Proceedings of the 12th Web as Corpus Workshop, pages 50–56
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging

Shabnam Behzad, Amir Zeldes
Corpling Lab
Georgetown University
shabnam@cs.georgetown.edu, amir.zeldes@georgetown.edu

Abstract

Part of speech tagging is a fundamental NLP task often regarded as solved for high-resource languages such as English. Current state-of-the-art models have achieved high accuracy, especially on the news domain. However, when these models are applied to other corpora with different genres, and especially user-generated data from the Web, we see substantial drops in performance. In this work, we study how a state-of-the-art tagging model trained on different genres performs on Web content from unfiltered Reddit forum discussions. More specifically, we use data from multiple sources: OntoNotes, a large benchmark corpus with 'well-edited' text, the English Web Treebank with 5 Web genres, and GUM, with 7 further genres other than Reddit. We report the results when training on different splits of the data, tested on Reddit. Our results show that even small amounts of in-domain data can outperform the contribution of data an order of magnitude larger coming from other Web domains. To make progress on out-of-domain tagging, we also evaluate an ensemble approach using multiple single-genre taggers as input features to a meta-classifier. We present state of the art performance on tagging Reddit data, as well as error analysis of the results of these models, and offer a typology of the most common error types among them, broken down by training corpus.

Keywords: POS, tagger, genre, domain adaptation, ensemble

1. Introduction

With the rapid growth of social media platforms and general public participation on the Internet, user-generated content has become one of the main data resources for different applications (Sanguinetti et al., 2020). Textual data from these platforms are being used in many NLP tasks even though they are often not well-structured, and deviate from prescriptive language norms. The combination of such heterogeneous data and differences with typical kinds of training data (often newswire language) are hence challenging to work with: different types of noise are introduced into these datasets because of non-standard lexical items, spelling inconsistencies, informal abbreviations, and linguistic errors (Gui et al., 2017; Meftah and Semmar, 2018).

Part of speech tagging is a fundamental NLP task which has long been studied, and, based on standard benchmarks, now seems nearly solved: for example, recent approaches have reached an accuracy of 97.85% (Akbik et al., 2018) on the Wall Street Journal corpus, essentially approaching human levels of accuracy. However, when state-of-the-art models are evaluated on out of domain data, we observe a drop in performance (Derczynski et al., 2013); this could be the result of differences in topic, writing style and epoch between training and testing data (Manning, 2011). At the same time, high quality POS tagging is particularly pertinent for non-standard language, since exposing parts of speech in unusual text types gives access to underlying categories (e.g. proper nouns, predicates) which are difficult to recognize on a textual basis when they have unusual forms. Because the resulting POS tags are frequently used as part of downstream NLP tasks, errors caused by the tagger can propagate and affect the results of these downstream tasks as well (Foster et al., 2011). Thus, NLP tasks can benefit substantially from high accuracy, domain-robust POS tagging.

Enhancing the performance of taggers for social media data in particular has been studied before. Most of these studies, however, have focused on data from Twitter, which diverges from standard language strongly, but also represents a very narrow subdomain of user-generated content. In this work, we focus on a different platform, Reddit, which has a more heterogeneous text structure from Twitter. We compare the performance of the state-of-the-art tagging framework Flair (Akbik et al., 2018) trained on different genres and tested on Reddit data, and provide a deep analysis of the errors produced in each of the models. Our initial results suggest that even small amounts of in-domain data used in training can outperform the contribution of data an order of magnitude larger but from other domains, despite the fact that most of the data sources used in this paper come from a range of Web genres themselves. In order to achieve progress on generalization to new domains, we also evaluate an ensemble model which uses the predictions of multiple models trained on different genres as features. We observe the effectiveness of these features by an ablation study and report the results.

2. Related Work

Over the past decades there has been a growing body of work focusing on POS tagging and domain adaptation. Many approaches have been proposed to improve tagging performance using different models such as Conditional Random Fields, Hidden/Maximum Entropy Markov Models, linear classifiers and neural architectures (Mueller et al., 2013; Sun, 2014; Huang et al., 2015; Choi, 2016; Qi et al., 2018; Akbik et al., 2018).

With the growth of social media and the tremendous amount of user-generated textual data available, researchers now analyze and use these data in many different NLP tasks (Liu et al., 2018). Studies show that the performance of NLP tools including POS taggers typically degrades when the models are tested on unedited text such as

tweets (Ritter et al., 2011), however, retraining the models on in-domain data can improve performance (Neunerdt et al., 2013). Giesbrecht and Evert (2009) presented an evaluation of various POS taggers in German when trained on newspaper corpora and then tested on less standardized text genres such as Web corpora and observed a drop in performance. They also analyzed how tagging different web genres could present different levels of difficulty for the trained models. More specifically, they found TV episode guide, online forum, conference information and news report data to be harder than other text genres.

Studies of tagging specifically for the heterogeneous space of Reddit text remain outstanding. Previous research has studied the problem of POS tagging on social media data primarily by targeting Twitter. Some have proposed new tagging schemes and released new annotated datasets. Ritter et al. (2011) added new tags for Twitter specific phenomena such as #hashtags and @usernames. Gimpel et al. (2011), developed a POS tagset specifically for English Twitter and a new dataset of manually tagged tweets. Owoputi et al. (2013) released a new manually annotated dataset for English Twitter POS tagging along with a part of speech tagger for online conversational text. There have also been efforts on POS tagging for other languages such as Irish (Lynn et al., 2015) and Italian (Bosco et al., 2016). A shared task on the Automatic Linguistic Annotation of Computer-Mediated Communication (CMC) and Web Corpora for German was also organized by Beißwenger et al. (2016) to observe whether both CMC and Web corpora can be processed using the same methodologies and whether improved models can be introduced for tokenization and POS tagging of German computer-mediated communication using the new annotated data and other techniques, such as domain adaptation. Domain adaptation and regularization are helpful techniques when dealing with low-resource text types and many studies have focused on enhancing POS tagging using such methods (Meftah et al., 2019; März et al., 2019).

Some studies have conducted error analysis of social media taggers, though not yet on Reddit. Derczynski et al. (2013) evaluated the performance of existing POS taggers on Twitter datasets. They also provide an in-depth analysis of the errors on the tokens that were not seen during training. They report gold standard error, slang, genre-specific tokens and unseen proper nouns among the common error categories. Albogamy and Ramsay (2016) also evaluate state-of-the-art POS taggers on Arabic tweets. They categorize errors into 2 groups: errors on Arabic words and errors on non-Arabic tokens. Each of these groups includes subcategories such as named entities that were not seen during training, concatenation of multiple words, emoticons, foreign words, and others.

To the best of our knowledge, such in-depth studies have not yet been done on Reddit even though it is widely used as a data source for different NLP tasks. In this paper, we study genre effects on POS tagging accuracy for Reddit text when training data itself comes entirely from the Web (but not from Reddit), from other large benchmark resources such as OntoNotes (Hovy et al., 2006)) or both. We provide a detailed error analysis of different models, which suggests that some of the difficulties in tagging Reddit are not only due to the noisy nature of text online, but also to specific language use in Reddit as a genre. We also present an ensemble tagging approach that has a higher accuracy than the best single training genre baseline.

3. Approach

3.1. Data

In this study, we use three different corpora with different genres. The main corpus we used is GUM (the Georgetown University Multilayer corpus (Zeldes, 2017)) which was chosen because it contains gold standard tagged Reddit data. The corpus has manual annotation for different tasks such as POS tagging, lemmatization, dependency parses, discourse parses and entity and coreference resolution (Zeldes, 2017), though the latter layers are not used in this study. Currently, GUM comprises about 130,000 tokens with data from 8 different genres in English, which, aside from Reddit, include creative commons licensed Academic papers and Fiction, Biographies (Bio) from Wikipedia, WikiNews Interviews and News stories, Wikivoyage travel guides and Wikihow how-to guides (Whow). Importantly, all of the data in the corpus was harvested from the Web, meaning that even when training on other genres and testing on Reddit, only data which is encountered on the Internet is involved. In order to get comparable numbers for models trained on other popular benchmark resources, we also use larger corpora such as EWT (Bies et al., 2012) (about 250,000 tokens of data from the Web) and English OntoNotes (Weischedel et al., 2013) (about 2.6 million tokens, mostly from edited print texts and spoken data) in our experiments which are mainly used for POS tagging evaluations.

We have 12 different training splits for this task; for every GUM genre except for Reddit, we use all available data as the training set. Reddit has the smallest training set since we need some of the documents for development and test sets. Out of 11,182 annotated Reddit tokens in GUM version 5, we use 5,727 tokens for training, and 2,489 tokens for development and 2,966 tokens for the test set. The Reddit documents are from different discussion threads which makes evaluation more realistic.

We also create a split that contains the training data of multiple genres (Reddit, Academic, Bio, Fiction, Interview, News, Voyage, Whow) (*Multiple Genres*) and another one which contains training data from the same genres except for Reddit (*Multiple Genres w/o Reddit*). The size of all these training sets is shown in Table 1. For all of our models, we use the same Reddit development and test sets mentioned above.

For OntoNotes and EWT we use the entire corpora as datasets, without considering sub-genres within those resources, as most papers using them for training employ the entire corpus training set without internal distinctions. Although we recognize it would be interesting to analyze the contents of these data sets further, we leave that task open for future studies.

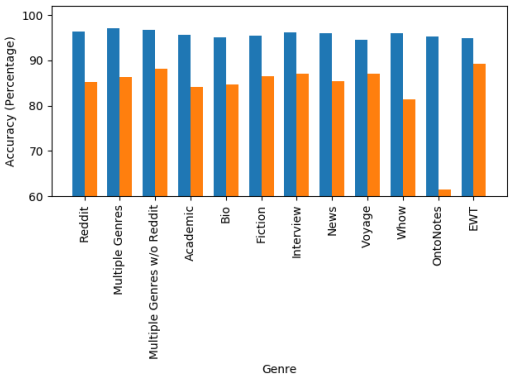

Figure 1: Accuracy on *Known* and *Unknown* tokens per genre; blue bars (left) correspond to accuracy of Known tokens and orange bars (right) correspond to accuracy of Unknown tokens.

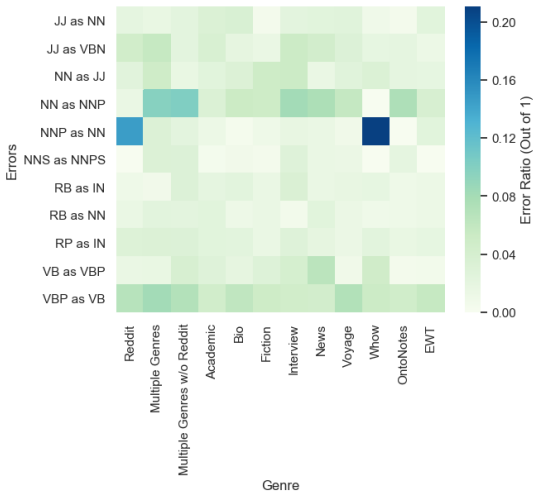

Figure 2: Most common prediction errors on the test set across models trained on different genres. Y-axis labeling: *A as B* means A incorrectly tagged as B.

3.2. Tagger

For tagging, we choose a state-of-the-art neural sequence tagger, Flair (Akbik et al., 2018) and retrain it on our splits. We used the sequence tagger model with the default 256 hidden unit bi-directional LSTM and trained with contextualized pre-trained Flair embeddings and uncontextualized pre-trained character embeddings, then evaluated performance by accuracy per token and also full-sentence accuracy (proportion of perfectly tagged sentences), since "a single bad mistake in a sentence can greatly throw off the usefulness of a tagger to downstream tasks such as dependency parsing" (Manning, 2011).

Finally, we use an ensemble approach to combine results from multiple models and study how much each of these sources contributes to the results of the ensemble model [1]. We use all the retrained Flair models on single genres except *Reddit*, and then make predictions on the Reddit training set. We then use these predictions as training features for our ensemble model, which uses XGBoost as a meta-learner. Based on the analysis described in section 4 and to help the model better distinguish between NN and NNP, we also incorporate three other features to help the ensemble classifier; for each token, we check if 1) the token itself, 2) the lower-cased version of the token and 3) the token starting with a capital letter, exists in a knowledge base taken from (Zeldes and Zhang, 2016) and add any entity types (e.g. Person, Organization etc.) which this token might have as n-hot encoded features. We then evaluate the classifier on the Reddit test set and perform an ablation study by removing the predictions of each genre and observing the changes in the accuracy.

4. Results and Analysis

The results of our experiments are shown in Table 1. As expected, the highest per token and also full-sentence accuracy belong to the model trained on multiple genres since there is more training data available and we are also including in-domain Reddit data in training. The model trained on multiple genres without Reddit gets only slightly lower accuracy per token, however, we observe a 3.2% drop on full-sentence accuracy. Even though the Reddit model has the smallest amount of data for training (less than half of most other genres, due to the held out dev and test sets), it performs better than almost all models trained on different genres, which shows the high importance of even a small amount of in-domain data. Interestingly, the model trained on interviews works slightly better than the model trained on Reddit, probably because interviews published online are the most similar to the largely first and second person interactions found in Reddit forum discussions, and because the interview dataset is substantially larger than the Reddit training data.

In Table 2, we can compare errors made by different models on the same sentences. None of the models can predict correct POS tags for the whole sentences. Surprisingly, even though *Multiple Genres* has more data than *Multiple Genres w/o Reddit* including in-domain data, it performs worse in the first sentence; it cannot predict the tag NNP for the token 'Wild'.

The most common error among all of the models is mistagging the token 'b.' in the first sentence, which is indicating an item of a list and should get the tag LS. The second most common mistake seems to be the emoticon :) in the second sentence. Reddit and EWT are the only models predicting correct tags for this token. 'love' has the gold label VBP but it is predicted as VB or NN by different models, due to the low frequency of subjectless sentences, which resemble imperatives or fragments if the missing 'I' is not recognized. NNP tokens such as 'Boo' or 'Wild' are incorrectly tagged as NN by many of the models, mirroring findings on NNP tagging problems in previous studies.

We also look at the accuracy of models on *Unknown* tokens (not seen during training) and *Known* tokens separately. Figure 1 shows these results. Except for models trained on Academic, Bio, Whow and OntoNotes data, all other models perform better than the Reddit model on Un-

[1] Setup details are available at `https://github.com/shabnam-b/reddit-pos-ensemble.git`

	Reddit	Multiple Genres	Multiple Genres w/o Reddit	Academic	Bio	Fiction
Training Set Size (Tokens)	5,727	107,004	101,277	11,868	12,562	12,843
Per Token	93.53	**95.89**	95.72	91.81	91.77	93.29
Full-sentence	36.08	**53.16**	49.37	29.75	25.32	37.97
	Interview	**News**	**Voyage**	**Whow**	**OntoNotes**	**EWT**
Training Set Size (Tokens)	18,037	14,092	14,955	16,920	2,442,000	204,609
Per Token	94.23	93.26	92.48	92.95	93.73	94.81
Full-sentence	39.87	31.01	30.38	31.01	41.14	48.10

Table 1: Accuracy scores calculated for tokens and full-sentences when trained on different genres individually and tested on Reddit.

Model	Example
Reddit	*b./: Using these to release Boo/NN into "The Wild/NN"*
	love/VB when I see people/places from Austin on FN :)
Multiple Genres	*b./FW Using these to release Boo/NN into "The Wild/NN"*
	love/VB when I see people/places from Austin on FN :)/-RRB-
Multiple Genres w/o Reddit	*b./FW Using these to release Boo/NN into "The Wild"*
	love/VB when I see people/places from Austin on FN :)/:
Academic	*b./NN Using these to release Boo into "/DT The Wild"/CC*
	love/NN when I see people/places from Austin on FN :)/:
Bio	*b./FW Using these to release Boo into "The Wild"*
	love/NN when I see people //CC places from Austin on FN :)/:
Fiction	*b./“ Using these to release Boo into "/NNP The Wild"*
	love/NN when I see people //: places from Austin on FN :)/:
Interview	*b./: Using these to release Boo into "The/NNP Wild"*
	love/VB when I see people //CC places from Austin on FN :)/:
News	*b./NNP Using these to release Boo/NN into "The Wild"*
	love/NN when I see people/places from Austin on FN :)/:
Voyage	*b./RB Using these to release Boo/NN into "The Wild"*
	love/VB when I see people/places from Austin on FN :)/:
Whow	*b./: Using these to release Boo/NN into "The Wild/NN"*
	love/VB when I see people/places from Austin on FN :)/:
OntoNotes	*b./NN Using these to release Boo into "The Wild"*
	love/VB when I see people/places from Austin on FN :)/.
EWT	*b./RB Using these to release Boo/NN into "The Wild"*
	love/VB when I see people//,places from Austin on FN :)

Table 2: Errors made by different models on two example sentences from Reddit posts.

known tokens, but this again could be the result of Reddit having a very small training set compared to other genres.

To further analyze the results, we looked at misclassifications which were common among multiple genres. The results are shown in Figure 2. The most common errors across all genres are VBP predicted as VB and NN predicted as NNP. The latter can stem from looser capitalization distinctions online, while the former can result when subject pronouns are dropped in informal English (e.g. ‘want to come?’ or ‘need this right now’). Comparing *Multiple Genres* and *Multiple Genres w/o Reddit*, we can observe that adding the Reddit data results in more accurate RB, RP, and VB tagging. We can also observe that a huge proportion of the *Whow* model’s errors belongs to mistagging NNP as NN, which is probably the result of fewer proper nouns appearing in Wikihow articles since they are sets of instructions for various tasks.

Furthermore, we manually observed 50 of the errors that the *Multiple Genres w/o Reddit* model made. The most common errors were 1) *Emoticons:* Emoticons such as :) , :(or others which are gold labeled as SYM are labeled with different tags such as ”, : or even NNP in cases where they contain an alphabetical character such as in D:>. 2) *Interjections* and mostly swear words appear in social media text more than other genres, as well as phonetic elongation or representations with repeated characters such as ‘NANANANA’, which do not appear in formal written text, but are common among users in social media (Sanguinetti et al., 2020). Some of these tokens were tagged as NNP instead of gold standard UH. Some other errors were 3) *Proper nouns not starting with a capital letter* (e.g. bobby/NN), 4) *Foreign words* (e.g. etcetera/NNP) and 5) *Abbreviations* (e.g. BTW/NNP).

Finally, in order to harness the increased stability offered by consulting multiple models and different features, Table 3 shows the results of the ensemble model described in Section 3. Except for Interview, all models positively contribute to the overall accuracy. Only the Interview model’s

Model	Per Token	Full-sentence
StanfordNLP	94.81	46.84
TreeTagger	92.08	28.48
Ensemble	95.99	53.80
- (Academic)	95.92	53.80
- (Bio)	95.89	53.16
- (Fiction)	95.95	54.43
- (Interview)	**96.12**	**56.96**
- (News)	95.89	53.80
- (Voyage)	95.92	55.06
- (Whow)	95.82	51.90
- (OntoNotes)	95.55	50.00
- (EWT)	95.62	50.63

Table 3: Accuracy score of StanfordNLP, TreeTagger and ensemble XGBoost when using the prediction of all trained models, and when each model is removed.

removal from the ensemble improves upon the results in Table 1 both in terms of per token accuracy and full sentence accuracy, which suggests that, at least for the test set at hand, other genres combined do a better job of predicting correct tags, despite the usefulness of interviews in a single genre model. The final model without 'interview' thus represents our best results and a new state-of-the-art score on Reddit tagging using the GUM benchmark, with 96.12% accuracy despite not including any Reddit data in training the features for the meta-classifier. Furthermore, we compare these results with two pretrained off-the-shelf taggers: TreeTagger trained on Penn treebank and StanfordNLP (Qi et al., 2018) trained on GUM. We also looked at the effect of removing the named entity features on the results; without the named entities, the best model's accuracy (Ensemble-Interview) drops to 95.89% per token and 55.06% for full-sentence. Comparing these numbers to the best single model in Table 1, the ensemble approach without any extra features is resulting in the same token accuracy, but we get almost 2% increase in full-sentence accuracy.

5. Conclusion

In this work, we looked at the effect of genre on Reddit POS tagging. We analyzed the results of the same tagger, trained on 10 different sources with multiple genres, including Reddit itself, and also trained on the combination of all genres. We observed that within single genre models, the Interview model has the highest accuracy on Reddit, which might be the result of comparatively much training data (Interview has somewhat more tokens than the other single genre datasets), or the nature of some of the Reddit documents which are back and forth conversations between different users and is similar to the nature of interviews. However, in combination with other models in the ensemble approach, removing the Interview model seemed to increase the performance slightly, and OntoNotes predictions seem to have the most positive contributions to the accuracy of the ensemble model, possibly because of the wide coverage of rare items resulting from the large corpus size. Finally, the results of our error analysis were in line with prior studies on Web text-types other than Reddit. The most important problem in this task using deep learning models with word/character embeddings is when we have unseen data in the test set; this unseen data could be in the form of creative emoticons, repeated characters in a word, abbreviations, etc. To improve the performance on user-generated content in domains such as social media, we either need to collect sufficient in-domain data to train a genre-specific model, or find other ways of addressing unseen tokens such as using lexical resources or doing specific preprocessing to normalize the tokens before testing. The present paper demonstrates that, in the absence of substantial amounts of in-domain data, ensembling the outputs of multiple tagging models with different training datasets can lead to very good results, in this case giving a new SOA score of 96.12% token accuracy on the Reddit data. At the same time, full-sentence accuracy remains below 57%, suggesting that there is still room for substantial improvements.

6. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Albogamy, F. and Ramsay, A. (2016). Fast and robust POS tagger for Arabic tweets using agreement-based bootstrapping. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1500–1506, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 44–56, Berlin, August. Association for Computational Linguistics.

Bosco, C., Fabio, T., Andrea, B., and Mazzei, A. (2016). Overview of the evalita 2016 part of speech on twitter for italian task. In *Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749.

Choi, J. D. (2016). Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 271–281, San Diego, California, June. Association for Computational Linguistics.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and Van Genabith, J. (2011). # hardtoparse: POS tagging and parsing the Twitterverse.

In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In *Proceedings of the 5th Web as Corpus Workshop*.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA, June. Association for Computational Linguistics.

Gui, T., Zhang, Q., Huang, H., Peng, M., and Huang, X. (2017). Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420, Copenhagen, Denmark, September. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into universal dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana, June. Association for Computational Linguistics.

Lynn, T., Scannell, K., and Maguire, E. (2015). Minority language twitter: Part-of-speech tagging and analysis of Irish tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8, Beijing, China, July. Association for Computational Linguistics.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

März, L., Trautmann, D., and Roth, B. (2019). Domain adaptation for part-of-speech tagging of noisy user-generated text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3415–3420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Meftah, S. and Semmar, N. (2018). A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Meftah, S., Tamaazousti, Y., Semmar, N., Essafi, H., and Sadat, F. (2019). Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4107–4112, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Neunerdt, M., Trevisan, B., Reyer, M., and Mathar, R. (2013). Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web*, pages 139–150. Springer.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Sanguinetti, M., Bosco, C., Cassidy, L., Özlem Çetinoğlu, Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2020). Treebanking user-generated content: A proposal for a unified representation in universal dependencies. In *Proceedings of LREC 2020*, Marseille, France.

Sun, X. (2014). Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems 27*, pages 2402–2410. Curran Associates, Inc.

Zeldes, A. and Zhang, S. (2016). When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the NAACL2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, pages 92–101, San Diego, CA.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

7. Language Resource References

Ann Bies and Justin Mott and Colin Warner and Seth Kulick. (2012). *English Web Treebank*. LDC, ISLRN

230-396-178-102-3.

Ralph Weischedel and Martha Palmer and Mitchell Marcus and Eduard Hovy and Sameer Pradhan and Lance Ramshaw and Nianwen Xue and Ann Taylor and Jeff Kaufman and Michelle Franchini and Mohammed El-Bachouti and Robert Belvin and Ann Houston. (2013). *OntoNotes Release 5.0*. LDC, 5.0, ISLRN 151-738-649-048-2.

Amir Zeldes. (2017). *The Georgetown University Multilayer Corpus*. Georgetown University.