Parsing Typologically Diverse Languages

Miryam de Lhoneux
@mdlhx

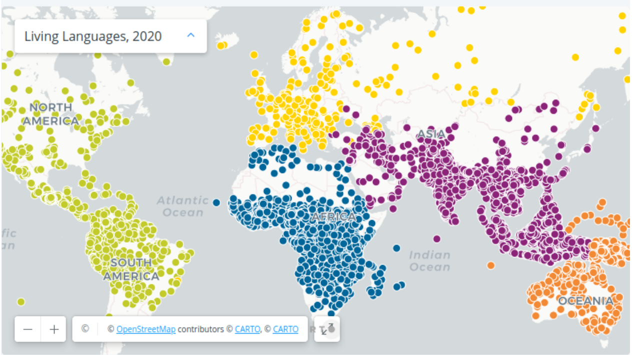
UNIVERSITY OF COPENHAGEN

27-28 October 2020

International Workshop on Treebanks and Linguistic Theories

Outline for section 1

The bad news: world's languages



https://www.ethnologue.com/guides/how-many-languages

The good news

🔥 Cross-lingual learning is on the rise 🔥

Figure from Plank (2019)

The bad news

XNLI

The bad news

XQUAD

The good news: UD

UD v1.0

Figure adapted from Nivre et al. (2020)

The good news: UD

UD v2.0

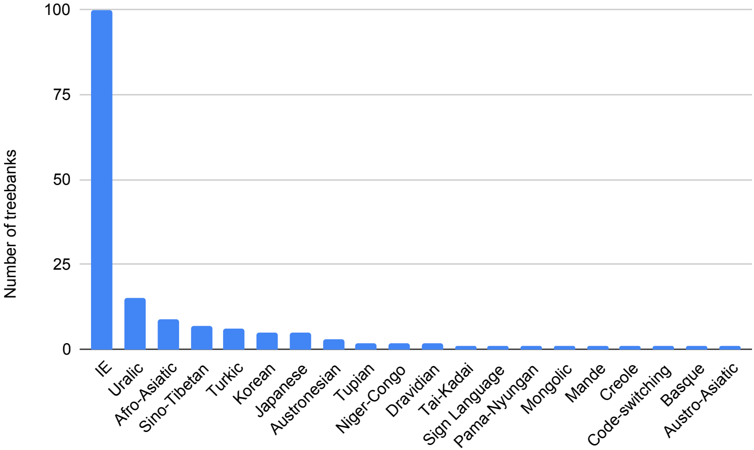
Figure adapted from Nivre et al. (2020)

The good news: UD

UD v2.5

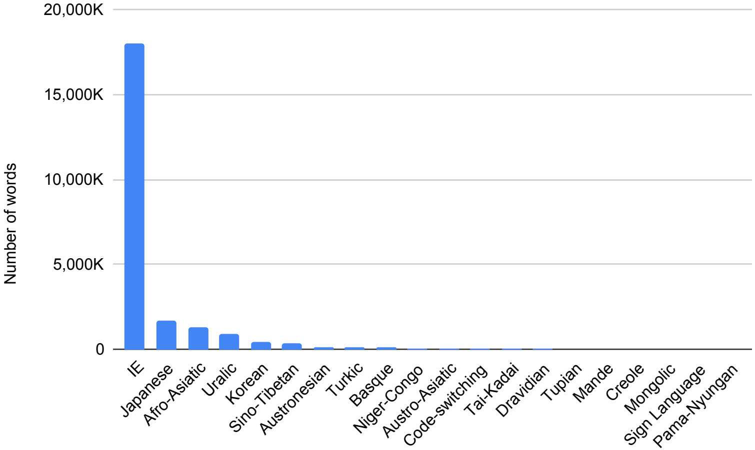
Figure adapted from Nivre et al. (2020)

The bad news



Number of treebanks
100
75
50
25
0
IE
Uralic
Afro-Asiatic
Sino-Tibetan
Turkic
Korean
Japanese
Austronesian
Tupian
Niger-Congo
Dravidian
Tai-Kadai
Sign Language
Pama-Nyungan
Mongolic
Mande
Creole
Code-switching
Basque
Austro-Asiatic

The bad news



Number of words
20,000K
15,000K
10,000K
5,000K
0
IE
Japanese
Afro-Asiatic
Uralic
Korean
Sino-Tibetan
Austronesian
Turkic
Basque
Niger-Congo
Austro-Asiatic
Code-switching
Tai-Kadai
Dravidian
Tupian
Mande
Creole
Mongolic
Sign Language
Pama-Nyungan

Universal Dependencies: opportunities

Universal Dependencies: opportunities

- Do our parsing systems generalize to typologically diverse languages?

Universal Dependencies: opportunities

- Do our parsing systems generalize to typologically diverse languages?
- How well do parsers work in the truly low-resource scenario?

Universal Dependencies: opportunities

- Do our parsing systems generalize to typologically diverse languages?
- How well do parsers work in the truly low-resource scenario?

Important questions for multilingual NLP
Parsing as a test case

Outline for section 2

Parser evaluation

CoNLL 2018 results

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8
- min LAS best: Thai PUD: 13.7

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8
- min LAS best: Thai PUD: 13.7

How should we evaluate?

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8
- min LAS best: Thai PUD: 13.7

How should we evaluate?

- average per treebank, language, or language family?

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8
- min LAS best: Thai PUD: 13.7

How should we evaluate?

- average per treebank, language, or language family?
- evaluate a representative sample?

Parser evaluation

CoNLL 2018 results

- best average LAS: 75.8
- max LAS best: Polish LFG: 94.8
- min LAS best: Thai PUD: 13.7

How should we evaluate?

- average per treebank, language, or language family?
- evaluate a representative sample?
- report worst case?

The case for not averaging per treebank

Anastasopoulos (2019)

Treebank	UDify	UDPipe
UD_Czech-CAC	**92.4**	90.7
UD_Czech-CLTT	**89.9**	84.3
UD_Czech-FicTree	**92.8**	89.8
UD_Czech-PDT	**92.9**	91.3
UD_Czech-PUD	**88.0**	85.0
UD_North_Sami-Giella	67.1	**74.5**
UD_Polish-LFG	**94.6**	94.2
UD_Polish-SZ	89.2	**91.2**
Average LAS	**88.4**	87.6

The case for not averaging per treebank

Anastasopoulos (2019)

Language	UDify	UDPipe
Czech (avg)	**91.2**	88.2
North Sami	67.1	**74.5**
Polish (avg)	91.9	**92.7**
Average LAS	83.4	**85.1**

The case for not averaging per treebank

Anastasopoulos (2019)

Language Family	**UDify**	**UDPipe**
Slavic	**91.6**	90.5
Uralic	67.1	**74.5**
Average LAS	79.3	**82.5**

Representative sample

Representative sample (de Lhoneux et al., 2017b)

1. Typological variety
    1. Only different genera and as many families as possible
    2. Diversity in morphological complexity
    3. One treebank with high non-projective arcs ratio
2. Variety in treebank sizes and domains
3. High annotation quality

Representative sample

Representative sample (de Lhoneux et al., 2017b)

Treebank Sample.

language	family	genus
Czech	IE	Slavic
Chinese	Sino-tibetan	Sinitic
Finnish	Uralic	Finnic
English	IE	Germanic
Ancient Greek	IE	Hellenic
Kazakh	Turkic	N.western
Tamil	Dravidian	Southern
Hebrew	Afro-Asiatic	Semitic

Representative sample

- Bulgarian, Catalan, Czech, Dutch, English, French, German, Italian, Norwegian, Romanian, Russian and Spanish

Representative sample

- Bulgarian, Catalan, Czech, Dutch, English, French, German, Italian, Norwegian, Romanian, Russian and Spanish
- 12 IE; 5 Romance, 4 Germanic, 3 Slavic

Representative sample

- Bulgarian, Catalan, Czech, Dutch, English, French, German, Italian, Norwegian, Romanian, Russian and Spanish
- 12 IE; 5 Romance, 4 Germanic, 3 Slavic
- This ! is ! not ! a ! representative ! sample !

How accurate are current parsers?

Post CoNLL 2018

How accurate are current parsers?

Post CoNLL 2018

- (m)BERT: +3-4 LAS (Kulmizev et al., 2019; Kondratyuk and Straka, 2019)

How accurate are current parsers?

Post CoNLL 2018

- (m)BERT: +3-4 LAS (Kulmizev et al., 2019; Kondratyuk and Straka, 2019)
- Finnish: 93 LAS (Virtanen et al., 2019)

How accurate are current parsers?

Post CoNLL 2018

- (m)BERT: +3-4 LAS (Kulmizev et al., 2019; Kondratyuk and Straka, 2019)
- Finnish: 93 LAS (Virtanen et al., 2019)
- Thai PUD: 26 LAS (Kondratyuk and Straka, 2019)

Outline for section 3

Low-resource parsing: brief history

Low-resource parsing: brief history

- Annotation projection

Low-resource parsing: brief history

- Annotation projection
- Treebank translation

Low-resource parsing: brief history

- Annotation projection
- Treebank translation
- Delexicalized parsing

Low-resource parsing: brief history

- Annotation projection
- Treebank translation
- Delexicalized parsing
- Typological information

Low-resource parsing: brief history

Low-resource parsing: brief history

Most of this work is evaluated in a *simulated* low-resource setting.

Low-resource parsing: brief history

Most of this work is evaluated in a *simulated* low-resource setting.
And relies on

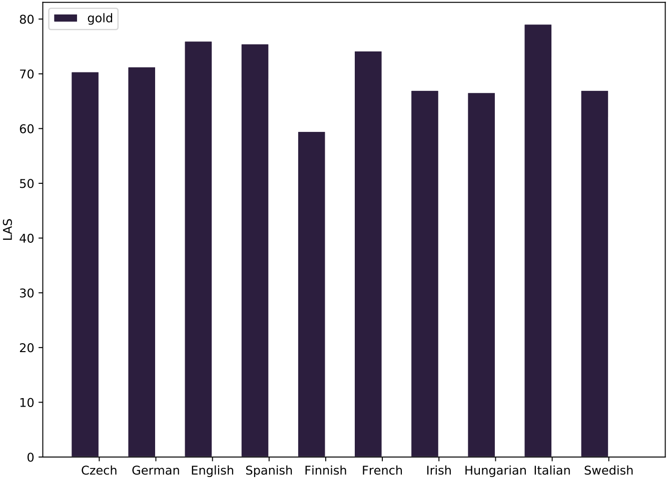
- parallel data

Low-resource parsing: brief history

Most of this work is evaluated in a *simulated* low-resource setting.
And relies on

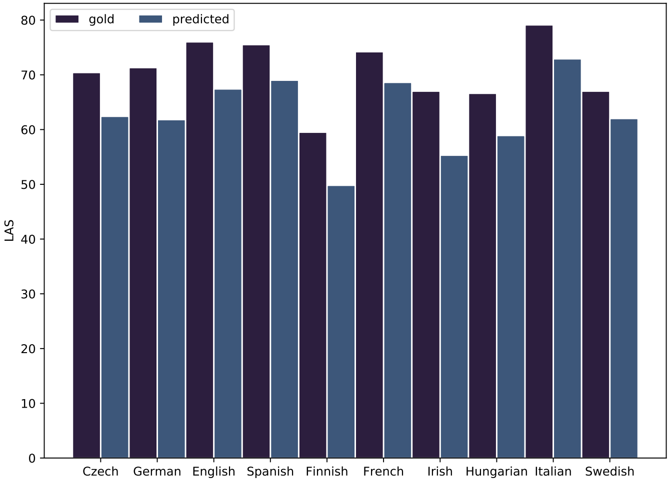
- parallel data
- (gold) POS tags

Low-resource parsing: brief history



Tiedemann (2015)

Low-resource parsing: brief history



Tiedemann (2015)

Low-resource parsing: brief history

POS-taggers for truly low-resource languages are not nearly as accurate! (Kann et al., 2020)

Low-resource parsing: recent developments

Polyglot training:

Low-resource parsing: recent developments

Polyglot training:

- Treebank concatenation

Low-resource parsing: recent developments

Polyglot training:

- Treebank concatenation
- Treebank/Language embedding

Low-resource parsing: recent developments

Polyglot training:

- Treebank concatenation
- Treebank/Language embedding
- Typology vector (e.g. WALS)

Polyglot parsing with treebank embeddings

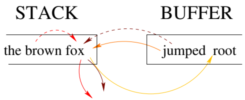
Configuration:

STACK	BUFFER
the brown fox	jumped root

Kiperwasser and Goldberg (2016); de Lhoneux et al. (2017a)

Polyglot parsing with treebank embeddings

Configuration:



Transitions:

LEFT−ARC

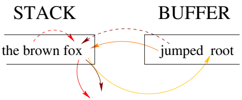
RIGHT−ARC

SHIFT

SWAP

Kiperwasser and Goldberg (2016); de Lhoneux et al. (2017a)

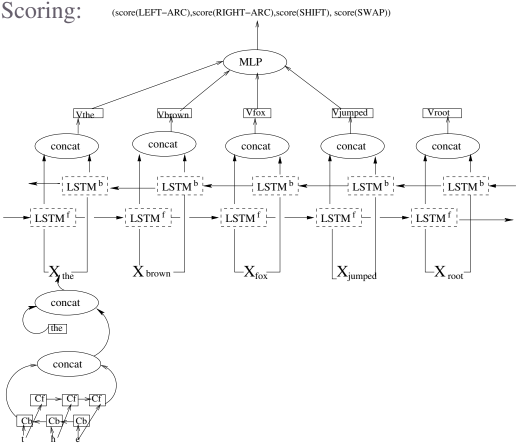
Configuration:



Transitions:

- LEFT–ARC
- RIGHT–ARC
- SHIFT
- SWAP

Scoring:

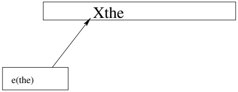


Kiperwasser and Goldberg (2016); de Lhoneux et al. (2017a)

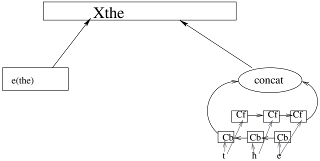
Polyglot parsing with treebank embeddings

Xthe

Polyglot parsing with treebank embeddings



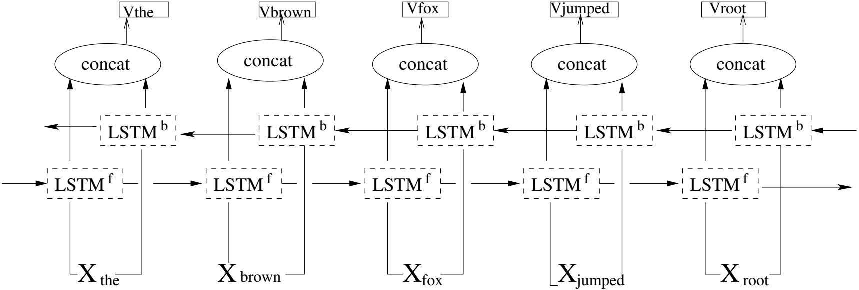
Polyglot parsing with treebank embeddings



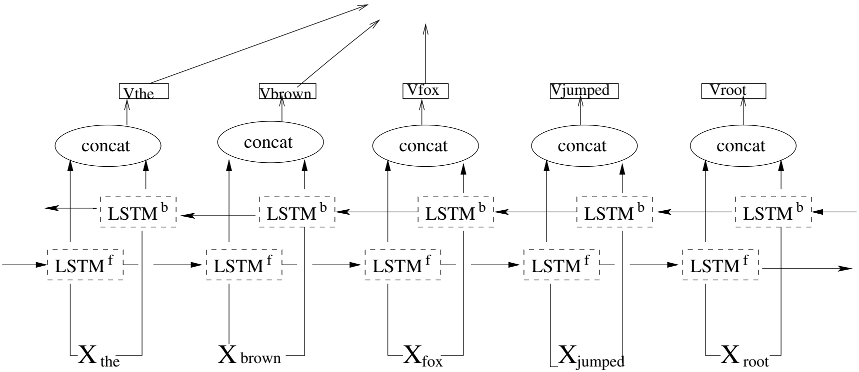
Polyglot parsing with treebank embeddings

X_{the} X_{brown} X_{fox} X_{jumped} X_{root}

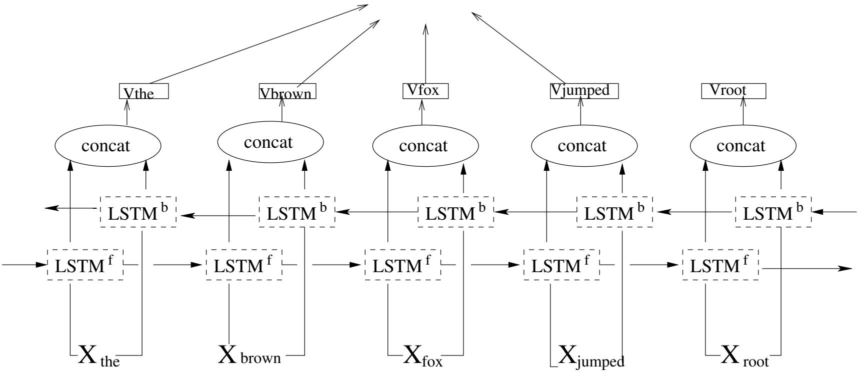
Polyglot parsing with treebank embeddings



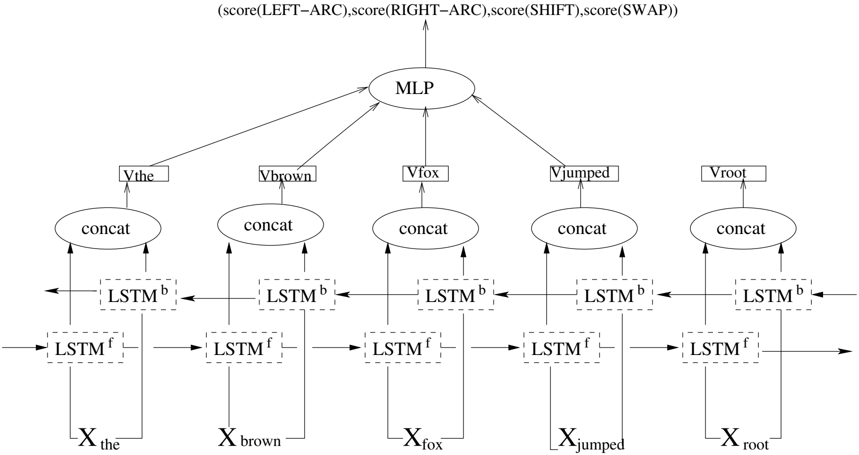
Polyglot parsing with treebank embeddings



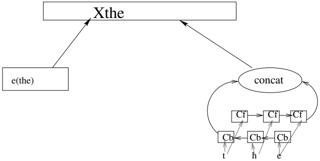
Polyglot parsing with treebank embeddings



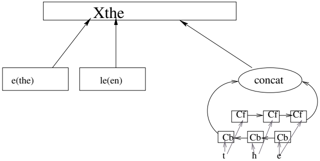
Polyglot parsing with treebank embeddings



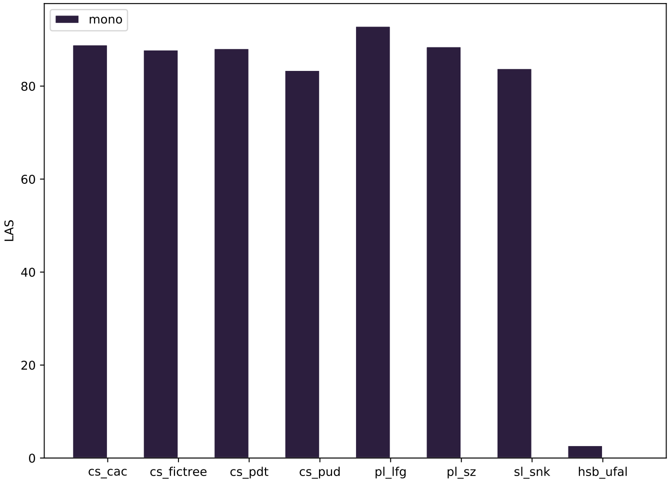
Polyglot parsing with treebank embeddings



Polyglot parsing with treebank embeddings

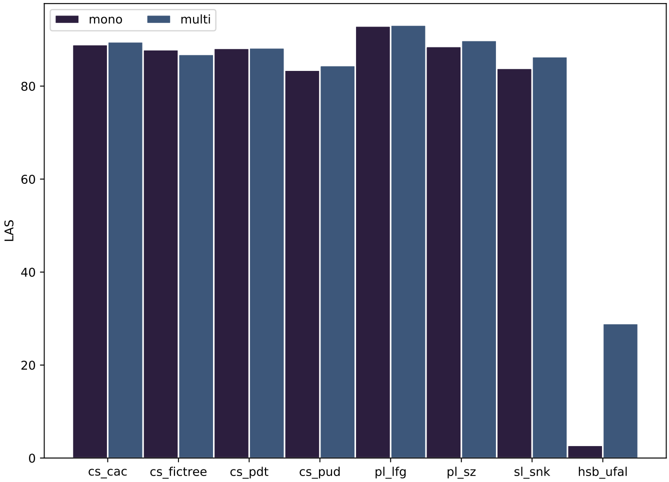


Polyglot parsing with treebank embeddings



Smith et al. (2018)

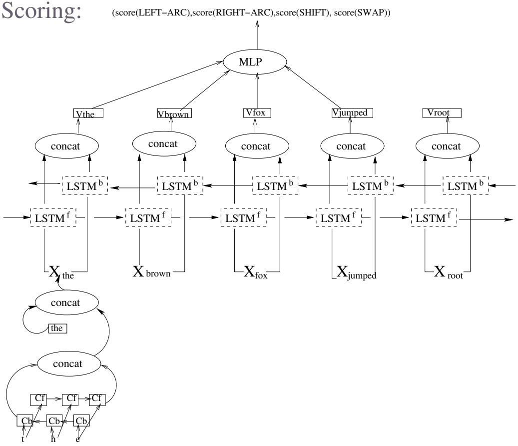
Polyglot parsing with treebank embeddings



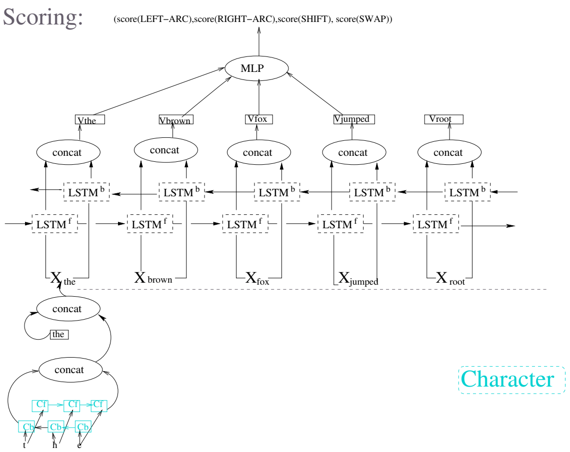
Smith et al. (2018)

Parameter sharing

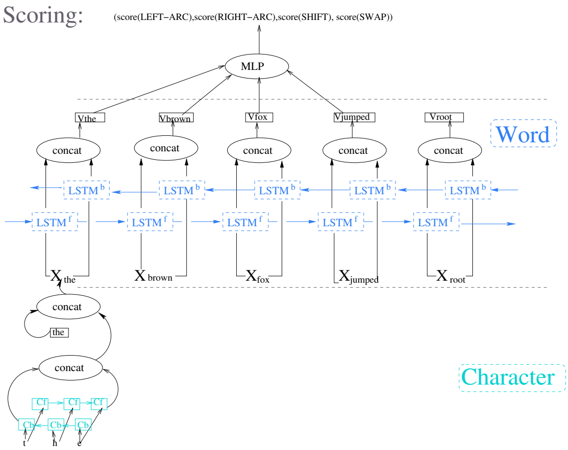
Scoring:



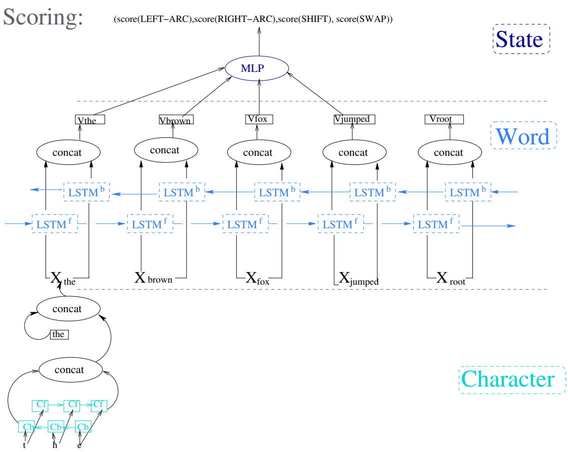
Parameter sharing



Parameter sharing



Parameter sharing



Parameter sharing

- 3 types of sharing: hard, soft, not

Parameter sharing

- 3 types of sharing: hard, soft, not
- 3 sets of parameters: MLP, word, char

Parameter sharing

- 3 types of sharing: hard, soft, not
- 3 sets of parameters: MLP, word, char
- $3^3 = 27$ combinations

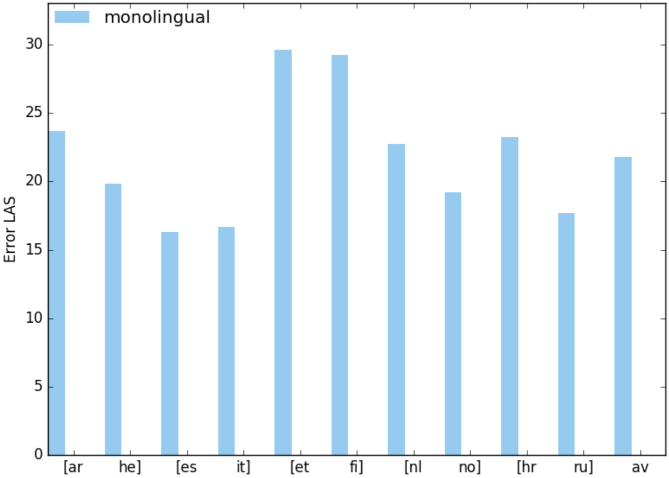
Parameter sharing

Lang	#sen	#tok	Group	Word order
Arabic	5,000	208,932	Semitic	VSO
Hebrew	5,000	161,685	Semitic	SVO
Estonian	5,000	60,393	Finnic	SVO
Finnish	5,000	67,258	Finnic	SVO
Croatian	5,000	109,965	Slavic	SVO
Russian	5,000	90,170	Slavic	SVO
Italian	5,000	113,825	Romance	SVO
Spanish	5,000	154,844	Romance	SVO
Dutch	5,000	75,796	Germanic	No dom. order
Norwegian	5,000	76,622	Germanic	SVO

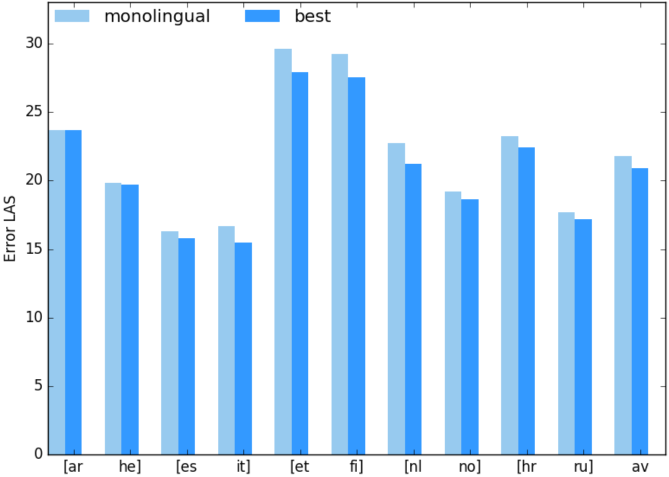
Dataset characteristics

Parameter sharing

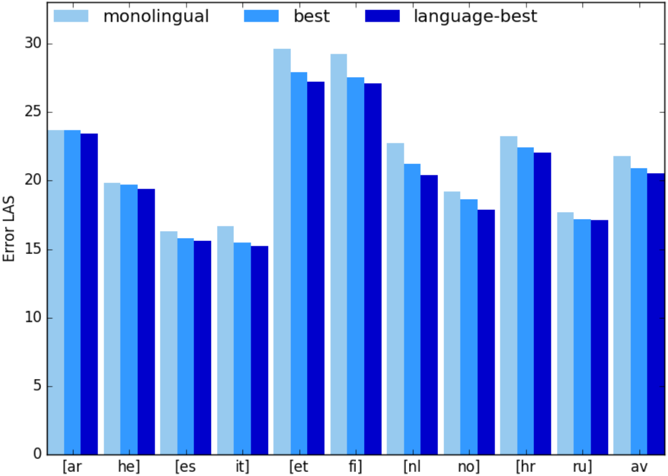
Parameter sharing



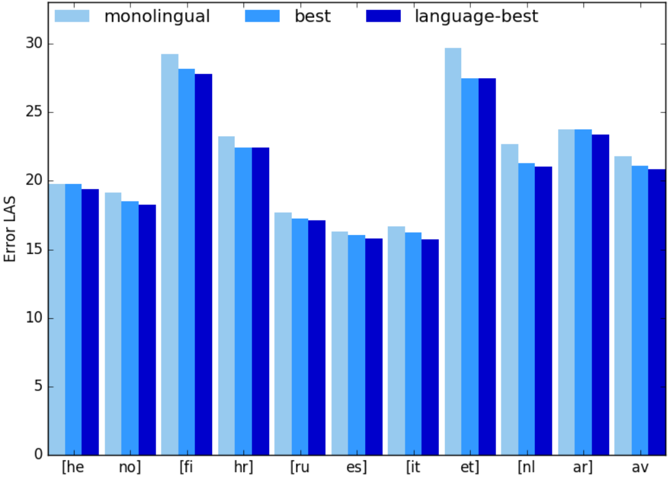
Parameter sharing



Parameter sharing



Parameter sharing



Parameter sharing

Findings

Parameter sharing

Findings

- Sharing MLP helps for all pairs

Parameter sharing

Findings

- Sharing MLP helps for all pairs
- Sharing word and character depends on language pair

Polyglot parsing

Polyglot parsing exploiting language relatedness:

- Vania et al. (2019)
- Meechan-Maddon and Nivre (2019)
- Barry et al. (2019)

Polyglot parsing

Polyglot parsing exploiting language relatedness:

- Vania et al. (2019)
- Meechan-Maddon and Nivre (2019)
- Barry et al. (2019)

We can do cross-lingual without using POS tags!
We can improve parsing for low-resource languages if there is a related high-resource language.

Polyglot parsing without using relatedness

Basirat et al. (2019)

Training			Test	
Afrikaans	Finnish	Russian	Arabic	OCSlavonic
Bulgarian	Hungarian	Slovenian	Belarusian	Serbian
Catalan	Indonesian	Spanish	Coptic	Telugu
Danish	Italian	Swedish	Gothic	Urdu
English	Polish	Turkish	Hindi	Uyghur
Estonian	Portuguese	Ukrainian	Marathi	Vietnamese

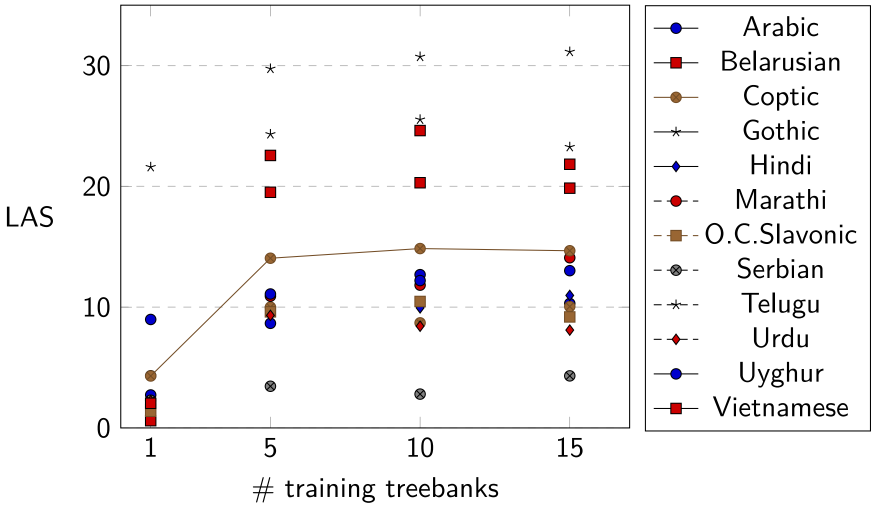
Polyglot parsing without using relatedness

Basirat et al. (2019)

Training			Test	
Afrikaans	Finnish	Russian	Arabic	OCSlavonic
Bulgarian	Hungarian	Slovenian	Belarusian	Serbian
Catalan	Indonesian	Spanish	Coptic	Telugu
Danish	Italian	Swedish	Gothic	Urdu
English	Polish	Turkish	Hindi	Uyghur
Estonian	Portuguese	Ukrainian	Marathi	Vietnamese

- Pretrained language embeddings
- Multilingual word embeddings

Polyglot parsing without using relatedness



Mind the relatedness gap!

Mind the relatedness gap!

- LSTMs vs Transformers (Ahmad et al., 2019)

Mind the relatedness gap!

- LSTMs vs Transformers (Ahmad et al., 2019)
- Transfer works best between typologically similar languages in mBERT (Pires et al., 2019)

Typology

Typological features in WALS: cover many languages

Typology

Typological features in WALS: cover many languages

Mixed results with gold POS tags

- Ammar et al. (2016)
- Scholivet et al. (2019)
- Fisch et al. (2019)

Typology

Typological features in WALS: cover many languages

More promising

Üstün et al. (2020)

	be	br*	bxr*	cy	fo*	gsw*	hsb*	kk	koi*	krl*	mdf*	mr	olo*	pcm*	sa*	tl	yo*	yue*	AVG
multi-udify	**80.1**	**60.5**	26.1	53.6	68.6	43.6	53.2	**61.9**	20.8	**49.2**	24.8	**46.4**	42.1	36.1	19.4	62.7	41.2	30.5	45.2
udapter-proxy	69.9	-	-	-	64.1	23.7	44.4	45.1	-	45.6	-	29.6	41.1	-	15.1	-	-	24.5	-
udapter	79.3	58.5	**28.9**	**54.4**	**69.2**	**45.5**	**54.2**	60.7	**23.1**	48.4	**26.6**	44.4	**43.3**	**36.7**	**22.2**	**69.5**	**42.7**	**32.8**	**46.2**

Conclusion

Take-away

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation
- Gold POS are not realistic for truly low-resource - but also not needed

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation
- Gold POS are not realistic for truly low-resource - but also not needed
- Current cross-lingual transfer methods rely on structural and lexical similarities between languages

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation
- Gold POS are not realistic for truly low-resource - but also not needed
- Current cross-lingual transfer methods rely on structural and lexical similarities between languages
- We are not good at parsing low-resource languages for which there is no high-resource related language

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation
- Gold POS are not realistic for truly low-resource - but also not needed
- Current cross-lingual transfer methods rely on structural and lexical similarities between languages
- We are not good at parsing low-resource languages for which there is no high-resource related language

Open questions

- Are we reaching the limits of implicit cross-lingual transfer?

Conclusion

Take-away

- UD parsing is at the forefront of multilingual NLP
- Think carefully about treebanks used for evaluation
- Gold POS are not realistic for truly low-resource - but also not needed
- Current cross-lingual transfer methods rely on structural and lexical similarities between languages
- We are not good at parsing low-resource languages for which there is no high-resource related language

Open questions

- Are we reaching the limits of implicit cross-lingual transfer?
- Can we explicitly transfer structure? With typological features?

Thanks to my collaborators!

Joakim Nivre

Sara Stymne

Aaron Smith

Ali Basirat

Anders Soegaard

Johannes Bjerva

Isabelle Augenstein

And more!

Questions?

Thanks for listening! Questions?

References I

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing. In *NAACL*.

Waleed Ammar, Phoebe Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many Languages, One Parser. *TACL*, 4:431–444.

Antonis Anastasopoulos. 2019. A note on evaluating multilingual benchmarks. blogpost.

James Barry, Joachim Wagner, and Jennifer Foster. 2019. Cross-lingual parsing with polyglot training and multi-treebank learning: A Faroese case study. In *DeepLo 2019*.

Ali Basirat, Miryam de Lhoneux, Artur Kulmizev, Murathan Kurfal, Joakim Nivre, and Robert Östling. 2019. Polyglot parsing for one thousand and one languages (and then some). In *Typology for Polyglot NLP workshop*.

Adam Fisch, Jiang Guo, and Regina Barzilay. 2019. Working hard or hardly working: Challenges of integrating typology into neural dependency parsers. In *EMNLP*.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *AAAI*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *TACL*, 4:313–327.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *EMNLP-IJCNLP*.

Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano, and Joakim Nivre. 2019. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited. In *EMNLP-IJCNLP*.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *EMNLP*.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017a. From Raw Text to Universal Dependencies - Look, No Tags! In *CoNLL 2017 Shared Task*.

References II

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017b. Old School vs. New School: Comparing Transition-Based Parsers with and without Neural Network Enhancement. In *TLT*.

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Depling, SyntaxFest*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *LREC*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Barbara Plank. 2019. Transferring NLP models across languages and domains. Invited talk at SyntaxFest.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *NAACL*.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *CoNLL 2018 Shared Task*.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Depling*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *EMNLP*.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *EMNLP*.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.