Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets

Željko Agić, Jörg Tiedemann
Kaja Dobrovoljc, Simon Krek, Danijela Merkler, Sara Može

LT4CloseLang / Doha, Qatar, 2014-10-29

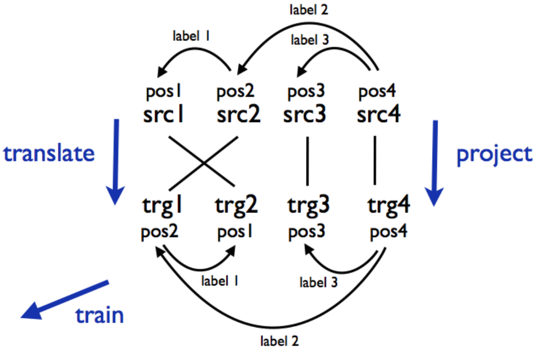
Introduction

How to parse a language for which no treebanks exist?

- Cross-lingual parsing
  - Source language: well-resourced, target: under-resourced
  - Annotation projection
    Parse the source side of the parallel corpus, project the annotations to the target side. If the source side is a treebank, even better.
  - Model transfer
    Apply the source side parser to the target side, possibly with adaptations.

Introduction

- Pros and cons?
  - Projection: parser noise, annotation transfer, corpora availability
  - Transfer: shared feature representation
- Alternatives
  - Lexical features via bilingual dictionaries (Durrett et al., 2012)
  - Synthesize treebanks via full-scale SMT (Tiedemann et al., 2014)



Introduction

- Best of both worlds?
  - ✓ Manual annotations get projected
  - ✓ No need for a shared feature representation
  - ✓ Word alignments from SMT are possibly more reliable
  - ✗ Projection heuristics
  - ✗ Availability and quality of SMT
- Beats baseline model transfer substantially
  - Compared to McDonald et al. (2013)
  - Improvements up to 7 points LAS, all language pairs improved
- New test case
  - Actual under-resourced languages
  - Closely related
  - Rich feature representations

Datasets

- Croatian – *hr*, Slovene – *sl* and Serbian – *sr*
  - Slavic languages
  - Rich inflectional morphology, relatively free word order
  - Different stages of under-resourcedness
- The four treebanks

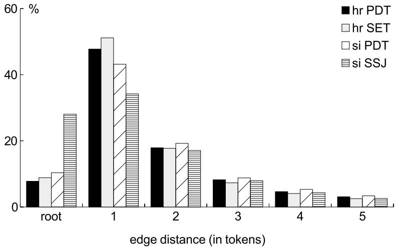
pdt Prague-style annotation scheme

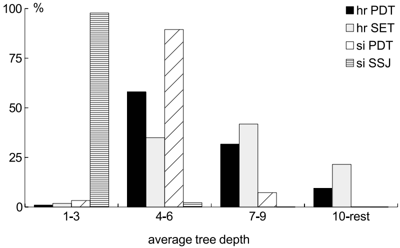
set, ssj Simplified schemes adapted from pdt for Croatian and Slovene

Feature	*hr* pdt	*hr* set	*sl* pdt	*sl* ssj
Sentences	4,626	8,655	1,534	11,217
Tokens	117,369	192,924	28,750	232,241
Types	25,038	37,749	7,128	48,234
POS tags	13	13	12	13
MSD tags	821	685	725	1,142
Syntactic tags	26	15	26	10

Datasets

- Treebank diversity





- Morphosyntactic tagset

Language	MSD tag	Attribute-value pairs
hr	`Vmn`	Category = **V**erb, Type = **m**ain, Vform = i**n**finitive
sl	`Vmen`	Category = **V**erb, Type = **m**ain, Aspect = p**e**rfective, VForm = i**n**finitive
sr	`Vmn--an-n--e`	Category = **V**erb, Type = **m**ain, VForm = i**n**finitive, Voice = **a**ctive, Negative = **n**o, Clitic = **n**o, Aspect = p**e**rfective

Experiment

- Monolingual parsing

  Train parsers for Croatian and Slovene, apply respectively.
- Direct cross-lingual parsing

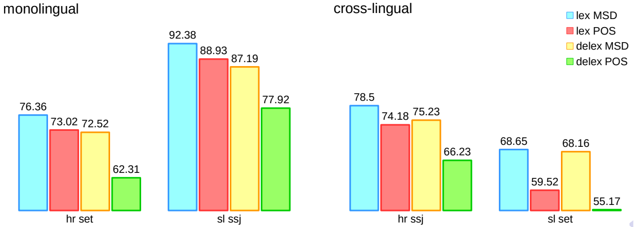
  Apply Croatian parsers on Serbian and Slovene, and Slovene parsers on Croatian and Serbian.
- Cross-lingual parsing with treebank translation

  Translate between Croatian and Slovene, apply to all.
- Test sets
  - 200 sentences per language
  - Each annotated with all three schemes: pdt, set and ssj
  - Language-specific MSD annotations
- Used `mate-tools` graph-based parser (Bohnet, 2010)

Monolingual and direct cross-lingual parsing

- Lots of state-of-the-art scores: bigger treebanks, new parser, etc.
- Feature importance
  - MSD, then lexicalization
  - Dropping both and leaving POS only = substantial decrease
  - Decreases closely resemble those of McDonald et al. (2013)
  - Applies for monolingual and direct cross-lingual scenario

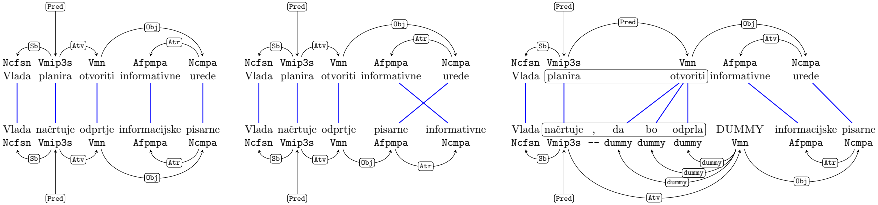
		lexicalized						*delexicalized*					
		hr		*sl*		*sr*		*hr*		*sl*		*sr*	
		MSD	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*
hr	pdt	69.45	66.95	60.09	50.19	69.42	66.96	66.03	57.79	57.98	42.66	66.79	57.41
	set	76.36	73.02	68.65	59.52	76.08	73.37	72.52	62.31	68.16	55.17	72.71	62.04
sl	pdt	51.19	47.99	76.46	73.33	52.46	49.64	49.58	42.59	71.96	62.99	50.41	44.11
	ssj	78.50	74.18	92.38	88.93	78.94	75.96	75.23	66.23	87.19	77.92	75.25	67.47



Cross-lingual parsing with treebank translation

- Standard SMT components: GIZA++, KenLM, Moses
- Four approaches to translation and projection

lookup	Word-to-word translation with no reordering
char	Same as lookup, but character-based SMT
word	Word-to-word translation with reordering
phrase	Full phrase-based reordering with projection of Hwa et al. (2005)


Pred
Sb
Atv
Obj
Atr
Ncfsn Vmip3s Vmn Afpmpa Ncmpa
Vlada planira otvoriti informativne urede
Vlada načrtuje odprtje informacijske pisarne
Ncfsn Vmip3s Vmn Afpmpa Ncmpa
Sb
Atv
Atr
Obj
Pred
Pred
Sb
Atv
Obj
Atr
Ncfsn Vmip3s Vmn Afpmpa Ncmpa
Vlada planira otvoriti informativne urede
Vlada načrtuje odprtje pisarne informativne
Ncfsn Vmip3s Vmn Afpmpa Ncmpa
Sb
Atv
Obj
Atr
Pred
Pred
Pred
Obj
Sb
Atv
Atr
Ncfsn Vmip3s Vmn Afpmpa Ncmpa
Vlada planira otvoriti informativne urede
Vlada načrtuje , da bo odprla DUMMY informacijske pisarne
Ncfsn Vmip3s -- dummy dummy dummy Vmn Afpmpa Ncmpa
Sb
dummy
dummy
dummy
Atr
Obj
Atv
Pred

Cross-lingual parsing with treebank translation

			lexicalized						*delexicalized*					
			hr		*sl*		*sr*		*hr*		*sl*		*sr*	
			MSD	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*	*MSD*	*POS*
char	$hr \mapsto sl$	pdt	66.92	60.25	61.49	55.57	67.83	62.04	66.56	57.63	58.34	43.04	66.89	57.65
		set	73.65	64.64	70.52	66.11	72.95	64.44	72.98	62.98	69.03	54.81	72.74	62.73
	$sl \mapsto hr$	pdt	51.96	48.14	72.35	63.71	53.11	49.47	49.58	42.59	71.96	62.99	50.41	44.11
		ssj	78.69	75.45	88.21	78.88	79.25	77.09	75.23	66.23	87.19	77.92	75.25	67.47
lookup	$hr \mapsto sl$	pdt	67.55	59.96	60.81	56.54	67.78	61.41	66.56	57.63	58.34	43.04	66.89	57.65
		set	73.58	64.98	69.93	68.09	73.70	64.25	72.52	62.72	68.47	55.27	72.71	62.73
	$sl \mapsto hr$	pdt	51.74	49.15	72.02	63.08	53.49	51.33	49.58	42.59	71.96	62.99	50.41	44.11
		ssj	79.25	77.06	88.10	78.53	79.81	77.23	75.23	66.23	87.19	77.92	75.25	67.47
word	$hr \mapsto sl$	pdt	67.33	59.24	61.80	57.14	68.11	61.13	65.84	57.12	58.17	42.99	67.12	57.70
		set	73.26	65.87	69.98	68.98	73.63	65.85	72.71	62.29	68.50	55.06	73.14	62.40
	$sl \mapsto hr$	pdt	51.67	49.58	71.47	63.51	54.62	51.82	50.25	43.17	71.27	62.79	50.79	44.07
		ssj	79.51	76.89	88.71	79.69	79.81	78.03	75.95	67.19	86.92	77.28	75.89	68.18
phrase	$hr \mapsto sl$	pdt	67.28	58.90	60.53	56.79	67.92	61.36	65.77	55.06	58.18	45.41	66.16	55.79
		set	74.68	65.29	69.42	68.55	74.31	65.17	73.36	60.77	68.16	58.42	72.15	61.55
	$sl \mapsto hr$	pdt	49.92	46.82	68.18	58.18	52.15	49.42	47.73	41.08	68.51	55.29	48.93	42.59
		ssj	79.29	78.09	88.24	78.75	79.32	78.85	75.33	68.10	86.59	75.66	75.91	68.67

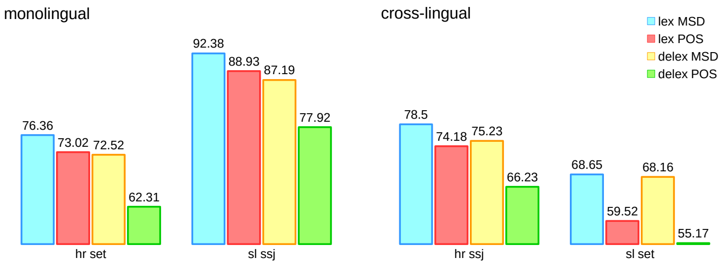
Cross-lingual parsing with treebank translation

- All the best models are lexicalized and use full MSD
- Top-performing SMT approach: word (1:1 with reordering)
- SMT improves where monolingual unavailable
- *sr* is very closely related to *hr*, direct transfer from *hr* practically equals monolingual *hr* scores

Target	Approach	pdt	set	ssj
hr	monolingual	69.45	76.36	–
	direct	51.19	–	78.50
	translated	67.55 ♡	74.68 ♢	79.51 ♣
sl	monolingual	76.46	–	92.38
	direct	60.09	68.65	–
	translated	72.35 ♠	70.52 ♠	88.71 ♣
sr	monolingual	–	–	–
	direct	69.42	76.08	78.94
	translated	68.11 ♣	74.31 ♢	79.81 ♡♣
Legend:	♠ char ♡ lookup ♢ phrase ♣ word			

Conclusions

- Cross-lingual parsing using SMT works for closely-related languages
- Rich morphosyntactic tagsets very beneficial, as well as lexical features, esp. when provided by SMT
- Should we consider using them wherever applicable?



Thank you for your attention. ☺