Tagging Performance Correlates with Author Age

Dirk Hovy[1] and Anders Søgaard[1]
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
{dirk.hovy,soegaard}@hum.ku.dk

Abstract

Many NLP tools for English and German are based on manually annotated articles from the Wall Street Journal and Frankfurter Rundschau. The average readers of these two newspapers are middle-aged (55 and 47 years old, respectively), and the annotated articles are more than 20 years old by now. This leads us to speculate whether tools induced from these resources (such as part-of-speech taggers) put older language users at an advantage. We show that this is actually the case in both languages, and that the cause goes beyond simple vocabulary differences. In our experiments, we control for gender and region.

1 Introduction

One of the main challenges in natural language processing (NLP) is to correct for biases in the manually annotated data available to system engineers. Selection biases are often thought of in terms of textual domains, motivating work in domain adaptation of NLP models (Daume III and Marcu, 2006; Ben-David et al., 2007; Daume III, 2007; Dredze et al., 2007; Chen et al., 2009; Chen et al., 2011, inter alia). Domain adaptation problems are typically framed as adapting models that were induced on newswire to other domains, such as spoken language, literature, or social media.

However, newswire is not just a domain with particular conventions. It is also a source of information written by and for particular people. The reader base of most newspapers is older, richer, and more well-educated than the average population. Also, many newspapers have more readers in some regions of their country. In addition, newswire text is much more canonical than other domains, and includes fewer neologisms and non-standard language. Both, however, are frequent in the language use of young adults, who are the main drivers of language change (Holmes, 2013; Nguyen et al., 2014).

In this paper, we focus on the most widely used manually annotated resources for English and German, namely the English Penn Treebank and the TIGER Treebank for German. The English treebank consists of manually annotated Wall Street Journal articles from 1989. The TIGER Treebank consists of manually annotated Frankfurter Rundschau articles from the early 1990s. Both newspapers have regionally and demographically biased reader bases, e.g., with more old than young readers. We discuss the biases in §2.

In the light of recent research (Volkova et al., 2013; Hovy, 2015; Jørgensen et al., 2015), we explore the hypothesis that these biases transfer to NLP tools induced from these resources. As a result, these models perform better on texts written by certain people, namely those whose language is closer to the training data. Language dynamics being what they are, we expect English and German POS taggers to perform better on texts written by older people. To evaluate this hypothesis, we collected English and German user reviews from a user review site used by representative samples of the English and German populations. We annotated reviews written by users whose age, gender, and location were known with POS tags. The resulting data set enables us to test whether there are significant performance differences between ages, genders, and regions, while controlling for the two respective other, potentially confounding, factors.

Contribution We show that age bias leads to significant performance differences in off-the-shelf POS taggers for English and German. We also analyze the relevant linguistic differences between the age groups, and show that they are *not*

[1] Both authors contributed equally to the paper, and flipped a heavily biased coin until they were both satisfied with the order.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 483–488,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

solely lexical, but instead extend to the grammatical level. As a corollary, we also present several new evaluation datasets for English and German that allow us to control for age, gender, and location.

2 Data

2.1 Wall Street Journal and Frankfurter Rundschau

The *Wall Street Journal* is a New York City-based newspaper, in print since 1889, with about two million readers. It employs 2,000 journalists in 85 news bureaus across 51 countries. Wall Street Journal is often considered business-friendly, but conservative. In 2007, Rupert Murdoch bought the newspaper. The English Penn Treebank consists of manually annotated articles from 1989, including both essays, letters and errata, but the vast majority are news pieces.[1]

Frankfurter Rundschau is a German language newspaper based in Frankfurt am Main. Its first issue dates back to 1945, shortly after the end of the second world war. It has about 120,000 readers. It is often considered a left-wing liberal newspaper. According to a study conducted by the newspaper itself,[2] its readers are found in “comfortable” higher jobs, well-educated, and on average in their mid-forties. While the paper is available internationally, most of its users come from the Rhine-Main region.

2.2 The Trustpilot Corpus

The Trustpilot Corpus (Hovy et al., 2015a) consists of user reviews scraped from the multilingual website trustpilot.com. The reviewer base has been shown to be representative of the populations in the countries for which large reviewer bases exist, at least wrt. age, gender, and geographical spread (Hovy et al., 2015a). The language is more informal than newswire, but less creative than social media posts. This is similar to the language in the reviews section of the English Web Treebank.[3] For the experiments below, we annotated parts of the British and German sections of the Trustpilot Corpus with the tag set proposed in Petrov et al. (2011).

2.3 POS annotations

We use an in-house interface to annotate the English and German data. For each of the two languages, we annotate 600 sentences. The data is sampled in the following way: we first extract all reviews associated with a location, split and tokenize the review using the NLTK tokenizer for the respective language, and discard any sentences with fewer than three or more than 100 tokens. We then map each review to the NUTS region corresponding to the location. If the location name is ambiguous, we discard it.

We then run two POS taggers (TreeTagger[4], and a model implemented in CRF++[5]) to obtain log-likelihoods for each sentence in the English and German sub corpora. We normalize by sentence length and compute the average score for each region under each tagger.

We single out the two regions in England and Germany with the highest, respectively lowest, average log-likelihoods from both taggers. We do this to be able to control for dialectal variation. In each region, we randomly sample 200 reviews written by women under 35, 200 reviews written by men under 35, 200 reviews written by women over 45, and 200 reviews written by men over 45. This selection enables us to study and control for gender, region, and age.

While sociolinguistics agrees on language change between age groups (Barke, 2000; Schler et al., 2006; Barbieri, 2008; Rickford and Price, 2013), it is not clear where to draw the line. The age groups selected here are thus solely based on the availability of even-sized groups that are separated by 10 years.

3 Experiments

3.1 Training data and models

As training data for our POS tagging models, we use manually annotated data from the Wall Street Journal (English Penn Treebank) and Frankfurter Rundschau (TIGER). We use the training and test sections provided in the CoNLL 2006–7 shared tasks, but we convert all tags to the universal POS tag set (Petrov et al., 2011).

[1] http://www.let.rug.nl/~bplank/metadata/genre_files_updated.html

[2] http://www.fr-online.de/wir-ueber-uns/studie-wer-sind-unsere-leser-,4353508,4356262.html

[3] https://catalog.ldc.upenn.edu/LDC2012T13

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[5] http://taku910.github.io/crfpp/

Our POS taggers are trained using TreeTagger with default parameters, and CRF++ with default parameters and standard POS features (Owoputi et al., 2013; Hovy et al., 2015b). We use two different POS tagger induction algorithms in order to be able to abstract away from their respective inductive biases. Generally, TreeTagger (TreeT) performs better than CRF++ on German, whereas CRF++ performs best on English.

3.2 Results

country	group	TreeT	CRF++	avg.
DE	U35	87.42	85.93	86.68
	O45	**89.39**	87.04	88.22
	male	88.53	86.11	87.32
	female	88.21	**86.78**	87.50
	highest reg.	88.46	86.49	87.48
	lowest reg.	88.85	87.41	88.13
EN	U35	87.92	88.23	88.08
	O45	**88.26**	**88.40**	88.33
	male	88.19	88.55	88.37
	female	87.97	88.08	88.03
	highest reg.	88.27	88.57	88.42
	lowest reg.	88.24	88.52	88.38

Table 1: POS accuracy on different demographic groups for English and German. Significant differences per tagger in bold

Table 1 shows the accuracies for both algorithms on the three demographic groups (age, gender, region) for German and English. We see that there are some consistent differences between the groups. In both languages, results for both taggers are better for the older group than for the younger one. In three out of the four cases, this difference is statistically significant at $p < 0.05$, according to a bootstrap-sample test. The difference between the genders is less pronounced, although we do see CRF++ reaching a significantly higher accuracy for women in German. For regions, we find that while the models assign low log-likelihood scores to some regions, this is not reflected in the accuracy.

As common in NLP, we treat American (training) and British English (test data) as variants. It is possible that this introduces a confounding factor. However, since we do not see marked effects for gender or region, and since the English results closely track the German data, this seems unlikely. We plan to investigate this in future work.

4 Analysis

The last section showed the performance differences between various groups, but it does not tell us where the differences come from. In this section, we try to look into potential causes, and analyze the tagging errors for systematic patterns. We focus on age, since this variable showed the largest differences between groups.

Holmes (2013) argues that people between 30 and 55 years use standard language the most, because of societal pressure from their workplace. Nguyen et al. (2014) made similar observations for Twitter. Consequently, both young and retired people often depart from the standard linguistic norms, young people because of innovation, older people because of adherence to previous norms. Our data suggests, however, that young people do so in ways that are more challenging for off-the-shelf NLP models induced on age-biased data. But what exactly are the linguistic differences that lead to lower performance for this group?

The obvious cause for the difference between age groups would be *lexical* change, i.e., the use of neologisms, spelling variation, or linguistic change at the structural level in the younger group. The resulting vocabulary differences between age groups would result in an increased out-of-vocabulary (OOV) rate in the younger group, which in turn negatively affects model performance.

While we do observe an unsurprising correlation between sentence-level performance and OOV-rate, the young reviewers in our sample do *not* use OOV words more often than the older age group. Both groups differ from the training data roughly equally. This strongly suggests that age-related differences in performance are *not* a result of OOV items.

In order to investigate whether the differences extend beyond the vocabulary, we compare the *tag* bigram distributions, both between the two age groups and between each group and the training data. We measure similarity by KL divergence between the distributions, and inspect the 10 tag bigrams which are most prevalent for either group. We use Laplace smoothing to

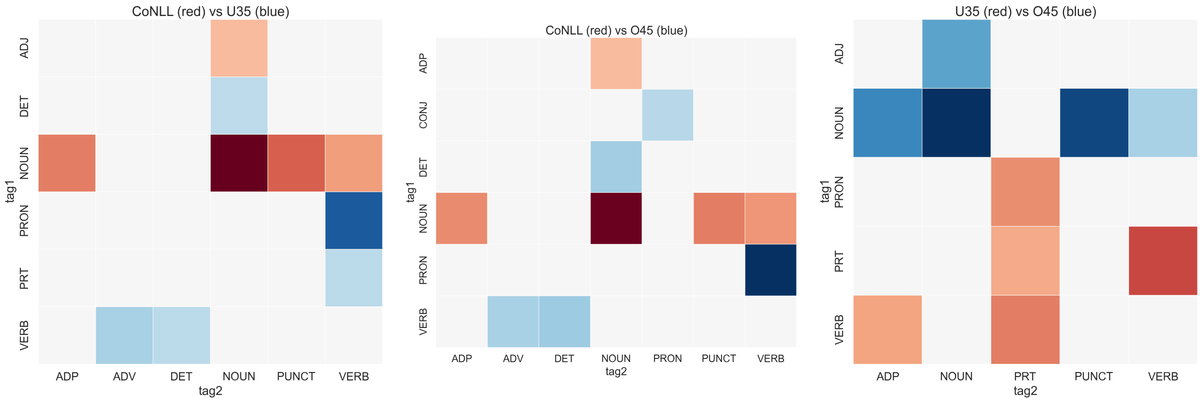

Figure 1: Tag bigrams with highest differences between distributions in English data.

account for missing bigrams and ensure a proper distribution.

For the English age groups, we find that a) the two Trustpilot data sets have a smaller KL divergence with respect to each other ($1.86e-6$) than either has with the training data (young: $3.24e-5$, old.: $2.36e-5$, respectively). We do note however, b), that the KL divergence for the older groups is much smaller than for the younger group. This means that there is a cross-domain effect, which is bigger, measured this way, than the difference in age groups. The age group difference in KL divergence, however, suggests that the two groups use different syntactic constructions.

Inspecting the bigrams which are most prevalent for each group, we find again that a) the Trustpilot groups show more instances involving verbs, such as PRON–VERB, VERB–ADV, and VERB–DET, while the English Penn Treebank data set has a larger proportion of instances of nominal constructions, such as NOUN–VERB, NOUN–ADP, and NOUN–NOUN.

On the other hand, we find that b) the younger group has more cases of verbal constructions and the use of particles, such as PRT–VERB, VERB–PRT, PRON–PRT, and VERB–ADP, while the older group–similar to the treebank–shows more instances of nominal constructions, i.e., again NOUN–VERB, ADJ–NOUN, NOUN–ADP, and NOUN–NOUN.

The heatmaps in Figure 1 show all pairwise comparisons between the three distributions. In the interest of space and visibility, we select the 10 bigrams that differ most from each other between the two distributions under comparison. The color indicates in which of the two distributions a bigram is more prevalent, and the degree of shading indicates the size of the difference.

For German, we see similar patterns. The Trustpilot data shows more instances of ADV–ADV, PRON–VERB, and ADV–VERB, while the TIGER treebank contains more NOUN–DET, NOUN–ADP, and NOUN–NOUN.

Again, the younger group is more dissimilar to the CoNLL data, but less so than for English, with CONJ–PRON, NOUN–VERB, VERB–VERB, and PRON–DET, while the older group shows more ADV–ADJ, ADP–NOUN, NOUN–ADV, and ADJ–NOUN.

In all of these cases, vocabulary does *not* factor into the differences, since we are at the POS level. The results indicate that there exist fundamental *grammatical* differences between the age groups, which go well beyond mere lexical differences. These findings are in line with the results in Johannsen et al. (2015), who showed that entire (delexicalized) dependency structures correlate with age and gender, often across several languages.

4.1 Tagging Error Analysis

Analyzing the tagging errors of our model can give us an insight into the constructions that differ most between groups.

In German, most of the errors in the younger group occur with adverbs, determiners, and verbs. Adverbs are often confused with adjectives, because adverbs and adjectives are used as modifiers in similar ways. The taggers also frequently confused adverbs with nouns, especially sentence-initially, presumably largely because they are capitalized. Sometimes, such errors are also due to spelling mistakes and/or English loanwords. Determiners are often incorrectly predicted to be pronouns, presumably due to homography: in German, *der*, *die*, *das*, etc. can be used as determiners, but also as relative pronouns, depending on the position. Verbs are often incorrectly predicted to be nouns. This last error is again mostly due to capitalization, homographs, and, again, English loanwords. Another interesting source is sentence-initial use of verbs, which is unusual in canonical German declarative sentences, but common in informal language, where pronouns are dropped, i.e, "[Ich] Kann mich nicht beschweren" (*[I] Can't complain*).

Errors involving verbs are much less frequent in the older group, where errors with adjectives and nouns are more frequent.

For English, the errors in the younger and older group are mostly on the same tags (nouns, adjectives, and verbs). Nouns often get mis-tagged as VERB, usually because of homography due to null-conversion (*ordering*, *face*, *needs*). Adjectives are also most commonly mis-tagged as VERB, almost entirely due to homography in participles (*–ed*, *–ing*). We see more emoticons (labeled X) in the younger group, and some of them end up with incorrect tags (NOUN or ADV). There are no mis-tagged emoticons in the older group, who generally uses fewer emoticons (see also Hovy et al. (2015a)).

5 Conclusion

In this position paper, we show that some of the common training data sets bias NLP tools towards the language of older people. I.e., there is a statistically significant correlation between tagging performance and age for models trained on CoNLL data. A study of the actual differences between age groups shows that they go beyond the vocabulary, and extend to the grammatical level.

The results suggest that NLP's focus on a limited set of training data has serious consequences for model performance on new data sets, but also demographic groups. Due to language dynamics and the age of the data sets, performance degrades significantly for younger speakers. Since POS tagging is often the first step in any NLP pipeline, performance differences are likely to increase downstream. As a result, we risk disadvantaging younger groups when it comes to the benefits of NLP.

The case study shows that our models are susceptible to the effects of language change and demographic factors. Luckily, the biases are not *inherent* to the models, but reside mostly in the data. The problem can thus mostly be addressed with more thorough training data selection that takes demographic factors into account. It does highlight, however, that we also need to develop more robust technologies that are less susceptible to data biases.

Acknowledgements

We would like to thank the anonymous reviewers for their comments, which helped improve the paper. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Federica Barbieri. 2008. Patterns of age-based linguistic variation in American English. *Journal of sociolinguistics*, 12(1):58–88.

Andrew J Barke. 2000. The Effect of Age on the Style of Discourse among Japanese Women. In *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, pages 23–34.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NIPS*.

Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. 2009. Extracting discriminative concepts for domain adaptation in text mining. In *KDD*.

Minmin Chen, Killiang Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. In *NIPS*.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca, and Fernando Pereira. 2007. Frustratingly Hard Domain Adaptation for Dependency Parsing. In *EMNLP-CoNLL*.

Janet Holmes. 2013. *An introduction to sociolinguistics*. Routledge.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015a. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*.

Dirk Hovy, Barbara Plank, Héctor Martínez Alonso, and Anders Søgaard. 2015b. Mining for unambiguous instances to adapt pos taggers to new domains. In *Proceedings of NAACL-HLT*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Workshop on Noisy User-generated Text (W-NUT)*.

Dong Nguyen, Dolf Trieschnigg, A. Seza Dogruöz, Rilana Gravel, Mariet Theune, Theo Meder, and Franciska De Jong. 2014. Predicting Author Gender and Age from Tweets: Sociolinguistic Theories and Crowd Wisdom. In *Proceedings of COLING 2014*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.

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

John Rickford and Mackenzie Price. 2013. Girlz ii women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics*, 17(2):143–179.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, pages 1815–1827.