Data Augmentation for Context-Sensitive Neural Lemmatization Using Inflection Tables and Raw Text

Toms Bergmanis, Sharon Goldwater

Lemmatization

	Sing	Plural
NOM	ceļš	ceļi
GEN	ceļa	ceļu
DAT	ceļam	ceļiem
ACC	ceļu	ceļus
INST	ar ceļu	ar ceļiem
LOC	ceļā	ceļos
VOC	ceļ	ceļi

ceļš

Latvian: ceļš (English: *road*)

Previous work:

*"sentence context helps to lemmatize **ambiguous** and **unseen** words"*

Bergmanis and Goldwater, 2018

Ambiguous words: **ceļu**

Lemma could be:

A. ***ceļš*** (*road*): NOUN, sing., ACC
B. ***celis*** (*knee*): NOUN, plur., DAT
C. ***celt*** (*to lift*):VERB, 1st p., sing., pres.

Latvian examples

Learning from sentences

1. Lemma annotated sentences are scarce for low resource languages
2. annotating sentences is slow
3. *N* types > *N* (contiguous) tokens

Learning from sentences

1. Lemma annotated sentences are scarce for low resource languages
2. annotating sentences is slow
3. *N* types > *N* (contiguous) tokens

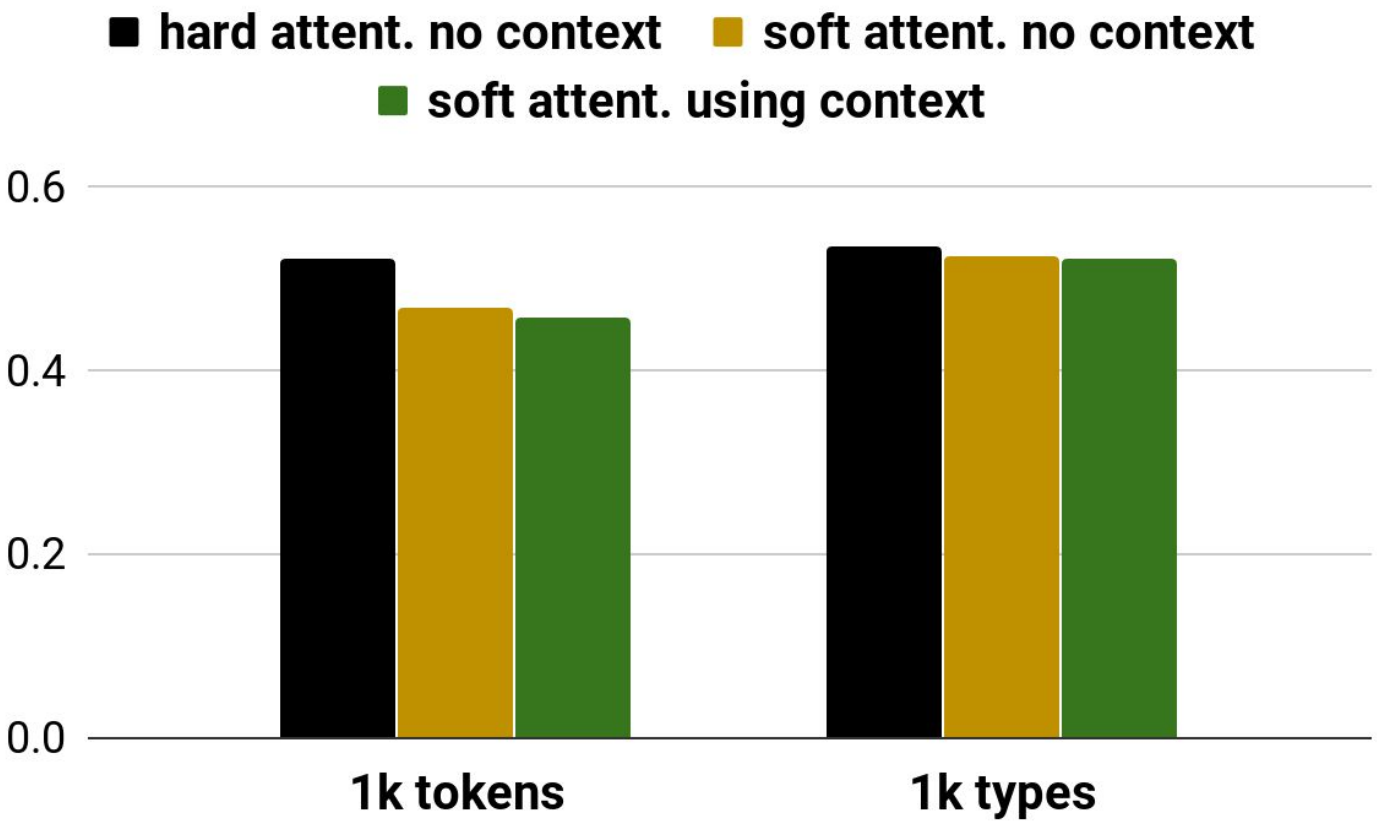
Chakrabarty et al., 2017

Learning from sentences

1. Lemma annotated sentences are scarce for low resource languages
2. annotating sentences is slow
3. *N* types > *N* (contiguous) tokens

Garrette et al., 2013

N types > *N* tokens



Training on 1k UDT tokens/types

Types in context

*algorithms get **smarter** , computers faster*

smart

Bergmanis and Goldwater, 2018

Proposal: Data Augmentation

Combine...

UniMorph
Inflection tables

+

...to get types in context

Method: Data Augmentation

WIKIPEDIA
The Free Encyclopedia

Inflection

ceļā

UniMorph
Inflection tables:

...		
ceļš	ceļš	N;NOM;SG
ceļš	**ceļā**	N;LOC;SG
...		

Method: Data Augmentation

Dzīves pēdējā ***ceļā*** *pavadot mūsu*

WIKIPEDIA
The Free Encyclopedia

Context

UniMorph
Inflection tables:

...

ceļš ceļš N;NOM;SG

ceļš **ceļā** N;LOC;SG

...

Method: Data Augmentation

Dzīves pēdējā ***ceļā*** *pavadot mūsu* → **ceļš**

WIKIPEDIA
The Free Encyclopedia

UniMorph
Inflection tables:

...

ceļš ceļš N;NOM;SG

ceļš **ceļā** N;LOC;SG

...

Lemma

Inflection Tables:

	Sing	Plural
NOM	ceļš	ceļi
GEN	ceļa	ceļu
DAT	ceļam	ceļiem
ACC	ceļu	ceļus
INST	ar ceļu	ar ceļiem
LOC	ceļā	ceļos
VOC	ceļ	ceļi

Latvian: ceļš (English: *road*)

Inflection Tables:

	Sing	Plural
NOM	ceļš	ceļi
GEN	ceļa	ceļu
DAT	ceļam	ceļiem
ACC	ceļu	ceļus
INST	ar ceļu	ar ceļiem
LOC	ceļā	~~ceļos~~
VOC	ceļ	ceļi

ceļot (travel)

Inflection Tables:

	Sing	Plural
NOM	ceļš	ceļi
GEN	ceļa	~~ceļu~~
DAT	~~ceļam~~	ceļiem
ACC	~~ceļu~~	ceļus
INST	ar ceļu	ar ceļiem
LOC	ceļā	~~ceļos~~
VOC	~~ceļ~~	ceļi

celt (build) ceļot (travel)

Inflection Tables:

	Sing	Plural
NOM	ceļš	~~ceļi~~
GEN	~~ceļa~~	~~ceļu~~
DAT	~~ceļam~~	~~ceļiem~~
ACC	~~ceļu~~	~~ceļus~~
INST	ar ceļu	~~ar ceļiem~~
LOC	ceļā	~~ceļos~~
VOC	~~ceļ~~	~~ceļi~~

celt (build) ceļot (travel) celis (knee)

Key question:

If ambiguous words “enforce” the use of context:

Is context still useful in the absence of ambiguous forms?

Experiments

Train: 1k types from universal dependency corpus

Augment: 1k, 5k, 10k types of UniMorph in Wikipedia contexts

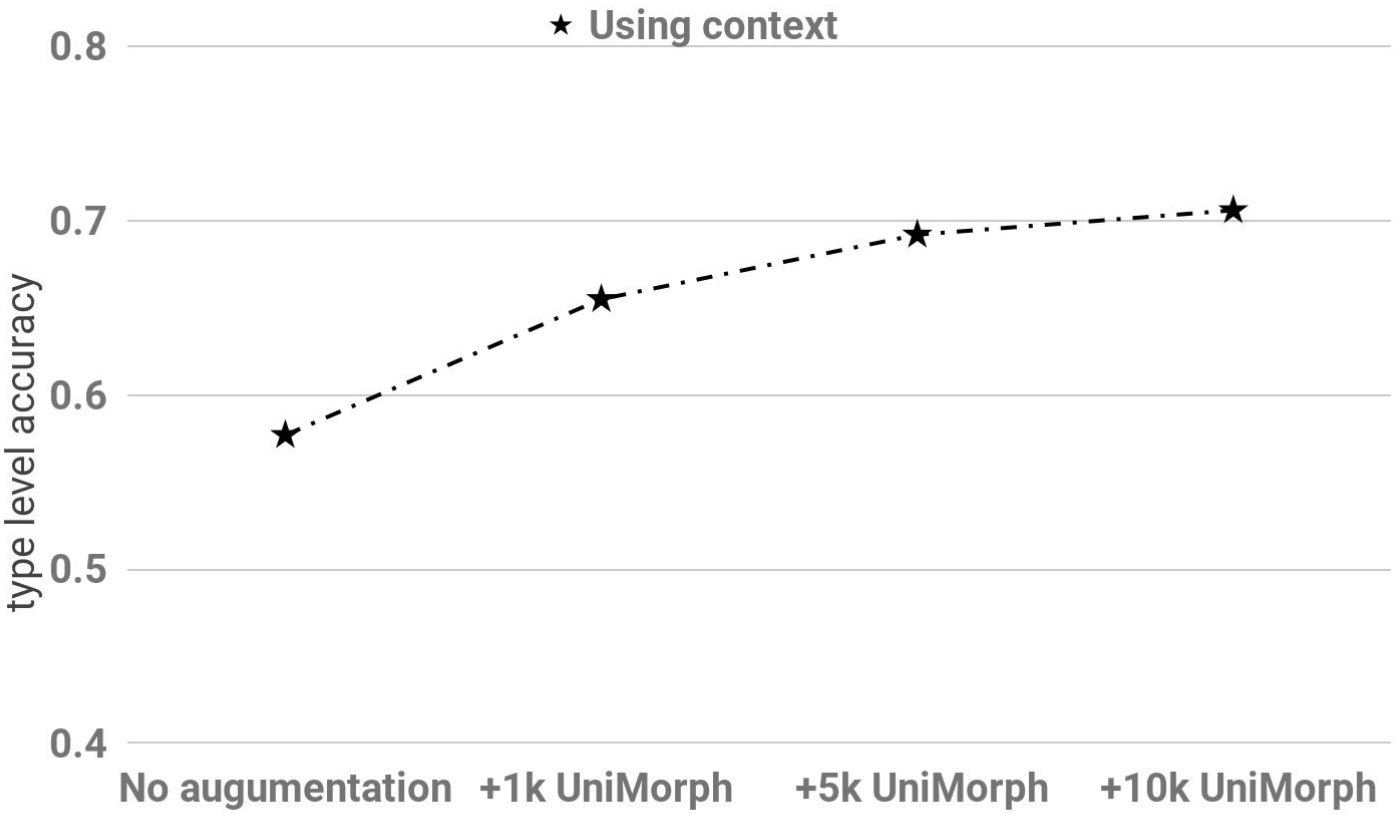
Languages: Bulgarian, Czech, Estonian, Finnish, Latvian, Polish, Romanian, Russian, Swedish, Turkish

Experiments

Metric: type level macro average accuracy

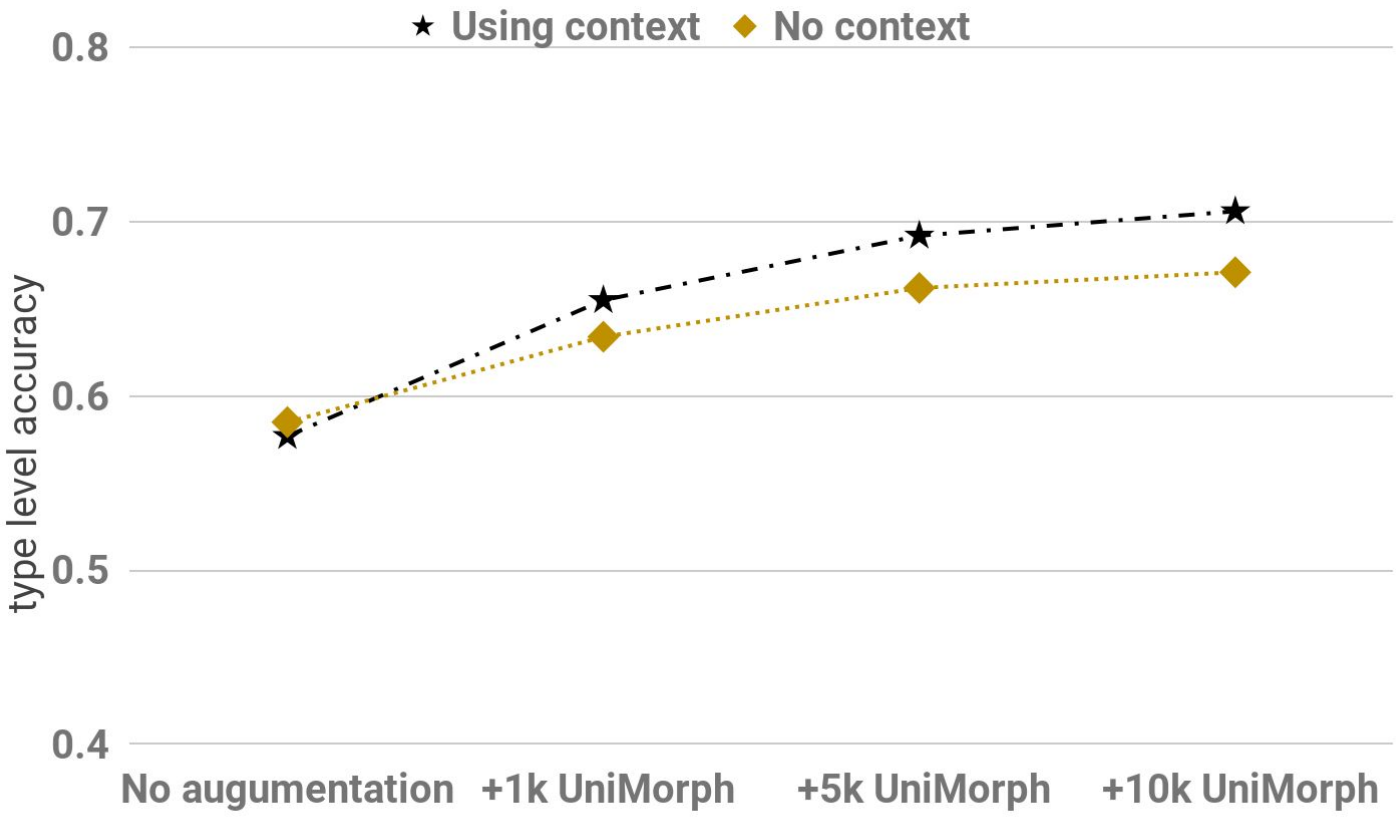
Test: on standard splits of universal dependency corpus

Results: Data augmentation



using context

Does model learn from context?



context vs no context

Afix ambiguity: **wuger**

Lemma depends on context:

A. if **wuger** is **adjective** then lemma could be **wug**

B. if **wuger** is **noun** then lemma could be **wuger**

English examples

Takeaways/conclusions:

Despite biased data and divergent lemmatization standards

Type based data augmentation helps (+14% accuracy)

Takeaways/conclusions:

Even without the ambiguous types that "enforce" the use of context

Model use context to disambiguate affixes of unseen words (+5% accuracy)

THE UNIVERSITY *of* EDINBURGH

Data Augmentation for Context-Sensitive Neural Lemmatization Using Inflection Tables and Raw Text

toms.bergmanis@gmail.com

https://bitbucket.org/tomsbergmanis/data_augumentation_um_wiki