Producing Unseen Morphological Variants in Statistical Machine Translation

Matthias Huck[1], Aleš Tamchyna[1,2], Ondřej Bojar[2], Alexander Fraser[1]

[1]LMU Munich
[2]Charles University in Prague

4 April 2017

Introduction

Translating into morphologically rich languages is difficult.

- Permissible morphological variants remain unseen in training
- SMT systems fail at producing them

Our approach: Providing full coverage of morphological variants.

- Missing variants are synthesized
- The decoder can choose freely amongst all inflected forms

Challenge: How to score unseen morphological variants?

- Additional features in a phrase-based model
- A discriminative classifier that is designed to generalize to unseen morphological variants

Outline

Related Work

Two-step (Toutanova et al., 2008; Bojar and Kos, 2010; Fraser et al., 2012; Burlot et al., 2016)

- Lexical choice (of the lemma) carried out in a separate step from morphological prediction

Factored MT with separate translation and generation (Koehn and Hoang, 2007)

- Too many options blow up the search space
- Useful information is dropped due to separate modeling

Back-translation of lemmatized monolingual data (Bojar and Tamchyna, 2011)

- Computationally expensive
- Back-translated text may contain errors

Synthetic phrases in Hiero (Chahuneau et al., 2013)

- No dynamically-generated target context taken into account

Motivation: English→Czech

case	surface form	50K	500K	5M	50M
1	čéšky	●	●	●	●
2	čéšek	–	●	●	●
3	čéškám	–	–	●	●
4	čéšky	○	○	●	●
5	čéšky	○	○	○	○
6	čéškách	–	●	●	●
7	čéškami	–	–	–	●

Morphological variants of the Czech lemma “čéška”. For differently sized corpora (50K/500K/5M/50M), “●” indicates that the variant is present, and “○” that the same surface form realization occurs, but in a different syntactic case.

Generating Unseen Variants

baseline phrase		lemma		morphological generator
kneecaps ||| čéšky	→	čéška	→	čéšky čéšek čéškám čéšky čéšky čéškách čéškami

Missing variants are added as new translation options.

Settings:

- **word** Phrases of length 1 on both source and target side
- **mtu** Arbitrary length of the phrase source side
  - ⋆ Force some attributes to match the original word form (“tag template”, e.g. for number, tense, . . .)

Czech morphological generator: **MorphoDiTa** (Straková et al., 2014)

Scoring Unseen Variants

In decoding, a “flat” factored phrase-based model provides factors such as morphosyntactic tag and lemma.

Extra features to **independently model word sense and morphological attributes**:

- *n-gram LMs over lemmas and over morphosyntactic tags*
- *OSM over target lemmas*

(included in baseline already)

Specific to synthesized entries:

- *Phrase translation and lexical translation scores over target lemmas*

Discriminative Classifier

morph-vw (*Vowpal Wabbit for Morphology*):

- A decoder-integrated classifier that generalizes to unseen morphological variants.

Feature templates:

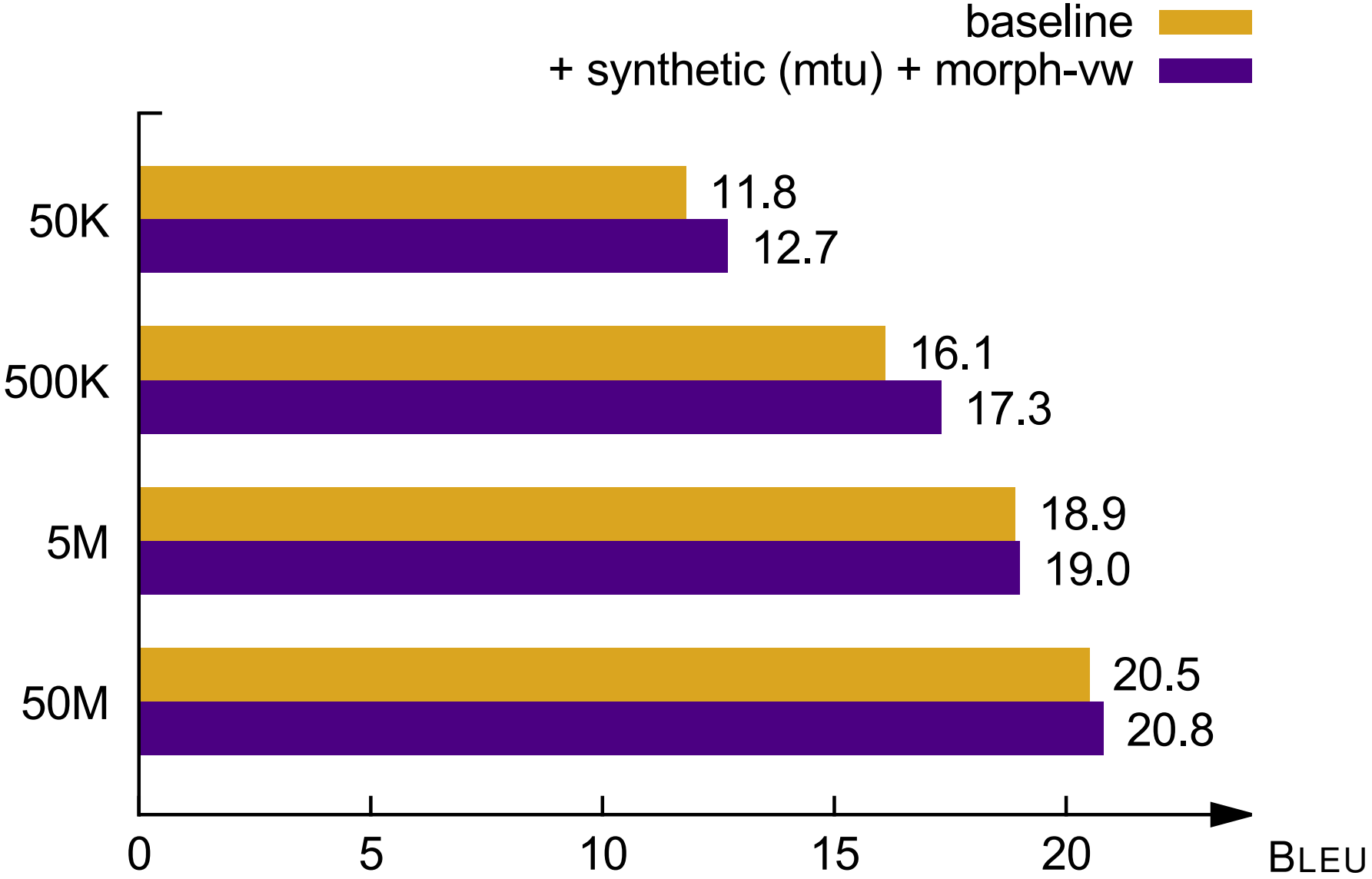
feature type	configurations
source indicator	l, t
source internal	l, l+r, l+p, t, r+p
source context	l (-3,3), t (-5,5)
target indicator	l, t
target internal	l, t
target context	l (-2), t (-2)

l (lemma), **t** (morphosyntactic tag), **r** (syntactic role), **p** (lemma of dependency parent). Numbers in parentheses indicate context size.

Experimental Results (500K)

system \ newstest	2014 BLEU	2015 BLEU	2016 BLEU
baseline 500K	17.7	14.4	16.1
+ morph-vw-500K	17.6	14.4	16.5
+ synthetic (word)	18.1	14.7	16.4
+ morph-vw-500K	18.4	15.2	17.3
+ synthetic (word⋆)	18.0	14.8	16.6
+ morph-vw-500K	18.2	14.9	17.0
+ synthetic (mtu)	18.1	14.8	16.6
+ morph-vw-500K	18.5	15.3	17.3
+ synthetic (mtu⋆)	18.3	15.0	16.9
+ morph-vw-500K	**18.6**	**15.4**	**17.4**

Experimental Results: Scaling Up/Down



Analysis: Phrase Tables & Vocabularies

setup	#phrases full	#phrases filtered	OOV (target) types	OOV (target) tokens
baseline 50K	1.6 M	0.2 M	**45.8 %**	**16.6 %**
+ synthetic (word)	7.8 M	3.9 M	26.7 %	9.9 %
+ synthetic (word⋆)	2.1 M	0.5 M	35.0 %	12.5 %
+ synthetic (mtu)	19.0 M	5.7 M	**26.2 %**	**9.7 %**
+ synthetic (mtu⋆)	3.0 M	0.7 M	34.5 %	12.3 %
baseline 500K	14.5 M	1.4 M	**21.0 %**	**7.1 %**
+ synthetic (word)	44.3 M	16.0 M	11.9 %	4.2 %
+ synthetic (word⋆)	16.9 M	2.5 M	15.2 %	5.2 %
+ synthetic (mtu)	134.4 M	25.8 M	**11.6 %**	**4.1 %**
+ synthetic (mtu⋆)	24.0 M	3.3 M	14.9 %	5.1 %

Analysis: Phrase Tables & Vocabularies

setup	#phrases full	#phrases filtered	OOV (target) types	OOV (target) tokens
baseline 5M	126.6 M	7.4 M	**9.1 %**	**3.1 %**
+ synthetic (word)	254.4 M	58.0 M	5.8 %	2.2 %
+ synthetic (word⋆)	137.1 M	11.4 M	6.7 %	2.4 %
+ synthetic (mtu)	953.3 M	105.9 M	**5.7 %**	**2.1 %**
+ synthetic (mtu⋆)	192.1 M	15.0 M	6.6 %	2.4 %
baseline 50M	996.5 M	23.4 M	**4.9 %**	**1.7 %**
+ synthetic (word)	1 415.2 M	122.2 M	3.6 %	**1.3 %**
+ synthetic (word⋆)	1 030.7 M	30.4 M	4.0 %	1.4 %
+ synthetic (mtu)	6 256.2 M	287.4 M	**3.5 %**	**1.3 %**
+ synthetic (mtu⋆)	1 414.1 M	426 M	3.9 %	1.4 %

Analysis: Translation Example (500K)

input:

now , six in 10 Republicans have a favorable view of Donald Trump .

baseline:

teď , šest **v** 10 **republikáni** mají příznivý **výhled Donald Trump** .

now, six ***in***$_{location}$ *10* ***Republicans***$_{nom}$ *have a_favorable* ***outlook Donald***$_{nom}$ ***Trump***$_{nom}$.

+ synthetic (mtu) + morph-vw:

teď , šest **do** deseti republikánů má příznivý názor na Donalda Trumpa .

now, six **into** *ten*$_{gen}$ *Republicans*$_{gen}$ *have a_favorable opinion of Donald*$_{acc}$ *Trump*$_{acc}$.

Experimental Results: HimL

setup \ corpus size	En→Cs HimL Test 50K BLEU	500K BLEU	5M BLEU	50M BLEU
baseline	14.6	18.4	20.8	23.6
+ morph-vw	14.7	19.6	21.7	23.9
+ synthetic (word)	14.9	18.5	20.9	23.3
+ morph-vw	15.1	19.5	**21.9**	23.9
+ synthetic (word⋆)	15.1	18.3	20.8	23.4
+ morph-vw	15.4	19.5	21.7	24.0
+ synthetic (mtu)	15.1	18.6	20.7	23.7
+ morph-vw	15.2	**19.7**	21.8	**24.1**
+ synthetic (mtu⋆)	15.3	18.6	20.8	23.3
+ morph-vw	**15.6**	19.6	21.7	**24.1**

Summary

Full coverage of all valid inflected target word forms

- for each known lemma
- with direct integration into phrase-based search

Effective scoring of unseen morphological variants

- utilizing a tailored discriminative classifier
- taking both source and target context into account

Substantial BLEU score improvements,
particularly on small to medium resource translation tasks

Questions?

Health in my Language

HimL

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements № 644402 (*HimL*) and № 645452 (*QT21*), from the European Research Council (ERC) under grant agreement № 640550, and from the DFG grant *Models of Morphosyntax for Statistical Machine Translation (Phase Two)*. This work has been using language resources and tools developed and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References I

Bojar, O. and Kos, K. (2010). 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden. Association for Computational Linguistics.

Bojar, O. and Tamchyna, A. (2011). Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Burlot, F., Knyazeva, E., Lavergne, T., and Yvon, F. (2016). Two-Step MT: Predicting Target Morphology. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Seattle, Washington, USA.

References II

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687, Seattle, Washington, USA. Association for Computational Linguistics.

Fraser, A., Weller, M., Cahill, A., and Cap, F. (2012). Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674, Avignon, France. Association for Computational Linguistics.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.

References III

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio. Association for Computational Linguistics.