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

Toms Bergmanis
School of Informatics
University of Edinburgh
T.Bergmanis@sms.ed.ac.uk

Sharon Goldwater
School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

Abstract

Lemmatization aims to reduce the sparse data problem by relating the inflected forms of a word to its dictionary form. Using context can help, both for unseen and ambiguous words. Yet most context-sensitive approaches require full lemma-annotated sentences for training, which may be scarce or unavailable in low-resource languages. In addition (as shown here), in a low-resource setting, a lemmatizer can learn more from n labeled examples of distinct words (types) than from n (contiguous) labeled tokens, since the latter contain far fewer distinct types. To combine the efficiency of type-based learning with the benefits of context, we propose a way to train a context-sensitive lemmatizer with little or no labeled corpus data, using inflection tables from the UniMorph project and raw text examples from Wikipedia that provide sentence contexts for the unambiguous UniMorph examples. Despite these being unambiguous examples, the model successfully generalizes from them, leading to improved results (both overall, and especially on unseen words) in comparison to a baseline that does not use context.

1 Introduction

Many lemmatizers work on isolated wordforms (Wicentowski, 2002; Dreyer et al., 2008; Rastogi et al., 2016; Makarov and Clematide, 2018b,a). Lemmatizing in context can improve accuracy on ambiguous and unseen words (Bergmanis and Goldwater, 2018), but most systems for context-sensitive lemmatization must train on complete sentences labeled with POS and/or morphological tags as well as lemmas, and have only been tested with 20k-300k training tokens (Chrupała et al., 2008; Müller et al., 2015; Chakrabarty et al., 2017).[1]

Intuitively, though, sentence-annotated data is inefficient for training a lemmatizer, especially in low-resource settings. Training on (say) 1000 word types will provide far more information about a language's morphology than training on 1000 contiguous tokens, where fewer types are represented. As noted above, sentence data can help with ambiguous and unseen words, but we show here that when data is scarce, this effect is small relative to the benefit of seeing more word types.[2] Motivated by this result, we propose a training data augmentation method that combines the efficiency of type-based learning and the expressive power of a context-sensitive model.[3] We use *Lematus* (Bergmanis and Goldwater, 2018), a state-of-the-art lemmatizer that learns from lemma-annotated words in their N-character contexts. No predictions about surrounding words are used, so fully annotated training sentences are not needed. We exploit this fact by combining two sources of training data: 1k lemma-annotated types (with contexts) from the Universal Dependency Treebank (UDT) v2.2[4] (Nivre et al., 2017), plus examples obtained by finding *unambiguous* word-lemma pairs in inflection tables from the Universal Morphology (UM) project[5] and collecting sentence contexts for them from Wikipedia. Although these examples are noisy and biased, we show that they improve lemmatization accuracy in experiments on 10 languages, and that the use of context helps, both overall and especially on unseen words.

2 Method

Lematus (Bergmanis and Goldwater, 2018) is a neural sequence-to-sequence model with attention

[1]The smallest of these corpora contains 20k tokens of Bengali annotated only with lemmas, which Chakrabarty et al. (2017) reported took around two person months to create.

[2]Garrette et al. (2013) found the same for POS tagging.

[3]Code and data: https://bitbucket.org/tomsbergmanis/data_augumentation_um_wiki

[4]http://hdl.handle.net/11234/1-2837

[5]http://unimorph.org

inspired by the re-inflection model of Kann and Schütze (2016), which won the 2016 SIGMORPHON shared task (Cotterell et al., 2016). It is built using the Nematus machine translation toolkit,[6] which uses the architecture of Sennrich et al. (2017): a 2-layer bidirectional GRU encoder and a 2-layer decoder with a conditional GRU (Sennrich et al., 2017) in the first layer and a GRU in the second layer.

Lematus takes as input a character sequence representing the wordform in its N-character context, and outputs the characters of the lemma. Special input symbols are used to represent the left and right boundary of the target wordform (`<lc>`, `<rc>`) and other word boundaries (`<s>`). For example, if $N = 15$, the system trained on Latvian would be expected to produce the characters of the lemma *ceļš* (meaning *road*) given input such as:

```
  s a k a <s> p a š v a l d ī b u
          <lc> c e ļ u <rc>
u n <s> i e l u <s> r e ģ i s t r
```

When $N = 0$ (**Lematus 0-ch**), no context is used, making Lematus 0-ch comparable to other systems that do not model context (Dreyer et al., 2008; Rastogi et al., 2016; Makarov and Clematide, 2018b,a). In our experiments we use both Lematus 0-ch and **Lematus 20-ch** (20 characters of context), which was the best-performing system reported by Bergmanis and Goldwater (2018).

2.1 Data Augmentation

Our data augmentation method uses UM inflection tables and creates additional training examples by finding Wikipedia sentences that use the inflected wordforms in context, pairing them with their lemma as shown in the inflection table. However, we cannot use all the words in the tables because some of them are ambiguous: for example, Figure 1 shows that the form `ceļi` could be lemmatized either as `ceļš` or `celis`. Since we don't know which would be correct for any particular Wikipedia example, we only collect examples for forms which are unambiguous according to the UM tables. However, this method is only as good as the coverage of the UM tables. For example, if UM doesn't include a table for the Latvian verb *celt*, then the underlined forms in Table 1 would be incorrectly labeled as unambiguous.

[6]Code for Nematus: `https://github.com/EdinburghNLP/nematus`, Code for Lematus: `https://bitbucket.org/tomsbergmanis/lematus.git`

	noun: ceļš		**noun: celis**	
	SG	**PL**	**SG**	**PL**
NOM	ceļš	~~ceļi~~	celis	~~ceļi~~
GEN	~~ceļa~~	~~ceļu~~	~~ceļa~~	~~ceļu~~
DAT	ceļam	~~ceļiem~~	celim	~~ceļiem~~
ACC	ceļu	~~ceļus~~	celi	~~ceļus~~
INS	ceļu	~~ceļiem~~	celi	~~ceļiem~~
LOC	ceļā	~~ceļos~~	celī	~~ceļos~~
VOC	ceļ	~~ceļi~~	celi	~~ceļi~~

Table 1: Example UM inflection tables for Latvian nouns *ceļš* (*road*) and *celis* (*knee*). The ~~crossed out forms~~ are examples of evidently ambiguous forms that are not used for data augmentation because of being shared by the two lemmas. The underlined forms appear unambiguous in this toy example but actually conflict with inflections of the verb *celt* (*to lift*).

There are several other issues with this method that could potentially limit its usefulness. First, the UM tables only include verbs, nouns and adjectives, whereas we test the system on UDT data, which includes all parts of speech. Second, by excluding ambiguous forms, we may be restricting the added examples to a non-representative subset of the potential inflections, or the system may simply ignore the context because it isn't needed for these examples. Finally, there are some annotation differences between UM and UDT.[7] Despite all of these issues, however, we show below that the added examples and their contexts do actually help.

3 Experimental Setup

Baselines and Training Parameters We use four baselines: (1) **Lemming**[8] (Müller et al., 2015) is a context-sensitive system that uses log-linear models to jointly tag and lemmatize the data, and is trained on sentences annotated with both lemmas and POS tags. (2) The hard monotonic attention model (**HMAM**)[9] (Makarov and Clematide, 2018b) is a neural sequence-to-sequence model with a hard attention mechanism that advances through the sequence monotonically. It is trained on word-lemma pairs (without context)

[7]Recent efforts to unify the two resources have mostly focused on validating dataset schema (McCarthy et al., 2018), leaving conflicts in word lemmas unresolved. We estimated (by counting types that are unambiguous in each dataset but have different lemmas across them) that annotation inconsistencies affect up to 1% of types in the languages we used.

[8]`http://cistern.cis.lmu.de/lemming`

[9]`https://github.com/ZurichNLP/coling2018-neural-transition-based-morphology`

with character-level alignments learned in a preprocessing step using an alignment model, and it has proved to be competitive in low resource scenarios. (3) Our naive **Baseline** outputs the most frequent lemma (or one lemma at random from the options that are equally frequent) for words observed in training. For unseen words it outputs the wordform itself. (4) We also try a baseline data augmentation approach (**AE Aug Baseline**) inspired by Bergmanis et al. (2017) and Kann and Schütze (2017), who showed that adding training examples where the network simply learns to auto-encode corpus words can improve morphological inflection results in low-resource settings. The AE Aug Baseline is a variant of Lematus 0-ch which augments the UDT lemmatization examples by auto-encoding the inflected forms of the UM examples (i.e., it just treats them as corpus words). Comparing AE Aug Baseline to Lematus 0-ch augmented with UM lemma-inflection examples tells us whether using the UM lemma information helps more than simply auto-encoding more inflected examples.

To train the models we use the default settings for Lemming and the suggested lemmatization parameters for HMAM. We mainly follow the hyperparameters used by Bergmanis and Goldwater (2018) for Lematus; details are in Appendix B.

Languages and Training Data We conduct preliminary experiments on five development languages: Estonian, Finnish, Latvian, Polish, and Russian. In our final experiments we also add Bulgarian, Czech, Romanian, Swedish and Turkish. We vary the amount and type of training data (types vs. tokens, UDT only, UM only, or UDT plus up to 10k UM examples), as described in Section 4.

To obtain N UM-based training examples, we select the first N unambiguous UM types (with their sentence contexts) from shuffled Wikipedia sentences. For experiments with $j > 1$ examples per type, we first find all UM types with at least j sentence contexts in Wikipedia and then choose the N distinct types and their j contexts uniformly at random.

Evaluation To evaluate models' ability to lemmatize wordforms in their sentence context we follow Bergmanis and Goldwater (2018) and use the full UDT development and test sets. Unlike Bergmanis and Goldwater (2018) who reported token level lemmatization exact match accuracy, we report *type-level* micro averaged lemmatization exact match accuracy. This measure better reflects improvements on unseen words, which tend to be rare but are more important (since a most-frequent-lemma baseline does very well on seen words, as shown by Bergmanis and Goldwater (2018)).

		Ambig.	**Unseen**	**All**
Tokens	*Baseline*	41.0	26.6	31.0
	Lemming	38.2	48.3	50.6
	HMAM	41.4	50.2	52.1
	Lematus 0-ch	39.9	43.7	46.8
	Lematus 20-ch	38.4	42.8	45.8
Types	*Baseline*	45.0	26.6	32.4
	Lemming	N/A	N/A	N/A
	HMAM	41.8	53.7	56.3
	Lematus 0-ch	42.5	53.7	55.1
	Lematus 20-ch	43.1	51.7	54.9

Table 2: Average type level lemmatization exact match accuracy on five development languages in type and token based training data scenarios. Colour-scale is computed over the whole *Ambig.* column and over all but *Baseline* rows for the other columns.

We separately report performance on unseen and ambiguous tokens. For a fair comparison across scenarios with different training sets, we count as unseen only words that are not ambiguous and are absent from *all* training sets/scenarios introduced in Section 4. Due to the small training sets, between 70-90% of dev set types are classed as unseen in each language. We define a type as ambiguous if the empirical entropy over its lemmas is greater than 0.1 in the full original UDT training splits.[10] According to this measure, only 1.2-5.3% of dev set types are classed as ambiguous in each language.

Significance Testing All systems are trained and tested on ten languages. To test for statistically significant differences between the results of two systems we use a Monte Carlo method: for each set of results (i.e. a set of 10 numerical values) we generate 10000 random samples, where each sample swaps the results of the two systems for each language with a probability of 0.5. We then obtain a p-value as the proportion of samples for which the difference on average was at least as large as the difference observed in our experiments.

4 Experiments, Results, and Discussion

Types vs. Tokens and Context in Very Low Resource Settings We compare training on the first

[10]This measure, *adjusted ambiguity*, was defined by Kirefu (2018), who noticed that many frequent wordforms appear to have multiple lemmas due to annotation errors. The adjusted ambiguity filters out these cases.

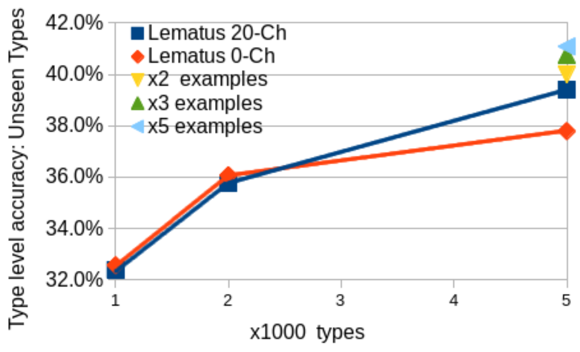

Figure 1: Average type level lemmatization exact match accuracy on unseen words of five development languages. X-axis: thousands of types in training data.

1k tokens vs. first 1k distinct types of the UDT training sets. Table 2 shows that if only 1k examples are available, using types is clearly better for all systems. Although Lematus does relatively poorly on the token data, it benefits the most from switching to types, putting it on par with HMAM and suggesting is it likely to benefit more from additional type data. Lemming requires token-based data, but does worse than HMAM (a context-free method) in the token-based setting, and we also see no benefit from context in comparing Lematus 20-ch vs Lematus 0-ch. So overall, in this very low-resource scenario with no data augmentation, context does not appear to help.

Using UM + Wikipedia Only We now try training only on UM + Wikipedia examples, rather than examples from UDT. We use 1k, 2k or 5k unambiguous types from UM with a single example context from Wikipedia for each. With 5k types we also try adding more example contexts (2, 3, or 5 examples for each type).

Figure 1 presents the results (for unseen words only). As with the UDT experiments, there is little difference between Lematus 20-ch and Lematus 0-ch in the smallest data setting. However, when the number of training types increases to 5k, the benefits of context begin to show, with Lematus 20-ch yielding a 1.6% statistically significant ($p < 0.001$) improvement over Lematus 0-ch. The results for increasing the number of examples per type are numerically higher than the one-example case, but the differences are not statistically significant.

It is worth noting that the accuracy even with 5k UM types is considerably lower than the accuracy of the model trained on only 1k UDT types (see Table 2). We believe this discrepancy is due to the issues of biased/incomplete data noted above.

	DEVELOPMENT			TEST
	Ambig.	*Unseen*	*All*	*All*
1k UDT types (No augmentation)				
Baseline	**49.1**	30.8	36.7	-
HMAM	46.3	**58.9**†‡	**61.5**†‡	**62.6**†‡
Lematus 0-ch	46.5	55.0	58.5	59.1‡
Lematus 20-ch	45.0	54.3	57.7	57.7
1k UDT types + 1k UM types				
Baseline	45.9	30.8	38.4	-
AE Aug Baseline	45.6	57.5	60.4	60.8
HMAM	45.9	60.2	64.2	64.3
Lematus 0-ch	46.6	59.0	63.4	63.6
Lematus 20-ch	**49.8***	**61.7***†	**65.5***†	**65.3**†
1k UDT types + 5k UM types				
Baseline	**55.4***†‡	30.7	41.7	-
AE Aug Baseline	46.0	58.8	61.3	61.6
HMAM	46.7	60.8	65.7	65.7
Lematus 0-ch	46.2	61.5	66.1	66.4
Lematus 20-ch	48.6	**65.4***†	**69.2***†	**69.5** *†
1k UDT types + 10k UM types				
Baseline	**54.9***†	31.2	43.5	-
AE Aug Baseline	46.3	58.6	61.2	61.7
HMAM	45.4	60.8	65.5	65.3
Lematus 0-ch	45.5	62.1	66.4	66.4
Lematus 20-ch	49.5*	**66.7***†	**70.6***†	**70.9***†

Table 3: Average lemmatization accuracy for all 10 languages, trained on 1k UDT types (No aug.), or 1k UDT plus 1k, 5k, or 10k UM types with contexts from Wikipedia. The numerically highest scores in each data setting are bold; *, †, and ‡ indicate statistically significant improvements over HMAM (Makarov and Clematide, 2018b), Lematus 0-ch and 20-ch, respectively (all $p < 0.05$; see text for details). Colour-scale is computed over the whole *Ambig.* column and over all but *Baseline* rows for the other columns.

For example, we analyzed the Latvian data and found that the available tables for nouns, verbs, and adjectives give rise to 78 paradigm slots. The 17 POS tags in UDT give rise to about 10 times as many paradigm slots, although only 448 are present in the unseen words of the dev set. Of these, 197 are represented amongst the 1k UDT training types, whereas only 25 are included in the 1k UM training types. As a result, about 72% of the unseen types of dev set have no representative of their paradigm slot in 1k types of UM, whereas this figure is only 17% for the 1k types of UDT.

Data Augmentation Although UM + Wikipedia examples alone are not sufficient to train a good lemmatizer, they might improve a low-resource baseline trained on UDT data. To see, we augmented the 1k UDT types with 1k, 5k or 10k UM

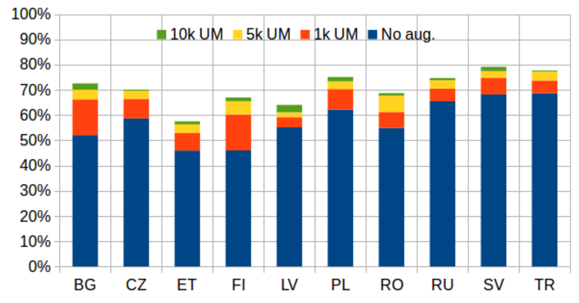

Figure 2: Lematus 20-ch lemmatization accuracy for each language on all types in the dev sets.

types with contexts from Wikipedia.

Table 3 summarizes the results, showing that despite the lower quality of the UM + Wikipedia examples, using them improves results of all systems, and more so with more examples. Improvements are especially strong for unseen types, which constitute more than 70% of types in the dev set. Furthermore, the benefit of the additional UM examples is above and beyond the effect of auto-encoding (AE Aug Baseline) for all systems in all data scenarios.

Considering the two context-free models, HMAM does better on the un-augmented 1k UDT data, but (as predicted by our results above) it benefits less from data augmentation than does Lematus 0-ch, so with added data they are statistically equivalent ($p = 0.07$ on the test set with 10k UM).

More importantly, Lematus 20-ch begins to outperform the context-free models with as few as 1k UM + Wikipedia examples, and the difference increases with more examples, eventually reaching over 4% better on the test set than the next best model (Lematus 0-ch) when 10k UM + Wikipedia examples are used ($p < 0.001$) This indicates that the system can learn useful contextual cues even from unambiguous training examples.

Finally, Figure 2 gives a breakdown of Lematus 20-ch dev set accuracy for individual languages, showing that data augmentation helps consistently, although results suggest diminishing returns.

Data Augmentation in Medium Resource Setting To examine the extent to which augmented data can help in the medium resource setting of 10k continuous tokens of UDT used in previous work, we follow Bergmanis and Goldwater (2018) and train Lematus 20-ch models for all ten languages using the first 10k tokens of UDT and compare them with models trained on 10k tokens of UDT augmented with 10k UM types. To provide a better comparison of our results, we report both the type and the token level development set accuracy. First

Type accuracy:	**Ambig.**	**Unseen**	**All**
1k UDT+10k UM	49.5	66.7	70.6
10k UDT tok.	59.6	71.4	76.6
10k UDT tok.+10k UM	**60.8**	**75.1**	**80.1**
Token accuracy:	**Ambig.**	**Uns.**	**All**
1k UDT+10k UM	55.5	66.5	77.0
10k UDT tok.	**72.4**	72.5	85.3
10k UDT tok.+10k UM	72.3	**75.3**	**87.3**

Table 4: Lematus 20-ch average lemmatization type and token accuracy for all 10 languages, trained on 1k UDT types, 1k UDT augmented with 10k UM types, 10k UDT continuous tokens, or 10k UDT continuous tokens augmented with 10k UM types. Unless specified otherwise data consists of distinct types.

of all, Table 4 shows that training on 10k continuous tokens of UDT yields a token level accuracy that is about 8% higher than when using the 1k types of UDT augmented with 10k UM types—the best-performing data augmentation systems (see Table 3). Again, we believe this performance gap is due to the issues with the biased/incomplete data noted above. For example, we analyzed errors that were unique to the model trained on the Latvian augmented data and found that 41% of the errors were due to wrongly lemmatized words other than nouns, verbs, and adjectives—the three POSs with available inflection tables in UM. For instance, improperly lemmatized pronouns amounted to 14% of the errors on the Latvian dev set. Table 4 also shows that UM examples with Wikipedia contexts benefit lemmatization not only in the low but also the medium resource setting, yielding statistically significant type and token level accuracy gains over models trained on 10k UDT continuous tokens alone (for both Unseen and All $p < 0.001$).

5 Conclusion

We proposed a training data augmentation method that combines the efficiency of type-based learning and the expressive power of a context-sensitive lemmatization model. The proposed method uses Wikipedia sentences to provide contextualized examples for unambiguous inflection-lemma pairs from UniMorph tables. These examples are noisy and biased, but nevertheless they improve lemmatization accuracy on all ten languages both in low (1k) and medium (10k) resource settings. In particular, we showed that context is helpful, both overall and especially on unseen words—the first work we know of to demonstrate improvements from context in a very low-resource setting.

References

Joakim Nivre et al. 2017. Universal dependencies 2.0 CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Toms Bergmanis and Sharon Goldwater. 2018. Context Sensitive Neural Lemmatization with Lematus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training Data Augmentation for Low-Resource Morphological Inflection. In *Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, Vancouver, Canada. Association for Computational Linguistics.

Abhisek Chakrabarty, Onkar Arun Pandit, and Utpal Garain. 2017. Context Sensitive Lemmatization Using Two Successive Bidirectional Gated Recurrent Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491, Vancouver, Canada. Association for Computational Linguistics.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning Morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared taskmorphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Markus Dreyer, Jason R Smith, and Jason Eisner. 2008. Latent-Variable Modeling of String Transductions with Finite-State Methods. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1080–1089. Association for Computational Linguistics.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of postaggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 583–592.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of ACL*. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2017. Unlabeled Data for Morphological Generation With Character-Based Sequence-to-Sequence Models. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 76–81.

Faheem Kirefu. 2018. *Exploring Context Representations for Neural Lemmatisation*. Master's thesis, University of Edinburgh.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Peter Makarov and Simon Clematide. 2018a. Imitation Learning for Neural Morphological String Transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882.

Peter Makarov and Simon Clematide. 2018b. Neural Transition-based String Transduction for Limited-Resource Setting in Morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93.

Arya D McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Lutz Prechelt. 1998. Early Stopping-but When? *Neural Networks: Tricks of the trade*, pages 553–553.

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 157–163.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting Finite-State Transductions With Neural Context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde.

2017. Nematus: a toolkit for neural machine translation. *CoRR*, abs/1703.04357.

Richard Wicentowski. 2002. *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. Ph.D. thesis, Johns Hopkins University.

A Lematus Training

Lematus is implemented using the *Nematus* machine translation toolkit[11]. We use default training parameters of Lematus as specified by Bergmanis and Goldwater (2018) except for early stopping with patience (Prechelt, 1998) which we increase to 20. Similar to Bergmanis and Goldwater (2018) we use the first epochs as a burn-in period, after which we validate the current model by its lemmatization exact match accuracy on the *first 3k instances* of development set and save this model if it performs better than the previous best model. We choose a burn-in period of 20 and validation interval of 5 epochs for models that we train on datasets up to 2k instances and a burn-in period of 10 and validation interval of 2 epochs for others. As we work with considerably smaller datasets than Bergmanis and Goldwater (2018) we reduce the effective model size and increase the rate of convergence by tying the input embeddings of the encoder, the decoder and the softmax output embeddings (Press and Wolf, 2017).

B Data Preparation

Wikipedia database dumps contain XML structured articles that are formatted using the *wikitext* markup language. To obtain wordforms in their sentence context we 1) use *WikiExtractor*[12] to extract plain text from Wikipedia database dumps, followed by scripts from *Moses* statistical machine translation system[13] (Koehn et al., 2007) to 2) split text into sentences (*split-sentences.perl*), and 3) extract separate tokens (*tokenizer.perl*). Finally, we shuffle the extracted sentences to encourage homogeneous type distribution across the entire text.

[11]https://github.com/EdinburghNLP/nematus

[12]https://github.com/attardi/wikiextractor

[13]https://github.com/moses-smt/mosesdecoder

C Result Breakdown by Language

	Type accuracy:	**Ambig.**	**Unseen**	**All**
Bulgarian	*Baseline*	**63.5**	39.3	45.0
	AE Aug Baseline	-	-	-
	HMAM	50.7	**61.0**	**63.5**
	Lematus 0-ch	45.9	51.3	55.7
	Lematus 20-ch	41.6	47.2	52.1
Czech	*Baseline*	38.1	31.2	33.0
	AE Aug Baseline	-	-	-
	HMAM	**45.2**	**66.8**	**66.7**
	Lematus 0-ch	40.7	59.9	60.1
	Lematus 20-ch	40.1	58.3	58.6
Estonian	*Baseline*	**51.0**	24.1	32.0
	AE Aug Baseline	-	-	-
	HMAM	39.9	41.2	46.2
	Lematus 0-ch	38.0	**42.8**	**47.6**
	Lematus 20-ch	47.8	39.9	45.9
Finnish	*Baseline*	46.4	21.3	26.1
	AE Aug Baseline	-	-	-
	HMAM	**44.7**	**48.0**	**50.4**
	Lematus 0-ch	44.4	41.5	44.9
	Lematus 20-ch	44.6	43.0	46.0
Latvian	*Baseline*	42.4	25.6	31.6
	AE Aug Baseline	-	-	-
	HMAM	44.0	**52.6**	**55.6**
	Lematus 0-ch	**47.1**	51.8	55.2
	Lematus 20-ch	43.1	52.1	55.2
Polish	*Baseline*	**42.9**	26.6	33.3
	AE Aug Baseline	-	-	-
	HMAM	41.2	**60.5**	62.4
	Lematus 0-ch	40.9	60.4	**62.6**
	Lematus 20-ch	35.5	59.7	62.2
Romanian	*Baseline*	27.6	34.9	40.0
	AE Aug Baseline	-	-	-
	HMAM	38.8	**55.1**	**57.9**
	Lematus 0-ch	**44.6**	50.2	54.5
	Lematus 20-ch	40.7	50.9	54.9
Russian	*Baseline*	43.0	34.9	39.0
	AE Aug Baseline	-	-	-
	HMAM	39.3	**66.4**	**67.0**
	Lematus 0-ch	42.3	63.4	65.4
	Lematus 20-ch	**44.6**	63.7	65.5
Swedish	*Baseline*	77.8	42.8	52.7
	AE Aug Baseline	-	-	-
	HMAM	58.5	**67.7**	**72.6**
	Lematus 0-ch	**59.5**	64.1	70.1
	Lematus 20-ch	54.0	62.6	68.1
Turkish	*Baseline*	58.8	26.6	33.6
	AE Aug Baseline	-	-	-
	HMAM	60.2	**69.6**	**72.3**
	Lematus 0-ch	**61.8**	64.7	68.4
	Lematus 20-ch	58.2	65.4	68.6

Table 5: Individual type level lemmatization accuracy for all 10 languages on development set, trained on **1k UDT types (no augmentation)** with contexts from Wikipedia. The numerically highest scores for each language are bold. For the summary of results see Table 3.

	Type accuracy:	**Ambig.**	**Unseen**	**All**
Bulgarian	*Baseline*	**64.3**	39.3	47.2
	AE Aug Baseline	49.2	60.3	63.3
	HMAM	41.4	**63.8**	**67.4**
	Lematus 0-ch	49.2	59.2	64.2
	Lematus 20-ch	53.3	61.7	66.2
Czech	*Baseline*	40.3	31.2	34.2
	AE Aug Baseline	42.5	63.6	63.6
	HMAM	42.5	**64.9**	**66.6**
	Lematus 0-ch	38.9	58.7	60.9
	Lematus 20-ch	**53.3**	61.7	66.2
Estonian	*Baseline*	**58.1**	24.1	34.4
	AE Aug Baseline	41.4	42.6	47.4
	HMAM	47.9	43.8	51.4
	Lematus 0-ch	48.1	45.2	52.5
	Lematus 20-ch	45.4	**46.3**	**52.9**
Finnish	*Baseline*	46.3	21.3	27.4
	AE Aug Baseline	44.3	43.3	45.5
	HMAM	45.6	55.7	59.3
	Lematus 0-ch	47.1	55.2	59.3
	Lematus 20-ch	**49.2**	**56.6**	**60.3**
Latvian	*Baseline*	45.4	25.6	33.7
	AE Aug Baseline	38.8	52.2	54.9
	HMAM	42.7	52.7	56.9
	Lematus 0-ch	45.2	51.8	56.4
	Lematus 20-ch	**48.6**	**56.3**	**59.2**
Polish	*Baseline*	**46.3**	26.6	35.4
	AE Aug Baseline	37.5	62.7	64.8
	HMAM	37.4	62.0	66.4
	Lematus 0-ch	45.3	62.3	67.1
	Lematus 20-ch	38.1	**66.9**	**70.3**
Romanian	*Baseline*	37.7	34.9	43.3
	AE Aug Baseline	42.6	53.2	57.0
	HMAM	48.3	**57.3**	62.7
	Lematus 0-ch	**51.1**	57.0	**62.8**
	Lematus 20-ch	49.0	55.9	61.2
Russian	*Baseline*	44.4	34.7	40.6
	AE Aug Baseline	42.6	65.7	67.1
	HMAM	43.2	66.1	68.3
	Lematus 0-ch	38.7	64.6	67.3
	Lematus 20-ch	**50.3**	**67.6**	**70.4**
Swedish	*Baseline*	**78.4**	42.8	54.2
	AE Aug Baseline	58.4	64.8	70.3
	HMAM	53.5	68.9	73.8
	Lematus 0-ch	48.8	**70.9**	**75.4**
	Lematus 20-ch	56.4	69.7	74.8
Turkish	*Baseline*	59.9	26.6	35.5
	AE Aug Baseline	58.6	66.9	69.9
	HMAM	56.1	67.0	69.4
	Lematus 0-ch	54.1	65.2	67.8
	Lematus 20-ch	**62.9**	**70.6**	**73.7**

Table 6: Individual type level lemmatization accuracy for all 10 languages on development set, trained on **1k UDT types plus 1k UM types** with contexts from Wikipedia. The numerically highest scores for each language are bold. For the summary of results see Table 3.

	Type accuracy:	Ambig.	Unseen	All
Bulgarian	*Baseline*	**67.2%**	39.3%	50.0%
	AE Aug Baseline	47.9%	62.6%	65.0%
	HMAM	44.3%	**68.2%**	**72.1%**
	Lematus 0-ch	43.1%	67.0%	71.1%
	Lematus 20-ch	50.4%	65.9%	70.0%
Czech	*Baseline*	43.0%	31.2%	36.8%
	AE Aug Baseline	43.2%	66.9%	66.6%
	HMAM	41.0%	61.9%	64.7%
	Lematus 0-ch	39.5%	61.6%	64.4%
	Lematus 20-ch	**42.6%**	**68.4%**	**69.7%**
Estonian	*Baseline*	62.9%	24.1%	37.1%
	AE Aug Baseline	43.1%	40.3%	45.3%
	HMAM	48.0%	44.9%	53.3%
	Lematus 0-ch	51.3%	45.2%	53.5%
	Lematus 20-ch	**48.6%**	**49.7%**	**56.3%**
Finnish	*Baseline*	49.4%	21.3%	30.3%
	AE Aug Baseline	42.5%	44.9%	47.6%
	HMAM	44.0%	58.4%	62.5%
	Lematus 0-ch	45.9%	60.8%	64.7%
	Lematus 20-ch	**52.5%**	**61.9%**	**65.5%**
Latvian	*Baseline*	**45.6%**	25.6%	35.9%
	AE Aug Baseline	39.6%	53.4%	55.3%
	HMAM	45.2%	52.3%	57.6%
	Lematus 0-ch	43.8%	54.5%	59.1%
	Lematus 20-ch	44.7%	**57.6%**	**61.1%**
Polish	*Baseline*	**50.4%**	26.6%	39.2%
	AE Aug Baseline	38.8%	64.1%	66.2%
	HMAM	41.6%	62.3%	68.4%
	Lematus 0-ch	43.3%	65.2%	70.7%
	Lematus 20-ch	40.3%	**69.7%**	**73.4%**
Romanian	*Baseline*	44.3%	34.9%	47.9%
	AE Aug Baseline	41.3%	54.9%	58.4%
	HMAM	50.2%	58.4%	65.6%
	Lematus 0-ch	51.4%	60.8%	67.2%
	Lematus 20-ch	**47.9%**	**62.6%**	**67.7%**
Russian	*Baseline*	48.5%	34.7%	44.4%
	AE Aug Baseline	42.1%	65.5%	66.5%
	HMAM	46.4%	65.5%	69.7%
	Lematus 0-ch	40.5%	64.4%	68.5%
	Lematus 20-ch	**42.7%**	**71.1%**	**73.8%**
Swedish	*Baseline*	**80.6%**	42.8%	58.0%
	AE Aug Baseline	58.7%	67.3%	71.4%
	HMAM	51.9%	**72.6%**	**77.7%**
	Lematus 0-ch	49.1%	71.4%	76.2%
	Lematus 20-ch	49.5%	72.3%	77.4%
Turkish	*Baseline*	61.8%	26.6%	37.9%
	AE Aug Baseline	62.8%	68.5%	71.2%
	HMAM	54.2%	63.6%	65.7%
	Lematus 0-ch	53.6%	63.9%	65.5%
	Lematus 20-ch	**67.1%**	**74.6%**	**77.2%**

Table 7: Individual type level lemmatization accuracy for all 10 languages on development set, trained on **1k UDT types plus 5k UM types** with contexts from Wikipedia. The numerically highest scores for each language are bold. For the summary of results see Table 3.

	Type accuracy:	Ambig.	Unseen	All
Bulgarian	*Baseline*	**66.2**	39.7	51.2
	AE Aug Baseline	48.1	62.8	65.4
	HMAM	42.7	**70.6**	**74.2**
	Lematus 0-ch	44.8	67.4	71.4
	Lematus 20-ch	50.6	68.4	72.5
Czech	*Baseline*	43.2	31.5	38.2
	AE Aug Baseline	44.9	68.0	68.1
	HMAM	41.1	61.7	64.7
	Lematus 0-ch	38.1	61.9	64.6
	Lematus 20-ch	**42.7**	**68.7**	**70.1**
Estonian	*Baseline*	**62.0**	24.3	37.8
	AE Aug Baseline	45.8	41.0	45.8
	HMAM	48.9	45.4	53.7
	Lematus 0-ch	46.6	45.7	53.8
	Lematus 20-ch	44.9	**51.3**	**57.5**
Finnish	*Baseline*	49.3	21.9	32.7
	AE Aug Baseline	41.6	46.0	48.5
	HMAM	45.0	59.2	62.7
	Lematus 0-ch	43.2	62.8	66.2
	Lematus 20-ch	**49.4**	**63.8**	**67.0**
Latvian	*Baseline*	46.7	25.8	36.9
	AE Aug Baseline	41.6	52.0	54.6
	HMAM	44.6	53.8	59.0
	Lematus 0-ch	42.6	55.3	59.7
	Lematus 20-ch	**47.7**	**60.6**	**64.0**
Polish	*Baseline*	**48.7**	27.0	42.1
	AE Aug Baseline	36.8	64.3	65.4
	HMAM	44.0	60.9	66.4
	Lematus 0-ch	46.2	67.2	72.4
	Lematus 20-ch	42.0	**71.2**	**75.1**
Romanian	*Baseline*	43.7	35.5	49.6
	AE Aug Baseline	41.0	54.3	57.2
	HMAM	45.6	56.8	63.7
	Lematus 0-ch	**50.3**	61.7	67.8
	Lematus 20-ch	49.5	**63.4**	**68.7**
Russian	*Baseline*	**50.2**	35.4	47.1
	AE Aug Baseline	46.8	65.6	67.0
	HMAM	39.1	64.6	68.2
	Lematus 0-ch	38.7	64.6	67.3
	Lematus 20-ch	47.3	**71.2**	**74.7**
Swedish	*Baseline*	**77.3**	43.0	59.9
	AE Aug Baseline	58.8	66.9	71.7
	HMAM	47.3	73.0	77.7
	Lematus 0-ch	55.5	74.1	78.6
	Lematus 20-ch	55.6	**75.1**	**79.1**
Turkish	*Baseline*	62.4	27.1	39.5
	AE Aug Baseline	57.9	67.7	70.3
	HMAM	55.9	62.4	64.8
	Lematus 0-ch	49.1	60.7	62.5
	Lematus 20-ch	**65.8**	**73.6**	**76.8**

Table 8: Individual type level lemmatization accuracy for all 10 languages on development set, trained on **1k UDT types plus 10k UM types** with contexts from Wikipedia. The numerically highest scores for each language are bold. For the summary of results see Table 3.

		Type level accuracy:			Token level accuracy:		
	Training data	**Ambig.**	**Unseen**	**All**	**Ambig.**	**Unseen**	**All**
Bulgarian	10k UDT tok.	**62.3**	75.7	80.1	72.3	75.7	89.5
	10k UDT tok. + 10k UM types	62.2	**78.7**	**83.6**	**73.3**	**78.1**	**91.0**
Czech	10k UDT tok.	49.7	76.4	77.8	**80.7**	77.7	88.3
	10k UDT tok. + 10k UM types	**52.4**	**78.3**	**80.4**	80.0	**80.0**	**89.6**
Estonian	10k UDT tok.	65.3	54.0	64.5	80.1	54.3	76.8
	10k UDT tok. + 10k UM types	**65.9**	**63.4**	**72.6**	**81.5**	**64.2**	**82.4**
Finnish	10k UDT tok.	**60.7**	60.1	66.5	**73.8**	62.4	78.2
	10k UDT tok. + 10k UM types	57.8	**63.7**	**69.4**	70.3	**66.0**	**79.8**
Latvian	10k UDT tok.	57.5	70.9	75.6	69.2	70.5	82.6
	10k UDT tok. + 10k UM types	**58.9**	**73.6**	**77.8**	**70.2**	**73.8**	**84.4**
Polish	10k UDT tok.	**59.8**	78.7	83.6	**76.5**	78.8	89.5
	10k UDT tok. + 10k UM types	57.4	**81.2**	**86.1**	71.3	**81.4**	**90.9**
Romanian	10k UDT tok.	51.7	61.1	66.6	54.1	60.6	79.1
	10k UDT tok. + 10k UM types	**57.1**	**68.2**	**74.2**	**60.7**	**68.2**	**83.9**
Russian	10k UDT tok.	**64.4**	80.5	83.5	**65.9**	80.8	88.5
	10k UDT tok. + 10k UM types	61.1	**82.6**	**85.9**	59.9	**82.7**	**89.8**
Swedish	10k UDT tok.	63.2	74.9	80.9	78.5	73.6	89.6
	10k UDT tok. + 10k UM types	**65.1**	**78.4**	**83.7**	**79.0**	**75.9**	**90.4**
Turkish	10k UDT tok.	64.2	82.1	87.1	73.1	81.8	91.2
	10k UDT tok. + 10k UM types	69.9	**82.9**	**87.3**	**76.9**	**82.7**	**91.5**

Table 9: Individual type and token level lemmatization accuracy for all 10 languages on development set for Lematus 20-ch models trained on **10k UDT tokens** and **10k UDT tokens plus 10k UM types** with contexts from Wikipedia. The numerically highest scores for each language are bold. For the summary of results see Table 4.