Learning from Within? Comparing PoS Tagging Approaches for Historical Text

Sarah Schulz, Jonas Kuhn
Institute for Natural Language Processing (IMS)
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart
firstname.lastname@ims.uni-stuttgart.de

Abstract

In this paper, we investigate unsupervised and semi-supervised methods for part-of-speech (PoS) tagging in the context of historical German text. We locate our research in the context of Digital Humanities where the non-canonical nature of text causes issues facing an Natural Language Processing world in which tools are mainly trained on standard data. Data deviating from the norm requires tools adjusted to this data. We explore to which extend the availability of such training material and resources related to it influences the accuracy of PoS tagging. We investigate a variety of algorithms including neural nets, conditional random fields and self-learning techniques in order to find the best-fitted approach to tackle data sparsity. Although methods using resources from related languages outperform weakly supervised methods using just a few training examples, we can still reach a promising accuracy with methods abstaining additional resources.

Keywords: PoS tagging, historical language, self-taught learning, neural nets, stacking, Digital Humanities, low resource languages

1. Introduction

PoS (part-of-speech) tagging is a preprocessing step which is indispensable for many Natural Language Processing (NLP) tasks and needs to be done with high accuracy to ensure success in the subsequent task. Therefore, it is a well understood field offering a variety of techniques suitable for different languages. Schmid (1995) reports PoS tagging accuracy of 97.5% on German newspaper text.

However, most of these approaches use a large number of training examples. Even if it comes to unsupervised methods, the unlabeled amount of data has to be satisfactorily large. Yet, there are scenarios in which neither labeled nor unlabeled data is sufficiently available.

Digital Humanities (DH) is a field of research that is developing fast. It holds its very own kind of challenges and scientific issues both interesting for Humanities and Computer Science. One of the problems for NLP is the non-canonical nature of text especially found in projects dealing with historical text. These texts deviate from text that has been the main focus of NLP so far by lacking standardized orthography and grammar. This can often lead to a decrease in performance of tools trained on standard text (Melero et al., 2012; Eisenstein, 2013). Data needed for training dedicated tools for these texts is often not available. However, in case the non-canonical text is related to another language, e.g., the modern stage of the language, we can exploit this relatedness to facilitate tagging.

To illustrate this, we investigate PoS tagging of a unique late Middle High German (MHG) text in the transition period between MHG and Early New High German (ENHG). This leads to a text with mixed features of two historical stages of German.

In this paper, we investigate PoS tagging for historical texts sharing a lot of the challenges of PoS tagging for low resourced languages. We do this by means of what we call expanding exploration. We compare different approaches towards boosting performance of PoS tagging of text for which no suitable PoS tagger is available and no or really limited annotated data is available. Departing from the assumption that we have no text-external resources to our disposal, we experiment with unsupervised and weakly supervised learning methods. Moreover, we follow experiments performed by Garrette and Baldridge (2013) who describe PoS tagging research for low resourced languages using really small amounts of annotated data. Expanding to text-external resources, we include taggers that have been developed for related languages into our experiments. The aim of this study is to evaluate the performance of PoS tagging considering different supply conditions of data and related external resources. This can serve as a reference point for DH projects. We strive for a better idea of how one can gain performance in such a context by investing time in developing resources. The approximation of such a gain is an important consideration given that those texts often have very specific characteristics and developed resources might not be reused in another context.

2. Related Work

Completely unsupervised PoS tagging is still in its very early stages. Biemann (2006) relies on a graph clustering method. Unlike in current state-of-the-art approaches, the kind and number of different tags are generated by the method itself. Contrary to this, Haghighi and Klein (2006) use distributional prototypes in the learning process of their log-linear model. This way they inform the algorithm indirectly about the PoS classes. These unsupervised or semi-supervised approaches make use of distributional semantics (Turian et al., 2010). In this context, the use of word embeddings have to be mentioned. Their ability to capture syntactic and semantic regularities (Mikolov et al., 2013) can be utilized to compensate for the high number of hapax legomena in sparse data. Word embeddings have been used by Lin et al. (2015) for unsupervised PoS induction.

Weakly supervised techniques can involve supervision of different degrees and of different kinds. There exist approaches using parallel data like Moon and Baldridge

set	# sentences	av. # tokens
train	100	1374
dev	100	1372
test	50	688

Table 1: Average number of sentences and tokens in train, development and test set of our gold standard.

(2007) who use aligned text to compensate for the lack of annotated data in the language under investigation. Sánchez-Martínez et al. (2007) unsupervisedly train a HMM-based Occitan part-of-speech tagger used within an MT system using translation probabilities given tag assignments to inform the HMM. Agic et al. (2015) introduce an approach using the bible as a parallel corpus aggregating over the tags from annotated languages. This way, they train PoS taggers for 100 languages like Cakchiquel and Akawaio. Das and Petrov (2011) locate their approach on the unsupervised side, however, they use translated text in a resource-rich language for cross-lingual knowledge transfer. Several other approaches utilize lexicons providing the learning algorithm with possible valid PoS for a part of the vocabulary (Ravi and Knight, 2009). Garrette and Baldridge (2013) show that there is no need for huge annotated corpora but that reasonable results can be achieved by generalizing from just a little amount of annotated data.
Moreover, simple PoS taggers developed for closely-related languages can be applied as done in Zeman and Resnik (2008). This requires a proper mapping from one tag set to another.
In the field of low resource language processing, not just parallel data of closely-related languages is used, but the task is often tackled as domain-adaptation of tools developed for a related language. Blitzer et al. (2006) introduce structural correspondence learning for domain adaptation from newspaper text to the biomedical domain also for the setting when there is no labeled data from the target domain.

Dipper (2010) and Bollmann (2013) concentrate on PoS tagging of historical German text using normalization to modern German as a preprocessing step before applying a PoS tagger for standard German. This approach requires the availability of a normalizer and tagging quality is highly dependent on the normalization success. Moreover, the heterogeneity of texts even within the same time period is a crucial issue for an approach using text normalization.
Being confronted with a diversity of methods to tackle PoS tagging for underresourced languages, we investigate those being feasible regarding our data situation. Therefore, we focus on weak supervision following Garrette and Baldridge (2013), the unsupervised approach by Biemann (2006), model transfer similar to Zeman and Resnik (2008) and fathom out the opportunities that word embeddings (Mikolov et al., 2013) and combinations of methods hold.

2.1. Data

Middle High German texts are characterized by their high degree of diversity with respect to graphematic realization and choice of vocabulary (Dipper, 2010). Depending on the exact period and point of origin, the author and even the printer, a text may or may not be readable even for native speakers of modern German. In fact, even though MHG constitutes an early phase of nowadays German, it differs significantly with respect to different linguistic features. These characteristics make it impossible to directly use any off-the-shelf tool for automatic processing of this kind of text and moreover complicate the development of domain specific tools. We work on Heinrich von Neustadt's *Apollonius von Tyrland*[1], a 20,645 verses long opus containing approximately 180,000 types and 800,000 tokens. Heinrich von Neustadt lived in the 13th century and just two writings, the other one being *Gottes Zukunft*, can be attributed to him. Considering these two texts as an independent text domain, this leaves us with a quite limited amount of data. Moreover, the language he uses can be located in an intermediate phase between Middle High German and Early High German. This is crucial to know since this means that neither tools developed for MHG (Dipper, 2010; Bollmann, 2013) nor for standard German will work reliably. However, its relative closeness to both can nevertheless be beneficial.
We annotated 250 sentences comprising about 3625 tokens with a simplified version of the HiTS for historical German (Dipper et al., 2013). We use train and development sets of 100 sentences each and a test set of 50 sentences (Table 1).

3. Learning from Within the Text

In the first phase of our experimentation we are evaluating different techniques using nothing but the text at hand. We use unsupervised methods as well as weakly supervised techniques.
Training a tagger from scratch, we are confronted with the issue of extreme data sparsity. Different from a low resourced language, our text at hand provides us with just some thousand sentences in total (including a high number of hapax legomena) and considerably less annotated data. Thus, abstraction from the surface form is preferable. In the context of language modeling with the help of neural networks, it has been shown helpful to train so called word embeddings (e.g., Mikolov et al 2013, Lebret et al 2013). These embeddings are high dimensional vectors representing features of words in a high feature space and are able to capture syntactic and semantic regularities (Mikolov et al., 2013). These characteristics make them a good departure point for a scenario in which one faces data sparsity.
We train 64-dimensional word embeddings using word2embeddings (Al-Rfou et al., 2013) and a window size of five tokens on our entire corpus. Although this oversteps the clear division between training and test data because those vectors summerize the context of

[1]Based on the Gotha manuscript edited by Samuel Singer, Berlin 1906. Digitalized version from `http://www.mhgta.uni-trier.de` (Gärtner, 2002).

STTS tags	HiTS tags	sHiTS tag
ADJA, ADJD	ADJA, ADJD, ADJN, ADJS	ADJ
APPR, APPRART, APPO	APPO, APPR	APP
ADV, PWAV	AVD, AVG, AVNEG, AVW	AV
CARD	CARDA, CARDD, CARDN, CARDS	CARD
ART, PDAT, PDS, PIAT, PIDAT	DDA, DDART, DDD, DDN, DDS,DIA, DIART, DID, DIN, NIS	DD, DI
PIAT	DNEGA, DNEGD, DNEGN, DNEGS	DNEG
APZR, TRUNC, PTK, PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA	PRKA, PTKANT, PTKINT, PTKNEG, PTKREL, PTKVZ	PTK
DRELS, PRELAT, PRELS	DRELS	DRELS
PPOSAT, PPOSS	DPOSA, DPOSD, DPOSGEN, DPOSN, DPOSS	DPOS
PIS, PPER, PRF	PRF,PG, PI, PNEG, PPER	PR
PWAT, PWS	DGA, DGD, DGN, DGS	DG
PWAT, PWS	DWA, DWD, DWN, DWS	DW
KON, KOUS, KOKOM, KOUI	KO*, KOKOM, KON, KOUS	KO
NN	NA	NA
NE	NE	NE
PAV, PWAV	PAVAP, PAVD, PAVG, PAVREL, PAVW	PAV
VVFIN, VVIMP, VVINF, VVIZU, VVPP	VVFIN, VVIMP, VVINF, VVPP, VVPS	VV
VAFIN, VAIMP, VAINF, VAPP	VAFIN, VAIMP, VAINF, VAPP, VAPS	VA
VMFIN, VMINF, VMPP	VMFIN, VMIM, VNINF, VMPP, VMPS	VM
PWS, PWAT, PWAV	PW	PW
ITJ	ITJ	ITJ
FM	FM	FM
$, $., $(	$,, $(	$

Table 2: Mapping between STTS for German, HiTS for historical German and sHiTS.

the words in the entire corpus, we consider this a valid approach since we can assume the same treatment during application to rest of the corpus. Moreover, we do not claim generalizability of our tagger to other data but are driven by the goal to tag in-domain text.

Word embeddings are used as a way to abstract from surface form in two of our approaches: in an unsupervised clustering approach and for training a multilayer perceptron neural net (MNN). We compare these approaches to a sequence labeling approach (CRF) (Lafferty et al., 2001) using just surface forms. To compare the performance of different neural net architectures, we additionally experiment with a long short-term memory neural net. Moreover, we investigate self-learning for the MNN and the CRF training aiming at further improvement.

3.1. K-Means Clustering

We experiment with **k-means clustering** informing the cluster analysis (CA) algorithm with the number of PoS we have annotated in our gold standard. Moreover, we initialize our cluster centroids with prototypical words from the training data for each PoS inspired by Haghighi and Klein (2006).[2] This rather simple approach does not take the sequence in which words appear in the text into account but relies only on the context information encoded in the word embeddings. This means that each token can only be assigned to one PoS.

3.2. Neural Net

We train both a **multi-layer perceptron (MLP) neural net** using nlpnet (Fonseca et al., 2013) on our word embeddings and a **long-short-term-memory (LSTM) neural net** using an integrated compositional character to word (C2W) model based on a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) using the Java Neural Network (JNN) Toolkit (Ling et al., 2015). Since the JNN toolkit allows to extend the feature space by additional features, we add information on the word, suffix of 3 and prefix of 3 to the training process. A comparison of the performance of these two architectures is interesting since LSTMs are known to capture long term dependencies and and could therefore perform better in learning the structure of sentences.

3.3. Conditional Random Field

We train a **CRF** tagger (Lafferty et al., 2001) using a context window of 5 tokens and 6 features. We include the following features for each token:

- token is punctuation or not
- word length
- character prefix of length 2
- character prefix of length 3
- character suffix of length 2
- character suffix of length 3

3.4. Self-Learning

With the intention to overcome the sparsity of training data, we apply self-learning. We tag the unannotated part of our corpus with the CRF tagger and the neural net tagger, respectively. Subsequently, we sort the automatically tagged sentences by tagging confidence (Viterbi scores for the neural net and the conditional probability for the CRF) and add the best 200 sentences to our training data and retrain the tagger. We evaluate the performance before and after extension of the training data on the development set. In case the performance increases after extension, we keep the new classifier and start the next iteration by tagging the unannotated data anew. In case the performance decreases, we discard the new classifier and append the next 200 sentences of the automatically tagged data. This way we extend our training set by in average 6 times[3] for neural net training. Surprisingly, we cannot improve the CRF tagger, at all. To make sure that the batch size of 200 sentences is not too big, we experiment with 100, 50 and 1. However, we consistently experience a decrease in performance even when

[2]This also facilitates the evaluation because clusters can be mapped to PoS more easily.

[3]In randomized sub-sampling setting.

just adding one automatically tagged sentence from our raw corpus to the training data.

4. Stretching Out: Including Text-External Resources

Following the assumption that closely related languages have similar features, applying taggers trained for those languages is promising. We use the TreeTagger for German (Schmid, 1994) and a TreeTagger model trained on Middle High German data (Dipper, to appear; Rothmund, 2015). We map the STTS (Schiller et al., 1995) and the HiTS (Dipper et al., 2013) to a simplified version of the HiTS developed by Rothmund (2015).

Suspecting that different models have different strengths, we use the meta-learning method of stacking (Wolpert, 1992) to combine these advantages. We use the predictions of the weakly-supervised CRF classifier and the neural net classifier along with the predictions of the tree tagger models for MHG and NHG on the development for training a meta-learner. The meta-learner we use is a CRF classifier (Lafferty et al., 2001).

Moreover, we implement tritraining (Zhou and Li, 2005).

4.1. Model Transfer

Working on text with characteristics from both NHG and MHG, we use a tagger built for German and Middle High German respectively for our data. An issue arising from this otherwise simple approach of transferring the model of a related language to another is the mapping of the tag sets. This process of mapping one tag set to the other is accompanied by a loss of information considering that even though languages might be related, they rarely cover exactly the same space of grammatical features.

The mapping of the different tag sets is described in Table 2.

4.2. Stacking

To combine the knowledge of the MNN tagger, the CRF classifier introduced in Section 3 and the two taggers for closely-related languages, we implement stacking (Wolpert, 1992) using a CRF meta-learner. The meta-learner uses the surface form of the word and the PoS tags attributed by each of the four taggers as features and a context window of 5 on all these features.

4.3. Tritraining

As another implementation of self-learning, this time using external classifiers, we use tritraining (Zhou and Li, 2005). We use two classifiers, our external taggers for NHG and MHG, to inform our third classifier about which sentence from the unlabeled data set to add to the training process. For this decision, we choose simple agreement of both classifiers. If they agree on the tagging of an entire sentence, this sentence is added to the training data.

5. Evaluation

5.1. Results

It is a challenge to evaluate the clustering performance and not a combination between clustering and mapping induction to the PoS classes. Moreover, evaluation on a gold standard for PoS tagging seems counter-intuitive given that the clustering is not informed about the task at hand. Vlachos (2011) advocates the evaluation as clustering-based word representation induction. Extrinsic evaluation is suggested as a solution to this problem. Having all these drawbacks in mind, we evaluate the overlap between the clustering results and the gold standard data without drawing strict conclusions about the usefulness of the clustering results for downstream tasks. To facilitate the mapping and weakly inform the clustering about the task at hand, we use a typical word for each PoS as seed for each cluster inspired by prototype learning introduced by Haghighi and Klein (2006). This leaves us with four clusters in which none of the prototype words can be found and four clusters containing two of them. In favor of the clustering method, we assume a clusters containing two prototype words to cover both their PoS classes. Those not containing any of the prototype words are analyzed with the help of the data in our gold standard. The PoS most often found in the gold standard for the words in the cluster is attributed to it.

We evaluate our experiments in a 10-fold Monte Carlo cross-validation setting. Accuracy scores for all experiments averaged over all 10 samples are given in Table 3 along with the standard deviation for the 10 samples. The stacking approach outperforms all of the other approaches, the best single-handed classification can be achieved by applying the MHG tagger to our data. Statistical significance is calculated using McNemar's test (McNemar, 1947). We compare differences in accuracy between all methods. The LSTM neural net performs better than all other methods without external resources with an accuracy score of 0.72. Model transfer from MHG and stacking outperform all approaches without external resources with accuracy scores of 0.73 and 0.77, respectively. However, the LSTM can outperform two of the extended approaches, namely model transfer from NHG and tritraining. Least convincing are unsupervised clustering, the MNN neural net and MNN neural net self-learning.

5.2. Discussion

Non-surprisingly, approaches using external resources perform generally better than approaches without external resources. However, also approaches only relying on a few annotated sentences achieve results that can serve as basis for the investigation of many research questions in DH projects. Especially results achieved using an LSTM neural net are convincing. We want to emphasize that our weakly supervised methods make use of about 2000 tokens opposed to e.g. 410,000 tokens used for training of the MHG tagger. However, self-training approaches do not show any improvement but rather lower accuracy. Clustering in turn has to be evaluated in an extrinsic setting in order to make reliable statements about the usefulness.

6. Conclusion

In this paper, we give directions towards the tagging of languages or domains for which no labeled data is available. We can show that even for very specific texts we can successfully apply semi-supervised methods. Already using a

	Without external resources					With external resources			
Method	**CA**	**CRF**	**MNN**	**LSTMNN**	**MNN self**	**NHG**	**MHG**	**ST**	**TRI**
CA	0.18 0.015	α_2	α_2	α_2	α_2	α_2	α_2	α_2	α_2
CRF		0.69 0.022		α_2			α_2	α_2	
MNN		α_2	0.66 0.015	α_2			α_2	α_2	α_2
LSTMNN				**0.72** 0.024			α_1	α_2	
MNN self		α_2		α_2	0.66 0.015		α_2	α_2	α_2
NHG		α_2	α_2	α_2	α_2	0.57 0.024	α_2	α_2	α_2
MHG							0.73 0.011	α_2	
ST								**0.77** 0.014	
TRI				α_1			α_2	α_2	0.70 0.024

Table 3: On the diagonal, accuracies of all PoS tagging approaches evaluated in a 10-fold Monte Carlo cross-validation setting along with the standard deviation for the 10 samples are reported. Accuracy is given in the top field, standard deviation is given in the bottom field. On the left-hand side the results for experiments without external resources are listed in the following order: clustering (CA), conditional random fields classifier (CRF), MLP neural net (MNN), LSTM neural net (LSTMNN) and MNN self-learning (MNN self). On the right-hand side the results for experiments with external resources are listed in the following order: model transfer from New High German (NHG) and Middle High German (MHG), stacking (ST) and tritraining (TRI). Label α_1 and α_2 indicate that the p-value for the improvement of the classifier in the column over the classifier in the row is significant on the significance level $alpha_1 = 0.05$ and $alpha_2 = 0.001$, respectively. The shades of gray emphasize the respective groups (with and without external resources).

small amount of annotated data can lead to reasonable results using LSTM neural nets. Adding resources developed for related languages boosts results even further. In the future, we plan to perform experiments for other sorts of texts, in order to see whether our results can be generalized beyond one domain and serve as a guideline for PoS tagging of low resource languages in general. Moreover, we aim at the evaluation of our results within the context of a DH project in need of PoS-tagged text.

7. Acknowledgements

This research has been done in the context of the Center for Reflected Text Analytics (CRETA) and has been funded by the German Ministry of Education and Research (BMBF). Moreover, we want to thank Manuel Braun and his research group for annotations and helpful suggestions with respect to the Apollonius text.

Zeljko Agic, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 268–272.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chris Biemann. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, COLING ACL '06, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcel Bollmann. 2013. Pos tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability in Discourse (LAW7-ID)*, pages 11–18.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. Hits: ein tagset für historische sprachstufen des deutschen. *JLCL*, 28(1):85–137.

Stefanie Dipper. 2010. POS-tagging of historical language data: First experiments. In Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde, and Angelika Storrer, editors, *Semantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2010 (KONVENS)*, pages 117–121, Saarbrücken, Germany. Universaar.

Stefanie Dipper. to appear. Annotierte korpora für die historische syntaxforschung: Anwendungsbeispiele anhand des referenzkorpus mittelhochdeutsch. *Zeitschrift für Germanistische Linguistik*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. Association for Computational Linguistics.

Erick Rocha Fonseca, João Luís, and G. Rosa. 2013. Macmorpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 98–107.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-13)*, pages 138–147, Atlanta, GA, June.

Kurt Gärtner. 2002. Comprehensive digital text archives: A digital middle high german text archive and its perspectives. In *First EU/NSF Digital Libraries All Projects Meeting*.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 320–327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori S. Levin. 2015. Unsupervised POS induction with word embeddings. *CoRR*, abs/1503.06760.

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. *CoRR*, abs/1508.02096.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Maite Melero, Marta R. Costa-Jussà, Judith Domingo, Montse Marquina, and Martí Quixal. 2012. Holaaa!! writin like u talk is kewl but kinda hard 4 nlp. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3794–3800, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 390–399.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 504–512, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beate Rothmund. 2015. Automatic extracton of keyphrases from middle high german texts. Bachelor thesis, Universitä Stuttgart.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Felipe Sánchez-Martínez, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. *Procesamiento del Lenguaje Natural*, pages –1–1.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for

semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Vlachos. 2011. Evaluating unsupervised learning for natural language processing tasks. In *Proceedings of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 35 – 42, Edinburgh, UK.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5:241–259.

D. Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. *NLP for Less Privileged Languages*, pages 35 – 35.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.