE v4 and MTE v5 morphosyntactic tagset
- HOBS, HOBS + Sub and SETimes.HR syntactic tagset
- language difference
  - measured using a Croatian inflectional lexicon
  - Croatian: 4% OOV, Serbian: 12% OOV
  - evaluated by native speakers of Croatian and Serbian

Test sets

Test set statistics

	set.test		wiki.test	
Features	hr	sr	hr	sr
Sentences	100	100	100	100
Tokens	2 285	2 308	1 878	1 947
Types	1 265	1 246	1 027	1 055
Lemmas	989	979	803	797
MSD tags				
MTE v4 tags	236	237	189	193
MTE v5 tags	233	234	192	195
Syntactic tags				
HOBS	22 (37)	23 (37)	22 (41)	22 (44)
HOBS + Sub	22 (46)	24 (49)	23 (49)	22 (50)
SETimes.HR	15	15	15	15

Experiment setup

- experiments
  - comparison of syntactic formalism
  - influence of lemmatization and morphosyntactic tagging
  - impact of morphosyntactic features
- parser
  - MSTParser system
  - non-projective MST parsing (CLE)
  - setup imposed by previous experiments with HOBS
  - not a parser evaluation

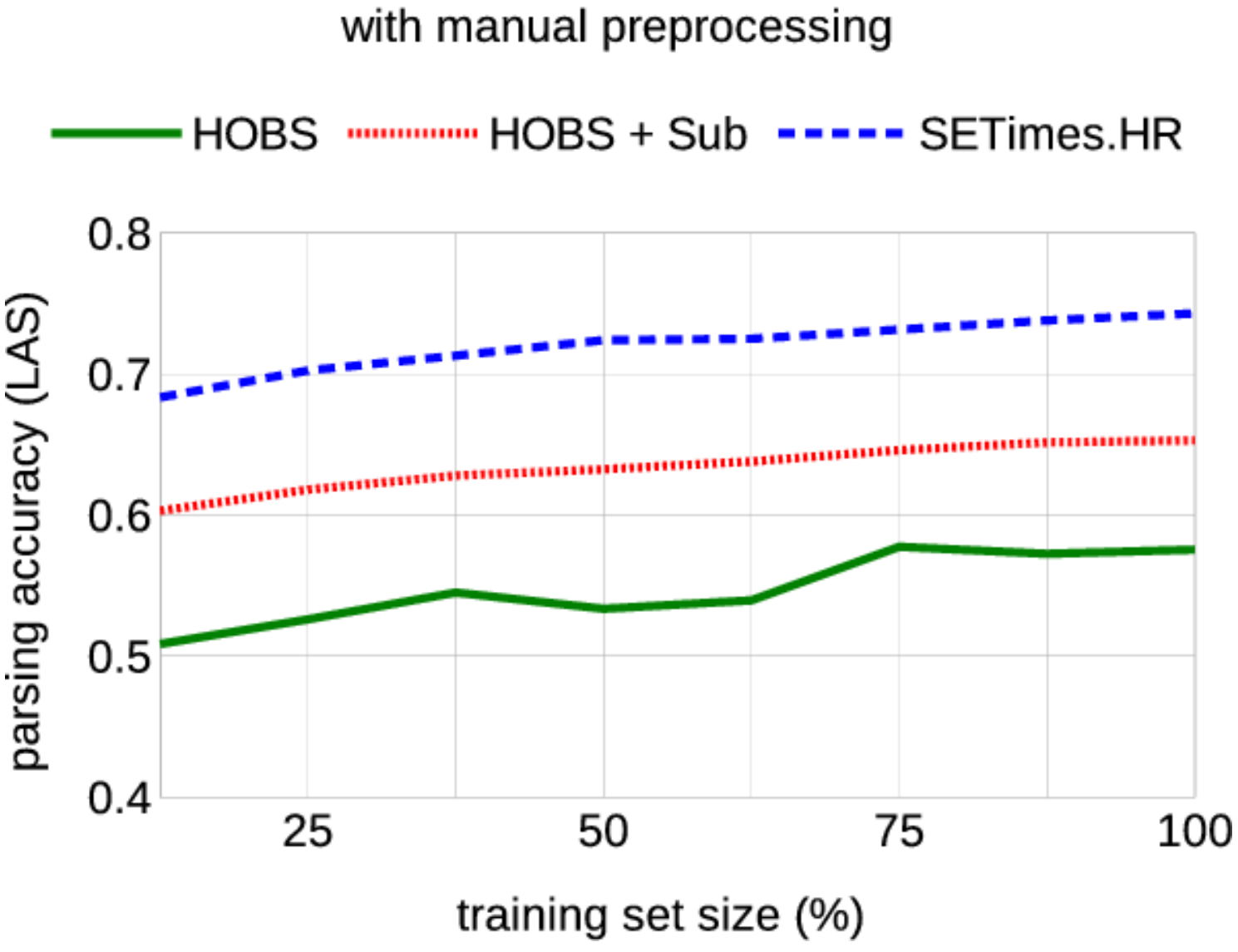
Results

Parsing accuracy with perfect preprocessing

	set.test		wiki.test		
LAS	hr	sr	hr	sr	overall
HOBS	59.9	58.7	55.5	55.4	57.6
HOBS + Sub	68.3	66.9	62.4	62.7	65.3
SETimes.HR	**76.7**	**75.4**	**71.9**	**72.4**	**74.3**
UAS					
HOBS	73.7	75.9	72.3	72.6	73.8
HOBS + Sub	78.1	79.0	76.5	76.5	77.6
SETimes.HR	**81.6**	**80.6**	**80.0**	**80.6**	**80.8**

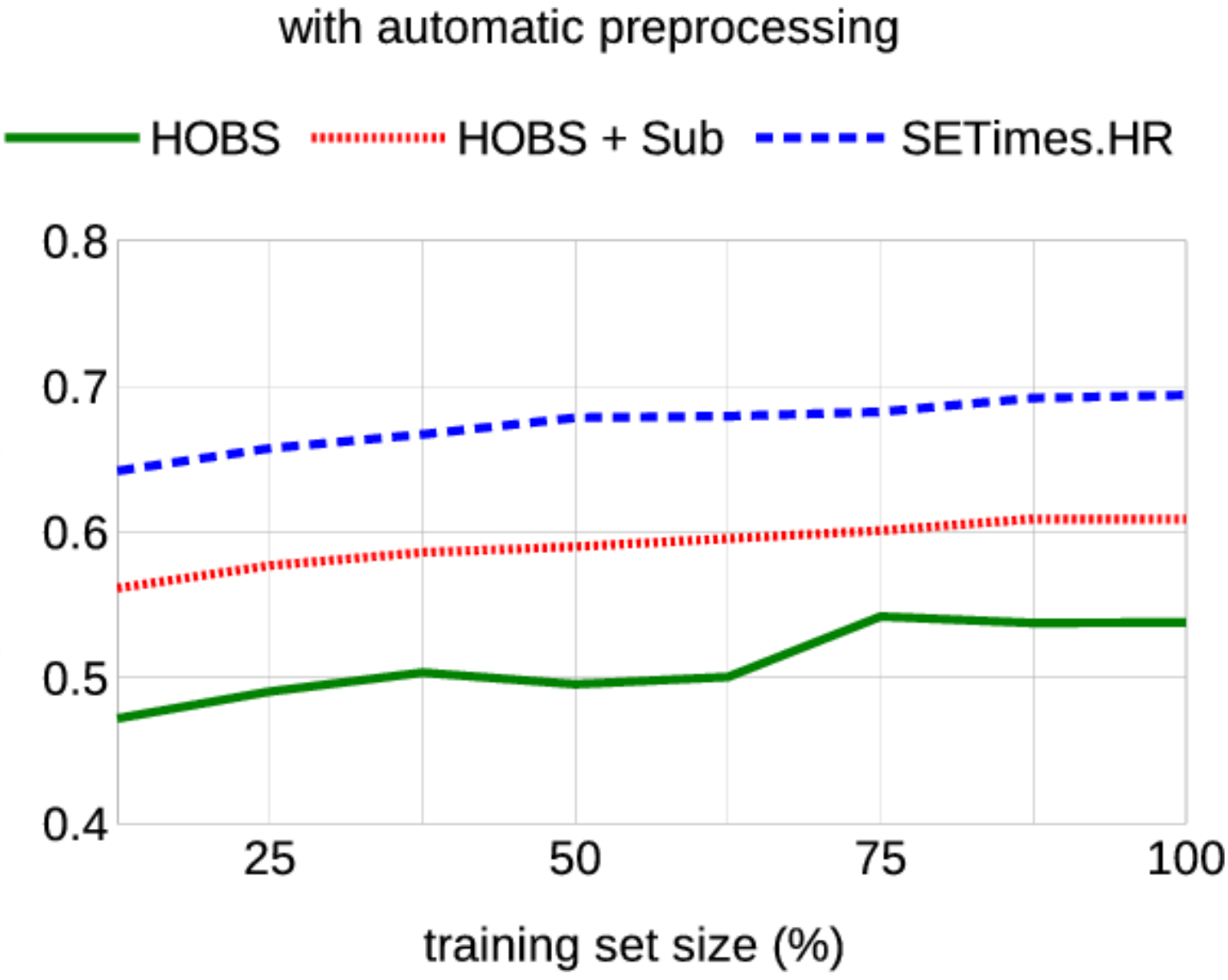
Results

LAS learning curves



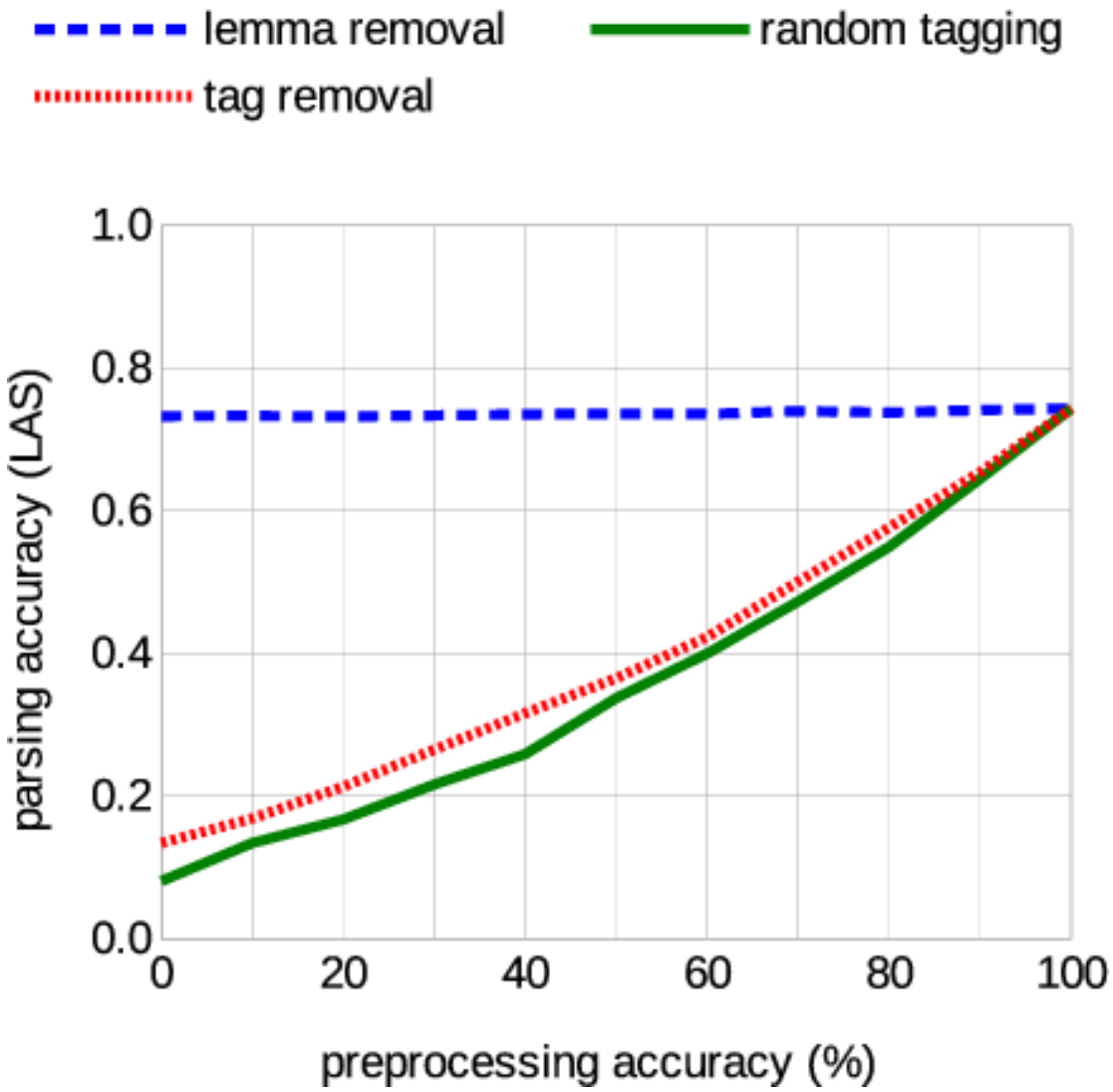
Results

LAS learning curves



Results

LAS as function of preprocessing accuracy for SETimes.HR



Results

LAS for main syntactic categories

	HOBS + Sub		SETimes.HR	
Syntactic tag	hr	sr	hr	sr
Adverb	**50.4**	46.6	**50.4**	**47.2**
Attribute	81.4	82.3	**87.9**	**88.4**
Object	56.4	51.3	**68.9**	**70.2**
Predicate	75.1	71.9	**80.7**	**81.2**
Preposition	65.5	**66.4**	**66.4**	64.0
Subject	70.3	71.3	**74.8**	**77.6**

Results

Impact of morphosyntactic features

	Croatian		Serbian	
Features	LAS	UAS	LAS	UAS
Adjective				
Type	74.3	80.7	**74.6**	**81.2**
Degree	74.3	80.7	73.7	80.2
Gender	74.1	80.7	**74.5**	**81.0**
Number	74.5	**81.0**	**74.3**	**80.8**
Case	**75.0**	**81.5**	**74.4**	**81.1**
Noun				
Type	74.3	80.8	72.9	80.0
Gender	74.4	80.8	74.1	**80.7**
Number	74.1	80.7	74.0	**80.7**
Case	73.3	**81.0**	72.3	80.0
Verb				
Type	**74.6**	**81.3**	**74.3**	**80.8**
Form	74.3	80.9	**74.3**	**81.0**
Person	74.3	**81.0**	73.5	80.0
Number	74.4	80.8	74.1	80.6
Gender	74.4	80.8	**74.4**	**81.0**
Full feature set	74.5	80.9	74.1	80.6

Results

LAS confusion matrix for Croatian (bottom left) and Serbian (top right)

	Adv	Ap	Atr	Atv	Aux	Co	Elp	Obj	Oth	Pnom	Pred	Prep	Punc	Sb	Sub
Adv		0	15	1	0	2	2	5	13	2	1	3	0	2	2
Ap	1		10	0	0	0	2	3	0	1	0	0	0	5	0
Atr	23	9		6	1	0	14	23	3	3	3	0	0	25	2
Atv	0	1	6		0	0	0	0	0	1	26	0	0	1	0
Aux	0	0	0	0		1	0	0	0	0	28	0	0	0	1
Co	0	0	1	0	0		0	0	5	0	0	2	11	0	0
Elp	1	2	12	0	0	0		0	4	3	2	0	0	4	0
Obj	6	3	16	3	0	0	1		0	1	1	0	0	2	0
Oth	14	4	3	0	0	12	1	1		0	0	1	0	1	24
Pnom	3	0	8	0	0	0	3	0	0		24	1	0	3	0
Pred	1	0	2	5	26	0	0	1	1	23		0	0	0	0
Prep	1	0	0	0	1	1	0	0	2	0	0		0	0	0
Punc	0	0	0	0	0	17	0	0	0	0	0	0		1	1
Sb	2	11	26	1	0	0	5	1	4	4	1	0	0		1
Sub	1	0	0	0	0	0	0	0	2	0	0	0	0	0	

Conclusions

- parsed Croatian and Serbian text by using dependency parsing models trained on Croatian data
  - Croatian and Serbian mutually parseable
  - LAS: 74.5% and 74.1%
- domain influence more substantial than language difference
  - 2-5 points LAS across the two domains
  - need for domain adaptation
- parsing remains robust in light of preprocessing
  - 3 points LAS decrease across the domains and languages
- morphosyntactic tagset requires adaptation for parsing
  - observed hindering features
  - adjective number and case, verb type
  - noun subtags favor parsing
- our work downloadable at `http://nlp.ffzg.hr/`

Future work

- treebank enlargement
  - added 800 new sentences after this experiment
  - new domains: business, IT
- domain adaptation
  - try at least the baseline approaches
- better parsers
  - many good parsers out there, do the parser evaluation
  - recently switched to `mate-tools`: +3 LAS points
- using all treebanks
  - experiments in combining diverse treebanks
  - use HOBS and HOBS + Sub as additional features for parsing using SETimes.HR formalism
- tagset design
  - the search for the best MTE v5 subset
  - underway: conversion to (Universal) Stanford Dependencies
- get included in SPMRL shared tasks

Thank you for your attention. ☺