

Robustness to Capitalization Errors in Named Entity Recognition

Sravan Bodapati

Amazon.com

sravanb@amazon.com

Hyokun Yun

Amazon.com

yunhyoku@amazon.com

Yaser Al-Onaizan

Amazon.com

onaizan@amazon.com

Abstract

Robustness to capitalization errors is a highly desirable characteristic of named entity recognizers, yet we find standard models for the task are surprisingly brittle to such noise. Existing methods to improve robustness to the noise completely discard given orthographic information, which significantly degrades their performance on well-formed text. We propose a simple alternative approach based on data augmentation, which allows the model to *learn* to utilize or ignore orthographic information depending on its usefulness in the context. It achieves competitive robustness to capitalization errors while making negligible compromise to its performance on well-formed text and significantly improving generalization power on noisy user-generated text. Our experiments clearly and consistently validate our claim across different types of machine learning models, languages, and dataset sizes.

1 Introduction

In the last two decades, substantial progress has been made on the task of named entity recognition (NER), as it has enjoyed the development of probabilistic modeling (Lafferty et al., 2001; Finkel et al., 2005), methodology (Ratinov and Roth, 2009), deep learning (Collobert et al., 2011; Huang et al., 2015; Lample et al., 2016) as well as semi-supervised learning (Peters et al., 2017, 2018). Evaluation of these developments, however, has been mostly focused on their impact on global average metrics, most notably the micro-averaged F1 score (Chinchor, 1992).

For practical applications of NER, however, there can be other considerations for model evaluation. While standard training data for the task consists mainly of well-formed text (Tjong Kim Sang, 2002; Pradhan and Xue, 2009),

models trained on such data are often applied on a broad range of domains and genres by users who are not necessarily NLP experts, thanks to the proliferation of toolkits (Manning et al., 2014) and general-purpose machine learning services. Therefore, there is an increasing demand for the strong robustness of models to unexpected noise.

In this paper, we tackle one of the most common types of noise in applications of NER: unreliable capitalization. Noisiness in capitalization is a typical characteristic of user-generated text (Ritter et al., 2011; Baldwin et al., 2015), but it is not uncommon even in formal text. Headings, legal documents, or emphasized sentences are often capitalized. All-lowercased text, on the other hand, can be produced in large scale from upstream machine learning models such as speech recognizers and machine translators (Kubala et al., 1998), or processing steps in the data pipeline which are not fully under the control of the practitioner. Although a text without correct capitalization is perfectly legible for human readers (Cattell, 1886; Rayner, 1975) with only a minor impact on the reading speed (Tinker and Paterson, 1928; Arditì and Cho, 2007), we show that typical NER models are surprisingly brittle to all-uppercasing or all-lowercasing of text. The lack of robustness these models show to such common types of noise makes them unreliable, especially when characteristics of target text are not known a priori.

There are two standard treatments on the problem in the literature. The first is to train a case-agnostic model (Kubala et al., 1998; Robinson et al., 1999), and the second is to explicitly correct the capitalization (Srihari et al., 2003; Lita et al., 2003; Ritter et al., 2011). One of the main contributions of this paper is to empirically evaluate the effectiveness of these techniques across models, languages, and dataset sizes. How-

| Annotation | O | O | O | B-ORG | I-ORG | E-ORG |
|--------------------------|---|------|----|-------|-------|-------|
| (a) Original Sentence | I | live | in | New | York | City |
| (b) Lower-cased Sentence | i | live | in | new | york | city |
| (c) Upper-cased Sentence | I | LIVE | IN | NEW | YORK | CITY |

Table 1: Example of Data Augmentation

ever, both approaches have clear conceptual limitations. Case-agnostic models discard orthographic information (how the given text was capitalized), which is considered to be highly useful (Robinson et al., 1999); our experimental results also support this. The second approach of correcting the capitalization of the text, on the other hand, requires an access to a high-quality truecasing model, and errors from the truecasing model would cascade to final named entity predictions.

We argue that an ideal approach should take a full advantage of orthographic information when it is correctly present, but rather than assuming the information to be always perfect, the model should be able to *learn* to ignore the orthographic information when it is unreliable. To this end, we propose a novel approach based on data augmentation (Simard et al., 2003). In computer vision, data augmentation is a highly successful standard technique (Krizhevsky et al., 2012), and it has found adoptions in natural language processing tasks such as text classification (Zhang and LeCun, 2015), question-answering (Yu et al., 2018) and low-resource learning (Sahin and Steedman, 2018). Consistently across a wide range of models (linear models, deep learning models to deep contextualized models), languages (English, German, Dutch, and Spanish), and dataset sizes (CoNLL 2003 and OntoNotes 5.0), the proposed method shows strong robustness while making little compromise to the performance on well-formed text.

2 Formulation

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be a sequence of words in a sentence. We follow the standard approach of formulating NER as a sequence tagging task (Rabiner, 1989; Lafferty et al., 2001; Collins, 2002). That is, we predict a sequence of tags $\mathbf{y} = (y_1, y_2, \dots, y_n)$ where each y_i identifies the type of the entity the word x_i belongs to, as well as the position of it in the surface form according to IOBES scheme (Uchimoto et al., 2000). See Table 1 (a) for an example annotated sentence. We train probabilistic models under the maximum

likelihood principle, which produce a probability score $\mathbb{P}[\mathbf{y} | \mathbf{x}]$ for any possible output sequence \mathbf{y} .

All-uppercasing and all-lowercasing are common types of capitalization errors. Let $\text{upper}(x_i)$ and $\text{lower}(x_i)$ be functions that lower-cases and upper-cases the word x_i , respectively. Robustness of a probabilistic model to these types of noise can be understood as the quality of scoring function $\mathbb{P}[\mathbf{y} | \text{upper}(x_1), \dots, \text{upper}(x_n)]$ and $\mathbb{P}[\mathbf{y} | \text{lower}(x_1), \dots, \text{lower}(x_n)]$ in predicting the correct annotation \mathbf{y} , which can still be quantified with standard evaluation metrics such as the micro-F1 score.

3 Prior Work

There are two common strategies to improve robustness to capitalization errors. The first is to completely ignore orthographic information by using case-agnostic models (Kubala et al., 1998; Robinson et al., 1999). For linear models, this can be achieved by restricting the choice of features to case-agnostic ones. On the other hand, deep learning models without hand-curated features (Lample et al., 2016; Chiu and Nichols, 2016) can be easily made case-agnostic by lower-casing every input to the model. The second strategy is to explicitly correct the capitalization by using another model trained for this purpose, which is called “truecasing” (Srihari et al., 2003; Lita et al., 2003). Both methods, however, have the common limitation that they discard orthographic information in the target text, which can be correct; this leads to degradation of performance on well-formed text.

4 Data Augmentation

Data augmentation refers to a technique of increasing the size of training data by adding label-preserving transformations of them (Simard et al., 2003). For example, in image classification, an object inside of an image does not change if the image is rotated, translated, or slightly skewed; most people would still recognize the same object they would find in the original image. By training a model on transformed versions of training

| Model | Method | CoNLL-2003 English | | | OntoNotes 5.0 English | | | Transfer to Twitter | | |
|--------|------------|--------------------|-------|-------|-----------------------|-------|-------|---------------------|-------|-------|
| | | Original | Lower | Upper | Original | Lower | Upper | Original | Lower | Upper |
| Linear | Baseline | 89.2 | 57.8 | 75.2 | 81.7 | 37.4 | 15.1 | 24.4 | 6.9 | 20.2 |
| | Caseless | 83.7 | 83.7 | 83.7 | 75.5 | 75.5 | 75.5 | 20.3 | 20.3 | 20.3 |
| | Truecasing | 83.8 | 83.8 | 83.8 | 76.6 | 76.6 | 76.6 | 24.0 | 24.0 | 24.0 |
| | DA | 88.2 | 85.6 | 86.1 | - | - | - | 28.2 | 26.4 | 27.0 |
| BiLSTM | Baseline | 90.8 | 0.4 | 52.3 | 87.6 | 38.9 | 15.5 | 18.1 | 0.1 | 7.9 |
| | Caseless | 85.7 | 85.7 | 85.7 | 83.2 | 83.2 | 83.2 | 20.3 | 20.3 | 20.3 |
| | Truecasing | 84.6 | 84.6 | 84.6 | 81.7 | 81.7 | 81.7 | 18.7 | 18.7 | 18.7 |
| | DA | 90.4 | 85.3 | 83.8 | 87.5 | 83.2 | 82.6 | 21.2 | 17.7 | 18.4 |
| ELMo | Baseline | 92.0 | 34.8 | 71.6 | 88.7 | 66.6 | 48.9 | 31.6 | 1.5 | 19.6 |
| | Caseless | 89.1 | 89.1 | 89.1 | 85.3 | 85.3 | 85.3 | 31.8 | 31.8 | 31.8 |
| | Truecasing | 86.2 | 86.2 | 86.2 | 83.2 | 83.2 | 83.2 | 28.8 | 28.8 | 28.8 |
| | DA | 91.3 | 88.7 | 87.9 | 88.3 | 85.8 | 83.6 | 34.6 | 31.7 | 30.2 |

Table 2: F1 scores on original, lower-cased, and upper-cased test sets of English Datasets. Stanford Core NLP could not be trained on the augmented dataset even with 512GB of RAM.

images, the model becomes invariant to the transformations used (Krizhevsky et al., 2012).

In order to improve the robustness of NER models to capitalization errors, we appeal to the same idea. When a sentence is all-lowercased or all-uppercased as in Table 1 (b) and (c), each word would still correspond to the same entity. This implies such transformations are also label-preserving ones: for a sentence x and its ground-truth annotation y , y would still be a correct annotation for the all-uppercased sentence ($\text{upper}(x_1), \dots, \text{upper}(x_n)$) as well as the all-lowercased version ($\text{lower}(x_1), \dots, \text{lower}(x_n)$). Indeed, all three sentences (a), (b) and (c) in Table 1 would share the same annotation.

5 Experiments

We consider following three models, each of which is state-of-the-art in their respective group: **Linear**: Linear CRF model (Finkel et al., 2005) from Stanford Core NLP (Manning et al., 2014), which is representative of feature engineering approaches. **BiLSTM**: Deep learning model from Lample et al. (2016) which uses bidirectional LSTM for both character-level encoder and word-level encoder with CRF loss. This is the state-of-the-art supervised deep learning approach (Reimers and Gurevych, 2017). **ELMo**: Bidirectional LSTM-CRF model which uses contextualized features from deep bidirectional LSTM language model (Peters et al., 2018). For all models, we used hyperparameters from original papers.

We compare four strategies: **Baseline**: Models are trained on unmodified training data. **Caseless**: We lower-case input data both at the training time and at the test time. **Truecasing**: Models are still

trained on unmodified training data, but every input to test data is “truecased” (Lita et al., 2003) using CRF truecasing model from Stanford Core NLP (Manning et al., 2014), which ignores given orthographic information in the text. Due to the lack of access to truecasing models in other languages, this strategy was used only on English. **DA (Data Augmentation)**: We augment the original training set with upper-cased and lower-cased versions of it, as discussed in Section 4.

We evaluate these models and methods on three versions of the test set for each dataset: **Original**: Original test data. **Upper**: All words are upper-cased. **Lower**: All words are lower-cased. Note that both Caseless and Truecasing method perform equally on all three versions because they ignore any original orthographic information in the *test* dataset. We focus on micro-averaged F1 scores.

We use CoNLL-2002 Spanish and Dutch (Tjong Kim Sang, 2002) and CoNLL-2003 English and German (Sang and De Meulder, 2003) to cover four languages, all of which orthographic information is useful in identifying named entities, and upper or lower-casing of text is straightforward. We additionally evaluate on OntoNotes 5.0 English (Pradhan and Xue, 2009), which is about five times larger than CoNLL datasets and contains more diverse genres. F1 scores are shown in Table 2 and 3.

Question 1: How robust are NER models to capitalization errors? Models trained with the standard Baseline strategy suffer from significant loss of performance when the test sentence is upper/lower-cased (compare ‘Original’ column with ‘Lower’ and ‘Upper’). For example, F1 score of BiLSTM on lower-cased CoNLL-2003 English is abysmal 0.4%, completely losing any predic-

| Model | Method | CoNLL-2002 Spanish | | | CoNLL-2002 Dutch | | | CoNLL-2003 German | | |
|--------|----------|--------------------|-------|-------|------------------|-------|-------|-------------------|-------|-------|
| | | Original | Lower | Upper | Original | Lower | Upper | Original | Lower | Upper |
| Linear | Baseline | 80.7 | 1.1 | 22.1 | 79.1 | 9.8 | 9.7 | 68.4 | 11.8 | 11.3 |
| | Caseless | 69.9 | 69.9 | 69.9 | 63.9 | 63.9 | 63.9 | 53.3 | 53.3 | 53.3 |
| | DA | 77.3 | 70.9 | 73.2 | 74.4 | 68.5 | 68.5 | 61.8 | 57.8 | 62.8 |
| BiLSTM | Baseline | 85.4 | 1.0 | 26.8 | 87.3 | 2.0 | 15.8 | 79.5 | 6.5 | 9.8 |
| | Caseless | 77.8 | 77.8 | 77.8 | 77.7 | 77.7 | 77.7 | 69.8 | 69.8 | 69.8 |
| | DA | 85.3 | 78.4 | 76.5 | 84.8 | 75.0 | 75.9 | 76.8 | 69.7 | 69.7 |

Table 3: F1 scores on original, lower-cased, and upper-cased test sets of Non-English Datasets

tive power. Linear and ELMo are more robust than BiLSTM thanks to smaller capacity and semi-supervision respectively, but the degradation is still strong, ranging 20pp to 60pp loss in F1.

Question 2: How effective Caseless, Truecasing, and Data Augmentation approaches are in improving robustness of models? All methods show similar levels of performance on lower-cased or upper-cased text. Since Caseless and Data Augmentation strategy do not require additional language-specific resource as truecasing does, they seem to be superior to the truecasing approach, at least on CoNLL-2003 English and OntoNotes 5.0 datasets with the particular truecasing model used. Across all datasets, the performance of Linear model on lower-cased or upper-cased test set is consistently enhanced with data augmentation, compared with caseless models.

Question 3: How much performance on well-formed text is sacrificed due to robustness? Caseless and Truecasing methods are perfectly robust to capitalization errors, but only at the cost of significant degradation on well-formed text: caseless and truecasing strategy lose 5.1pp and 6.2pp respectively on the original test set of CoNLL-2003 English compared to Baseline strategy, and on non-English datasets the drop is even bigger. On the other hand, data augmentation preserves most of the performance on the original test set: with BiLSTM, its F1 score drops by only 0.4pp and 0.1pp respectively on CoNLL-2003 and OntoNotes 5.0 English. On non-English datasets, the drop is bigger (0.1pp on Spanish but 2.5pp on Dutch and 2.7pp on German) but still data augmentation performs about 7pp higher than Caseless on original well-formed text across languages.

Question 4: How do models trained on well-formed text generalize to noisy user-generated text? The robustness of models is especially important when the characteristics of target text are not known at the training time and can deviate significantly from those of training data. To

this end, we trained models on CoNLL 2003-English, and evaluated them on annotations of Twitter data from Fromreide et al. (2014), which exhibits natural errors of capitalization common in user-generated text. ‘Transfer to Twitter’ column of Table 2 reports results. In this experiment, Data Augmentation approach consistently and significantly improves upon Baseline strategy by 3.8pp, 3.1pp, and 3.0pp with Linear, BiLSTM, and ELMo models respectively on Original test set of Twitter, demonstrating much strengthened generalization power when the test data is noisier than the training data.

In order to understand the results, we examined some samples from the dataset. Indeed, on a sentence like ‘OHIO IS STUPID I HATE IT’, BiLSTM model trained with Baseline strategy was unable to identify ‘OHIO’ as a location although the state is mentioned fifteen times in the training dataset of CoNLL 2003-English as ‘Ohio’. BiLSTM models trained with all other strategies correctly identified the state. On the other hand, on another sample sentence ‘Someone come with me to Raging Waters on Monday’, BiLSTM models from Baseline and Data Augmentation strategies were able to correctly identify ‘Raging Waters’ as a location thanks to the proper capitalization, while the model from Caseless strategy failed on the entity due to its ignorance of orthographic information.

6 Conclusion

We proposed a data augmentation strategy for improving robustness of NER models to capitalization errors. Compared to previous methods, data augmentation provides competitive robustness while not sacrificing its performance on well-formed text, and improving generalization to noisy text. This is consistently observed across models, languages, and dataset sizes. Also, data augmentation does not require additional language-specific resource, and is trivial to implement for

many natural languages. Therefore, we recommend to use data augmentation by default for training NER models, especially when characteristics of test data are little known a priori.

References

- Aries Arditi and Jianna Cho. 2007. Letter case and text legibility in normal and low vision. *Vision research*, 47(19):2499–2505.
- Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.
- James McKeen Cattell. 1886. The time it takes to see and name objects. *Mind*, 11(41):63–65.
- Nancy Chinchor. 1992. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, pages 22–29. Association for Computational Linguistics.
- Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.
- Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.
- Hege Fromreide, Dirk Hovy, and Anders Sjøgaard. 2014. Crowdsourcing and annotating ner for twitter# drift. In *LREC*, pages 2544–2547.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. Named entity extraction from speech. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292. Citeseer.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.
- Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. Truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 152–159. Association for Computational Linguistics.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- Matthew E Peters, Waleed Ammar, Chandra Bhagavathula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Sameer S Pradhan and Nianwen Xue. 2009. Ontonotes: The 90% solution. In *HLT-NAACL (Tutorial Abstracts)*, pages 11–12.
- Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.
- Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1):65–81.
- Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.
- Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental

- study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.
- Patricia Robinson, Erica Brown, John Burger, Nancy Chinchor, Aaron Douthat, Lisa Ferro, and Lynette Hirschman. 1999. Overview: Information extraction from broadcast news. In *Proceedings of DARPA Broadcast News Workshop*, pages 27–30.
- Gozde Gul Sahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009.
- Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Patrice Y Simard, Dave Steinkraus, and John C Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE.
- Rohini K Srihari, Cheng Niu, Wei Li, and Jihong Ding. 2003. A case restoration approach to named entity tagging in degraded documents. In *null*, page 720. IEEE.
- Miles A Tinker and Donald G Paterson. 1928. Influence of type form on speed of reading. *Journal of Applied Psychology*, 12(4):359.
- Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.
- Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. 2000. Named entity extraction based on a maximum entropy model and transformation rules. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 326–335. Association for Computational Linguistics.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.
- Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.