Semi-supervised part-of-speech tagging with eye gaze features

Melanie Platt
Northeastern University
CS 6120 Natural Language Processing
April 29, 2021

Abstract

This study explores the integration of eye gaze data as features in semi-supervised part-of-speech (PoS) tagging. Eye gaze has proven effective in several NLP tasks, given its informative value for human processing load. Driven by the need for better PoS taggers that do not require ample annotated data, and the comparative ease of collecting eye gaze data via reading, eye gaze is a promising feature for improving performance in PoS tagging.

1 Introduction

Part-of-speech (PoS) tagging has been widely researched for its ubiquitous role in Natural Language Processing (NLP) tasks, such as machine translation, question-answering, and information extraction. There are many state-of-the-art PoS taggers that use supervised learning methods. These models achieve $>$ 97% accuracy (e.g., Akbik et al. 2018; Bohnet et al. 2018; Ling et al. 2015), however, require large amounts of annotated data. Such data is expensive and not readily available in all languages. For this reason, there is an ongoing search for ways to train improved PoS taggers with limited-to-no annotated data.

One proposed method of improvement draws from eye gaze data. This type of human cognitive processing has proven informative for several NLP tasks (for an overview, see Hollenstein et al. 2020) and can also be collected without the training that annotation requires. The present study aims to bring these two ideas together, exploring improvements to PoS tagging for an existing semi-supervised learning model, utilizing both minimal annotated data and novel eye gaze features.

2 Previous work

2.1 Semi- and un-supervised PoS taggers

Eye gaze aside, various studies have explored semi-supervised and unsupervised learning methods for PoS tagging. Lin et al. 2015 proposed unsupervised PoS induction using a conditional random field autoencoder and word embedding features, showing that embeddings were an effective predictor of syntactic categories even with no labeled data for training. Another study by Stratos and Collins 2015 used active learning to produce partially-labeled datasets and trained a classifier using features such as word embeddings and Brown cluster bitstrings. By a similar token, Li et al. 2012 trained what they called a "weakly-supervised" model, in which they had no annotated data, but use a tagging dictionary and HMM with maximum entropy emissions to learn PoS tags. Earlier methods use mainly various versions of the HMM with differing features (for an overview, see Christodoulopoulos et al. 2010). The proposed models achieve moderate levels of accuracy, the highest at 92%, but still not reaching that of supervised learning methods. A gap thus remains in the field of NLP to discover ways to improve accuracy, while still avoiding the need for fully annotated datasets.

2.2 Eye gaze in PoS tagging

Recent work attempts to ameliorate PoS taggers with features from eye gaze data. Eye gaze is the information collected while tracking one's eye movements as they perform a task – in this case, reading. It is well-known in psycholinguistics that fixation duration reflects the processing load of the reader. Specifically, readers fixate longer on open class words, like nouns and verbs, compared to closed class words, like determiners (Rayner 1998). It stands then that if eye gaze reflects classes of words, eye gaze in turn can be used to inform which class a word belongs to.

Most importantly, eye gaze is cheaper than annotation, inasmuch that collecting eye gaze data only requires that the participant be a reader of the language. Annotation, on the other hand, requires the annotator be linguistically trained to preform the time-consuming task of labeling words. Eye gaze can be conducted with relatively inexpensive software and a webcam or smartphone (Skovsgaard et al. 2013; Xu et al. 2015).

Barrett, Bingel, et al. 2016 build upon the "weakly-supervised" model posed by Li et al. 2012 with eye gaze features (introduced in Section 1), improving performance by 3%, a modest, yet significant improvement. This study extracted gaze features per word including fixation duration, number of fixations, and fixation regressions (back to the word after the first look). Features were extracted from the Dundee corpus (Kenney et al. 2003) which is composed of readings from the newspaper *The Independent*. These features were then incorporated into a type-constrained second-order HMM with maximum entropy emissions. Type-constraining meant that the emissions were confined for any given word to tags specified by a tagging dictionary (using the crowed-sourced Wikitonary, [1]). Barrett, Bingel, et al. 2016 trained their model using features at both the token and type levels for each word, finding better performance overall with type level features. This finding bodes well for eye gaze usage in NLP as it means that token level features are not needed at test time.

Another study by Barrett, González-Garduño, et al. 2018 incorporated eye gaze (as well as key stroke and prosodic features) into unsupervised PoS induction. They found these features led to error reductions of 1.5-4.5% on various datasets compared to models using statistical and word embedding features. Other studies integrate eye gaze into models with various levels of supervision (Barrett, Keller, et al. 2016; Barrett and Søgaard 2015; Klerke and Plank 2019). Performance, however, remains under 90% accuracy when full annotated data is not available.

3 The present study

The goal of the present study is to chose a semi-supervised learning model that has not yet been attempted with eye gaze features. Using a corpus that reports eye gaze data from reading, the model will first be re-run in effort to replicate previous findings and determine how well this model can generalize to a new domain (Experiment 1). Second, the model's features will be replaced with eye gaze features to compare performance (Experiment 2).

3.1 The GECO Corpus

Eye gaze features were extracted from the English portion of the Ghent Eye-Tracking Corpus (GECO) which is available free for use (Cop et al. 2017). This corpus contains reading data from 14 monolingual English speakers reading the full Agatha Christie novel *The Mysterious Affair at Styles*. The text was presented to readers one paragraph at a time (max 145 characters) and reading was self-paced. In four equal partitions of the novel, readers took breaks and answered comprehension questions about the novel's content. Eye tracking equipment was re-calibrated approximately every 10 minutes, as participants tend to drift from their original positions. The corpus contains over 50,000 words and over 5,000 distinct word types.

A total of 11 eye gaze features were extracted for each word w. These included:

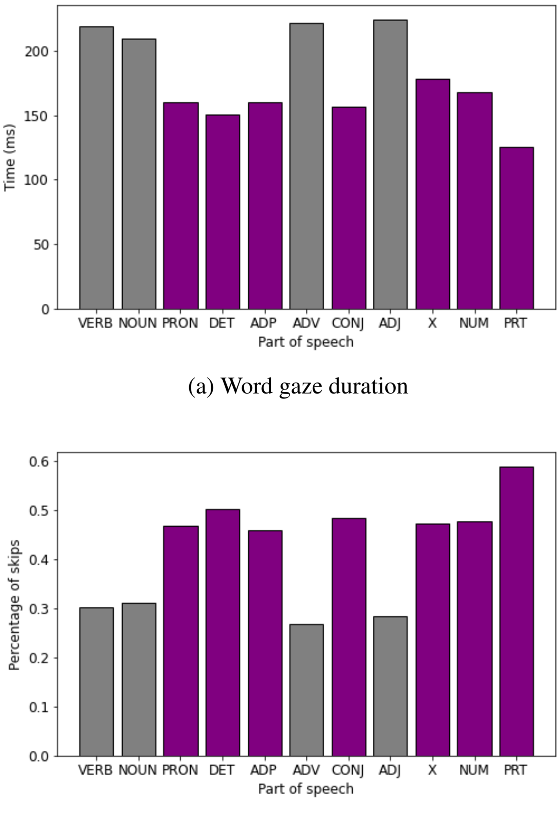
1. Fixation count for w
2. Fixation % for w
3. Gaze duration for w
4. First fixation duration on w
5. Second fixation duration on w
6. Third fixation duration on w
7. Total reading time of w
8. Total reading time % for w
9. Whether w was skipped
10. Word spillover for w
11. Word run count for w

Features did not directly match those 22 eye gaze features of Barrett, Bingel, et al. 2016 from the Dundee Corpus; however, eye gaze data from the GECO corpus has proven useful in other PoS studies (Barrett, González-Garduño, et al. 2018; Klerke and Plank 2019). The Dundee corpus was

[1] https://en.wiktionary.org/wiki/Wiktionary:MainPage

not used here, as it was only available by request. Type-level features were achieved by averaging over word tokens to avoid the need for eye gaze data at test time. Words with a frequency under 13 tokens were replaced with an <UNK> token to account for unknown words appearing in the test set.

The GECO corpus was previously annotated for PoS tags. In the present study, these tags were mapped to the set of 11 Universal PoS tags (Petrov et al. 2012) to match those used by Barrett, Bingel, et al. 2016 and the model by Stratos and Collins 2015 later described in Section 3.2. Averaging each word type over its given label indicated that this gaze data followed previous psycholinguistics findings for gaze patterns and word classes. Figure 1 shows, overall, readers looked at open class words longer than closed class words, but were more likely to skip over closed class words than open class words.



(a) Word gaze duration

(b) Word skip percentage

Figure 1: Eye gaze features from the GECO corpus by PoS tag (gray indicates open class words; purple indicates closed class words)

3.2 The model

The model used in this study, comes from research by Stratos and Collins 2015 who used semi-supervised learning to train a PoS tagger with little annotated data (first introduced in Section 1).

The model begins with the assumption that PoS tagging is deterministic: each word type w maps to a single tag t. This assumption is more restrictive than that of Barrett, Bingel, et al.'s 2016 assumption that each tag can be mapped to a tag available in the Wikitonary, and is not necessarily true in practice (e.g., *fly* could be a verb or noun), but allows for the use of type-level features.

The problem is formulated such that a feature vector $\phi(x, i)$ is extracted for each sentence x and position i in x. Then a multi-class classifier is trained to map $\phi(x, i)$ to a single tag t. The classifier used here is a support vector machine (SVM) implemented with the liblinear library (Fan et al. 2008). This model has many benefits: it can be implemented via out-of-the-box libraries and it is easily amenable to new features. The implementation of Stratos and Collins 2015 [2] was used with modifications to accept the new features. The code for the present study is available at https://github.com/mplatt27/POS-Tagger-w-Eye-Gaze-Data.

Active learning: Active learning with margin sampling was used to obtain partially-labeled datasets to feed the classifier. The goal was to see how little annotated data was needed in order to train a classifier on par with those trained through supervised learning. The active learning process looped for a set number of iterations, until the desired number of labels was achieved. During each iteration, the model selected a new word (or set of words) to label by choosing those with the least confident predicted tag from a pool of unlabeled words.

In order to obtain a dataset with M labeled examples, the initial seed size was set to $k \geq M$ and a step size ξ. On each iteration, until the desired number of annotated words were obtained, the first k most frequent word types were selected for labeling. Then, for $M - k/\xi$ iterations, the model was trained with the current number of labeled words. After which, ξ new examples (x, i) that had the least confident values were given labels.

The seed and step sizes were set to 1, as was done with Stratos and Collins 2015. Partially-

[2] https://github.com/karlstratos/minitagger

labeled datatets of sizes ranging from 100 to 1000 were obtained. These were each used separately to train the classifier, and then test on a dev set, in order to determine the accuracy produced with each size labeled set. Because the GECO corpus was already annotated, active learning was simulated with the provided labels (i.e., no human annotation in the active learning phase).

Features: A set of baseline features was used, matching the baseline set in Stratos and Collins 2015. The set included prefixes and suffixes for the word of lengths one through four, and whether or not the word was capitalized, numeric, or alphanumeric.

In their study, Stratos and Collins 2015 had two conditions: the set of baseline features alone, and the baseline set plus the addition of a feature using Brown cluster bitstrings.[3] These bitstrings come from the Brown clustering algorithm (Brown et al. 1992), which takes in a sequence of text and outputs a tree hierarchy that assigns each word type a bitstring. Branches of the tree can be grouped together to create various clusters. The clustering itself is a hierarchical clustering method in which semantically related words are grouped together because they are embedded in similar contexts. These clusters have demonstrated success as features in various NLP tasks, including PoS tagging (e.g., Kashefi 2018; Miller et al. 2004).

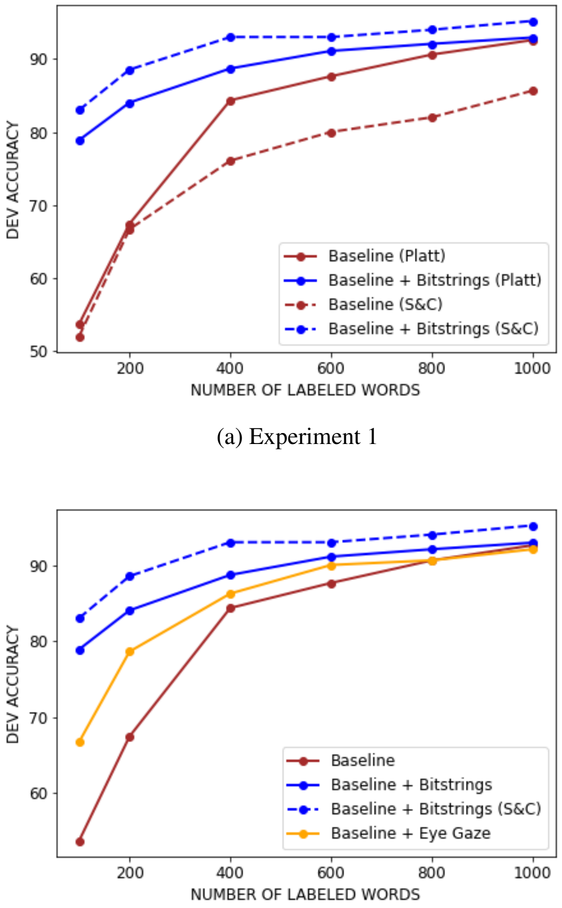
Experiment 1 in the present study aims to replicate Stratos and Collins 2015 using the baseline features alone and the baseline + bitstring features. Bitstring features were re-generated from the GECO corpus using the Brown clustering implementation by Liang 2005 [4]. Experiment 2 replaces the bitstring features with the eye gaze features discussed previously in Section 2.2.

For each word w, the feature vector consisted of features for w, as well as $w+1$, $w+2$, $w-1$, and $w-2$. All features were normalized before being passed to the model.

4 Results

4.1 Experiment 1

The goal of Experiment 1 was to replicate Stratos and Collins 2015 with the GECO corpus. The model was trained using both baseline features alone and baseline + Brown cluster bitstring features. Figure 2 (a) shows the accuracy of the dev set at varying levels of labeled training data [5].



(a) Experiment 1

(b) Experiment 2

Figure 2: Accuracy of the dev set when trained with various amounts of labeled words in Experiments 1 and 2.

Using training data with only 400 labeled words, the model achieves 84.33% accuracy with baseline features alone. Adding bitstring features, accuracy increases to 89.69%. This is the same pattern demonstrated by Stratos and Collins 2015 with the universal treebank dataset. At 400 labeled words, Stratos and Collins 2015 report lower accuracy with baseline features alone, but a higher accuracy when adding the bitstrings.

As the number of labeled words approaches 1000, the accuracy converges around 92% for the baseline and bitstring features combined, indicating the the bitstrings are more informative with lower amounts of labeled training data.

[3] Another condition was included with word embeddings, but will not be considered here.

[4] https://github.com/percyliang/brown-cluster

[5] Exact percentage values were only reported by Stratos and Collins (2015) for 200, 400, and 1000 labeled words, so other values are estimated from the figure they provide.

	Baseline	Baseline + Bitstrings	Baseline + Eye gaze
200	67.38	84.02	78.59
400	84.33	88.69	86.23
1000	92.59	92.94	92.31

Table 1: Dev accuracy percentages for each set of features, reported for 200, 400, and 1000 labeled words in the training data.

4.2 Experiment 2

In Experiment 2, the bitstring features were removed and replaced by the 11 eye gaze features detailed in Section 2.2.

Figure 2 (b) plots again the accuracy of the dev set for various levels of labeled training data. The addition of eye gaze features increased accuracy compared to baseline features alone from 84.33% to 86.23% with only 400 labeled words. This modest increase did not quite match the performance of the baseline + bitstring feature set from Experiment 1 (88.69%). Again, as the number of labels approaches 1000, the features differences are less important for performance. The accuracy percentages are reported in Table 1.

A third condition was run that combined all three features sets: baseline + bitstrings + eye gaze. This grouping, however, produced accuracy similar to the baseline, likely due to the model being overfit from the large number of features. (200 labeled words: 78.88% accuracy; 400 labeled words: 84.11% accuracy; 1000 labeled words: 89.23% accuracy).

5 Discussion

This study set out to explore improvements to a semi-supervised learning model with eye gaze features. The first goal was to replicate the findings from Stratos and Collins 2015 using the GECO corpus, and baseline and bitstring features. The second goal was to swap the bitstring features for eye gaze features from the GECO corpus and compare performance. Ultimately, the overall objective was to match the performance observed in supervised learning models.

Experiment 1 replicated the pattern that Stratos and Collins 2015 observed, in that accuracy increased from baseline features alone compared to the use of baseline + bitstring features. This finding differed in that Stratos and Collins's 2015 dataset performed worse with the baseline features alone, but better with baseline + bitstring features. This may be due to domain differences across datasets.

Experiment 2 also found that the baseline + eye gaze features increased accuracy from the baseline features alone. However, the baseline + eye gaze features did not out-perform the baseline + bitstring features. The differences in performance were modest, but numerically, this suggests that the bitstrings are more indicative of syntactic categories than eye gaze. Nonetheless, before the use of eye gaze can be fully understood additional tests are needed, such as using data from another domain or using another semi-supervised learning model. However, because bitstrings are cheaper to produce than eye gaze data, if they are truly more effective, then they would be a better feature selection. More research is needed to elucidate this relationship.

Overall, the model was able to achieve moderate levels of accuracy, the highest being ≈92% with 1000 labeled words in the training data. However, with this number of labeled words, there were minimal differences between the performance of the different feature sets. With just 400 labeled words in the training set, it is clear that the features chosen do make a difference in performance. The highest accuracy achieved was ≈89% with the baseline + bitstring features. There is still work to be done in order to match the performance of supervised learning models. Even so, this is a relatively high level of accuracy with such minimal labeled training data.

6 Future Research

Future studies could explore eye gaze in another semi-supervised or unsupervised model, or in combination with other features (statistical or cognitive-based).

Other human processing signals may be promising features in PoS tagging, and can be collected with similar ease to that of eye gaze. These include prosody in speech, key strokes in typing, or fMRI and EEG signals (e.g., Barrett, González-Garduño, et al. 2018; Hollenstein et al. 2020). While each of these features would still need to be collected by humans, there is no annotation training required. Likewise, the signals can be collected at type-level, so they are not needed at test time. Future research could determine if these would be effective (and, more effective than eye gaze or bitstrings) in improving PoS tagging performance.

Outside the realm of PoS tagging, the use of eye gaze should be further explored in other NLP tasks. While the model in this study found bitstrings to be more effective than eye gaze features, it is possible that other NLP tasks may not have a competing feature that can be collected with the same ease as bitstrings. These tasks could include (but are not limited to) sarcasm detection (Mishra et al. 2016), referential/non-referential it (Yaneva et al. 2020), and predicting language proficiency (Kunze et al. 2013). More studies are needed to fully understand the impact that eye gaze can make in NLP.

References

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649.

Barrett, M., Bingel, J., Keller, F., & Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 579–584.

Barrett, M., González-Garduño, A. V., Frermann, L., & Søgaard, A. (2018). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2028–2038.

Barrett, M., Keller, F., & Søgaard, A. (2016). Cross-lingual transfer of correlations between parts of speech and gaze features. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1330–1339.

Barrett, M., & Søgaard, A. (2015). Reading behavior predicts syntactic categories. *CoNLL*.

Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., & Maynez, J. (2018). Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2642–2652.

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, *18*(4), 467–480.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 575–584.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*, 602–615.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, *9*(Aug), 1871–1874.

Hollenstein, N., Barrett, M., & Beinborn, L. (2020). Towards best practices for leveraging human language processing signals for natural language processing. *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, 15–27.

Kashefi, O. (2018). Unsupervised part-of-speech induction.

Kenney, A., Hill, R., & Pynte, J. (2003). The dundee corpus. *Proceedings of the 12th European conference on eye movement*.

Klerke, S., & Plank, B. (2019). At a glance: The impact of gaze aggregation views on syntactic tagging. *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, 51–61.

Kunze, K., Kawaichi, H., Yoshimura, K., & Kise, K. (2013). Towards inferring language expertise using eye tracking. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 217–222.

Li, S., Graça, J. V., & Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging, 1389–1398.

Liang, P. (2005). Semi-supervised learning for natural language. *Masters thesis*.

Lin, C.-C., Ammar, W., Dyer, C., & Levin, L. (2015). Unsupervised POS induction with word embeddings. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1311–1316.

Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., & Luıs, T. (2015). Finding function in form: Compositional character models for open vocabulary word representation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1520–1530.

Miller, S., Guinness, J., & Zamanian, A. (2004). Name tagging with word clusters and discriminative training. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 337–342.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., & Bhattacharyya, P. (2016). Harnessing cognitive features for sarcasm detection. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1095–1104.

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2089–2096.

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

Skovsgaard, H., Hansen, J., & Møllenbach, E. (2013). Gaze tracking through smartphones [Gaze Interaction in the Post-WIMP World : CHI 2013 One-day Workshop, CHI 2013 ; Conference date: 27-04-2013 Through 27-04-2013]. *Proceedings. CHI 2013 Workshop on "Gaze Interaction in the Post-WIMP World"*.

Stratos, K., & Collins, M. (2015). Simple semi-supervised POS tagging. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 79–87.

Wiktionary, the free dictionary. (n.d.). https://en.wiktionary.org/wiki/Wiktionary:Main_Page (accessed: 04.24.2021)

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *CoRR*, *abs/1504.06755*.

Yaneva, V., Ha, L. A., Evans, R., & Mitkov, R. (2020). Classifying referential and non-referential it using gaze.