Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data

Maria Barrett[†] **Joachim Bingel**[†] **Frank Keller**[*] **Anders Søgaard**[†]
[†]Centre for Language Technology, University of Copenhagen
Njalsgade 140, 2300 Copenhagen S, Denmark
{barrett, bingel, soegaard}@hum.ku.dk
[*]School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

Abstract

For many of the world's languages, there are no or very few linguistically annotated resources. On the other hand, raw text, and often also dictionaries, can be harvested from the web for many of these languages, and part-of-speech taggers can be trained with these resources. At the same time, previous research shows that eye-tracking data, which can be obtained without explicit annotation, contains clues to part-of-speech information. In this work, we bring these two ideas together and show that given raw text, a dictionary, and eye-tracking data obtained from naive participants reading text, we can train a weakly supervised PoS tagger using a second-order HMM with maximum entropy emissions. The best model use type-level aggregates of eye-tracking data and significantly outperforms a baseline that does not have access to eye-tracking data.

1 Introduction

According to Ethnologue, there are around 7,000 languages in the world.[1] For most of these languages, no or very little linguistically annotated resources are available. This is why over the past decade or so, NLP researchers have focused on developing unsupervised algorithms that learn from raw text, which for many languages is widely available on the web. An example is part-of-speech (PoS) tagging, in which unsupervised approaches have been increasingly successful (see Christodoulopoulos et al. (2010) for an overview). The performance of unsupervised PoS taggers can be improved further if dictionary information is available, making it possible to constrain the PoS tagging process. Again, dictionary information can be harvested readily from the web for many languages (Li et al., 2012).

In this paper, we show that PoS tagging performance can be improved further by using a weakly supervised model which exploits eye-tracking data in addition to raw text and dictionary information. Eye-tracking data can be obtained by getting native speakers of the target language to read text while their gaze behavior is recorded. Reading is substantially faster than manual annotation, and competent readers are available for languages where trained annotators are hard to find or non-existent. While high quality eye-tracking equipment is still expensive, $100 eye-trackers such as the EyeTribe are already on the market, and cheap eye-tracking equipment is likely to be widely available in the near future, including eye-tracking by smartphone or webcam (Skovsgaard et al., 2013; Xu et al., 2015).

Gaze patterns during reading are strongly influenced by the parts of speech of the words being read. Psycholinguistic experiments show that readers are less likely to fixate on closed-class words that are predictable from context. Readers also fixate longer on rare words, on words that are semantically ambiguous, and on words that are morphologically complex (Rayner, 1998). These findings indicate that eye-tracking data should be useful for classifying words by part of speech, and indeed Barrett and Søgaard (2015) show that word-type-level aggregate statistics collected from eye-tracking corpora can be used as features for supervised PoS tagging, leading to substantial gains in accuracy across domains. This leads us to hypothesize that gaze data should also improve weakly supervised PoS tagging.

In this paper, we test this hypothesis by experimenting with a PoS tagging model that uses raw text, dictionary information, and eye-tracking

[1] http://www.ethnologue.com/world

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Short Papers, 579–584.
Berlin, 2016.

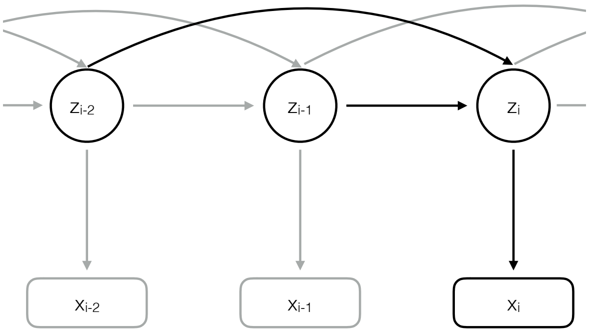

Figure 1: Second-order HMM. In addition to the transitional probabilities of the antecedent state z_{i-1} in first-order HMMs, second-order models incorporate transitional probabilities from the second-order antecedent state z_{i-2}.

data, but requires no explicit annotation. We start with a state-of-the-art unsupervised PoS tagging model, the second-order hidden Markov model with maximum entropy emissions of Li et al. (2012), which uses only textual features. We augment this model with a wide range of features derived from an eye-tracking corpus at training time (type-level gaze features). We also experiment with token-level gaze features; the use of these features implies that eye-tracking is available both at training time and at test time. We find that eye-tracking features lead to a significant increase in PoS tagging accuracy, and that type-level aggregates work better than token-level features.

2 The Dundee Treebank

The Dundee Treebank (Barrett et al., 2015) is a Universal Dependency annotation layer that has recently been added to the world's largest eye-tracking corpus, the Dundee Corpus (Kennedy et al., 2003). The English portion of the corpus contains 51,502 tokens and 9,776 types in 2,368 sentences. The Dundee Corpus is a well-known and widely used resource in psycholinguistic research. The corpus enables researchers to study the reading of contextualized, running text obtained under relatively naturalistic conditions. The eye-movements in the Dundee Corpus were recorded with a high-end eye-tracker, sampling at 1000 Hz. The corpus contains the eye-movements of ten native English speakers as they read the same twenty newspaper articles from *The Independent*. The corpus was augmented with Penn Treebank PoS annotation by Frank (2009). When constructing the Dundee Treebank, this PoS annotation was checked and corrected if necessary. In the present paper, we use Universal PoS tags (Petrov et al., 2011), which were obtained by automatically mapping the original Penn Treebank annotation of the Dundee Treebank to Universal tags.

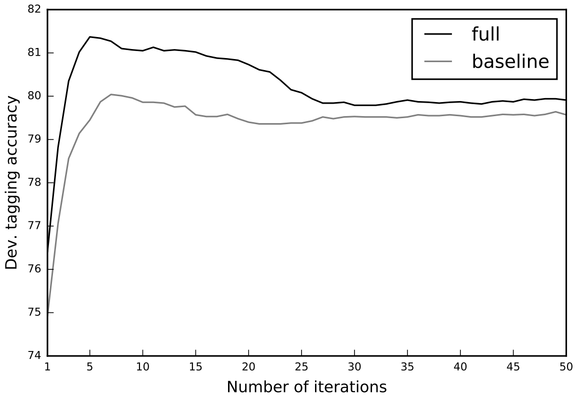

Figure 2: Tagging accuracy on development data (token-level) as a function of number of iterations on baseline and full model.

3 Type-constrained second-order HMM PoS tagging

We build on the type-constrained second-order hidden Markov model with maximum entropy emissions (SHMM-ME) proposed by Li et al. (2012). This model is an extension of the first-order max-ent HMM introduced by Berg-Kirkpatrick et al. (2010). Li et al. (2012) derive type constraints from crowd-sourced tag dictionaries obtained from Wiktionary. Using type constraints means confining the emissions for a given word to the tags specified by the Wiktionary for that word. Li et al. (2012) report a considerable improvement over state-of-the-art unsupervised PoS tagging models by using type constraints. In our experiments, we use the tag dictionaries they made available[2] to facilitate comparison. Li et al.'s model was evaluated across nine languages and outperformed a model trained on the Penn Treebank tagset, as well as a models that use parallel text. We follow Li et al.'s approach, including the mapping of the Penn Treebank tags to

[2] https://code.google.com/archive/p/wikily-supervised-pos-tagger/

Early	First fixation duration
	w-1 fixation probability
	w-1 fixation duration
	First pass duration
Late	Total regression-to duration
	n long regressions to *w*
	n refixations
	Re-read probability
	n regressions to *w*
Basic	Total fixation duration
	Mean fixation duration
	n fixations
	Fixation probability
RegFr.	*n* regressions from *w*
	n long regressions from *w*
	Total regression-from duration
Context	*w*+1 fixation probability
	w+1 fixation duration
	w+2 fixation probability
	w+2 fixation duration
	w-2 fixation probability
	w-2 fixation probability
NoGazeB.	Word length
	BNC log frequency
	w-1 BNC log frequency
	BNC forward transitional log probability
	BNC backward transitional log probability
NoGazeD.	Word length
	Dundee log frequency
	w-1 Dundee log frequency
	Dundee forward transitional log probability
	Dundee backward transitional log probability

Table 1: Features in feature selection groups.

the Universal PoS tags (Petrov et al., 2011). Figure 1 shows a graphical representation of a second-order hidden Markov model.

Li et al. explore two aspects of type-constrained HMMs for unsupervised PoS tagging: the use of a second-order Markov model, and the use of textual features modeled by maximum entropy emissions. They find that both aspects improve tagging accuracy and report the following results for English using Universal PoS tags on the Penn Treebank: first-order HMM 85.4, first-order HMM with max-ent emissions 86.1, second-order HMM 85.0, and second-order HMM with max-ent emissions 87.1. Li et al. employ a set of basic textual features for the max-ent versions, which encode word identity, presence of a hyphen, a capital letter, or a digit, and word suffixes of two to three letters.

4 Experiments

Features Based on the eye-movement data in the Dundee Corpus, we compute token-level values for 22 features pertaining to gaze and complement them with another nine non-gaze features. Word length and word frequency are known to correlate and interact with gaze features. We use frequency counts from both a large corpus (the British National Corpus, BNC) and the Dundee Corpus itself. From these corpora, we also obtain forward and backward transitional probabilities, i.e., the conditional probabilities of a word given the previous or next word.

Features	**TA**
NoGazeDun	81.03
NoGazeBNC	80.69
Basic	80.30
Early	79.96
Late	79.87
RegFrom	79.62
Context	79.53
Best Group Comb (All)	81.37
Best Gaze-Only Comb (Basic-Late)	80.45

Table 2: Tagging accuracy on the development set (token-level) for all individual feature groups, for the best combination of groups and for the best gaze-only combination of groups.

All gaze features are averaged over the ten readers and normalized linearly to a scale between 0 and 1. We divide the set of 31 features, which we list in Table 1, into the following seven groups in order to examine for their individual contribution:

1. Early measures of processing such as first-pass fixation duration. Fixations on previous words are included in this group due to preview benefits. Early measures capture lexical access and early syntactic processing.
2. Late measures of processing such as number of regressions to a word and re-fixation probability. These measures reflect late syntactic processing and disambiguation in general.
3. Basic word-level features, e.g., mean fixation duration and fixation probability. These metrics do not belong explicitly to early or late processing measures.
4. RegFrom includes a small selection of measures based on regressions departing from a token. It also includes counts of long regressions[3]. The token of departure of a regression

[3] defined as saccades going further back than w_{i-2}

System	TA
Baseline (Li et al., 2012)	79.77
NoTextFeats	74.61
NoTextFeats + Best Group Comb (token)	79.56
NoTextFeats + Best Group Comb (type)	81.94*
Token-level features	
Best Gaze Group (Basic)	80.42*
Best Gaze-Only Comb (Basic+Late)	80.45*
Best Single Group (NoGazeDun)	80.61*
Best Group Comb (All)	81.00*
Type-averaged features	
Best Gaze Group (Basic)	81.28*
Best Gaze-Only Comb (Basic+Late)	81.38*
Best Group (NoGazeDun)	81.52*
Best Group Comb (All)	82.44*

Table 3: Tagging accuracy for the baseline, for models with no text features and for our gaze-enriched models using type and token gaze features. Significant improvements over the baseline marked by * ($p < 10^{-3}$, McNemar's test).

Feature groups	Accuracy	Δ
All groups	81.00	
−NoGazeBNC	80.80	−0.20
−NoGazeDun	80.28	−0.52*
−Basic	80.20	−0.08
−Early	79.78	−0.42*
−Late	79.53	−0.25
−RegFrom	79.24	−0.29*
−Context (Baseline)	79.77	+0.53*

Table 4: Results of an ablation study over feature groups on the test set on token-level features. Significant differences with previous model are marked by * ($p < 0.05$, McNemar's test).

can have syntactic relevance, e.g., in garden path sentences.

5. Context features of the surrounding tokens. This group contains features relating to the fixations of the words in near proximity of the token. The eye can only recognize words a few characters to the left, and seven to eight characters to the right of the fixation (Rayner, 1998). Therefore it is useful to know the fixation pattern around the token.

6. NoGazeBNC includes word length and word frequency obtained from the British National Corpus, as well as forward and backward transitional probabilities. These were computed using the KenLM language modeling toolkit (Heafield, 2011) with Kneser-Ney smoothing for unseen bigrams.

7. NoGazeDun includes the same features as NogazeBNC, but computed on the Dundee Corpus. They were extracted using CMU-Cambridge language modeling toolkit.[4]

Setup The Dundee Corpus does not include a standard train-development-test split, so we divided it into a training set containing 46,879 tokens/1,896 sentences, a development set containing 5,868 tokens/230 sentences, and a test set of 5,832 tokens/241 sentences.

To tune the number of EM iterations required for the SHMM-ME model, we ran several experiments on the development set using 1 through 50 iterations. The result is fairly consistent for both the baseline (the original model of Li et al. (2012)) and the full model (which includes all feature groups in Table 1). Tagging accuracy as a function of number of iterations is graphed in Figure 2. The best number of iterations on the full model is five, which we will use for the remaining experiments.

We perform a grid search over all combinations of the seven feature groups, using five EM iterations for training, evaluating the resulting models on token-level features of the development set. We observe that the best single feature group is NoGazeDUN, the best single group of gaze features is Basic, the best gaze-only group combination is basic-Late and the best group combination is obtained by including all seven feature groups. Using all feature groups outperforms any individual feature group on development data. The performance of all the individual groups and of the best group combinations can be seen in Table 2. We run experiments on the test set and report results using the best single group (NoGazeDun), the best single gaze group (Basic), the best gaze-only group combination (Basic-Late) and the best group combination (all features).

Following Barrett and Søgaard (2015), we contrast the token-level gaze features with features ag-

[4] http://www.speech.cs.cmu.edu/SLM/toolkit.html

gregated at the type level. Type-level aggregation was used by Barrett and Søgaard (2015) for supervised PoS tagging: A lexicon of word types was created and the features values were averaged over all occurrences of each type in the training data.

As our baseline, we train and evaluate the original model proposed by Li et al. (2012) on the train-test split described above, and compare it to the models that make use of eye-tracking measures.

To get an estimate of the effect of the textual features of Li et al., we train a model without these features, labeled NOTEXTFEATS. We also augment this model with the best combination of feature groups.

Results The main results are presented in Table 3. We first of all observe that both type- and token-level gaze features lead to significant improvements over Li et al. (2012), but type-level features perform better than token-level. We observe that the best individual feature group, NOGAZEDUN, performs better than the best individual gaze feature group, BASIC and the best gaze-only feature group, BASIC+LATE. This is true on both type and token-level. Using the best combination of feature groups (All features) works best for both type- and token-level features. Also when excluding the textual feature model gaze helps and type-level features also work better than token-level here.

A feature ablation study (see Table 4) supports the hierarchical ordering of the features based on the development set results (see Table 1).

5 Related Work

The proposed approach continues the work of Barrett and Søgaard (2015) by augmenting an unsupervised baseline PoS tagging model instead of a supervised model. Our work also explores the potentials of token-level features. Zelenina (2014) is the only work we are aware of that uses gaze features for unsupervised PoS tagging. Zelenina (2014) employs gaze features to re-rank the output of a standard unsupervised tagger. She reports a small improvement with gaze features when evaluating on the Universal PoS tagset, but finds no improvement when using the Penn Treebank tagset.

6 Discussion

The best individual feature group is NOGAZEDUN, indicating that just using word length and word frequency, as well as transitional probabilities, leads to a significant improvement in tagging accuracy. However, performance increases further when we add gaze features, which supports our claim that gaze data is useful for weakly supervising PoS induction.

Type-level features work noticeably better than token-level features, suggesting that access to eye-tracking data at test time is not necessary. On the contrary, our results support the more resource-efficient set-up of just having eye-tracking data available at training time. We assume that this finding is due to the fact that eye-movement data is typically quite noisy; averaging over all tokens of a type reduces the noise more than just averaging over the ten participants that read each token. Thus token-level aggregation leads to more reliable feature values.

Our finding that the best model includes all groups of gaze features, and that the best gaze-only group combination works better than the best individual gaze group suggest that different eye-tracking features contain complementary information. A broad selection of eye-movement features is necessary for reliably identifying PoS classes.

7 Conclusions

We presented the first study of weakly supervised part-of-speech tagging with eye-tracking data, using a type-constrained second-order hidden Markov model with max-ent emissions. We performed experiments adding a broad selection of eye-tracking features at training time (type-level features) and at test time (token-level features). We found significant improvements over the baseline in both cases, but type averaging worked better than token-level features. Our results indicate that using traces of human cognitive processing, such as the eye-movements made during reading, can be used to augment NLP models. This could enable us to bootstrap better PoS taggers for domains and languages for which manually annotated corpora are not available, in particular once eye-trackers become widely available through smartphones or webcams (Skovsgaard et al., 2013; Xu et al., 2015).

Acknowledgments

This research was partially funded by the ERC Starting Grant LOWLANDS No. 313695, as well as by Trygfonden.

References

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. *CoNLL 2015*, pages 345–349.

Maria Barrett, Željko Agić, and Anders Søgaard. 2015. The dundee treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 242–248.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, , and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of NAACL*, pages 582–590.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of EMNLP*, pages 575–584.

Stefan L. Frank. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual Conference of the Cognitive Science Society*, pages 1139–1144.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *EMNLP*, pages 1389–1398.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

Henrik Skovsgaard, John Paulin Hansen, and Emilie Møllenbach. 2013. Gaze tracking through smartphones. In *Gaze Interaction in the Post-WIMP World CHI 2013 One-day Workshop*.

P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, , and J. Xiao. 2015. TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. arXiv:1504.06755.

Maria Zelenina. 2014. Part of speech induction with gaze features. Master's thesis, University of Edinburgh, United Kingdom.