Joint Lemmatization and Morphological Tagging with LEMMING

Thomas Müller[1] **Ryan Cotterell**[1,2]
Center for Information and Language Processing[1]
University of Munich, Germany
muellets@cis.lmu.de

Alexander Fraser[1] **Hinrich Schütze**[1]
Department of Computer Science[2]
Johns Hopkins University, USA
ryan.cotterell@jhu.edu

Abstract

We present LEMMING, a modular log-linear model that jointly models lemmatization and tagging and supports the integration of arbitrary global features. It is trainable on corpora annotated with gold standard tags and lemmata and does not rely on morphological dictionaries or analyzers. LEMMING sets the new state of the art in token-based statistical lemmatization on six languages; e.g., for Czech lemmatization, we reduce the error by 60%, from 4.05 to 1.58. We also give empirical evidence that jointly modeling morphological tags and lemmata is mutually beneficial.

1 Introduction

Lemmatization is important for many NLP tasks, including parsing (Björkelund et al., 2010; Seddah et al., 2010) and machine translation (Fraser et al., 2012). Lemmata are required whenever we want to map words to lexical resources and establish the relation between inflected forms, particularly critical for morphologically rich languages to address the sparsity of unlemmatized forms. This strongly motivates work on language-independent token-based lemmatization, but until now there has been little work (Chrupała et al., 2008).

Many regular transformations can be described by simple replacement rules, but lemmatization of unknown words requires more than this. For instance the Spanish paradigms for verbs ending in *ir* and *er* share the same 3rd person plural ending *en*; this makes it hard to decide which paradigm a form belongs to.[1] Solving these kinds of problems requires global features on the lemma. Global features of this kind were not supported by previous work (Dreyer et al., 2008; Chrupała, 2006; Toutanova and Cherry, 2009; Cotterell et al., 2014).

There is a strong mutual dependency between (i) lemmatization of a form in context and (ii) disambiguating its part-of-speech (POS) and morphological attributes. Attributes often disambiguate the lemma of a form, which explains why many NLP systems (Manning et al., 2014; Padró and Stanilovsky, 2012) apply a pipeline approach of tagging followed by lemmatization. Conversely, knowing the lemma of a form is often beneficial for tagging, for instance in the presence of syncretism; e.g., since German plural noun phrases do not mark gender, it is important to know the lemma (singular form) to correctly tag gender on the noun.

We make the following contributions. (i) We present the first joint log-linear model of morphological analysis and lemmatization that operates at the *token* level and is also able to lemmatize unknown forms; and release it as open-source (`http://cistern.cis.lmu.de/lemming`). It is trainable on corpora annotated with gold standard tags and lemmata. Unlike other work (e.g., (Smith et al., 2005)) it does not rely on morphological dictionaries or analyzers. (ii) We describe a log-linear model for lemmatization that can easily be incorporated into other models and supports arbitrary global features on the lemma. (iii) We set the new state of the art in token-based statistical lemmatization on six languages (English, German, Czech, Hungarian, Latin and Spanish). (iv) We experimentally show that jointly modeling morphological tags and lemmata is mutually beneficial and yields significant improvements in joint (tag+lemma) accuracy for four out of six languages; e.g., Czech lemma errors are reduced by $>37\%$ and tag+lemma errors by $>6\%$.

[1] Compare *admiten* "they admit" → *admitir* "to admit", but *deben* "they must" → *deber* "to must".

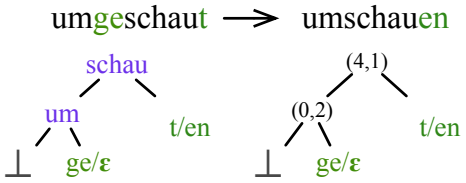

Figure 1: Edit tree for the inflected form *umgeschaut* "looked around" and its lemma *umschauen* "to look around". The right tree is the actual edit tree we use in our model, the left tree visualizes what each node corresponds to. The root node stores the length of the prefix *umge* (4) and the suffix *t* (1).

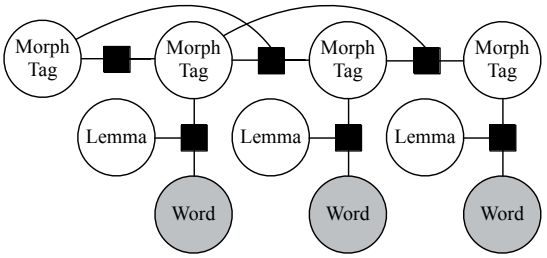

Figure 2: Our model is a 2nd-order linear-chain CRF augmented to predict lemmata. We heavily prune our model and can easily exploit higher-order (>2) tag dependencies.

2 Log-Linear Lemmatization

Chrupała (2006) formalizes lemmatization as a classification task through the deterministic pre-extraction of edit operations transforming forms into lemmata. Our lemmatization model is in this vein, but allows the addition of external lexical information, e.g., whether the candidate lemma is in a dictionary. Formally, lemmatization is a string-to-string transduction task. Given an alphabet Σ, it maps an inflected form $w \in \Sigma^*$ to its lemma $l \in \Sigma^*$ given its morphological attributes m. We model this process by a log-linear model:

$$p(l \mid w, m) \propto h_w(l) \cdot \exp\left(\boldsymbol{f}(l, w, m)^T \boldsymbol{\theta}\right),$$

where $\boldsymbol{f}$ represents hand-crafted feature functions, $\boldsymbol{\theta}$ is a weight vector, and $h_w : \Sigma^* \rightarrow \{0, 1\}$ determines the support of the distribution, i.e., the set of candidates with non-zero probability.

Candidate selection. A proper choice of the support function $h(\cdot)$ is crucial to the success of the model – too permissive a function and the computational cost will build up, too restrictive and the correct lemma may receive no probability mass. Following Chrupała (2008), we define $h(\cdot)$ through a deterministic pre-extraction of *edit trees*. To extract an edit tree e for a pair form-lemma $\langle w, l \rangle$, we first find the longest common substring (LCS) (Gusfield, 1997) between them and then recursively model the prefix and suffix pairs of the LCS. When no LCS can be found the string pair is represented as a substitution operation transforming the first string to the second. The resulting edit tree does not encode the LCSs but only the length of their prefixes and suffixes and the substitution nodes (cf. Figure 1); e.g., the same tree transforms *worked* into *work* and *touched* into *touch*.

As a preprocessing step, we extract all edit trees that can be used for more than one pair $\langle w, l \rangle$. To generate the candidates of a word-form, we apply all edit trees and also add all lemmata this form was seen with in the training set (note that only a small subset of the edit trees is applicable for any given form because most require incompatible substitution operations).[2]

Features. Our novel formalization lets us combine a wide variety of features that have been used in different previous models. All features are extracted given a form-lemma pair $\langle w, l \rangle$ created with an edit tree e.

We use the following three *edit tree features* of Chrupała (2008). (i) The edit tree e. (ii) The pair $\langle e, w \rangle$. This feature is crucial for the model to memorize irregular forms, e.g., the lemma of *was* is *be*. (iii) For each form affix (of maximum length 10): its conjunction with e. These features are useful in learning orthographic and phonological regularities, e.g., the lemma of *signalling* is *signal*, not *signall*.

We define the following *alignment features*. Similar to Toutanova and Cherry (2009) (TC), we define an alignment between w and l. Our alignments can be read from an edit tree by aligning the characters in LCS nodes character by character and characters in substitution nodes block-wise. Thus the alignment of *umgeschaut* - *umschauen* is: u-u, m-m, ge-ϵ, s-s, c-c, h-h, a-a, u-u, t-en. Each alignment pair constitutes a feature in our model. These features allow the model to learn that the substitution *t/en* is likely in German. We also concatenate each alignment pair with its form and lemma character context (of up to length 6) to learn, e.g., that *ge* is often deleted after *um*.

We define two simple *lemma features*. (i) We use the lemma itself as a feature, allowing us to learn which lemmata are common in the language. (ii) Prefixes and suffixes of the lemma (of maxi-

[2] Pseudo-code for edit tree creation and candidate lemma generation with examples can be found in the appendix (`http://cistern.cis.lmu.de/lemming/appendix.pdf`).

mum length 10). This feature allows us to learn that the typical endings of Spanish verbs are *ir*, *er*, *ar*.

We also use two *dictionary features* (on lemmata): Whether l occurs > 5 times in Wikipedia and whether it occurs in the dictionary ASPELL.[3] We use a similar feature for different capitalization variants of the lemma (lowercase, first letter uppercase, all uppercase, mixed). This differentiation is important for German, where nouns are capitalized and *en* is both a noun plural marker and a frequent verb ending. Ignoring capitalization would thus lead to confusion.

POS & morphological attributes. For each feature listed previously, we create a conjunction with the POS and each morphological attribute.[4]

3 Joint Tagging and Lemmatization

We model the sequence of morphological tags using MARMOT (Müller et al., 2013), a pruned higher-order CRF. This model avoids the exponential runtime of higher-order models by employing a pruning strategy. Its feature set consists of standard tagging features: the current word, its affixes and shape (capitalization, digits, hyphens) and the immediate lexical context. We combine lemmatization and higher-order CRF components in a tree-structured CRF. Given a sequence of forms $\boldsymbol{w}$ with lemmata $\boldsymbol{l}$ and morphological+POS tags $\boldsymbol{m}$, we define a globally normalized model:

$$p(\boldsymbol{l}, \boldsymbol{m} \mid \boldsymbol{w}) \propto \prod_i h_{w_i}(l_i) \exp(\boldsymbol{f}(l_i, w_i, m_i)^T \boldsymbol{\theta} + \boldsymbol{g}(m_i, m_{i-1}, m_{i-2}, \boldsymbol{w}, i)^T \boldsymbol{\lambda}),$$

where $\boldsymbol{f}$ and $\boldsymbol{g}$ are the features associated with lemma and tag cliques respectively and $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are weight vectors. The graphical model is shown in Figure 2. We perform inference with belief propagation (Pearl, 1988) and estimate the parameters with SGD (Tsuruoka et al., 2009). We greatly improved the results of the joint model by initializing it with the parameters of a pretrained tagging model.

4 Related Work

In functionality, our system resembles MORFETTE (Chrupała et al., 2008), which generates lemma candidates by extracting edit operation sequences between lemmata and surface forms (Chrupała, 2006), and then trains two maximum entropy Markov models (Ratnaparkhi, 1996) for morphological tagging and lemmatization, which are queried using a beam search decoder.

In our experiments we use the latest version[5] of MORFETTE. This version is based on structured perceptron learning (Collins, 2002) and edit trees (Chrupała, 2008). Models similar to MORFETTE include those of Björkelund et al. (2010) and Gesmundo and Samardzic (2012) and have also been used for generation (Dušek and Jurčíček, 2013). Wicentowski (2002) similarly treats lemmatization as classification over a deterministically chosen candidate set, but uses distributional information extracted from large corpora as a key source of information.

Toutanova and Cherry (2009)'s joint morphological analyzer predicts the set of possible lemmata and *coarse-grained POS* for a word *type*. This is different from our problem of lemmatization and *fine-grained* morphological tagging of *tokens in context*. Despite the superficial similarity of the two problems, *direct comparison is not possible*. TC's model is best thought of as inducing a tagging dictionary for OOV types, mapping them to a set of tag and lemma pairs, whereas LEMMING is a token-level, context-based morphological tagger.

We do, however, use TC's model of lemmatization, a string-to-string transduction model based on Jiampojamarn et al. (2008) (JCK), as a standalone baseline. Our tagging-in-context model is faced with higher complexity of learning and inference since it addresses a more difficult task; thus, while we could in principle use JCK as a replacement for our candidate selection, the edit tree approach – which has high coverage at a low average number of lemma candidates (cf. Section 5) – allows us to train and apply LEMMING efficiently.

Smith et al. (2005) proposed a log-linear model for the context-based disambiguation of a morphological dictionary. This has the effect of joint tagging, morphological segmentation and lemmatization, but, critically, is limited to the entries in the morphological dictionary (without which the approach cannot be used), causing problems of recall. In contrast, LEMMING can analyze any word,

[3] `ftp://ftp.gnu.org/gnu/aspell/dict`

[4] Example: for the Spanish noun *medidas* "measures" with attributes NOUN, COMMON, PLURAL and FEMININE, we conjoin each feature above with NOUN, NOUN+COMMON, NOUN+PLURAL and NOUN+FEMININE.

[5] `https://github.com/gchrupala/morfette/commit/ca886556916b6cc1e808db4d32daf720664d17d6`

				cs		de		en		es		hu		la	
				all	unk	all	unk	all	unk	all	unk	all	unk	all	unk
1	MARMOT		tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
2	JCK		lemma	95.95	81.28	96.63	85.84	99.08	94.28	97.69	87.19	96.69	88.66	90.79	58.23
3			tag+lemma	87.85	67.00	81.60	55.97	96.17	87.32	95.44	80.62	92.15	78.89	79.51	39.07
4	LEMMING-P	+dict	lemma	97.46	89.14	97.70	91.27	**99.21**	**95.59**	98.48	92.98	97.53	92.10	93.07	69.83
5			tag+lemma	88.86	72.51	82.27	59.42	**96.27**	**88.49**	96.12	85.80	92.59	80.77	80.49	44.26
6		+mrph	lemma	97.29	88.98	97.51	90.85	NA	NA	98.68	94.32	97.53	92.15	92.54	67.81
7			tag+lemma	89.23	74.24	82.49	60.42	NA	NA	96.35	87.25	93.11	82.56	80.67	45.21
8	LEMMING-J		tag	**90.34**$^{+}$	78.47	**83.10**$^{+}$	62.36	96.32	89.70	97.11	90.13	93.64	84.78	82.89	54.69
9		+dict	lemma	98.27	92.67	**98.10**$^{+}$	92.79	**99.21**	95.23	98.67	94.07	98.02	94.15	**95.58**$^{+}$	**81.74**$^{+}$
10			tag+lemma	89.69	75.44	82.64	60.49	96.17	87.87	96.23	86.19	92.84	81.89	81.92	49.97
11		+mrph	tag	90.20	**79.72***	**83.10**$^{+}$	**63.10***	NA	NA	**97.16**	**90.66**	**93.67**	**85.12**	**83.49***	**58.76***
12			lemma	**98.42***	**93.46***	**98.10**$^{+}$	**93.02**$^{+}$	NA	NA	**98.78***	**94.86***	**98.08**$^{+}$	**94.26**$^{+}$	95.36	80.94
13			tag+lemma	**89.90***	**78.34***	**82.84***	**62.10***	NA	NA	**96.41**$^{\times}$	**87.47**$^{\times}$	**93.40***	**84.15***	**82.57**$^{+}$	**54.63**$^{+}$

Table 2: *Test* results for LEMMING-J, the joint model, and pipelines (lines 2–7) of MARMOT and (i) JCK and (ii) LEMMING-P. In each cell, overall token accuracy is left (all), accuracy on unknown forms is right (unk). Standalone MARMOT tagging accuracy (line 1) is not repeated for pipelines (lines 2–7). The best numbers are bold. LEMMING-J models significantly better than LEMMING-P ($+$), or LEMMING models not using morphology (+*dict*) ($\times$) or both ($*$) are marked. More baseline numbers in the appendix (Table A2).

including OOVs, and only requires the same training corpus as a generic tagger (containing tags and lemmata), a resource that is available for many languages.

5 Experiments

Datasets. We present experiments on the joint task of lemmatization and tagging in six diverse languages: English, German, Czech, Hungarian, Latin and Spanish. We use the same data sets as in Müller and Schütze (2015), but do not use the out-of-domain test sets. The English data is from the Penn Treebank (Marcus et al., 1993), Latin from PROIEL (Haug and Jøhndal, 2008), German and Hungarian from SPMRL 2013 (Seddah et al., 2013), and Spanish and Czech from CoNLL 2009 (Hajič et al., 2009). For German, Hungarian, Spanish and Czech we use the splits from the shared tasks; for English the split from SANCL (Petrov and McDonald, 2012); and for Latin a 8/1/1 split into *train/dev/test*. For all languages we limit our training data to the first 100,000 tokens. Dataset statistics can be found in Table A4 of the appendix. The lemma of Spanish *se* is set to be consistent.

Baselines. We compare our model to three baselines. (i) MORFETTE (see Section 4). (ii) SIMPLE, a system that for each form-POS pair, returns the most frequent lemma in the training data or the form if the pair is unknown. (iii) JCK, our reimplementation of Jiampojamarn et al. (2008). Recall that JCK is TC's lemmatization model and that the full TC model is a type-based model that cannot be applied to our task.

As JCK struggles to memorize irregulars, we only use it for unknown form-POS pairs and use SIMPLE otherwise. For aligning the training data we use the edit-tree-based alignment described in the feature section. We only use output alphabet symbols that are used for ≥ 5 form-lemma pairs and also add a special output symbol that indicates that the aligned input should simply be copied. We train the model using a structured averaged perceptron and stop after 10 training iterations. In preliminary experiments we found type-based training to outperform token-based training. This is understandable as we only apply our model to unseen form-POS pairs. The feature set is an exact reimplementation of (Jiampojamarn et al., 2008), it consists of input-output pairs and their character context in a window of 6.

Results. Our candidate selection strategy results in an average number of lemma candidates between 7 (Hungarian) and 91 (Czech) and a coverage of the correct lemma on *dev* of >99.4 (except 98.4 for Latin).[6] We first compare the baselines to LEMMING-P, a pipeline based on Section 2, that lemmatizes a word given a predicted tag and is trained using L-BFGS (Liu and Nocedal, 1989). We use the implementation of MALLET (McCallum, 2002). For these experiments we train all models on gold attributes and test on attributes predicted by MORFETTE. MORFETTE's lemmatizer can only be used with its own tags. We thus use MORFETTE tags to have a uniform setup,

[6]Note that our definition of lemmatization accuracy and unknown forms ignores capitalization.

		cs	de	en	es	hu	la
baselines	SIMPLE	87.22	93.27	97.60	92.92	86.09	85.19
	JCK	96.24	97.67	98.71	97.61	97.48	93.26
	MORFETTE	96.25	97.12	98.43	97.97	97.22	91.89
LEMMING-P	edittree	96.29	97.84^+	98.71	97.91	97.31	93.00
	+align,+lemma	96.74^+	98.17^+	98.76^+	98.05	97.70^+	93.76^+
	+dict	$\mathbf{97.50}^+$	$\mathbf{98.36}^+$	$\mathbf{98.84}^+$	98.39^+	$\mathbf{97.98}^+$	$\mathbf{94.64}^+$
	+mrph	96.59^+	97.43^+	NA	$\mathbf{98.46}^+$	97.77^+	93.60

Table 1: Lemma accuracy on *dev* for the baselines and the different versions of LEMMING-P. POS and morphological attributes are predicted using MORFETTE. The best baseline numbers are underlined, the best numbers are bold. Models significantly better than the best baseline are marked (+).

which isolates the effects of the different taggers. Numbers for MARMOT tags are in the appendix (Table A1). For the initial experiments, we only use POS and ignore additional morphological attributes. We use different feature sets to illustrate the utility of our templates.

The first model uses the *edit tree features* (edittree). Table 1 shows that this version of LEMMING outperforms the baselines on half of the languages.[7] In a second experiment we add the *alignment* (+align) and *lemma features* (+lemma) and show that this consistently outperforms all baselines and edittree. We then add the *dictionary feature* (+dict). The resulting model outperforms all previous models and is significantly better than the best baselines for all languages.[8] These experiments show that LEMMING-P yields state-of-the-art results and that all our features are needed to obtain optimal performance. The improvements over the baselines are >1 for Czech and Latin and $\geq.5$ for German and Hungarian.

The last experiment also uses the additional morphological attributes predicted by MORFETTE (+mrph). This leads to a drop in lemmatization performance in all languages except Spanish (English has no additional attributes). However, preliminary experiments showed that correct morphological attributes would substantially improve lemmatization as they help in cases of ambiguity. As an example, number helps to lemmatize the singular German noun *Raps* "canola", which looks like the plural of *Rap* "rap". Numbers can be found in Table A3 of the appendix. This motivates the necessity of *joint tagging and lemmatization*.

For the final experiments, we run pipeline models on tags predicted by MARMOT (Müller et al., 2013) and compare them to LEMMING-J, the joint model described in Section 3. All LEMMING versions use exactly the same features. Table 2 shows that LEMMING-J outperforms LEMMING-P in three measures (see bold tag, lemma & joint (tag+lemma) accuracies) except for English, where we observe a tie in lemma accuracy and a small drop in tag and tag+lemma accuracy. Coupling morphological attributes and lemmatization (lines 8–10 vs 11–13) improves tag+lemma prediction for five languages. Improvements in lemma accuracy of the joint over the best pipeline systems range from .1 (Spanish), over $>.3$ (German, Hungarian) to $\geq.96$ (Czech, Latin).

Lemma accuracy improvements for our best models (lines 4–13) over the best baseline (lines 2–3) are >1 (German, Spanish, Hungarian), >2 (Czech, Latin) and even more pronounced on unknown forms: >1 (English), >5 (German, Spanish, Hungarian) and >12 (Czech, Latin).

6 Conclusion

LEMMING is a modular lemmatization model that supports arbitrary global lemma features and joint modeling of lemmata and morphological tags. It is trainable on corpora annotated with gold standard tags and lemmata, and does not rely on morphological dictionaries or analyzers. We have shown that modeling lemmatization and tagging jointly benefits both tasks, and we set the new state of the art in token-based lemmatization on six languages. LEMMING is available under an open-source licence (`http://cistern.cis.lmu.de/lemming`).

Acknowledgments

We would like to thank the anonymous reviewers for their comments. The first author is a recipient of the Google Europe Fellowship in Natural Language Processing, and this research is supported by this Google fellowship. The second author is supported by a Fulbright fellowship awarded by the German-American Fulbright Commission and the National Science Foundation under Grant No. 1423276. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL) and the DFG grant *Models of Morphosyntax for Statistical Machine Translation*. The fourth author was partially supported by Deutsche Forschungsgemeinschaft (grant SCHU 2246/10-1).

[7] Unknown word accuracies in the appendix (Table A1).

[8] We use the randomization test (Yeh, 2000) and $p = .05$.

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of COLING: Demonstrations*.

Grzegorz Chrupała, Georgiana Dinu, and Josef Van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of LREC*.

Grzegorz Chrupała. 2006. Simple data-driven contextsensitive lemmatization. *Procesamiento del Lenguaje Natural*.

Grzegorz Chrupała. 2008. *Towards a machine-learning architecture for lexical functional grammar parsing*. Ph.D. thesis, Dublin City University.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2014. Stochastic contextual edit distance and probabilistic FSTs. In *Proceedings of ACL*.

Markus Dreyer, Jason R Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of EMNLP*.

Ondřej Dušek and Filip Jurčíček. 2013. Robust multilingual statistical morphological generation models. In *Proceedings of ACL: Student Research Workshop*.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of EACL*.

Andrea Gesmundo and Tanja Samardzic. 2012. Lemmatisation as a tagging task. In *Proceedings of ACL*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European bible translations. In *Proceedings of LaTeCH*.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL*.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL: Demonstrations*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational linguistics*.

Andrew K McCallum. 2002. MALLET: A machine learning for language toolkit.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of NAACL*.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of EMNLP*.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of LREC*.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Proceedings of SANCL*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP*.

Djamé Seddah, Grzegorz Chrupała, Özlem Çetinoğlu, Josef Van Genabith, and Marie Candito. 2010. Lemmatization and lexicalized statistical parsing of morphologically rich languages: the case of French. In *Proceedings of SPMRL*.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: Cross-Framework evaluation of parsing morphologically rich languages. In *Proceedings of SPMRL*.

Noah A. Smith, David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of HLT-EMNLP*.

Kristina Toutanova and Colin Cherry. 2009. A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of ACL-IJCNLP*.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of ACL-IJCNLP*.

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

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING*.