Multilingual projection for parsing truly low-resource languages

Željko Agić Anders Johannsen Barbara Plank
Héctor Martínez Alonso Natalie Schluter Anders Søgaard

zeag@itu.dk

ACL 2016, Berlin, 2016-08-08

Motivation

Cross-lingual dependency parsing: **almost solved**?

Motivation

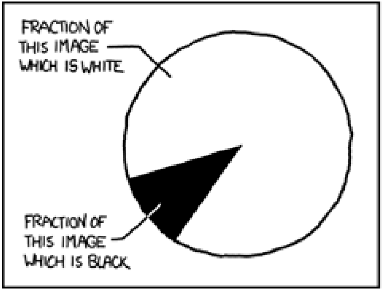
State of the art: +82% UAS on average,
using an annotation projection-based approach.

Motivation

(For German, Spanish, French, Italian, Portuguese, and Swedish.)

Motivation

Treebanks are only available for the 1%.
Cross-lingual learning aims at enabling the remaining 99%.



http://xkcd.com/688/

Motivation

The 1% is very cosy.
Limited evaluation spawns bias.

- POS tagger availability
- parallel corpora: coverage, size, quality of fit
- tokenization
- sentence and word alignment

Motivation

Cross-lingual dependency parsing: ~~almost solved~~ **a bit broken**.

Our approach

Start simple, but fair.

1. Low-resource languages are low-resource.
2. A handful of resource-rich source languages do exist.
3. Annotation projection seems to work.
4. Go for high coverage of the 99%, evaluate where possible.

Our approach

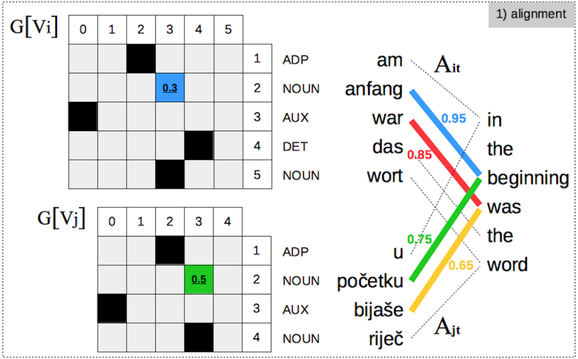
Projection of POS and dependencies

from multiple sources (the 1%)

to as many targets (the 99%) as possible.

Our approach

1. Tag and parse the source sides of parallel corpora.
2. For each source-target sentence pair,
   project POS tags and dependencies to the target tokens.
3. Decode the accumulated annotations, i.e.,
   select the best POS and head for each token among the candidates.
4. Train target-language taggers and parsers.



Our approach

What do we need for it to work?

Data

High-coverage parallel corpora.

- Bible: +1,600 languages online
- Watchtower: +300

- UN Declaration of Human Rights: +500
- OpenSubtitles

Tools

- source-side
  - POS tagger
  - arc-factored dependency parser
- no free preprocessing for parallel corpora
  - simplistic punctuation-based tokenization for all languages
  - automatic sentence and word alignment

Evaluation

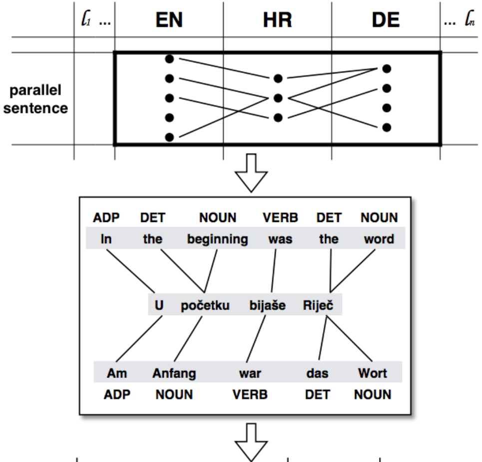
Generate models for the many, evaluate for the few.

21 sources, 6 + 21 targets (UD 1.2)
100 models, easily extends to +1000

Our approach

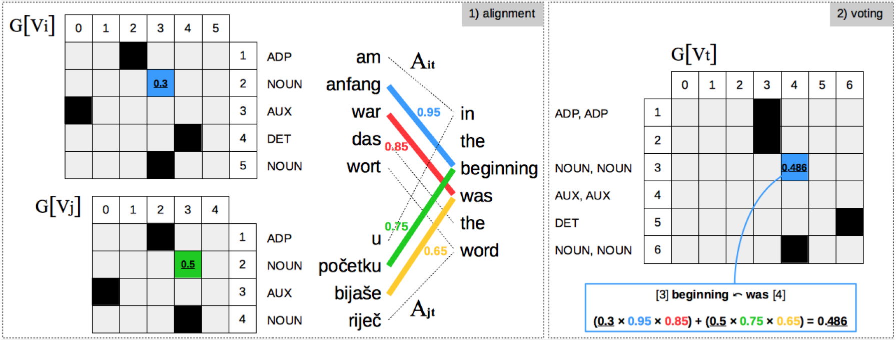
How exactly does our projection work?

Projecting POS

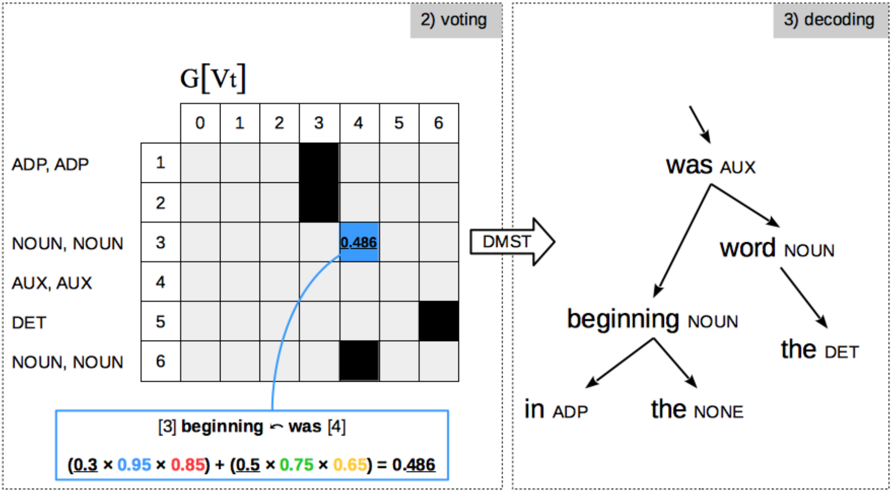


HR	EN	DE	...	voted	confidence
U	ADP	ADP	...	**ADP**	0.8667
početku	NOUN, DET	NOUN	...	**NOUN**	0.7448
bijaše	VERB	VERB	...	**VERB**	0.8560
Riječ	DET, NOUN	DET, NOUN	...	**NOUN**	0.6307

Projecting dependencies



Projecting dependencies



Our approach

Our models are built **from scratch**.
The parsers depend on the cross-lingual POS taggers.

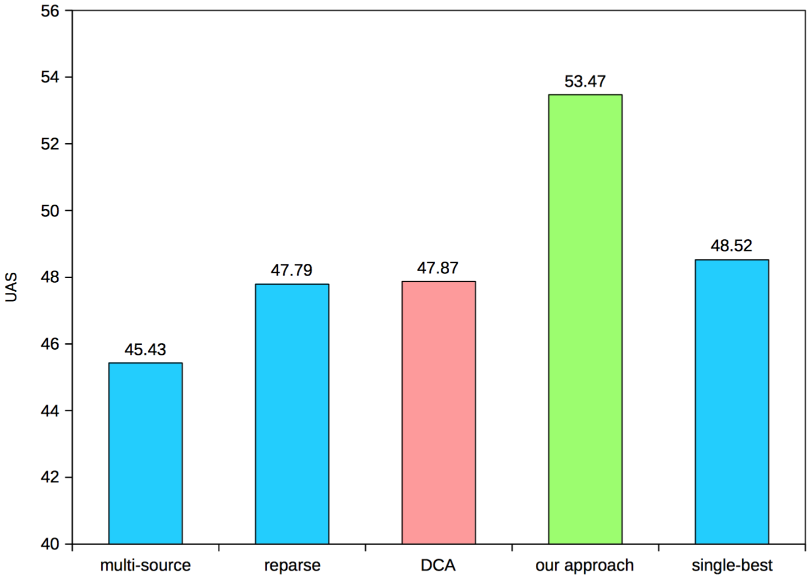
Experiment

- baselines
  - multi-source delexicalized transfer
  - DCA projection
  - voting multiple single-source delexicalized parsers
- upper bounds
  - single-best delexicalized parser
  - self-training
  - direct supervision
- parameters
  - parallel corpora: Bible vs. Watchtower
  - word alignment: IBM1 vs. IBM2

Results

Our approach vs. the rest:

Results

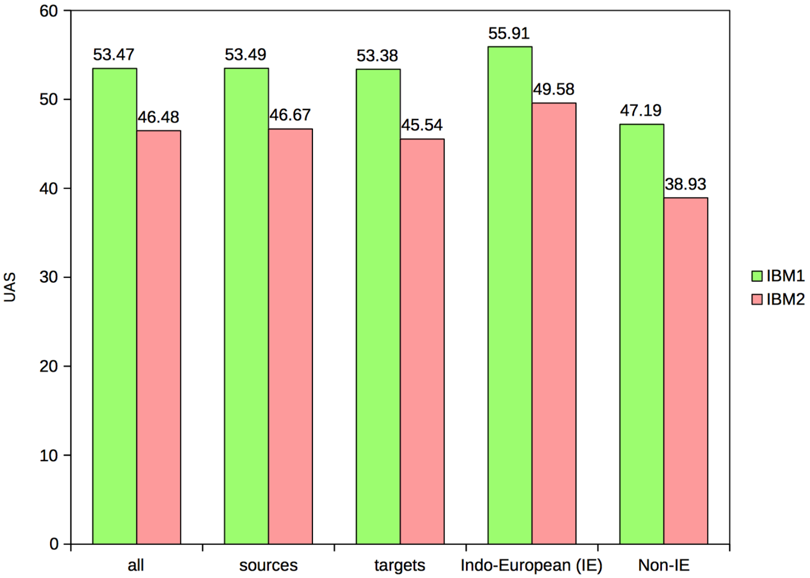


56
54
52
50
48
46
44
42
40
UAS
45.43
47.79
47.87
53.47
48.52
multi-source
reparse
DCA
our approach
single-best

Results

IBM1 vs. IBM2 at their best:

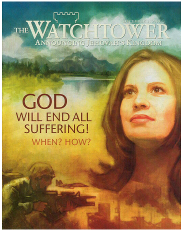
Results


60
50
40
30
20
10
0
UAS
53.47
46.48
53.49
46.67
53.38
45.54
55.91
49.58
47.19
38.93
all
sources
targets
Indo-European (IE)
Non-IE
IBM1
IBM2

Results

And the moment you've all been waiting for:

Results



parsing

$$53.47 > 49.57$$

tagging

$$70.56 > 65.18$$

Conclusions

Our approach is simple, and it works.

- Take-home messages

1. Limited evaluation spawns benchmarking bias.
2. Go for higher coverage, evaluate on a subset if need be.
3. Simple and generic beat complex and finely tuned.
    - IBM1 vs. IBM2
    - our projection vs. DCA
4. The baselines are better than credited for.

Follow-up work: Wednesday at 15:30 (Session 8D)

Joint projection of POS and dependencies from multiple sources!

Thank you for your attention. ☺

Data freely available at: https://bitbucket.org/lowlands/