JOHNS HOPKINS UNIVERSITY

A Rich Morphological Tagger for English:

Exploring the Cross-Linguistic Tradeoff Between Morphology and Syntax

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, Matt Post

JOHNS HOPKINS UNIVERSITY

Introduction

- Are all human languages equally expressive?
  - Czech makes many semantic distinctions using overt morphology
  - English makes same distinctions using syntactic structure and contextual cues
- We train an English tagger to supply Czech distinctions

JOHNS HOPKINS UNIVERSITY

Why This Matters

- Confirm equal expressivity assumption
- Machine Translation into morphologically rich languages:
  - Tagging the source (e.g., English) makes it more like the target (e.g., Czech)
  - Similar to intuition for 'de-compounding' German prior to MT into English (e.g., Niessen & Ney 2000)

JOHNS HOPKINS UNIVERSITY

Hand-Made Heuristics

- Drábek & Yarowsky (2005) and Avramidis & Koehn (2008) previously developed methods to provide rich morphological tags for English.
  - Manually-developed heuristics convert combinations of syntactic structures to morphological features.

- **Conceptual problem**: Morphological feature assignment based on analysts' intuitions.
- **Our Solution**: Learn correspondence of parse features to morphological features directly from projection across parallel text.

JOHNS HOPKINS UNIVERSITY

Projecting Morphological Features

- Prague Czech-English Dependency Treebank (PCEDT)
  - Corresponds to Wall Street Journal portion of Penn Treebank (PTB)
  - Native tagging converted to UniMorph Schema (Sylak-Glassman et al. 2015, Kirov et al. 2016, unimorph.org)
- Automatically aligned to English text using GIZA++

JOHNS HOPKINS UNIVERSITY

Projecting Morphological Features

- For each English word:
  - If aligned to single Czech word, take tag.
  - If aligned to multiple Czech words:
    - Take intersecting alignment point.
    - Take leftmost aligned word.
  - If not aligned, leave unannotated.

JOHNS HOPKINS UNIVERSITY

Tags, Subtags, and Values

Czech	PCEDT tag	UniMorph tag	=	English	PTB tag
je	VB-S---3P-AA---	V;ACT;POS;PRS;3;SG		is	VBZ

Subtag	Values
NUMBER	SG, DU, PL
CASE	NOM, GEN, DAT, ACC, VOC, ESS, INS
PERSON	1, 2, 3
TENSE	FUT, PRS, PST
GRADE	CMPR, SPRL
NEGATION	POS, NEG
VOICE	ACT, PASS

JOHNS HOPKINS UNIVERSITY

Validating Projections

- Do Czech features match English in the few cases where English expresses them?

PTB	Expected UM	Match %
NN	SG	87.8
NNP	SG	73.9
NNS	PL	83.3
NNPS	PL	65.1
JJR	CMPR	89.0
JJS	SPRL	79.3
RBR	CMPR	76.3
RBS	SPRL	68.7
VBZ	SG	91.3
VBZ	3	90.7
VBZ	PRS	89.4
VBG	PRS	55.9
VBP	PRS	87.2
VBD	PST	93.9
VBN	PST	78.7
Average Match %		80.7

JOHNS HOPKINS UNIVERSITY

English Features

- Contextual:
  - **WORD:** the English token in context
  - **NEIGHBORS**: left and right adjacent words
  - **NEIGHBOR POS**: POS of left and right adjacent words

JOHNS HOPKINS UNIVERSITY

English Features

- Dependency-based:
  - **HEAD**: the word's head word
  - **HEAD CHAIN POS**: the chain of POS tags up the dependency graph
  - **HEAD CHAIN LABELS**: the chain of dependency labels up the graph
  - **CHILD**: any child word with det or prep arc label

JOHNS HOPKINS UNIVERSITY

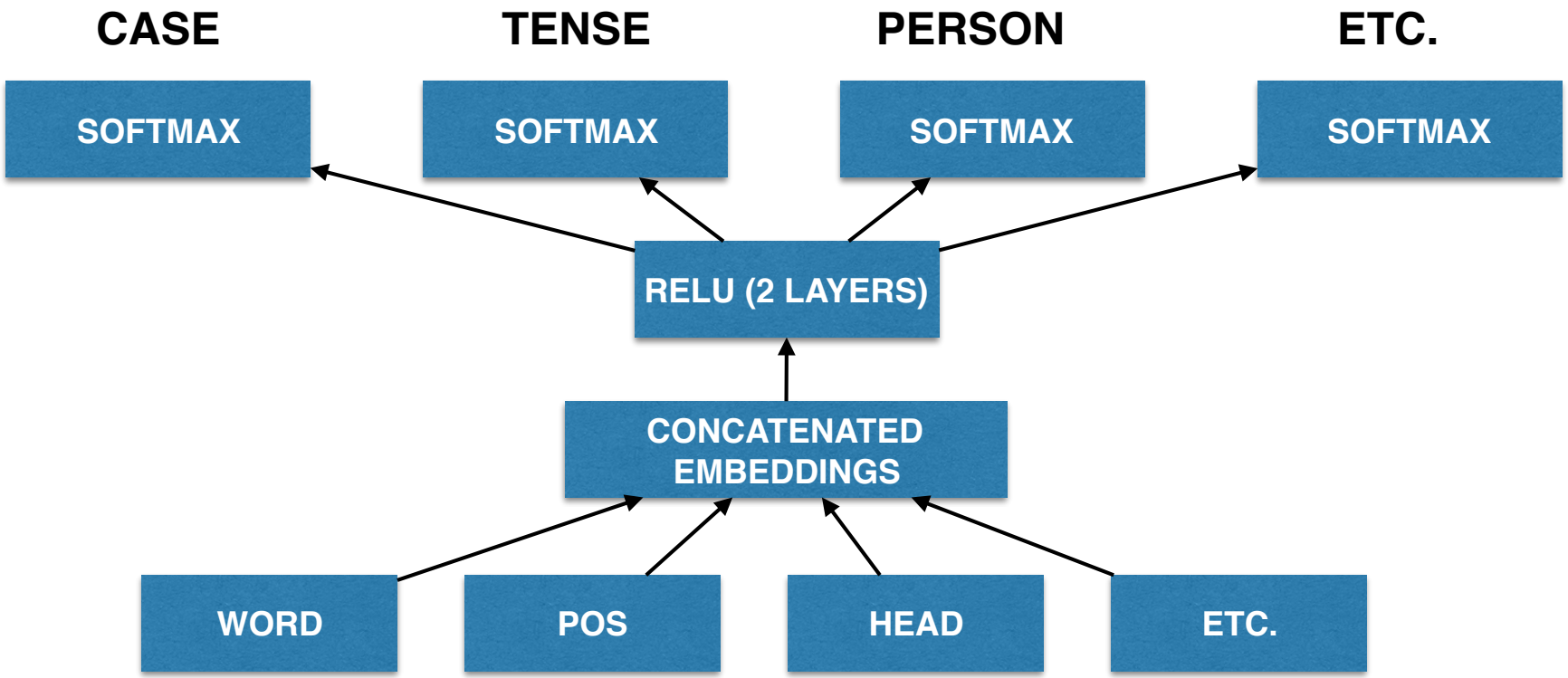
English Features

- Constituent (CFG)-based:
  - **POS**: the POS of the word, parent, and grandparent
  - **NODE CHAIN**: chain of tree nodes on path to root up to length 5 (NN.NP.S)
  - **ROOT DISTANCE**: distance from root

JOHNS HOPKINS UNIVERSITY

Neural Model

Word features jointly encoded and sent to individual subtag classifiers.



JOHNS HOPKINS UNIVERSITY

Experiment Setup

- All data from WSJ portion of PTB
  - Training Set: §02-§21
  - Dev Set: §00
  - Test Set: §22
- 39,832 training sentences (726,262 words)

JOHNS HOPKINS UNIVERSITY

Baseline Results

Other	*companies*	*are*	*introducing*	*related*	*products*
PL, NOM	PL, NOM	ACT, 3, PRS, PL	ACT, 3, PRS, PL	PL, ACC	PL, ACC

- Baseline 1: Most frequent value of subtag in Czech
- Baseline 2: English PTB -> UniMorph, then compare to projected subtag value from Czech; penalize non-matching and missing values in English UniMorph
- Baseline 3: Baseline 2, but penalize only non-matching values, not missing

source	case	tense	per	num	neg	grade	voice	all
Baseline 1	35.0	86.7	94.2	45.6	68.8	99.0	86.7	14.1
Baseline 2	0.7	61.5	29.3	46.0	—	62.6	—	4.3
Baseline 3	46.4	89.1	99.8	86.3	—	99.5	—	8.6
PCEDT	**69.1**	**93.3**	*96.5*	*78.3*	*89.4*	**99.5**	*93.7*	**54.7**

JOHNS HOPKINS UNIVERSITY

Feature Ablation Study

- Which syntactic features capture the most morphological information?

features	case	tense	person	num.	neg.	grade	voice
POS	46.4	91.2	95.3	68.7	84.2	99.3	91.8
Word	56.2	91.5	95.5	72.4	85.9	99.4	91.9
Word, POS	58.6	92.1	95.9	74.4	88.3	99.4	92.6
Word, POS, POS ctxt	63.8	92.7	96.1	77.5	89.1	**99.5**	93.2
CFG	65.0	92.7	96.2	77.5	88.8	99.4	93.1
dep	67.0	92.9	96.3	77.9	89.3	**99.5**	93.2
dep, CFG	**69.1**	92.9	96.4	78.0	89.2	**99.5**	93.2
dep, CFG, lex. ctxt	69.0	**93.2**	**96.6**	**79.1**	**89.8**	**99.5**	**93.7**

Bold = highest accuracy value for subtag

JOHNS HOPKINS UNIVERSITY

Conclusion

- Built ML Tagger to learn correlations between morphological and syntactic features derived from projection between parallel text
  - Avoids labor and pitfalls of manual heuristics
- We find empirical evidence for equal expressivity of languages, with a tradeoff between morphology and syntax
- Planned extension: Tagging in context

JOHNS HOPKINS UNIVERSITY

Thank You!

JOHNS HOPKINS UNIVERSITY

References

Paper URL: http://aclweb.org/anthology/E17-2018

- Eleftherios Avramidis and Philipp Koehn. 2008. "Enriching Morphologically Poor Languages for Statistical Machine Translation." In Proceedings of ACL-2008: HLT, 763-770. ACL: Columbus, OH.
- Elliott Franco Drábek and David Yarowsky. 2005. "Induction of Fine-grained Part-of-speech Taggers via Classifier Combination and Crosslingual Projection." Proceedings of the ACL Workshop on Building and Using Parallel Texts, 49-56. ACL: Ann Arbor, MI.
- Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. "Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms." Proceedings of LREC 2016, 3121-3126. ELRA: Portorož, Slovenia.
- John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. "A Language-Independent Feature Schema for Inflectional Morphology." In Proceedings of the ACL 53 - IJCNLP 7, Volume 2: Short Papers, 674-680. ACL: Beijing.

JOHNS HOPKINS UNIVERSITY

Acknowledgements

Paper URL: http://aclweb.org/anthology/E17-2018

- DARPA LORELEI
- NSF GRF
- DAAD
- NDSEG
- Amazon AARA

Christo Kirov ckirov@gmail.com
John Sylak-Glassman jcsg@jhu.edu
Rebecca Knowles rknowles@jhu.edu
Ryan Cotterell ryan.cotterell@jhu.edu
Matt Post post@cs.jhu.edu