One Sense Per Document: Improve Name Finding in the Wild with Document-wide Context

Bonan Min
Raytheon BBN Technologies
Cambridge, MA
bonan.min@raytheon.com

Marjorie Freedman
USC/Information Sciences Institute
Marina del Rey, CA
mrf@isi.edu

Ryan Gabbard
USC/Information Sciences Institute
Marina del Rey, CA
gabbard@isi.edu

ABSTRACT

Named entity recognition is fundamental to information extraction, knowledge base construction and many other tasks. Trained on annotated newswire documents, state-of-the-art sentence-level taggers do not perform well on newer documents nor informal text such that found on discussion fora. Frequently, names are missed due to insufficient context. We hypothesize that finding them benefits from using the context of the same string in other sentences within the document. We propose a simple joint decoding algorithm which enforces document-level tag consistency on top of sentence-level tagging decisions. Experiments on the challenging TAC-KBP Cold Start Entity Discovery dataset show that the proposed method improves performance in both news and discussion forum text, across all three entity types. End-to-end entity clustering is also improved because of more names found.

1 INTRODUCTION

Named Entity Recognition (NER) is a fundamental component in applications such as answering questions and web search. Named entities also serve as inputs to downstream Information Extraction (IE) algorithms such as relation detection, entity linking and knowledge base construction.

Despite extensive research in NER, the accuracy of state-of-the-art name taggers degrades on informal genres (e.g. discussion forum) and on more recent articles [15]. Name taggers age [24]–trained with news articles published years ago, name taggers see far more unrecognized words when presented with today's news. Frequently, there aren't sufficient context to identify names in informal text (e.g. online posts, or all-capitalized text) [1].

State-of-the-art name taggers are sequence tagging models such as CRF [17] and bidirectional LSTMs with an additional CRF layer layer [18]. Trained with labeled datasets, a name tagger decodes a sequence of tags for each sentence independent of the other sentences. While they model local context (e.g. previous and next words) within the sentence, they do not model shared information across the document. The scheme fails to recognize names if there aren't strong clues from surface forms nor context. Here are a few motivating examples [2] from New York Times and discussion forum posts in the 2015 TAC-KBP [3] evaluation dataset:

Example 1: new names
S_1: **SpaceX** *Completes First Mission...*
S_2: *Space Exploration Technologies (***SpaceX***) successfully completed its first ... mission*

A previously-unseen ORG *SpaceX* is missed in S_1 because context is ambiguous for recognizing it and typing it as an ORG.

Example 2: informal text or confusing capitalization
S_1 (question): *Any tripsters in the area?*
S_2 (reply1): **syracuse** *here*
S_3 (reply2): *...visit* **Syracuse** *to ...*
...
S_1 (title): **RBS** *TO SPLIT OFF $61 BILLION IN LOANS INTO INTERNAL 'BAD BANK'*
S_2: *Shares in* **RBS** *closed down 7.5 percent*

In the example above, the uncapitalized *syracuse* in the first S_2 is a missed GPE, but the following S_3 provides more context and the capitalized string is correctly tagged. We see a similar pattern for the acronym *RBS*.

Example 3: nested GPE
S_1: *No, i'm leaving* **boston** *tomorrow.*
S_2: *those on the* **boston** *pd who aren't working, probably will be protesting...*

Recent work in NER has attempted to find name-internal names [7]. Frequently, GPEs are nested in another name (especially ORGs located in the GPE). Here *boston pd* (Boston Police Department) is a GPE *boston* inside an ORG. Seeing both strings offers more confidence for tagging the nested GPE name.

These examples show that more context of the name strings from the rest of the document is extremely helpful for tagging difficult names. It is orthogonal to the information captured by the local sentence-level tagger. Inspired by *one-sense-per-discourse* [9], we propose *one-sense-per-document* (a.k.a., assuming all mentions of each name string to have the same tag sequence). We propose an algorithm which finds all the above-mentioned missing names and improves over a strong sequence-tagging model significantly. Our contributions are three-fold:

- An algorithm that combines a sentence-level sequence tagger and a document-level name-string-based tagger to enforce tag consistency.
- Demonstrated improvements on NER in the challenging TAC-KBP Cold Start Entity Discovery dataset.
- Improved name finding resulting in improvements on the entity clustering task.

[1] Some titles aren't full sentences. For example, *40 Minutes With Kellyanne Conway.*
[2] We show the names that are missed by our strong sequence tagging model in bold, and show the reference name that are tagged correctly with underline.
[3] https://tac.nist.gov/2015/KBP/

We will first describe related work, then the algorithm in details. We will present experimental results and conclusion in the end.

2 RELATED WORK

Named Entity Recognition (NER) was introduced as a separate task in Message Understanding Conference - 6 [11], and has been advanced through evaluations such as ACE [13], CoNLL [28, 29] and TAC [15]. NER has been studied extensively [26]. Notable supervised models for NER include: HMM [3, 6], Maximum Entropy model [5], Maximum Entropy Markov Model [20], Conditional Random Field [17], and Neural Network models [18]. [25] applied self-training with contemporary texts to update a name tagger. [30] and [2] proposed active learning for NER. [35] proposed expectation-driven learning for constructing a name tagger in a few hours for low-resource languages. There are also work on joint modeling of NER and linking [31], and NER with linked data [12]. [33] applied bootstrapping for domain adaptation for NER. Complementary to our work, [10] models coherence of entity mentions in a document for the task of entity resolution. It could be applied in conjunction with our method for further gain.

Most traditional NER tasks focus on newswire articles, with exceptions of biomedical [16, 32] and Tweets [27]. The 2015 TAC evaluation corpus [15] contains a significant number of discussion forum posts.

Previous works show that typing of names changes [34], new names emerges at a high rate and the NE tagger performance decreases through time [24]. [14] stressed that carefully selecting data is important for bootstrapping a name tagger. This is orthogonal to our observation that enforcing per-document tag consistency rather than corpus-level consistency is crucial to maintaining precision.

3 ALGORITHM

Given a sentence as a sequence of tokens $x_{1:T}$, the vanilla sequence tagging model aims at finding the best sequence of tags $\hat{y}_{1:T}$:

$$\hat{y}_{1:T} = \arg\max_{y_{1:T}} f(x_{1:T}, y_{1:T}; \theta)$$

in which f is a normalized joint probability in CRF [17], or a unnormalized measure in HMM [4]. The model is parameterized by θ which can be learnt from gradient descent [17] or structured perceptron [6].

This sentence-level model only captures local context. It fails to extract the difficult names in the examples in Section 1. We showed in Section 1 that tagging them benefits from more contexts around matching strings in the same document.

For many missing names, error analysis shows that missed forms were found elsewhere in the same document (e.g., examples in Section 1). We hypothesize that each name string most likely has one and only **one sense per-document** [4] and propose the following tag-consistency objective. The objective is to enforce that the tag sequences assigned to any pair of matching phrases [5] should have the same name tag sequences.

$$g(D) = \Sigma_{Y_i, Y_j \in D} \sigma(Y_i, Y_j)$$

$$\sigma(Y_i, Y_j) = \mathbb{1}[y_{m+k} = y_{n+k}, \forall k \in [0, |Y_i|]]$$

in which D is a document, $Y_i = y_{m:m+k}, Y_j = y_{n:n+k}$ are two tag sub-sequences for a pair of matching phrases (two sequences of ordered matching tokens of the same length).

We interpolate the two objectives to combine tag consistency in D with the sequence tagger. Let $\mathcal{Y} = \{Y\}$, the set of all Y in D,

$$\mathcal{L}(\mathcal{Y}; X, D) = \alpha \Sigma_{X \in D} f(X, Y) + (1 - \alpha) g(D)$$

The joint model is to find $\hat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} \mathcal{L}(\mathcal{Y})$.

Finding the exact solution $\hat{\mathcal{Y}}$ is computationally expensive since exponentially many Ys need to be evaluated. Furthermore, it becomes intractable because $g(D)$ introduces dependency between pairwise sub-sequences across the whole document. We seek to find an approximate solution by focusing on $g(D)$ [6] while using $\arg\max f(X, Y)$ to obtain an initial assignment of Ys. The heuristic algorithm consists of the following steps. We run these steps for 5 iterations.

- For all sentences in the document, compute

$$\hat{y}_{1:T} = \arg\max_{y_{1:T}} f(x_{1:T}, y_{1:T}; \theta)$$

- Build phrase table for all name strings found in the document. The tag sequence for each entry is the primary tag sequences (the type that appears most frequently).
- Re-tag each non-name string that matches an entry in the phrase table with the stored tag sequence. This guarantees to increase $g(D)$.
- Re-estimate θ with stochastic gradient descent.

4 EXPERIMENTS

Dataset We use the English subset of the Entity Discovery (ED) Evaluation dataset [7] from the 2015 TAC Cold Start Knowledge Base Population (KBP) [8] evaluation. The TAC-KBP evaluations are a series of evaluations organized by the U.S. National Institute of Standards and Technology. We choose the ED dataset for three reasons: First, we hypothesize that our approach will mitigate against trained systems' degradation over time as the real world drifts away from their training data (e.g. with the introduction of new names like *SpaceX*). Commonly used data sets (e.g. ACE, CONLL) are contemporaneous or overlap with our baseline system's training data. The TAC ED dataset consists primarily of recently (2014+) published documents. Second, this dataset represents texts "in the wild". It is quite realistic and challenging for name finding. In contrast to newswire datasets (e.g., CoNLL [28, 29]), it is a mix of newswire articles and discussion forum (DF) posts (> 50% are DF posts). Third, the dataset includes annotation of cross-document entity coreference allowing us to measure the impact of our approach on cross document entity coreference.

We use the official scoring software [9] from the TAC-KBP EDL evaluations with *strong_typed_mention_match* (scores by strictly

[4] Pushing it further, assuming one sense per-name-per-corpus will result in a significant drop in precision. We will show that in the experiment section.

[5] e.g., two named mentions of *XYZ LTD.*

[6] α is set to 0.01.

[7] The core document set is shared between the Entity Discovery and Linking (EDL) evaluation and the Entity Discovery Task. The EDL task includes some requirements that are not measured here thus making the reported scores compatible with ED but not EDL. EDL specific tasks include: the identification of specific individual nominal PER spans and linking to an external KB.

[8] https://tac.nist.gov/2015/KBP/ColdStart/

[9] https://github.com/wikilinks/neleval

Systems	strong_typed_mention_match			mention_ceaf			b_cubed		
	P	R	F1	P	R	F1	P	R	F1
Top-1	0.768	0.712	0.739	0.752	0.678	0.713	0.742	0.620	0.675
Baseline	0.770	0.712	0.740	0.742	0.686	0.713	0.745	0.623	0.679
This work	0.763	**0.726**	**0.744**	0.737	**0.701**	**0.718**	0.740	**0.644**	**0.689**

Table 1: Entity Discovery scores on the 2015 TAC-KBP Cold Start English ED task. "Top-1" shows the best performing system [23] in the 2015 TAC-KBP Cold Start English ED evaluation. "Baseline" shows our strong baseline system which uses a standard feature-rich discriminative sequence tagging approach. "This work" improves over the "baseline" system by adding *one-sense-per-document* constraint with the algorithm described in this paper.

Genre	Systems	P	R	F1
News	Baseline	0.804	0.829	0.816
	This work	0.801	**0.844**	**0.822**
DF	Baseline	0.689	0.511	0.587
	This work	0.674	**0.523**	**0.589**

Table 2: *strong_typed_mention_match* scores on news and discussion forum posts (DF) on the English ED task.

Type	Systems	P	R	F1
PER	Baseline	0.831	0.674	0.745
	This work	0.821	**0.684**	**0.746**
ORG	Baseline	0.625	0.643	0.634
	This work	0.622	**0.667**	**0.644**
GPE	Baseline	0.827	0.822	0.825
	This work	0.821	**0.833**	**0.827**

Table 3: *strong_typed_mention_match* scores on three entity types on the English ED task.

	Baseline	This work
PER	2706	2777(+2.6%)
ORG	2063	2152 (+4.3%)
GPE	2364	2413(+2.1%)

Table 4: Numbers of names found with both systems on the English ED dataset.

matching mention spans and their types to the ground truth) as the primary metric since it evaluates NER. We also include two clustering metrics *mention_ceaf [19]* and *b_cubed [1]* in Table 1 to show that changes in NER improves entity clustering.

Experimental setting As baseline, we use a standard feature-rich discriminative sequence tagging approach [21] [10] with LDC and internally annotated resources. Most of the training data was published prior to 2005 (with a small amount of 2010 data as a supplement). The system has been tuned for the ACE NER task and achieves state-of-the-art performance on the ED dataset (shown as baseline in Table 1). This system is reported as baseline and used as the basis of our extensions. The proposed approach applies to any sequence tagging model. Therefore it could be applied to other base NER approaches.

Our entity discovery system[22] works as follows: first it clusters names that can be linked to the same Freebase entry based on textual similarity (edit distance and contextual pattern matching), then it clusters the unlinkable mentions [11] into entities based on their pairwise textual similarity.

Results and discussion Table 1 shows that our baseline system outperforms the top system [23] ("top-1") in the TAC evaluation. TAC participants [15] use a diverse range of algorithms such as CRF, MEMM, with various background knowledge such as Wikilinking system output as features. This shows that our baseline is competitive. The proposed method improved our strong baseline in recall and F1 in all three metrics. It is able to find more name mentions without sacrificing much precision. Inspection of the results shows that the algorithm is able to transfer the knowledge from an "easy" context, e.g., in "**SpaceX** *CEO*", to more ambiguous contexts **SpaceX** *is the....* Table 2 shows that our system improves over the baseline in both newswire and discussion forums documents. More improvement is seen in newswire, where the baseline system's higher precision reduces the error introduced by "copying" names to a new context. Table 3 and 4 show the improvements and the count of new name mentions found broken down by entity type. Our algorithms finds more ORGs than PER and GPE, therefore the improvement on ORG is higher than the other two. As shown in Table 1, the improvements in base NER lead to a positive impact on cross-document clustering of mentions into entities.

We further extend our approach to enforcing the tag-sequence consistency of name strings to the entire corpus. This "*one-sense-per-corpus*" assumption, however, resulted in 8 points drop in precision with only 3 points improvement in recall. Error analysis showed that tagging errors from ambiguous contexts were propagated across the corpus, leading to the large precision drop. Thus the limited-scope *one-sense-per-document* assumption is more accurate and effective.

5 CONCLUSION

We present a novel name finding approach that combines a sentence-level sequence tagging model and a document-wide tag-consistency objective to improve the recall of NER. We show that it achieves state-of-the-art performance in both NER and mention clustering in the TAC-KBP Cold Start Entity Discovery task. Our next step is to improve the re-tagging scheme with fuzzy matching as well as using

[10] We also experimented with a CRF-based model [8] with a rich set of features. On ACE 2005 and TAC-KBP 2015 ED datasets, this model consistantly outperform the CRF model. Therefore, we choose this model as the baseline.

[11] mentions that couldn't be linked to a Freebase entry

world knowledge (e.g., name variants from Freebase). Since our algorithm is language-independent, it could be applied to Chinese, Spanish and low-resource languages.

ACKNOWLEDGMENTS

This work was supported by DARPA/I2O Contract No. FA8750-13-C-0008 under the DEFT program. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

REFERENCES

[1] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC*.

[2] Markus Becker, Ben Hachey, Beatrice Alex, and Claire Grover. 2005. Optimising Selective Sampling for Bootstrapping Named Entity Recognition. In *Proceedings of ICML Workshop on Learning with Multiple Views*.

[3] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP*.

[4] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An Algorithm that Learns What's in a Name. *Machine Learning* (1999).

[5] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In *Proceedings of WVLC*.

[6] Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of ACL*.

[7] Jenny Rose Finkel and Christopher D Manning. 2009. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 141–150. http://www.aclweb.org/anthology/D/D09/D09-1015

[8] Ryan Gabbard, Jay DeYoung, Constantine Lignos, Marjorie Freedman, and Ralph Weischedel. 2017. Combining rule-based and statistical mechanisms for low-resource named entity recognition. *Machine Translation* (20 Dec 2017). https://doi.org/10.1007/s10590-017-9208-0

[9] William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense Per Discourse. In *Proceedings of the Workshop on Speech and Natural Language (HLT '91)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 233–237. https://doi.org/10.3115/1075527.1075579

[10] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. 2016. Collective Entity Resolution with Multi-Focal Attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 621–631. http://www.aclweb.org/anthology/P16-1059

[11] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History.. In *Proceedings of COLING*.

[12] Sherzod Hakimov, Salih Atilay Oto, and Erdogan Dogdu. 2012. Named entity recognition and disambiguation using linked data and graph-based centrality scoring. In *Proceedings of the 4th International Workshop on Semantic Web Information Management*.

[13] Shudong Huang, Stephanie Strassel, Alexis Mitchell, and Zhiyi Song. 2004. Shared Resources for Multilingual Information Extraction and Challenges in Named Entity Annotation. In *Proceedings of IJCNLP*.

[14] Heng Ji and Ralph Grishman. 2006. Data selection in semi-supervised learning for name tagging. In *Proceedings of the Workshop on Information Extraction Beyond The Document*. Association for Computational Linguistics, 48–55.

[15] Heng Ji, Joel Nothman, and Ben Hachey. 2015. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In *Proceedings of Text Analysis Conference*.

[16] Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2007. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics* 19 (2007), 180âĂŞ182.

[17] 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)*. 282–289.

[18] Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL*.

[19] Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP*.

[20] Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of ICML*.

[21] Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name Tagging with Word Clusters and Discriminative Training. In *HLT-NAACL 2004: Main Proceedings*, Daniel Marcu Susan Dumais and Salim Roukos (Eds.). Association for Computational Linguistics, Boston, Massachusetts, USA, 337–342.

[22] Bonan Min and Marjorie Freedman. 2014. BBN System for Cold Start Knowledge Base Population. In *Proceedings of Text Analysis Conference*.

[23] Bonan Min, Marjorie Freedman, and Constantine Lignos. 2015. BBN's 2015 System for Cold Start Knowledge Base Population. *Proceedings of the Text Analysis Conference (TAC)* (2015).

[24] Cristina Mota and Ralph Grishman. 2008. Is this NE tagger getting old?. In *Proceedings of LREC*.

[25] Cristina Mota and Ralph Grishman. 2009. Updating a name tagger using contemporary unlabeled data. In *Proceedings of ACL*.

[26] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30, 1 (2007), 3âĂŞ26.

[27] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of EMNLP*.

[28] Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL*.

[29] 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 CoNLL*.

[30] Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of ACL*.

[31] Avirup Sil and Alexander Yates. 2013. Re-ranking for Joint Named-Entity Recognition and Linking. In *Proceedings of CIKM*.

[32] Nichalin Suakkaphong, Zhu Zhang, and Hsinchun Chen. 2011. Disease named entity recognition using semisupervised learning and conditional random fields. *Journal of the American Society for Information Science and Technology* 62, 4 (apr 2011), 727–737. https://doi.org/10.1002/asi.21488

[33] Ang Sun and Ralph Grishman. 2011. Cross-Domain Bootstrapping for Named Entity Recognition. In *SIGIR 2011 Workshop on Entity-Oriented Search*. Beijing, China.

[34] Masatoshi Tsuchiya, Shoko Endo, and Seiichi Nakagawa. 2009. Analysis and Robust Extraction of Changing Named Entities. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*. Association for Computational Linguistics, Suntec, Singapore, 161–167. http://www.aclweb.org/anthology/W/W09/W09-3534

[35] Boliang Zhang, Xiaoman Pan, Tianlu Wang, Ashish Vaswani, Heng Ji, Kevin Knight, and Daniel Marcu. 2016. Name Tagging for Low-resource Incident Languages based on Expectation-driven Learning. In *Proceedings of NAACL*.