Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts

Helmut Schmid
Center for Information and Language Processing, Ludwig Maximilians University
Munich, Germany
schmid@cis.lmu.de

ABSTRACT

Part-of-speech tagging, morphological tagging, and lemmatization of historical texts pose special challenges due to the high spelling variability and the lack of large, high-quality training corpora. Researchers therefore often first map the words to their modern spelling and then annotate with tools trained on modern corpora. We show in this paper that high quality part-of-speech tagging and lemmatization of historical texts is possible while operating directly on the historical spelling. We use a part-of-speech tagger based on bidirectional long short-term memory networks (LSTMs) [11] with character-based word representations and lemmatize using an encoder-decoder system with attention. We achieve state-of-the-art results for modern German morphological tagging on the Tiger corpus and also on two historical corpora which have been used in previous work.

ACM Reference Format:
Helmut Schmid. 2019. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In *Proceedings of Digital Access to Cultural Heritage (DATeCH).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Automatic annotation of historical texts with part-of-speech (POS), lemma, and morphological information is a challenging task because the spelling of such texts often shows considerable variation due to dialectal influences and a general lack of spelling standardization. Another problem is that manually labeled training data is not available in similar quantity and quality as for modern texts. These problems result in lower tagging accuracies.

State-of-the-art POS taggers for modern languages are often based on recurrent neural nets. The tagger of Ling et al. [17], e.g., first processes the character sequence of each word with a bidirectional LSTM net (BiLSTM) to obtain a word representation. Another BiLSTM scans these word representations and computes a contextual representation for each word from which the POS tag is predicted. This approach is promising for processing historical texts because the character-based BiLSTM is able to learn systematic spelling variations, which improves tagging accuracy.

Recurrent neural networks have also been used for lemmatization and the more general task of morphological reinflection. Kann and Schütze [14] used an encoder-decoder system with attention [1], which has originally been developed for machine translation. We will use a similar system to lemmatize historical texts.

In this paper, we present experiments on POS and morphological tagging as well as lemmatization of historical texts. We mainly work on two German datasets, the GerManC corpus from the years 1650–1800 and the *Reference Corpus for Middle High German* (ReM) from the years 1050–1350. The latter is more challenging because the texts are older and differ much more from modern German.

We make the following contributions:

- We present a character-based Bi-LSTM tagger with state-of-the-art performance on German morphological tagging.
- We test this tagger to the GerManC and ReM corpora, outperforming previous systems by a wide margin.
- We evaluate the tagger on historical corpora for six other languages.
- We use the POS tagger to identify middle-high German words with negation clitics.
- We show that a standard character-based encoder-decoder architecture can reliably lemmatize historical corpora.

2 PREVIOUS WORK

Taggers based on BiLSTMs have been presented e.g. by Ling et al. [17] and Huang et al. [12]. Heigold et al. [10] applied a BiLSTM tagger to German and Czech POS and morphological tagging and outperformed the previous state of the art [19] on these datasets.

Hardmeier [9] used a BiLSTM tagger for POS tagging of historical texts. In his experiments, a modern language corpus annotated with POS tags and a historical corpus annotated with modern spelling of the words were used. The tagger is concurrently trained on both corpora with a combined objective which maximizes (i) the POS tagging accuracy on the modern corpus and (ii) the similarity of the representations computed by the BiLSTM for the original spelling and the modern spelling of the historical corpus. He compares his tagger with the HunPos tagger which was directly trained on the annotated GerManC-GS data.

DATeCH, May 2019, Brussels, Belgium

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Other work on POS tagging for historical texts [e.g. 3, 5, 20, 23, 26] is based on a 2-step approach: The historical text is first normalized to modern spelling and then annotated with a POS tagger for the modern language. Dipper [6] applied the same approach to morphological tagging of historical texts. Rögnvaldsson and Helgadóttir [21] trained the TnT tagger [4] on the combination of an Old Icelandic corpus and a Modern Icelandic corpus with good results. Yang and Eisenstein [31] treat historical texts as a special domain of modern language texts and apply domain adaptation techniques.

Hupkes and Bod [13], Moon and Baldridge [18] use an annotation projection approach to train a POS tagger for historical languages. A parallel corpus consisting of a text in historical and modern language (e.g. the bible) is automatically aligned and POS tagged on the modern language side. The POS tags are projected via the alignment links to the historical text. Finally a POS tagger is trained on the annotated historical text.

Previous work on the lemmatization of historical texts is scarce. Souvay and Pierrel [27] used a database of known lemmatizations and a set of graphical and morphological rules to lemmatize unseen words in a Middle French corpus. van Halteren and Rem [30] follow a similar approach. Kestemont et al. [15] tackle lemmatization as a classification problem. They apply a convolutional neural network to the character sequence of the word and concatenate its output with pretrained word embeddings of the current word and the neighboring words. The resulting vector is fed to a neural network whose output is a softmax over all known lemmas. Unseen lemmas cannot be predicted.

Our lemmatization approach is closely related to Kann and Schütze [14]'s system for morphological reinflection. Both use an encoder-decoder system with attention [1],

3 OUR METHODS

3.1 POS Tagging

Similar to [17], we use a bidirectional LSTM to process the character sequence of each input word, and the final states of both directions are concatenated to form the word representation. Another BiLSTM processes the sequence of word representations and generates a contextual representation for each word by concatenating the state of the forward and backward LSTM at the respective position of the word. Each contextual word representation is then passed through a fully connected linear layer followed by a softmax over all possible POS tags. Figure 1 shows a graphical representation of the neural net. The system is trained with stochastic gradient descent to maximize the log-likelihood of the goldstandard tags.

In order to simplify parallel computation of the character-level BiLSTM network, we run its forward LSTM over a fixed-length suffix of the word and the backward LSTM over a fixed-length prefix, truncating or padding the character sequences as necessary. In our experiments, a prefix/suffix length of 10 resulted in good performance.[1]

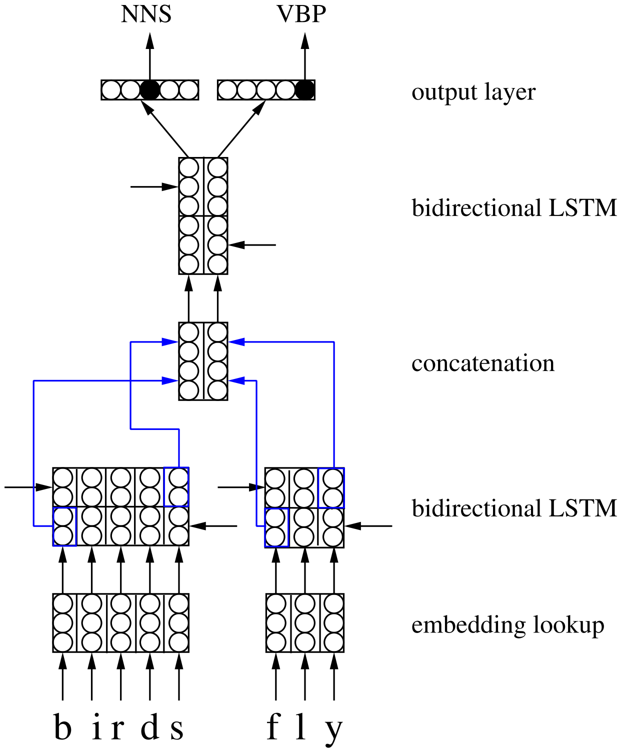

Figure 1: The neural network of the POS tagger

We apply dropout to the output representations of both BiLSTMs and to their recurrent connections. In the latter case, we use the same dropout mask in each time step [cmp. 8]. The tagger may use pretrained word embeddings which are concatenated with the character-based word representations and not fine-tuned. We also apply dropout to the word embeddings, but not to the character embeddings.

In order to verify that our system performs well on morphological tagging, we evaluated it on the German Tiger corpus and compared it to the state-of-the-art system [10]. We used character embeddings and LSTMs of size 400, a dropout rate of 0.5, prefixes and suffixes of length 10, and a learning rate of 0.03 which is multiplied by 0.95 after each epoch (after a burn-in period of 5 epochs) and early stopping. These metaparameters have been optimized on the development data. We used the same data split as Müller and Schütze [19] and Heigold et al. [10]. Characters occurring only once were replaced by UNK. Optionally, we used pretrained word embeddings[2] [2]. Table 1 shows that our tagger slightly outperforms Heigold et al. [10] both, with and without pretrained word embeddings.

[1] Only words longer than 20 characters are not unambiguously represented by the prefix and suffix, and the missing part in the middle is rarely important for POS tagging.

[2] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md.

	dev	test
Heigold		93.23
Heigold+emb		93.85
our tagger	94.80	93.42
our tagger+emb	94.90	93.88

Table 1: POS and morphological tagging accuracy on the German Tiger corpus

3.2 Lemmatizer

For lemmatization, we use the dl4mt system[3]. The input consists of the characters of the wordform, the POS tag, and the sequence of morphological features. The output are the lemma characters. A sample input from our Middle High German data is[4]

f r o g e t e VVFIN * Past Sg 3

A sample output is

v r â g e n

meaning "to ask".

4 EXPERIMENTS

4.1 Data

For our German experiments, we used the *Reference Corpus of Middle High German* (ReM corpus) and the GerManC corpus.

The GerManC corpus[5] [24] comprises more than 770,000 tokens from the period 1650–1800, which have been annotated with part of speech, morphological features, and lemmas. The GerManC-GS corpus is a strictly manually annotated subcorpus, which contains over 50,000 tokens. There are 62 different coarse POS tags and 2303 refined POS tags with morphological features. Following Hardmeier [9], we split this corpus into training, development, and test data, and use the same dev and test data also in our experiments on the larger GerManC data.

The ReM corpus[6] [16] contains texts from the period 1050–1350. They were diplomatically transcribed and semi-automatically annotated with POS, lemma, morphological features, and a normalized wordform. The number of word types outnumbers the number of *normalized* word types by a factor of 2.89 (2.75 for lower-cased forms). The morphological annotation uses the HiTS tagset [7], an extended version of the STTS tagset for modern German texts. The ReM corpus contains 72 different base POS labels and 2500 different refined POS tags with morphological features such as number, gender, case, strength, degree, tense, mood. There are many contracted forms such as *inhandon* (in hand). Since it is difficult to split such forms automatically, we kept them as a unit and concatenated their tags and lemmas. Including concatenated tags, the size of the tagset is 7830[7]. We used the files M351 through M358 of the more reliably annotated[8] subcorpus MiGraCo as development data, the files M359, M401, M404, M408–M411 as test data[9], and the rest as training data. Overall, he had 103,676 / 103,789 / 2,062,239 tokens as test/development/training data.

For the lemmatization experiments, we extracted all word-POS-morph triples and their most frequent lemma from the ReM corpus. We included contracted forms, but excluded forms which occurred only once, or contained characters other than letters, hyphens, and parentheses, or were tagged as foreign words[10] (POS tag *FM*). We randomly split the remaining data into 10,000 types each for testing and development and the remaining 149,337 for training.

4.2 Tagging Experiments on GerManC

We first trained our character-based BiLSTM tagger on the GerManC corpus and optimized the meta-parameters on the development data. We experimented with character embeddings sizes between 200 and 400 and RNN hidden state sizes between 400 and 600. Dropout rate was fixed at 0.5. Learning rate decay after each epoch was varied between 0.9 and 0.97.

We compare our accuracies with those of Hardmeier [9] who provides two results, one obtained with the HunPos tagger trained on GerManC, and another obtained with his own tagger[11] which is trained on a modern POS-tagged corpus and on a parallel historical/modern corpus. We outperform these taggers by 2 and 7 percentage points on test data (see Table 2). When we train on the full GerManC corpus (excluding dev and test data), we increase the test accuracy to 98.2%. We also tested our tagger on the extended POS tags with morphological features obtaining an accuracy of 87.7%.

When looking at the tagging errors, we found many cases where the goldstandard annotations seemed incorrect. An example is the phrase:

An/APPR die/ART.acc.pl.fem lieben/ADJA.acc.pl.fem.pos Landleute/NN.nom.pl. (to our dear compatriates).*

Gender and case should always be '*' and 'acc', here. We manually checked the differences between goldstandard annotations and automatically assigned tags (including morphological features) on the first 747 test tokens and found that 87 had been tagged differently. Both annotations were wrong in 23 cases, the goldstandard was wrong 38 times, the tagger was clearly wrong 20 times and possibly wrong 6 times. This result indicates that the accuracy of our tagger is comparable to that of the manual annotation.

[3] https://github.com/nyu-dl/dl4mt-tutorial/tree/master/session2

[4] '*' indicates that the verb mood is underspecified.

[5] available at http://www.llc.manchester.ac.uk/research/projects/germanc

[6] available at https://www.linguistics.rub.de/rem

[7] The original corpus contained a few misspelled POS labels which we mapped to their (likely) correct forms.

[8] Personal communication with Thomas Klein

[9] We omitted here files M402, M403, M405, M406, and M407 because they occurred in two versions.

[10] Foreign words lack reliable lemma information.

[11] These results are not fully comparable with the other results because the training data and/or process are different. We cite them nevertheless.

	dev	test
Hardmeier	83.82	86.15
HunPos (GerManC-GS)	90.82	91.54
our tagger (GerManC-GS)	91.58	93.58
our tagger (GerManC)	97.25	98.23
our tagger (GerManC+morph)	85.65	87.72

Table 2: POS and morphological tagging accuracies on the GerManC corpus

	dev	test
TreeTagger POS	85.28	90.40
our tagger POS	92.87	95.88
our tagger POS+morph	84.60	89.45

Table 3: POS and morphological tagging accuracies on the ReM corpus

	dev	test
Middle Dutch	91.10	91.01
Middle French	96.45	96.28
Old English	97.37	97.13
Old Greek	91.35	91.29
Old Icelandic	93.88	89.87
Old Italian	98.46	98.43

Table 4: Tagging accuracy on other historical corpora

4.3 Tagging Experiments on ReM

In a second series of tagging experiments, we evaluated our tagger on the ReM corpus obtaining a test accuracy of 95.9% in the POS tagging task and 89.5% for POS+morph tagging (see Table 3). The POS+morph accuracy on unseen words was 79% and the accuracy on seen words with an unseen tag was 57%. Table 3 also shows the accuracy[12] of the TreeTagger [25] which has been used in previous work on this corpus [6].

4.4 Experiments With Other Languages

We carried out experiments on further historical corpora using the optimal metaparameters from the ReM experiments. We used the middle-Dutch Gysseling corpus[13], the Syntactic Reference Corpus of Medieval French [28], the York-Toronto-Helsinki Parsed Corpus of Old English Prose [29], the Ancient Greek Dependency Treebank 2.1 [14], the Icelandic Parsed Historical Corpus [22], and the old-Italian Corpus Taurinese[15]. Table 4 shows the results obtained on these corpora.

4.5 Extraction of Negation Clitics

As a practical application, we used our POS tagger trained on the ReM corpus to extract words with clitic negation particles, such as *enfprecheft* (not speak+2sg) – where the prefix *en* is the negation particle. The corresponding POS tag is *PTKNEG-VVFIN*. We were able to identify such words based on the assigned POS tag with a precision of 96.4% and a recall of 97.8% in the ReM test data.

4.6 Lemmatization Experiments

For our lemmatization experiments, we trained the dl4mt system with character embeddings of size 300 and LSTM hidden states of size 600 with Adadelta and a learning rate of 0.001. On the test data, which consists of unseen word-tag pairs, we achieve an accuracy of 91.9% (cmp. Table 5). The accuracy on unseen words is 86.4%.

	dev	test
overall	92.52	91.87
unseen words	-	86.43
seen words with unseen tags	-	93.33

Table 5: Lemmatization accuracy on ReM

5 CONCLUSIONS

We presented a POS tagger based on bidirectional LSTMs which achieves state-of-the-art performance on modern German morphological POS tagging. We showed that the tagger is well suited for the POS tagging of historical texts and outperforms previous systems on the GerManC corpus with Early New High-German texts as well as the ReM corpus with Middle High-German texts. We also tested the tagger on six other historical corpora. Finally, we applied a standard encoder-decoder system to the lemmatization of the ReM texts with good results.

The code of our tagger and pretrained models for different languages will be made publicly available after acception of the paper.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (June 2017), 135–146.

[3] Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 11–18.

[4] Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP*. Seattle, WA.

[5] Stefanie Dipper. 2010. POS-Tagging of Historical Language Data: First Experiments. In *KONVENS*. 117–121.

[6] Stefanie Dipper. 2012. Morphological and part-of-speech tagging of historical language data: A comparison. In *Workshop on Annotation of Corpora*.

[7] Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics (JLCL)* 28, 1 (2013), 85–137.

[12] TreeTagger parameters were optimized on dev data.

[13] available at http://tst-centrale.org/nl/tst-materialen/corpora/corpus-gysseling-detail

[14] available at https://github.com/PerseusDL/treebank_data

[15] available at http://www.bmanuel.org/projects/ct-HOME.html

[8] Yarin Gal and Zoubin Ghahramani. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 1019–1027.

[9] Christian Hardmeier. 2016. A Neural Model for Part-of-Speech Tagging in Historical Texts. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING): Technical Papers*. Osaka, Japan, 922–931.

[10] Georg Heigold, Guenter Neumann, and Josef van Genabith. 2016. Neural Morphological Tagging from Characters for Morphologically Rich Languages. *CoRR* abs/1606.06640 (2016). http://arxiv.org/abs/1606.06640

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[12] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). http://arxiv.org/abs/1508.01991

[13] Dieuwke Hupkes and Rens Bod. 2016. POS-tagging of Historical Dutch. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*. Portoroz, Slovenia, 77–82.

[14] Katharina Kann and Hinrich Schütze. 2016. Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, 555–560. http://anthology.aclweb.org/P16-2090

[15] Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. 2016. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities* (2016).

[16] Thomas Klein and Stefanie Dipper. 2016. *Handbuch zum Referenzkorpus Mittelhochdeutsch* (Bochumer Linguistische Arbeitsberichte 19 ed.). Ruhr-Universität Bochum, Bochum, Germany.

[17] Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, 1520–1530.

[18] Taesun Moon and Jason Baldridge. 2007. Part-of-Speech Tagging for Middle English through Alignment and Projection of Parallel Diachronic Texts.. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learnin (EMNLP-CoNLL)*. 390–399.

[19] Thomas Müller and Hinrich Schütze. 2015. Robust Morphological Tagging with Word Representations.. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT)*. 526–536.

[20] Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics*. University of Birmingham, UK.

[21] Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In *Language Technology for Cultural Heritage*. Springer, 63–76.

[22] E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of LREC*. 1978–1984.

[23] Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*. Association for Computational Linguistics, 19–23.

[24] Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 124–128.

[25] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, Great Britain, 44–49.

[26] Gerold Schneider, Marianne Hundt, and Rahel Opplinger. 2016. Part-of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*. Bochum, Germany, 256–264.

[27] Gilles Souvay and Jean-Marie Pierrel. 2009. LGeRM Lemmatisation des mots en Moyen Français. *Traitement Automatique des Langues (ATALA)* 50, 2 (2009), 21.

[28] Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In *New Methods in Historical Corpora Corpus Linguistics and International Perspectives on Language*, Paul Bennett, Martin Durrell, Silke Scheible, and Richard Whitt (Eds.). CLIP, Vol. 3. Narr, Tübingen, 275–282.

[29] A. Taylor. 2007. The York-Toronto-Helsinki parsed corpus of old English prose. In *Creating and digitizing language corpora*, J. C. Beal, K. P. Corrigan, and H. L. Moisl (Eds.). Palgrave Macmillan, London.

[30] Hans van Halteren and Margit Rem. 2013. Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters. *Language resources and evaluation* 47, 4 (2013), 1233–1259.

[31] Yi Yang and Jacob Eisenstein. 2016. Part-of-Speech Tagging for Historical English. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics, San Diego, CA, 1318–1328.