HAL open science

Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, François Yvon

To cite this version:

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, François Yvon. Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning. Conference on Empirical Methods in Natural Language Processing, Oct 2014, Doha, Qatar. pp.1779 - 1785. hal-01908356

HAL Id: hal-01908356

https://hal.science/hal-01908356v1

Submitted on 19 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning

Guillaume Wisniewski **Nicolas Pécheux** **Souhir Gahbiche-Braham** **François Yvon**

Université Paris Sud
LIMSI-CNRS
91 403 ORSAY CEDEX, France
{wisniews,pecheux,souhir,yvon}@limsi.fr

Abstract

When Part-of-Speech annotated data is scarce, e.g. for under-resourced languages, one can turn to cross-lingual transfer and crawled dictionaries to collect partially supervised data. We cast this problem in the framework of *ambiguous learning* and show how to learn an accurate history-based model. Experiments on ten languages show significant improvements over prior state of the art performance.

1 Introduction

In the past two decades, supervised Machine Learning techniques have established new performance standards for many NLP tasks. Their success however crucially depends on the availability of annotated in-domain data, a not so common situation. This means that for many application domains and/or less-resourced languages, alternative ML techniques need to be designed to accommodate unannotated or partially annotated data.

Several attempts have recently been made to mitigate the lack of annotated corpora using parallel data pairing a (source) text in a resource-rich language with its counterpart in a less-resourced language. By transferring labels from the source to the target, it becomes possible to obtain noisy, yet useful, annotations that can be used to train a model for the target language in a *weakly supervised* manner. This research trend was initiated by Yarowsky et al. (2001), who consider the transfer of POS and other syntactic information, and further developed in (Hwa et al., 2005; Ganchev et al., 2009) for syntactic dependencies, in (Padó and Lapata, 2009; Kozhevnikov and Titov, 2013; van der Plas et al., 2014) for semantic role labeling and in (Kim et al., 2012) for named-entity recognition, to name a few.

Assuming that labels can actually be projected across languages, these techniques face the issue of extending standard supervised techniques with partial and/or uncertain labels in the presence of alignment noise. In comparison to the early approach of Yarowsky et al. (2001) in which POS are directly transferred, subject to heuristic filtering rules, recent works consider the integration of softer constraints using expectation regularization techniques (Wang and Manning, 2014), the combination of alignment-based POS transfer with additional information sources such as dictionaries (Li et al., 2012; Täckström et al., 2013) (Section 2), or even the simultaneous use of both techniques (Ganchev and Das, 2013).

In this paper, we reproduce the weakly supervised setting of Täckström et al. (2013). By recasting this setting in the framework of ambiguous learning (Bordes et al., 2010; Cour et al., 2011) (Section 3), we propose an alternative learning methodology and show that it improves the state of the art performance on a large array of languages (Section 4). Our analysis of the remaining errors suggests that in cross-lingual settings, improvements of error rates can have multiple causes and should be looked at with great care (Section 4.2).

All tools and resources used in this study are available at `http://perso.limsi.fr/wisniews/ambiguous`.

2 Projecting Labels across Aligned Corpora

Projecting POS information across languages relies on a rather strong assumption that morpho-syntactic categories in the source language can be directly related to the categories in the target language, which might not always be warranted (Evans and Levinson, 2009; Broschart, 2009). The universal reduced POS tagset proposed by Petrov et al. (2012) defines an operational, albeit rather empirical, ground to perform this mapping. It is made of the following 12 categories: NOUN (nouns), VERB (verbs), ADJ (ad-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1779–1785, October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

	ar	cs	de	el	es	fi	fr	id	it	sv
% of test covered tokens (type)	83.2	93.2	95.6	97.4	96.7	83.0	98.3	90.5	95.8	95.3
% of test correctly covered token (type)	72.9	94.2	93.7	92.9	93.8	93.6	92.1	89.6	93.6	94.1
avg. number of labels per token (type)	2.1	1.3	1.3	1.3	1.3	1.4	1.3	1.2	1.3	1.3
avg. number of labels per token (type+token)	1.7	1.1	1.1	1.1	1.1	1.2	1.2	1.1	1.1	1.1
% of aligned tokens	53.0	77.8	66.7	69.3	74.0	73.1	64.7	81.6	72.2	79.9
% of token const. violating type const.	2.5	16.0	15.8	21.4	16.9	14.3	16.1	19.3	17.5	13.6
% informative token const.	79.7	27.5	15.7	29.8	21.3	36.0	25.5	16.2	28.2	26.4

Table 1: Interplay between token and type constraints on our training parallel corpora. 'Informative' token constraints correspond to tokens for which (a) a POS is actually transfered and (b) type constraints do not disambiguate the label, but type+token constraints do.

jectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories). These labels have been chosen for their stability across languages and for their usefulness in various multilingual applications. In the rest of this work, all annotations are mapped to this universal tagset.

Transfer-based methods have shown to be very effective, even if projected labels only deliver a noisy supervision, due to tagging (of the source language) and other alignment errors (Yarowsky et al., 2001). While this uncertainty can be addressed in several ways, recent works have proposed to combine projected labels with monolingual information in order to filter out invalid label sequences (Das and Petrov, 2011; Täckström et al., 2013). In this work we follow Täckström et al. (2013) and use two families of constraints:

Token constraints rely on word alignments to project labels of source words to target words through alignment links. Table 1 shows that, dependening on the language, only $50-80\%$ of the target tokens would benefit from label transfer.

Type constraints rely on a tag dictionary to define the set of possible tags for each word type. Type constraints reduce the possible labels for a given word and help filtering out cross-lingual transfer errors (up to 20%, as shown in Table 1). As in (Täckström et al., 2013), we consider two different dictionaries. The first one is extracted automatically from `Wiktionary`,[1] using the method of (Li et al., 2012). The second tag dictionary is built by using for each word the two most frequently projected POS labels from the training data.[2] In contrast to Täckström et al. (2013) we use the intersection[3] of the two type constraints instead of their union. Table 1 shows the precision and recall of the resulting constraints on the test data.

These two information sources are merged according to the rules of Täckström et al. (2013). These rules assume that type constraints are more reliable than token constraints and should take precedence: by default, a given word is associated to the set of possible tags licensed type constraints; additionally, when a POS tag can be projected through alignment *and* also satisfies the type constraints, then it is actually projected, thereby providing a full (yet noisy) supervision.

As shown in Table 1, token and type constraints complement each other effectively and greatly reduce label ambiguity. However, the transfer method sketched above associates each target word with a set of possible labels, of which only one is true. This situation is less favorable than standard supervised learning in which one unique gold label is available for each occurrence. We describe in the following section how to learn from this *ambiguous supervision* information.

3 Modeling Sequences under Ambiguous Supervision

We use a history-based model (Black et al., 1992) with a LaSO-like training method (Daumé and Marcu, 2005). History-based models reduce structured prediction to a sequence of multi-class classification problems. The prediction of a complex structure (here, a sequence of POS tags) is thus modeled as a sequential decision problem: at each

[1] `http://www.wiktionary.org/`

[2] This heuristic is similar to the way Täckström et al. (2013) filter the tag distribution with a threshold to build the projected type constraints.

[3] If the intersection is empty we use the constraints from `Wiktionary` first, if also empty, the projected constraints then, and by default the whole tag set.

position in the sequence, a multiclass classifier is used to make a decision, using features that describe both the input structure and the history of past decisions (i.e. the partially annotated sequence).

Let $\mathbf{x} = (\mathbf{x}_i)_{i=1}^n$ denote the observed sequence and $\mathcal{Y}$ be the set of possible labels (in our case the 12 universal POS tags). Inference consists in predicting labels one after the other using, for instance, a linear model:

$$y_i^* = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w} | \phi(\mathbf{x}, i, y, h_i) \rangle \quad (1)$$

where $\langle \cdot | \cdot \rangle$ is the standard dot product operation, y_i^* the predicted label for position i, $\mathbf{w}$ the weight vector, $h_i = y_1^*, ..., y_{i-1}^*$ the history of past decisions and ϕ a joint feature map. Inference can therefore be seen as a greedy search in the space of the $\#\{\mathcal{Y}\}^n$ possible labelings of the input sequence. Trading off the global optimality of inference for additional flexibility in the design of features and long range dependencies between labels has proved useful for many sequence labeling tasks in NLP (Tsuruoka et al., 2011).

The training procedure, sketched in Algorithm 1, consists in performing inference on each input sentence and correcting the weight vector each time a wrong decision is made. Importantly (Ross and Bagnell, 2010), the history used during training has to be made of the previous predicted labels so that the training samples reflect the fact that the history will be imperfectly known at test time.

This *reduction* of sequence labeling to multiclass classification allows us to learn a sequence model in an ambiguous setting by building on the theoretical results of Bordes et al. (2010) and Cour et al. (2011). The decision about the correctness of a prediction and the weight updates can be adapted to the amount of supervision information that is available.

Full Supervision In a fully supervised setting, the correct label is known for each word token: a decision is thus considered wrong when this gold label is not predicted. In this case, a standard perceptron update is performed:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \phi(\mathbf{x}, i, y_i^*, h_i) + \phi(\mathbf{x}, i, \hat{y}_i, h_i) \quad (2)$$

where y_i^* and $\hat{y}_i$ are the predicted and the gold label, respectively. This update is a stochastic gradient step that increases the score of the gold label while decreasing the score of the predicted label.

Ambiguous Supervision During training, each observation i is now associated with a set of possible labels, denoted by $\hat{\mathcal{Y}}_i$. In this case, a decision is considered wrong when the predicted label is not in $\hat{\mathcal{Y}}_i$ and the weight vector is updated as follows:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \phi(\mathbf{x}, i, y_i^*, h_i) + \sum_{\hat{y}_i \in \hat{\mathcal{Y}}_i} \phi(\mathbf{x}, i, \hat{y}_i, h_i) \quad (3)$$

Compared to (2), this rule uniformly increases the scores of all the labels in $\hat{\mathcal{Y}}_i$.

It can be shown (Bordes et al., 2010; Cour et al., 2011), under mild assumptions (namely that two labels never systematically co-occur in the supervision information), that the update rule (3) enables to learn a classifier in an ambiguous setting, as if the gold labels were known. Intuitively, as long as two labels are not systematically co-occurring in $\hat{\mathcal{Y}}$, updates will reinforce the correct labels more often than the spurious ones; at the end of training, the highest scoring label should therefore be the correct one.

Algorithm 1 Training algorithm. In the ambiguous setting, $\hat{\mathcal{Y}}_i$ contains all possible labels; in the supervised setting, it only contains the gold label.

```
w_0 ← 0
for t ∈ [[1, T]] do
  Randomly pick example x, ŷ
  h ← empty list
  for i ∈ [[1, n]] do
    y*_i = arg max_{y∈Y} ⟨w_t | φ(x, i, y, h_i)⟩
    if y*_i ∉ Ŷ_i then
      w_{t+1} ← update(w_t, x, i, Ŷ_i, y*_i, h_i)
    end if
    push(y*_i, h)
  end for
end for
return (1/T) Σ_{t=1}^{T} w_t
```

4 Empirical Study

Datasets Our approach is evaluated on 10 languages that present very different characteristics and cover several language families.[4] In all our experiments we use English as the source language. Parallel sentences[5] are aligned with the standard

[4] Resources considered in the related works are not freely available, which prevents us from presenting a more complete comparison.

[5] All resources and features used in our experiments are thoroughly documented in the supplementary material.

	ar	cs	de	el	es	fi	fr	id	it	sv
HBAL	**27.9**	**10.4**	**8.8**	**8.1**	**8.2**	13.3	**10.2**	**11.3**	**9.1**	**10.1**
Partially observed CRF	33.9	11.6	12.2	10.9	10.7	**12.9**	11.6	16.3	10.4	11.6
HBSL	—	1.5	5.0	—	2.4	5.9	3.5	4.8	2.8	3.8
HBAL + matched POS	24.1	7.6	8.0	7.3	7.4	12.2	7.4	9.8	8.3	8.8
(Ganchev and Das, 2013)	49.9	19.3	9.6	9.4	12.8	—	12.5	—	10.1	10.8
(Täckström et al., 2013)	—	18.9	9.5	10.5	10.9	—	11.6	—	10.2	11.1
(Li et al., 2012)	—	—	14.2	20.8	13.6	—	—	—	13.5	13.9

Table 2: Error rate (in %) achieved by the method described in Sec. 3 trained in an ambiguous (HBAL) or in a supervised setting (HBSL), a partially observed CRF and different state-of-the-art results.

MOSES pipeline, using the intersection heuristic that only retains the most reliable alignment links.

The English side of the bitext is tagged using a standard linear CRF trained on the Penn Treebank. Tags are then transferred to the target language using the procedure described in Section 2. For each language, we train a tagger using the method described in Section 3 with $T = 100\,000$ iterations[6] using a feature set similar to the one of Li et al. (2012) and Täckström et al. (2013). The baseline system is our reimplementation of the partially observed CRF model of Täckström et al. (2013). Evaluation is carried out on the test sets of treebanks for which manual gold tags are known. For Czech and Greek, we use the CoNLL'07 Shared Task on Dependency Parsing; for Arabic, the Arabic Treebank; and otherwise the data of the Universal Dependency Treebank Project (McDonald et al., 2013). Tagging performance is evaluated with the standard error rate.

4.1 Results

Table 2 summarizes the performance achieved by our method trained in the ambiguous setting (HBAL) and by our re-implementation of the partially supervised CRF baseline. As an upper bound, we also report the score of our method when trained in a supervised (HBSL) settings considering the training part of the various treebanks, when it is available.[7] For the sake of comparison, we also list the *best scores* of previous studies. Note, however, that a direct comparison with these results is not completely fair as these systems were not trained and evaluated with the same exact resources (corpora,[8] type constraints, alignments, etc). Also note that the state-of-the-art scores have been achieved by different models, which have been selected based on their scores on the test set and not on a validation set.[9]

Experimental results show that HBAL significantly outperforms, on all considered languages but one, the partially observed CRF that was trained and tested in the same setting.

4.2 Discussion

The performance of our new method still falls short of the performance of a fully supervised POS tagger: for instance, in Spanish, full supervision reduces the error rate by a factor of 4. A fine-grained error analysis shows that many errors of HBAL directly result from the fact that, contrary to the fully supervised learner HBSL, our ambiguous setting suffers from a train/test mismatch, which has two main consequences. First, the train and test sets do not follow exactly the same normalization and tokenization conventions, which is an obvious source of mistakes. Second, and more importantly, many errors are caused by systematic differences between the test tags and the supervised tags (i.e. the English side of the bitext and `Wiktionary`). While some of these differences are linguistically well-justified and reflect fundamental differences in the language structure and usage, others seem to be merely due to arbitrary annotation conventions.

For instance, in Greek, proper names are labeled

[6] Preliminary experiments showed that increasing the number of iterations T in Algorithm 1 has no significant impact.

[7] In this setting, HBSL implements an averaged perceptron, and achieves results that are similar to those obtained with standard linear CRF.

[8] The test sets are only the same for Czech, Greek and Swedish.

[9] The partially observed CRF is the best model in (Täckström et al., 2013) only for German (de), Greek (el) and Swedish (sv), and uses only type constraints extracted from `Wiktionary`.

either as X (when they refer to a foreigner *and* are not transliterated) or as NOUN (in all other cases), while they are always labeled as NOUN in English. In French and in Greek, contractions of a preposition and a determiner such as '*στο*' ('σε το', meaning 'to the') or '*aux*' ('*à les*' also meaning 'to the') are labeled as ADP in the Universal Dependency Treebank but as DET in `Wiktionary` and are usually aligned with a determiner in the parallel corpora. In the Penn Treebank, quantifiers like 'few' or 'little' are generally used in conjunction with a determiner ('a few years', 'a little parable', ...) and labeled as ADJ; the corresponding Spanish constructions lack an article ('*mucho tempio*', '*pocos años*', ...) and the quantifiers are therefore labeled as DET. Capturing such subtle differences is hardly possible without prior knowledge and specifically tailored features.

This annotation mismatch problem is all the more important in settings like ours, that rely on several, independently designed, information sources, which follow contradictory annotation conventions and for which the mapping to the universal tagset is actually error-prone (Zhang et al., 2012). To illustrate this point, we ran three additional experiments to assess the impact of the train/test mismatch.

We first designed a control experiment in which the type constraints were manually completed with the gold labels of the most frequent errors of HBAL. These errors generally concern function words and can be assumed to result from systematic differences in the annotations rather than prediction errors. For instance, for French the type constraints for '*du*', '*des*', '*au*' and '*aux*' were corrected from DET to ADP. The resulting model, denoted 'HBAL + matched POS' in Table 2, significantly outperforms HBAL, stressing the divergence in the different annotation conventions.

Additionally, in order to approximate the ambiguous setting train/test mismatch, we learn two fully supervised Spanish taggers on the same training data as HBAL, using two different strategies to obtain labeled data. We first use HBSL (which was trained on the treebank) to automatically label the target side of the parallel corpus. In this setting, the POS tagger is trained with data from a different domain, but labeled with the same annotation scheme as a the test set. Learning with this fully supervised data yields an error rate of 4.2% for Spanish, almost twice as much as HBSL, bringing into light the impact of domain shift. We then use a generic tagger, FREELING,[10] to label the training data, this time with possible additional inconsistent annotations. The corresponding error rate for Spanish was 6.1%, to be compared with the 8.2% achieved by HBAL. The last two control experiments show that many of the remaining labeling errors seem to be due to domain and convention mismatches rather to the transfer/ambiguous setting, as supervised models also suffer from very similar conditions.

These observations show that the evaluation of transfer-based methods suffer from several biases. Their results must therefore be interpreted with great care.

5 Conclusion

In this paper, we have presented a novel learning methodology to learn from ambiguous supervision information, and used it to train several POS taggers. Using this method, we have been able to achieve performance that surpasses the best reported results, sometimes by a wide margin. Further work will attempt to better analyse these results, which could be caused by several subtle differences between HBAL and the baseline system. Nonetheless, these experiments confirm that cross-lingual projection of annotations have the potential to help in building very efficient POS taggers with very little monolingual supervision data. Our analysis of these results also suggests that, for this task, additional gains might be more easily obtained by fixing systematic biases introduced by conflicting mappings between tags or by train/test domain mismatch than by designing more sophisticated weakly supervised learners.

Acknowledgments

We wish to thank Thomas Lavergne and Alexandre Allauzen for early feedback and for providing us with the partially observed CRF implementation. We also thank the anonymous reviewers for their helpful comments.

References

Ezra Black, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer, and Salim Roukos. 1992. Towards history-based grammars: Using

[10] `http://nlp.lsi.upc.edu/freeling/`

richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 134–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, and Jason Weston. 2010. Label ranking under ambiguous supervision for learning semantic correspondences. In *ICML*, pages 103–110.

Jürgen Broschart. 2009. Why Tongan does it differently: Categorial distinctions in a language without nouns and verbs. *Linguistic Typology*, 1:123–166, 10.

Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, July.

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.

Hal Daumé, III and Daniel Marcu. 2005. Learning as search optimization: Approximate large margin methods for structured prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 169–176, New York, NY, USA. ACM.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–448, 10.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1996–2006, Seattle, Washington, USA, October. Association for Computational Linguistics.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. 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 369–377, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325, September.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 694–702, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shen Li, João V. Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1389–1398, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340, September.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Stéphane Ross and Drew Bagnell. 2010. Efficient reductions for imitation learning. In *AISTATS*, pages 661–668.

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Kazama. 2011. Learning with lookahead: Can history-based models rival globally optimized models? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 238–246, Portland, Oregon, USA, June. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Lonneke van der Plas, Marianna Apidianaki, and Chenhua Chen. 2014. Global methods for cross-lingual semantic role and predicate labelling. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1279–1290, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the ACL*, 2:55–66, February.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal pos tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1368–1378, Stroudsburg, PA, USA. Association for Computational Linguistics.