Context-Sensitive Recognition for Emerging and Rare Entities

Jake Ryland Williams and **Giovanni C. Santia**
College of Computing and Informatics, Drexel University
30 North 33rd Street, Philadelphia PA, 19104
{jw3477,gs495}@drexel.edu

Abstract

We present a novel named entity recognition (NER) system, and its participation in the Emerging and Rare Entity Recognition shared task, hosted at the 2017 EMNLP Workshop on Noisy User Generated Text (W-NUT). With a specialized evaluation highlighting performance on rare, and sparsely-occurring named entities, this task provided an excellent opportunity to build out a newly-developed statistical algorithm and benchmark it against the state-of-the-art. Powered by flexible context features of word forms, our system's capacity for identifying never-before-seen entities made it well suited for the task. Since the system was only developed to recognize a limited number of named entity types, its performance was lower overall. However, performance was competitive on the categories trained, indicating potential for future development.

1 Introduction

NER is a common foundational step for many pipelines that rely on natural language processing (NLP). The main goal is the identification of mentions of entities (e.g., persons or locations). As a pre-processing task for unstructured text, NER may, for example, provide index keywords for information retrieval systems (Tjong Kim Sang and De Meulder, 2003), or topic-rich features for machine learning (ML) applications (Kumaran and Allan, 2004; Vavliakis et al., 2013). Effective approaches to NER have long utilized conditional random fields (Lafferty et al., 2001), support vector machines (McCallum and Li, 2003), and perceptrons (Settles, 2004; Ju et al., 2011; Luo et al., 2015). In addition to relying on face-value, gold-standard data, systems may benefit from a variety of other data representations and sources (Strauss et al., 2016), including gazetteers, word classes (e.g, Brown clusters), orthographic features, and grammatical relations between types of words, such as part of speech. Large-scale annotated resources for NER have also been developed in semi-supervised fashions, constructed from online encyclopedias (Nothman et al., 2008, 2012) and refined by crowdsourcing (Bos et al., 2017).

While NER systems have been in development for some time, their applicability to noisy-text domains (i.e., unedited, user-generated content) is somewhat limited. This is a multi-faceted problem (Derczynski et al., 2015), involving grammatical inconsistency and rapidly-shifting domains, requiring specialized algorithms. While progress has been made through annotation and specialized systems development (Ritter et al., 2011), there are still large gains to be made for this domain (Augenstein et al., 2017), which is highlighted well by both the shared task at the W-NUT this year (Strauss et al., 2016), and that of the previous year.

Adaptation to the task domain's wide-range of writing styles and abundant grammatical inconsistencies presents the need for algorithmic flexibility. These properties make precision loss an issue, and the presence of rare and emerging entities makes recall an extreme challenge, too. Our participation in the present shared task relies on a novel approach: utilizing flexible "contexts" as features - derived from token forms - alone. We rely upon these features for their capacity to relate to never-before-seen tokens as potential entities, and incorporate them into a statistical model that can handle both gold-standard data and large, lexical resources.

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 172–176
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics

2 Approach

2.1 Shared Task Data

We began our approach by scoping the task data set composition. There were 6 named entity types: corporation, creative work, group, location, person, and product, which were a mapping down from 10 in the 2016 W-NUT Twitter NER Shared Task. A decomposition of the current shared task data (see Tab. 1) exhibits several important features. The proportion of unique entities out of all increased from about 80% to 90% from the training to the development and test sets. However, the training, development, and test sets all exhibited internal stability in the proportions of unique numbers for each type of named entity. In other words, no named entity type dropped out of proportion when considering unique forms. However, the focus on rare entities resulted in large increases in the percentage of the data occupied by the person category. These proportions and the availability of large-scale gazetteer data highlighted this type for the initial focus of our model's development.

2.2 System Design

2.2.1 Previous Work

Context models are conditional statistical models whose features are derived from the structural patterns surrounding or within written language. We refer to context models that rely on exterior information as **external** context models, and those that rely on interior information as **internal** context models. For example, word-level context models applied to the text: "Out to lunch in New York City." might place the entity "New York City" in the external context "Out to lunch in *.", or the internal context "New York *" (in each case reserving * as a wildcard).

Context models trace their roots to Shannon (1948), but have likewise seen recent attention (Piantadosi et al., 2011). They have been applied to both patterns of character appearance and word appearance, with the majority of attention directed towards word patterns and external models. In recent work by Williams et al. (2015a), an internal context model was used to identify missing multi-word dictionary entries. We utilize this model here, but apply it at the character level so as to be able to identify both single snd multi-word named entities.

2.2.2 Context-Sensitive NER

We represent a token, w, by its sequence of n characters:

$$w = (l_1, l_2, \cdots, l_n),$$

and define its set of 2^{n-1} contexts, C_w by the corresponding removal patterns of contiguous subsequences. The context, $c_{i\cdots j} \in C_w$, defined by the removal of characters i through j is:

$$c_{i\cdots j} = (l_1, \cdots, l_{i-1}, *, \cdots, *, l_{j+1}, \cdots, l_n).$$

Despite execution at the sub-word level, this is precisely the same construction as in Williams et al. (2015a), which was used to compute likelihoods of dictionary definition.

For a given word, weighting across its contexts is accomplished as in Williams et al. (2015a), induced by a partition process (Williams et al., 2015b). However instead of dictionary definition, we use the context conditional probabilities to determine the likelihoods of named entity tags. For any word, w, and positive tag, t (e.g., B-location, I-person, B-group, etc.), a computed likelihood, $L(t|C_w)$, can be interpreted as "the likelihood of drawing a t-tagged word from the contexts of w". Note that these likelihoods can be non-zero for words that were not present in training, and are higher for words that are similar to tagged words. For example, if w_1 = *Larry*, w_2 = *Harry*, and only w_1 appeared in a gold standard, with tag $t =$ B-person, $L(t|C_{w_2})$ would be elevated.

2.2.3 Entity Recognition

To handle entities composed of multiple words, e.g., $(w_1, w_2, \cdots, w_k)$, we assess a potential entity's membership to a particular type, e.g., "location", via the harmonic mean, $\overline{L}(t_1, t_2, \cdots, t_k | w_1, w_2, \cdots, w_k)$, of their component-word likelihood values, such that only the first word has the B-version tag (t_1) and all others have the I-version. A candidate is accepted if its likelihood mean is above a thresholds value, which is determined in optimization (see Sec. 4).

2.2.4 Conflict Resolution

A given word may fall within multiple predicted entities, both of different types and lengths. To resolve potential conflicts between predicted entities we establish precedence by accepting 1) predictions appearing first, over 2) longer predictions, over 3) predictions of higher likelihood.

Category	Training		Development		Test	
	Total (%)	Unique (%)	Total (%)	Unique (%)	Total (%)	Unique (%)
Corporation	221 (11.19)	140 (8.73)	34 (4.07)	32 (4.29)	69 (6.63)	63 (6.82)
Creative Work	140 (7.09)	127 (7.92)	105 (12.57)	101 (13.54)	141 (13.54)	135 (14.61)
Group	264 (13.37)	231 (14.4)	39 (4.67)	38 (5.09)	151 (14.51)	131 (14.18)
Location	538 (27.75)	434 (27.06)	74 (8.86)	68 (9.12)	138 (13.26)	114 (12.34)
Person	660 (33.42)	546 (34.04)	469 (56.17)	398 (53.35)	441 (39.77)	363 (39.29)
Product	142 (7.19)	126 (7.86)	114 (13.65)	109 (14.61)	128 (12.3)	118 (12.77)
All	1975	1604	835	746	1041	924

Table 1: Description of shared-task data. Each of the **Training**, **Development**, and **Test** data are broken down by types of named entities (**Corporation** , **Creative Work**, **Group**, **Location**, **Person**, and **Product**), with counts and percents for the **Unique** and **Total** named entity forms present, in addition to total numbers of **All** named entities present.

3 Materials

3.1 Gold-Standard Data

In addition to the gold-standard data provided for the shared task (see Sec. 2.1 and Tab. 1) we utilize 1) all components of the W-NUT 2016 Twitter NER shared task (Strauss et al., 2016), 2) all components of the 2003 CONLL NER shared task (Tjong Kim Sang and De Meulder, 2003), 3) the WikiNER annotations (Nothman et al., 2008, 2012), and 4) the Groningen Meaning Bank (Bos et al., 2017). Each corpus required mapping its entity types to the six 2017 shared task types, and for data sets (2), (3), and (4), only mappings for the location and person types were deemed appropriate (geo-loc, facility, and loc to location, and per to person). However for data set (1), additional mappings were accepted from tvshow and movie to creative-work, sportsteam to group, and company to corporation.

3.2 Supplemental Lexica

To extend model training to as many forms as possible, supplemental lexica were incorporated from the gazetteer materials provided alongside the gold data from the W-NUT 2016 Twitter NER shared task. Only several gazetteers were incorporated into the final model: automotive.model and business.consumer_product for the product type; firstname.5k, lastname.5000, people.family_name, and people.person.filtered for the person type; and location.country for the location type. Each entry in a given gazetteer was treated as a weighted instance of its named entity type. Weights offset the extreme size of gazetteers in comparison to the gold standard data, and were determined as follows. For a given entity type, let x be the number of typed named entities in the gold standard training data, and y be the number of gazetteer entries. The type's gazetteer entries were then incorporated with weight x/y, and all O-tagged tokens were counted with weight 2.

4 Optimization

Model development consisted of training on the gold-standard training data (see Sec. 2.1), in addition to the external gold standards (see Sec. 3.1), and the supplemental lexica (see Sec. 3.2). With the trained model, optimization was performed with respect to the development data set, which notably had a disproportionate representation of person entities. We determined thresholds for each of the entity types through separate optimizations. Given the brief timeline, these were conducted adaptively, optimizing thresholds for by-type F_1 values, honing in by step sizes of 0.1, 0.01, and finally 0.001. Note that the optimization procedure exhibited no predictive power on entity types creating work and corporation, leading us to restrain our model from predicting those types. After final threshold parameters were determined, a final combined model (see Sec. 2.2.4) was allowed to train additionally on the development data set before being applied to the final test data set.

5 Results

To understand our model's performance in the context of other systems, we provide a fine-grained system evaluation across the entity types (see Tab. 2). This follows the specialized shared-task evaluation method, focusing on precision, recall, and F_1 with respect to unique named entity surface forms. On the primary categories in which our model made predictions (location and person), our model's performance was reasonably competitive, with high levels of precision. At location, our system outperformed two other models by overall F_1, and was in range of the other models with respect to the person type. For all other entity types,

Category	Arcada	Drexel-CCI	FLYTXT	MIC-CIS	SJTU-Adapt	SpinningBytes	UH Ritual
Precision							
Corporation	20.37	0	25.00	12.86	26.67	10.59	38.89
Creative Work	37.21	0	33.33	23.64	60.00	26.03	35.71
Group	35.29	0	26.47	25.00	31.65	31.71	39.34
Location	39.04	58.21	33.33	36.21	34.48	61.05	52.34
Person	54.90	49.58	61.15	46.76	64.83	55.94	68.46
Product	25.00	25.00	16.67	18.03	21.15	16.67	28.21
All	43.93	50.64	43.52	36.82	46.36	45.33	55.18
Recall							
Corporation	32.83	0	12.70	14.29	19.05	14.29	22.22
Creative Work	11.85	0	12.59	9.63	2.22	14.07	7.41
Group	18.32	0	13.85	20.61	19.08	9.92	18.32
Location	49.57	33.91	44.35	54.78	52.17	50.43	48.70
Person	50.82	32.42	49.73	45.60	51.65	62.09	48.90
Product	9.32	0.85	6.78	9.32	9.32	4.24	9.32
All	32.83	17.06	30.45	31.21	32.29	35.64	31.64
F1							
Corporation	18.80	0	16.84	13.53	22.22	12.16	28.28
Creative Work	17.98	0	18.28	13.68	4.29	18.27	12.27
Group	24.12	0	17.09	26.87	23.81	15.12	25.00
Location	43.68	42.86	38.06	43.60	41.52	55.24	50.45
Person	52.78	39.2	54.85	46.18	57.49	58.85	57.05
Product	13.58	1.64	9.64	12.29	12.94	6.76	14.01
All	37.58	25.53	35.83	33.78	38.06.86	39.90	40.22

Table 2: Shared-task results. All precision, recall, and F_1 values are computed with respect to unique entity forms, in accordance with the task specific evaluation.

our system performed poorly (although no predictions were made for the corporation and creative work categories). Notably, the only categories at which other teams performed consistently well were the person and location categories, with the main observation being low recall, rarely above 20%.

6 Discussion

For this shared task we developed and evaluated a novel NER algorithm that relies only on features derived from word forms. Despite having the lowest task evaluation scores, this model exhibited competitive performance at two of the largest categories. These two categories (person and location) had significant external data availabile (both gold standards and supplemental lexica), and exhibited the most promise during model optimization. The system's ability to perform competitively at these entity types appears to suggest that increased performance at the other types may be possible with the availability of other, category-specific and large-scale external resources.

We note that our model's optimization exhibited an extreme lack of predictive power at the corporation and creative work categories, which, in addition to being affected by sparsity, may have also been affected by the lack of acceptable mappings from the external gold-standard resources into these categories. While lexical data were weighted to good effect (increased performance), the coverage of gold standard data only over the person and location entity types may have negatively impacted our system's ability to predict other types. Thus, a potential improvement for prediction of these types might be accomplished by applying a similar weighting scheme to the external gold-standard data. This leaves us with avenues for improvement, along with competitive, task-specific scores at the person and location categories; all of this, while relying on features derived only from word forms, points toward value in the continued development of context-sensitive NER for rare and emerging entities.

Acknowledgments

The authors thank the shared-task organizers for their efforts in running this event, the anonymous reviewers for their thoughtful comments, and greatfully acknowledge support from the Drexel University College of Computing and Informatics and Department of Information Science.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language* 44:61–83. https://doi.org/https://doi.org/10.1016/j.csl.2017.01.012.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. 2017. *The Groningen Meaning Bank*, Springer Netherlands, Dordrecht, pages 463–496. https://doi.org/10.1007/978-94-024-0881-2_18.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphal Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* 51(2):32 – 49. https://doi.org/http://dx.doi.org/10.1016/j.ipm.2014.10.006.

Zhenfei Ju, Jian Wang, and Fei Zhu. 2011. Named entity recognition from biomedical text using svm. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*. IEEE, pages 1–4.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '04, pages 297–304. https://doi.org/10.1145/1008992.1009044.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*. pages 282–289.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint named entity recognition and disambiguation. In *Proc. EMNLP*. pages 879–880.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. Association for Computational Linguistics, Stroudsburg, PA, USA, CONLL '03, pages 188–191. https://doi.org/10.3115/1119176.1119206.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *In Proceedings of the Australasian Language Technology Association Workshop 2008*. pages 124–132.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194:151–175. https://doi.org/10.1016/j.artint.2012.03.006.

S. T. Piantadosi, H. Tily, and E. Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9):3526. http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pages 1524–1534. http://dl.acm.org/citation.cfm?id=2145432.2145595.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*. Association for Computational Linguistics, Stroudsburg, PA, USA, JNLPBA '04, pages 104–107. http://dl.acm.org/citation.cfm?id=1567594.1567618.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 138–144. http://aclweb.org/anthology/W16-3919.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. Association for Computational Linguistics, Stroudsburg, PA, USA, CONLL '03, pages 142–147. https://doi.org/10.3115/1119176.1119195.

Konstantinos N. Vavliakis, Andreas L. Symeonidis, and Pericles A. Mitkas. 2013. Event identification in web social media through named entity recognition and topic modeling. *Data Knowl. Eng.* 88:1–24. https://doi.org/10.1016/j.datak.2013.08.006.

Jake Ryland Williams, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015a. Identifying missing dictionary entries with frequency-conserving context models. *Phys. Rev. E* 92:042808. https://doi.org/10.1103/PhysRevE.92.042808.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015b. Zipf's law holds for phrases, not words. *Nature Scientific Reports* 5:12209.