If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages

Željko Agić Dirk Hovy Anders Søgaard

Center for Language Technology, University of Copenhagen, Denmark

ACL 2015, Beijing, 2015-07-28

Motivation

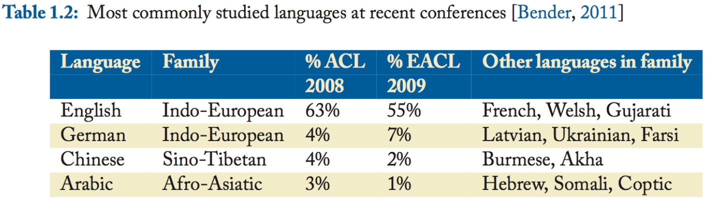


Table 1.2: Most commonly studied languages at recent conferences [Bender, 2011]

Language	Family	% ACL 2008	% EACL 2009	Other languages in family
English	Indo-European	63%	55%	French, Welsh, Gujarati
German	Indo-European	4%	7%	Latvian, Ukrainian, Farsi
Chinese	Sino-Tibetan	4%	2%	Burmese, Akha
Arabic	Afro-Asiatic	3%	1%	Hebrew, Somali, Coptic

Motivation

We want to process *all* languages.
Most of them are severely under-resourced.

How do we build POS taggers for those?

Motivation

- POS tagging for under-resourced languages
  - weak supervision

    (Li et al. 2012)
  - adding a couple of annotation hours

    (Garrette & Baldridge 2013)
  - leveraging parallel corpora

    (Yarowsky et al. 2001) (Das & Petrov 2011) (Täckström et al. 2013)
  - very exciting, high-quality work, lots of code & data made available
  - typically major Indo-European languages
  - high-quality corpora
    - amply sized
    - sentence splits, tokenization, alignment

Motivation

- stepping into a *truly* under-resourced environment
  - let's take nothing for granted
    - language relatedness
    - huge multi-parallel corpora such as Europarl
    - perfect (or any) preprocessing
  - and still try to provide and evaluate POS taggers for as many under-resourced languages as possible

Approach

- even for the most severely under-resourced languages, translations of parts of the Bible exist
  - Edinburgh Bible parallel corpus: 100 languages

    (Christodouloupoulos & Steedman 2014)
  - Parallel Bible corpus: 1,169 languages

    (Mayer & Cysouw 2014)
  - web sources: 1,646 languages

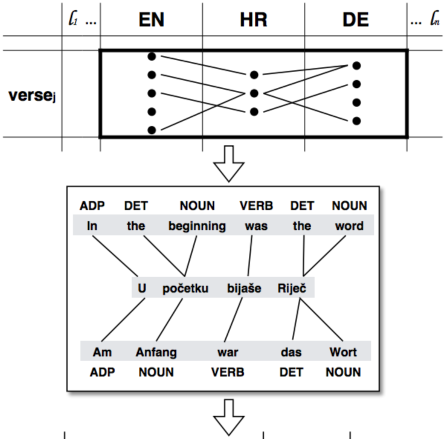
    `http://www.bible.is`

Approach

- "sentence" alignments come for free: verses ids
- multi-parallelism with 100+ languages enables the resource-rich *sources* vs. low-resource *targets* split

= multi-source annotation projection!

Approach



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

Approach

- two projection stages
  - sources to targets (k sources)
  - all to all (n-1 sources)

  - project, vote, train taggers on bibles

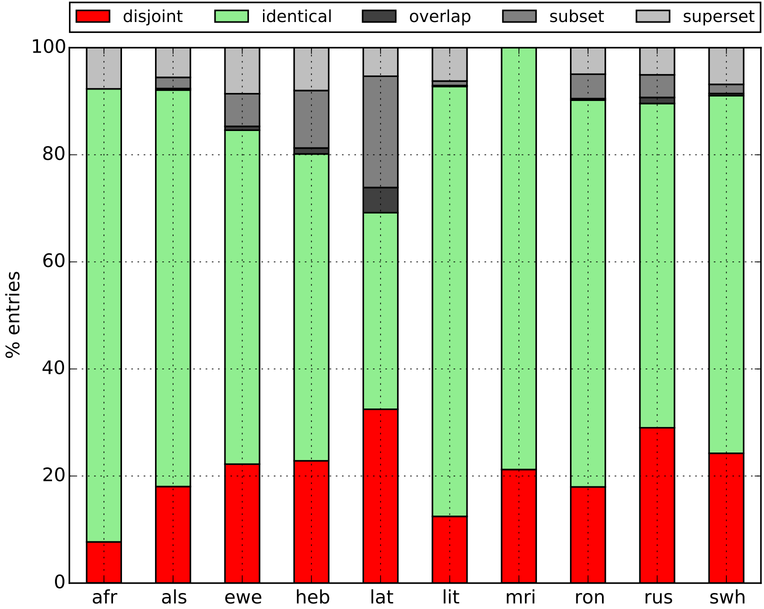
Setup

- the details
  - 18 source languages, 100 targets
  - CoNLL, Google treebanks, UD, HamleDT for training & test sets
  - evaluation
    - test set accuracy
    - do voted tags match Wiktionary tags?
    - do acquired dictionaries agree with wiktionaries?

Results

			Unsupervised						Upper bounds			
			Baselines		Our Systems				Weakly Sup		Supervised	
		OOV	Brown	2HMM	TnT-*k*-Src	TnT-*n*-1-Src	Gar-*k*-Src	Gar-*n*-1-Src	Das	Li	Gar	TnT
bul	YT	31.8	54.5	71.8	**78.0**	77.7	75.7	75.7	-	-	83.1	96.9
ces	YT	44.3	51.9	66.3	71.7	**73.3**	70.9	71.4	-	-	-	98.7
dan	YT	28.6	58.6	69.6	78.6	**79.0**	73.7	73.3	83.2	83.3	78.8	96.7
deu	YT	36.8	45.3	70.0	**80.5**	80.2	77.6	77.6	82.8	85.8	87.1	98.1
eng	YT	38.0	58.2	62.6	72.4	**73.0**	72.2	72.6	-	87.1	80.8	96.7
eus	NT	64.6	46.0	41.6	**63.4**	62.8	57.3	56.9	-	-	66.9	93.7
fra	YT	26.1	42.0	76.5	76.1	76.6	78.6	**80.2**	-	-	85.5	95.1
ell	YT	63.7	43.0	49.8	51.9	52.3	57.9	**59.0**	82.5	79.2	64.4	-
hin	Y	36.1	59.5	69.2	**70.9**	67.6	70.8	71.5	-	-	-	-
hrv	Y	34.7	52.8	65.6	**67.8**	67.1	67.2	66.7	-	-	-	-
hun	YT	41.2	45.9	57.4	70.0	70.4	71.3	**72.0**	-	-	77.9	95.6
isl	Y	19.7	42.6	65.9	**70.6**	69.0	68.7	68.3	-	-	-	-
ind	YT	29.4	52.6	73.1	76.6	**76.8**	74.9	76.0	-	-	87.1	95.1
ita	YT	24.0	45.1	78.3	76.5	76.9	78.5	**79.2**	86.8	86.5	83.5	95.8
plt	Y	35.0	48.9	44.3	56.4	56.6	62.0	**64.6**	-	-	-	-
mar	Y	33.0	**55.8**	45.8	52.0	52.9	52.8	52.3	-	-	-	-
nor	YT	27.5	56.1	73.0	**77.0**	76.7	75.4	76.0	-	-	84.3	97.7
pes	Y	33.6	57.9	**61.5**	59.3	59.6	59.1	60.8	-	-	-	-
pol	YT	36.4	52.2	68.7	**75.6**	75.1	70.8	74.0	-	-	-	95.7
por	YT	27.9	54.5	74.3	82.9	**83.8**	81.1	82.0	87.9	84.5	87.3	96.8
slv	Y	15.8	42.1	78.1	79.5	**80.5**	68.7	70.1	-	-	-	-
spa	YT	21.9	52.6	47.3	81.1	81.4	**82.6**	**82.6**	84.2	86.4	88.7	96.2
srp	Y	41.7	59.3	47.3	**69.6**	69.2	67.9	67.2	-	-	-	94.7
swe	YT	31.5	58.5	68.4	74.7	**75.2**	71.4	71.9	80.5	86.1	76.1	94.7
tur	YT	41.6	53.7	46.8	60.5	**61.3**	56.5	57.9	-	-	72.2	89.1
average		$\leq$ 50	52.2	64.4	72.1	72.2	70.8	71.5				

Results


disjoint
identical
overlap
subset
superset
% entries
0
20
40
60
80
100
afr
als
ewe
heb
lat
lit
mri
ron
rus
swh

Conclusions

- ingredients
  - 100 Bibles
  - 18 source languages with taggers
  - word aligner
  - very naïve tokenization: space, punctuation
- outcomes
  - created taggers for 100 languages
  - evaluated on 25 + 10 mostly under-resourced languages
  - simple approach, competitive performance
- different/more data sources, instance selection
- taking it beyond POS tagging

Thank you for your attention. ☺

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