The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

# Gazetteer Generation for Neural Named Entity Recognition

**Chan Hee Song**
University of Notre Dame
csong1@nd.edu

**Dawn Lawrie**
HLTCOE, JHU
lawrie@jhu.edu

**Tim Finin**
UMBC, HLTCOE
finin@umbc.edu

**James Mayfield**
HLTCOE, JHUAPL
mayfield@jhu.edu

## Abstract

We present a way to generate gazetteers from the Wikidata knowledge graph and use the lists to improve a neural NER system by adding an input feature indicating that a word is part of a name in the gazetteer. We empirically show that the approach yields performance gains in two distinct languages: a high-resource, word-based language, English and a high-resource, character-based language, Chinese. We apply the approach to a low-resource language, Russian, using a new annotated Russian NER corpus from Reddit tagged with four core and eleven extended types, and show a baseline score.

## 1 Introduction

Named entity recognition (NER) is an important task in natural language understanding that entails spotting mentions of conceptual entities in text and classifying them according to a given set of categories. It is particularly useful for downstream tasks such as information retrieval, question answering, and knowledge graph population. Developing a well-performing, robust NER system can facilitate more sophisticated queries that involve entity types in information retrieval and more complete extraction of information for knowledge graph population.

Various approaches exist to automated named entity recognition. Neural approaches to NER were introduced when Hammerton (2003) used Long Short-Term Memory (LSTM). LSTM was proposed by Hochreiter and Schmidhuber (1997), expanded by Gers, Schmidhuber, and Cummins (2000), and reached its modern form with Graves and Schmidhuber (2005). More recent approaches use deep neural models, beginning with Collobert et al. (2011). Recent NER systems have adopted a forward-backward LSTM or BiLSTM, mainly using the BiLSTM-CRF architecture first proposed by Huang, Xu, and Yu (2015), and now widely studied and augmented. For example, Chiu and Nichols (2016) and Ma and Hovy (2016) augmented the BiLSTM-CRF architecture with a character-level CNN to add additional features to the architecture.

Statistical approaches have benefited by using gazetteers as an additional source of information, often because the amount of labeled data for training an NER system tends to be small. A lack of training data is of particular concern when using neural architectures, which generally require large amounts of training data to perform well. Gazetteers are easier to produce than labeled training data and can be mined from existing sources. Therefore, it is important to know whether this rich source of information can be effectively integrated into a neural model.

We aim to provide additional external knowledge to neural systems similar to the way people use background knowledge to identify entities and their types. Adding gazetteer features (often called lexical features) to neural systems has been shown to improve performance on well-studied datasets like English OntoNotes and CONLL-2003 NER using a closed-world neural system (*i.e.,* BiLSTM-CRF) (Chiu and Nichols 2016). We extended this approach and validated that gazetteer features are still beneficial to datasets in a more diverse set of languages and with models that use a pretrained encoder. For generality, we applied and evaluated our approaches on datasets in three languages: a high-resource, word-based language, English; a high-resource, character-based language, Chinese; and a lower-resource, high morphology language, Russian.

## 2 Related Work

Adding lexical features to an NER system has been studied widely, mainly by matching words in the dataset to words in pre-gathered gazetteers. Passos, Kumar, and McCallum (2014) use gazetteers during embedding generation; Chiu and Nichols (2016) use gazetteers to generate a quinary encoded match of the words in the data to those in the gazetteers; and Ghaddar and Langlais (2016) generate gazetteer embeddings from Wikipedia. Ding et al. (2019) present an architecture incorporating gazetteer information for Chinese, a language that often has many false positive matches because it is logographic.

Our approach provides a simple augmentation to existing neural models and demonstrates that different types of language can benefit from gazetteer matches. We take the Chiu and Nichols (2016) approach to matching the gazetteer because of its simplicity and universality in application to many neural models. We also show that it is applicable to

|  |  | Coll. | Gazetteer Coverage |  |  |
|---|---|---|---|---|---|
| **Dataset** | **Type** | **Freq.** | **Train** | **Test** | **Dev** |
| **PER** | Person | 12.5% | 3.3k | 15.4% | 11.4% |
| **ORG** | Organization | 1.1k | 16.9% | 26.6% | 9.8% |
| COMM | Commercial Org. | 409 | 31.4% | 33.3% | 5.3% |
| POL | Political Org. | 174 | 23.1% | 15% | 3% |
| **GPE** | Geo-political Entity | 5.5k | 23.4% | 23.1% | 20.2% |
| **LOC** | Natural Location | 451 | 7.4% | 0% | 5.8% |
| FAC | Facility | 50 | 4.3% | 50.0% | 0% |
| GOVT | Government Building | 36 | 7.7% | 0% | 0% |
| AIR | Airport | 5 | 0% | 0% | 0% |
| EVNT | Named Event | 152 | 3.4% | 0% | 0% |
| VEH | Vehicle | 63 | 6.1% | 11.1% | 20.0% |
| COMP | Computer Hard/Software | 273 | 24.0% | 17.6% | 0% |
| WEAP | Weapon | 139 | 0% | 0% | 0% |
| CHEM | Chemical | 21 | 0% | 0% | 14.3% |
| MISC | Other named entity | 1.5k | 0% | 0% | 0% |

Table 1: Named entity types for Russian, with descriptions, collection frequency, and percentage of names in the train/test/dev that were found in the associated gazetteer.

neural models with a deep pre-trained encoder.

Transfer learning architectures have shown significant improvement in natural language processing (NLP) tasks such as understanding, inference, question answering, and machine translation. BERT (Devlin et al. 2018), uses stacked bi-directional transformer layers trained on masked word prediction and next sentence prediction tasks. BERT is trained on over 3.3 billion words gathered mainly from Wikipedia and Google Books. By adding a final output layer, BERT can be adapted to many different NLP tasks. In this work, we apply BERT to NER, using BiLSTM-CRF as the output layer of BERT. Our approach embodies a simple architecture that does not require a dataset-specific architecture or feature engineering.

Our NER architecture combines recent advances in transfer learning and a BiLSTM-CRF model, producing a BERT-BiLSTM-CRF model. We use a common baseline Bi-LSTM-CRF model which, like many sequence-to-sequence closed-world NER systems (Huang, Xu, and Yu 2015), includes a stacked bi-directional recurrent neural network with long short-term memory units and a conditional random field decoder, similar to Chiu and Nichols (2016) but without the character-level CNN. We combine this system with BERT, keeping the BERT model frozen during training and testing, feeding the text into BERT and concatenating its final four layers as an input to our Bi-LSTM-CRF. In addition, the features generated from gazetteers are concatenated with the outputs from BERT and are fed into the Bi-LSTM-CRF. An extended version of this paper has details about the hyperparameter settings (Song et al. 2020).

## 3 Gazetteer Creation

Our gazetteers were created by extracting canonical names (e.g., Manchester United F.C.) and aliases (e.g., Red Devil, Man U) of entities of a given type (e.g., ORG) from Wikidata (Vrandečić and Krötzsch 2014). Wikidata is a large, collaboratively edited knowledge graph with information drawn from and used by multiple Wikimedia projects, including 310 Wikipedia sites in different languages. Its goal is to integrate entities and knowable facts about them in a language-independent manner. It currently has about 900M statements about 77M entities, supported by an ontology with nearly 2.4 million types and moe than 7,250 properties. The data are exposed as RDF triples, queried using Wikimedia APIs or SPARQL queries sent to a public query service.

Our first step was to construct a mapping from our project's 16 target types, shown in Table 1, to Wikidata's fine-grained type system. The mapping for some types was simple: person corresponds to Wikidata's Q5 and vehicle to Q42889. Others had complex mappings that eliminate overly-specialized Wikidata subtypes (e.g., *lunar craters* from Wikidata's geographic object) or allow us to retrieve more entities within the server's query timeout.

Initial name lists were filtered by type-dependant regular expressions to delete unhelpful ones (e.g., *Francis of Assisi*), remove Wikipedia artifacts (e.g., parentheticals), and eliminate punctuation, unusually long or short names, and duplicates. We produced additional lists for Russian using a custom script that generates type-sensitive inflected and familiar forms of canonical names and aliases. For example, the Russian name for the person *Vladimir Vladimirovich Putin* produces more than 100 variants. The result is a collection of 96 gazetteer files with total 15.7M entity names (4.2M eng, 1.3M rus, 979K cmn, 8.7M rus inflected). The gazetteers and associated software are available at https://github.com/hltcoe/gazetteer-collection.

We also developed efficient mappings from Wikidata's immediate type assertions to our type system. The entity *MOMA*, for example, is identified as an instance of an *art museum*, an *art institution* and a *copyright holder's organisation*, which we mapped to *ORG*, *LOC*, and *FAC*.

## 4 Exploiting Gazetteers

To use a gazetteer as a feature in the NER system, words in the dataset are matched with a gazetteer token and turned into ternary vectors for each entity type. We use ternary vectors because our datasets are tagged with the BIO (Beginning-Inside-Outside) tagging scheme with three possible tag values. The ternary vectors are then concatenated with word embeddings generated from other sources. For example, a word embedding of size 768 from BERT is concatenated with the gazetteer ternary vectors sized to the number of entities, $x$. Although each gazetteer represents an entity type, no attempt is made to communicate that type to the Bi-LSTM layer.

For gazetteer matches, we use two matching schemes: full and partial match. Unlike Chiu and Nichols (2016), we found that matching the dataset's tagging scheme to the gazetteer tagging scheme yields the best performance. In a full match, a dataset $n$-gram matches an entire gazetteer entry. If there are multiple matches in same entity category, the longest is preferred. For a partial match, the $n$-gram matches part of a gazetteer entry. Only partial matches of length greater than one are accepted, except for the PER type, due to the frequency of one-word person names. For character-level tokenized text, like Chinese, we forgo partial matching because it produces too many false matches.

After matching, matches are ternary encoded with each tag type assigned a separate ternary vector. Therefore, for

| Dataset | Type | Train | Test | Dev |
|---|---|---|---|---|
| English OntoNotes | Sentences | 82.1k | 9.0k | 12.7k |
| | Tokens | 1644.2k | 172.1k | 251.0k |
| | Entities | 70.3k | 6.9k | 10.9k |
| Chinese OntoNotes | Sentences | 37.5k | 4.3k | 6.2k |
| | Tokens | 1241.1k | 149.7k | 178.4k |
| | Entities | 37.9k | 4.5k | 5.4k |
| Russian Reddit | Sentences | 22.8k | 3.2k | 3.1k |
| | Tokens | 281.7k | 39.3k | 37.9k |
| | Core Ent. | 8.1k | 1.1k | 1.0k |
| | Extended Ent. | 11.2k | 1.5k | 1.4k |

Table 2: Statistics of dataset sizes

| | | Collection | Gazetteer Coverage | | |
|---|---|---|---|---|---|
| Dataset | Type | Frequency | Train | Test | Dev |
| English OntoNotes | PER | 27.4k | 38.2% | 44.3% | 37.5% |
| | ORG | 30.0k | 19.3% | 17.2% | 19.0% |
| | GPE | 28.2k | 88.7% | 86.8% | 87.2% |
| | LOC | 2.7k | 26.3% | 23.7% | 30.0% |
| Chinese OntoNotes | PER | 14.1k | 24.0% | 21.2% | 21.6% |
| | ORG | 10.1k | 18.0% | 17.4% | 23.4% |
| | GPE | 20.2k | 76.2% | 75.4% | 77.2% |
| | LOC | 2.7k | 18.1% | 17.4% | 14.1% |

Table 3: Types for English and Chinese datasets, along with frequency of the type in data, and percentage of names in the train/test/dev that were found in the associated gazetteer.

each token in the text, it gets assigned a *number of tag types* of ternary vectors. These ternary vectors are concatenated to the other features, which are fed into the BiLSTM.

## 5 Results using Gazetteer Features

We use our models for NER tasks on the English OntoNotes, Chinese OntoNotes, and Russian Reddit datasets. For each, we ran the baseline and the model with added gazetteer features at least ten times, depending on the size of the collection. Performance is reported as precision (P), recall (R), and their harmonic mean (F1). Dataset statistics are shown in Tables 1, 2, and 3. Tables 1 and 3 include the statistics for gazetteer coverage. We use only the four core types for English and Chinese because our gazetteer tag types do not include the extended OntoNotes types. However, we experimented with both core (in bold in Table 1) and extended types for the Russian dataset. For each experiment, we trained for a fixed number of epochs and choose the model that shows the minimum loss on the development set.

### 5.1 English and Chinese OntoNotes Datasets

For English and Chinese, there are established datasets. We chose OntoNotes v5.0 (Pradhan et al. 2013) because it has a large number of labeled entities, as shown in Table 2, and followed its train/dev/test split. We use pre-trained Cased BERT-Base with 12-layer, 768-hidden, 12-heads, 110M parameters available on the Google GitHub version. The experiment is run for 10 trials and trained for 30 epochs. The model with the minimum dev set loss is selected and run on the test set. Table 4 shows our results. We compute the p-value of the distribution using a $t$-test. Adding gazetteer features increased the F1 score by $0.52$, an improvement that is statistically significant ($p < 0.001$). We attribute this to an even coverage of the percentage of entities across train, dev, and test sets, as seen in Table 3, as well as a high coverage (high 80s) for GPE entities, the entity type with the largest F1 gain.

| Dataset | Model | P | R | F1 |
|---|---|---|---|---|
| English | Baseline | 92.46 | 91.77 | 92.11 (SD: 0.10) |
| | Gazetteer | 92.82 | 92.44 | **92.63** (SD: 0.12) |
| | +Aliases | 92.69 | 92.50 | 92.59 (SD: 0.11) |
| Chinese | Baseline | 83.40 | 84.63 | 84.01 (SD: 0.16) |
| | Gazetteer | 83.91 | 84.72 | **84.31** (SD: 0.23) |
| | +Aliases | 83.84 | 84.76 | 84.30 (SD: 0.25) |

Table 4: Performance of BERT-BiLSTM-CRF baseline and + gazetteer features on English and Chinese OntoNotes, SD stands for standard deviation

We use the Chinese OntoNotes v5.0 dataset with four core types compiled for CoNLL-2013 and follow the standard train/dev/test split as before. We applied the pre-trained Chinese BERT-Base for simplified and traditional Chinese, which has 12-layer, 768-hidden, 12-heads, 110M parameters. The experiment was run for 10 trials and trained for 30 epochs. The model with minimum loss on the dev set was selected for testing. Using gazetteer features led to a statically significant improvement ($p = 0.003$), which we attribute to high GPE coverage and even coverage across dataset splits. However, the absolute increase in F1 score is around 0.3, which is lower than English dataset. We believe Chinese showed less improvement due to our decision to forgo partial matches because of the high frequency of partial matched n-grams stemming from the language's logographic nature.

### 5.2 Russian Reddit Dataset

Since Russian has little labeled NER data, we created our own dataset[1] using Russian informal text in comments collected from over 433 different Reddit forum threads. Our first step in building the collection was to identify Russian threads, which are connected to a submission posted to a channel. Annotators examined threads with at least ten comments in a majority of Cyrillic characters to identify Russian threads. We eliminated images and movies from the thread seeds, as well as seeds from sites primarily devoted to image content since they typically contain few named entities in their comments. Over 30,000 threads met these criteria, which were prioritized based on the source of the material in the submission, with newswire and blogs preferred.

Annotators examined about 800 of these threads and identified the comment language, 433 of which were in Russian. These were automatically sentence-segmented enabling annotators to perform sentence-level named entity tagging. The Dragonfly annotation tool (Lin et al. 2018) was used to

[1] Available at https://github.com/hltcoe/rus-reddit-ner-dataset

| Dataset | Tags-Model | P | R | F1 |
|---|---|---|---|---|
| Russian Reddit | C-Baseline | 80.21 | 72.12 | 75.95 (SD: 0.43) |
| | C-Gazetteer | 79.81 | 72.03 | 75.72 (SD: 0.44) |
| | C-Inflected | 79.75 | 72.01 | 75.68 (SD: 0.42) |
| | C-Alias | 79.68 | 72.05 | 75.67 (SD: 0.48) |
| | E-Baseline | 73.36 | 56.88 | 64.08 (SD: 0.44) |
| | E-Gazetteer | 73.33 | 57.05 | 64.17 (SD: 0.58) |
| | E-Inflected | 73.31 | 57.01 | 64.14 (SD: 0.51) |
| | E-Alias | 73.08 | 57.08 | 64.10 (SD: 0.48) |

Table 5: BERT-BiLSTM-CRF baseline and + gazetteer features for Russian with Core (C) and Extended (E) tag sets

record the entity tags through an in-house Mechanical Turk-like interface. One goal was to have a wider variety of entity types so that future research could investigate types that have varying attestation frequencies. Beyond the common core types, we chose types that were sufficiently attested in the data and some subtypes to facilitate experiments with hierarchical type relationships.

To assure quality annotations, each sentence was doubly-annotated with a third annotator reviewed disagreements or singly-annotated with a second annotator reviewed the annotations. Overall annotators agreed on the label for 97.8% of the tokens; however, when considering only tokens that were part of a name, annotators agreed on the label of a token for 53.8% of the tokens. The overall Cohen's Kappa statistic, a common measure of inner-annotator agreement, is 0.71. Finally the collection was split 80-10-10 into train, development, and test, respectively. Table 2 shows the size of the collection, which was labeled with 15 types. The frequency of each type is shown in Table 1.

We use the Russian Reddit dataset to evaluate the performance of our Russian NER system. We use the pre-trained multilingual-cased BERT-Base with 12-layer, 768-hidden, 12-heads, 110M parameters. We employed the same baseline BERT-BiLSTM-CRF model with gazetteer feature added. For the Russian dataset, the experiment is run for 20 trials with 30 epochs. The model with minimum loss on dev is selected for testing. The different number of trials stems from the smaller size of this dataset, as is shown in Table 2. We report experiments with both core types and extended types using the Russian dataset. Table 5 shows the Russian experiments with gazetteers and different tag types.

While the mean of the trials is slightly higher for those with gazetteer features, none of the results shows statistical significance. We attribute this to (a) lower coverage of our gazetteer for those in the dataset; and (b) uneven gazetteer coverage throughout train, dev, and test sets, as seen in Table 1. Table 5 reports additional results from using inflected and familiar forms of canonical names and aliases gazetteer described in Section 2. However, our takeaway here is that adding gazetteer feature does not hurt the performance of the neural systems, and improves it when the gazetteer has high coverage, as seen in English and Chinese experiments.

# 6 Conclusion

We described a simple way to generate a gazetteer, and show how it can be used in a neural NER systems. We also presented a new Russian NER corpus gathered from Reddit comments. We showed that with enough coverage on the dataset, gazetteer features improve neural NER systems, even systems using deep pre-trained models such as BERT. We believe gazetteer features should be a standard addition to any NER system and showed that even with low coverage, the gazetteer features do not hurt the performance of neural NER systems. Our gazetteer data and annotated Russian dataset are available on GitHub.

# References

Chiu, J. P., and Nichols, E. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Trans. ACL* 4:357–370.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *JMLR* 12(Aug).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, R.; Xie, P.; Zhang, X.; Lu, W.; Li, L.; and Si, L. 2019. A neural multi-digraph model for Chinese NER with gazetteers. In *Proc. ACL*, 1462–1467.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation* 12(10):2451–2471.

Ghaddar, A., and Langlais, P. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *LREC*.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *NEURAL NETWORKS* 5–6.

Hammerton, J. 2003. Named entity recognition with long short-term memory. In *NAACL*, CONLL '03, 172–175.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR* abs/1508.01991.

Lin, Y.; Costello, C.; Zhang, B.; Lu, D.; Ji, H.; Mayfield, J.; and McNamee, P. 2018. Platforms for non-speakers annotating names in any language. In *ACL Demonstrations*.

Ma, X., and Hovy, E. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proc. ACL*.

Passos, A.; Kumar, V.; and McCallum, A. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Conf. on Computational Natural Language Learning*.

Pradhan, S.; Moschitti, A.; Xue; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using OntoNotes. In *Conf. on Computational Natural Language Learning*.

Song, C. H.; Lawrie, D.; Finin, T.; and Mayfield, J. 2020. Improving neural named entity recognition with gazetteers. *arXiv preprint arXiv:2003.03072*.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57(10).