Improving Neural Named Entity Recognition with Gazetteers

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, JHU
mayfield@jhu.edu

Abstract

The goal of this work is to improve the performance of a neural named entity recognition system by adding input features that indicate a word is part of a name included in a gazetteer. This article describes how to generate gazetteers from the Wikidata knowledge graph as well as how to integrate the information into a neural NER system. Experiments reveal 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. Experiments were also performed in a low-resource language, Russian on a newly annotated Russian NER corpus from Reddit tagged with four core types and twelve extended types. This article reports a baseline score. It is a longer version of a paper in the 33rd FLAIRS conference (Song et al. 2020).

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. Older statistical methods use conditional random fields (Finkel, Grenager, and Manning 2005), perceptrons (Ratinov and Roth 2009), and support vector machines (Mayfield, McNamee, and Piatko 2003). More recent approaches have applied deep neural models, beginning with Collobert et al. (2011). Further advances came from the addition of a BiLSTM model with CRF decoding (Huang, Xu, and Yu 2015), which led to the current state-of-the-art model.

Our NER architecture combines recent advances in transfer learning (Devlin et al. 2018) and a BiLSTM-CRF model, producing a BERT-BiLSTM-CRF model. In our model, BERT generates an embedding for each word. This embedding is fed into a multi-layer BiLSTM, which is often jointly trained with a pre-trained encoder at training time. This fine-tunes the encoder to the NER task. At test time, the BiLSTM outputs are decoded using a CRF. Other approaches show similar results, such as a BiLSTM-CRF that uses a character-level CNN added to BiLSTM-CRF (Ma and Sun 2016; Chiu and Nichols 2016). We adopt a BERT-based model as the baseline system for comparison.

In the context of natural language understanding, a gazetteer is simply a collection of common entity names typically organized by their entity type. These have been widely used in natural language processing systems since the early 1990s, such as a large list of English place names provided by the MUC-5 Message Understanding Conference (Sundheim 1993) to support its TIPSTER information extraction task. Initially these lists were employed to help recognize and process mentions of entities that were places or geo-political regions, hence the name gazetteer. Their use quickly evolved to cover more entity types and subtypes, such as cities, people, organizations, political parties and religions.

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 much 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.

This paper first focuses on generating gazetteers from Wikidata, presenting a simple way to gather a large quantity of annotated entities from Wikidata. It then describes how to integrate the gazetteers with a neural architecture by generating features from gazetteers alongside the features from BERT as input to the BiLSTM. We aim to provide an additional external knowledge base to neural systems similar to the way people use external knowledge to determine what is an entity and to what category it belongs.

Adding gazetteer features (often called lexical features) to neural systems has been shown to improve performance on well-studied datasets like English OntoNotes and

Type	Description	Examples
PER	Person	Enrico Rastelli, Beyoncé
ORG	Organization	International Jugglers Association
COMM	Commercial Org.	Penguin Magic, Papermoon Diner
POL	Political Organization	Green Party, United Auto Workers
GPE	Geo-political Entity	Morocco, Carlisle, Côte d'Ivoire
LOC	Natural Location	Kartchner Caverns, Lake Erie
FAC	Facility	Stieff Silver Building, Hoover Dam
GOVT	Government Building	White House, 10 Downing St.
AIR	Airport	Ninoy Aquino International, BWI
EVNT	Named Event	WWII, Operation Desert Storm
VEH	Vehicle	HMS Titanic, Boeing 747
COMP	Computer Hard/Software	Nvidia GeForce RTX 2080 Ti, Reunion
MIL	Military Equip.	AK-47, Fat Man, cudgel
MIL_G	Generic Military Equip.	tank, aircraft carrier, rifle
MIL_N	Named Military Equip.	USS Nimitz, 13"/35 caliber gun
CHEM	Chemical	Iron, NaCl, hydrochloric acid
MISC	Other named entity	Dark Star, Lord of the Rings

Table 1: We worked with types where training data was available in several languages, including four core types (in bold) and twelve additional ones.

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 pre-trained 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 morphological language, Russian. We will first present how our gazetteer is generated from publicly available data source, Wikidata. Then we will analyze our experimental results.

2 Related Work

Our work builds on the neural approach to NER, which was introduced when Hammerton (2003) used Long Short-Term Memory (LSTM), achieving just above average performance for English and improvement for German. 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). 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.

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

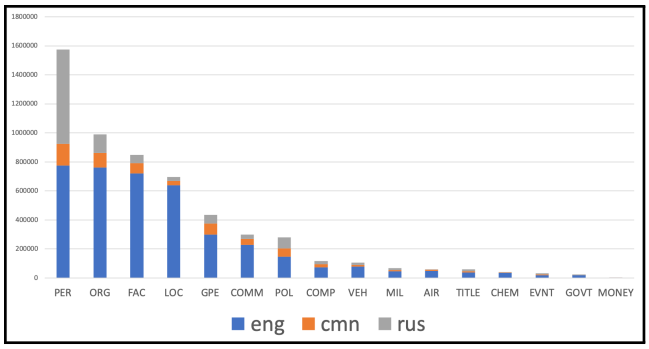

Figure 1: Statistics for canonical names for Wikidata entities for each type. Additional lists hold aliases and, for Russian, inflected forms.

However, our approach provides a simple augmentation to existing neural models and demonstrates that Chinese can benefit from gazetteer matches. We take the Chiu and Nichols (2016)'s approach to matching the gazetteer because of its simplicity and universality in application to many different neural models. We also show that it is applicable to neural models with a deep pre-trained encoder.

Transfer learning architectures have shown significant improvement in various natural language processing 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 natural language processing tasks. In this work, we apply BERT to NER and use 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.

3 Gazetteer Creation

We describe the knowledge source we use to create our gazetteers and outline the process we used to automatically produce cleaned gazetteers for the entity types of interest.

3.1 Wikidata Knowledge Graph

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 a number of Wikimedia projects, includ-

ing 310 Wikipedia sites in different languages. Its goal is to integrate entities and knowable facts about them for use in Wikimedia sites in a language-independent manner. Wikidata is multilingual, with all of its strings tagged with a two-letter ISO 639-1 language code.

Wikidata currently has more than 900 million statements about 77 million items, supported by an ontology with nearly 2.4 million fine-grained types and more than 7,250 properties. An example item is the entity Q7186 shown in Figure 2. Items have a canonical name, short description and set of aliases in one or more languages. Property statements encode relations between items or between an item and a literal value and can have metadata including qualifiers (e.g., a period of time during which the property held), provenance information (e.g., the URL of an attesting source), and a rank (e.g., to distinguish a preferred value from alternative or deprecated ones).

The data is exposed as RDF triples and can be queried using Wikimedia APIs or SPARQL queries sent to a public query service. We used the public SPARQL service to get both canonical names (e.g., Johns Hopkins University) and aliases (e.g., JHU, Johns Hopkins, Hopkins) in each of the languages studied for a given entity type (e.g., ORG). In addition, the Wikimedia community has developed many tools for searching for items, exploring the ontology, and updating entries.

3.2 Gazetteer Generation

The gazetteer is generated by searching Wikidata via SPARQL queries sent to the public query server to retrieve both canonical names (e.g., Johns Hopkins University) and aliases (e.g., JHU, Johns Hopkins, Hopkins)) in each of the languages studied. The 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 (Pellissier Tanon et al. 2016). Our types included four common core types (person, organization, geopolitical entity (GPE), location) and twelve additional types (airport, chemical, commercial organization, computer hardware/software, event, facility, government building, military equipment, money, political organization, title, vehicle).

The mapping for some types was simple: person corresponds to Wikidata's Q5 and vehicle to Q42889. Others had a complex mapping that eliminate Wikidata subtypes that seemed too specialized (e.g., *lunar craters* and *ice rumples* from Wikidata's geographic object) or allow us to retrieve more entity names given the public server's one-minute query timeout.

The initial name lists were filtered by type-dependant regular expressions to delete names we thought to be unhelpful (e.g., *Francis of Assisi* as a person because historical figures are unlikely to be mentioned in our targeted genres), remove Wikipedia artifacts (e.g., parentheticals), and eliminate punctuation, names that were too short or too long, and duplicate names. Although one could say that these changes bias the gazetteers, there is no reason not engineer a gazetteer in a way that is most helpful for the data. Wikidata is still being used in an automated way since we are relying on available labels.

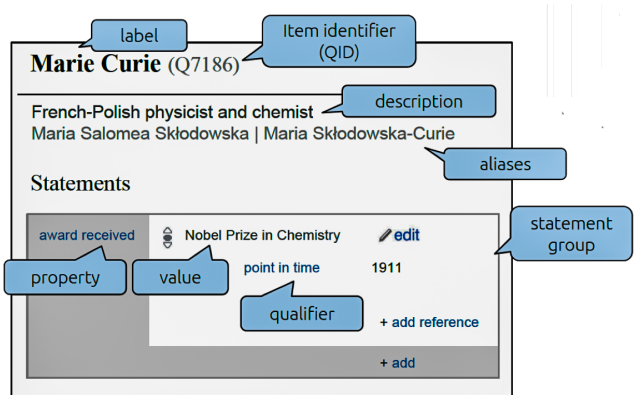

Figure 2: Wikidata entities have a unique ID, a canonical name, aliases and a short description in one or more languages along with any number of statements representing properties or relations and including qualifiers and provenance metadata.

We produced additional lists for Russian using a custom script that generates type-sensitive inflected and familiar forms of canonical names and aliases. For an extreme example, the Russian name for the person *Vladimir Vladimirovich Putin* (Владимир Владимирович Путин) produces more than 100 variations. The result is a collection of 96 gazetteer files with total 15.7M entity names, 4.2M for English, 2.1M for Russian and 584K for Chinese with an additional 8.7M Russian names produced by our morphological scripts. We kept the gazetteers for canonical names, aliases, and inflected forms separate to facilitate experimentation.

We also worked with data downloaded directly from Wikidata, which uses a JSON serialization in which only the immediate types of each entity is provided. The entity *Museum of Modern Art*, for example, is identified as an instance of an *art museum*, an *art institution* and a *copyright holder's organisation*. To decide which, if any, of our target types an entity belongs, we constructed a dictionary that maps relevant types to our 16 target types. For our example, this turns out to be three types: *ORG*, *LOC*, and *FAC*.

Although doing the mapping sounds daunting given an ontology with more than two million types, it is simplified by exploiting the fact that most of these types do not have any immediate instances. We developed SPARQL queries that identified for each of our 16 types all of their subtypes that had one or more immediate instances. The ORG type, for example has 15,904 subtypes but only 5,962 have immediate instances; the LOC type has only 1,181 subtypes with immediate instances. The resulting dictionary was thus relatively small without losing any information and was used to quickly recognize entities of interest in the Wikidata dumps as well as identify their target types.

4 Exploiting Gazetteers

4.1 Gazetteer Features

To use a gazetteer as a feature in the NER system, words in the dataset are matched with a gazetteer and turned into

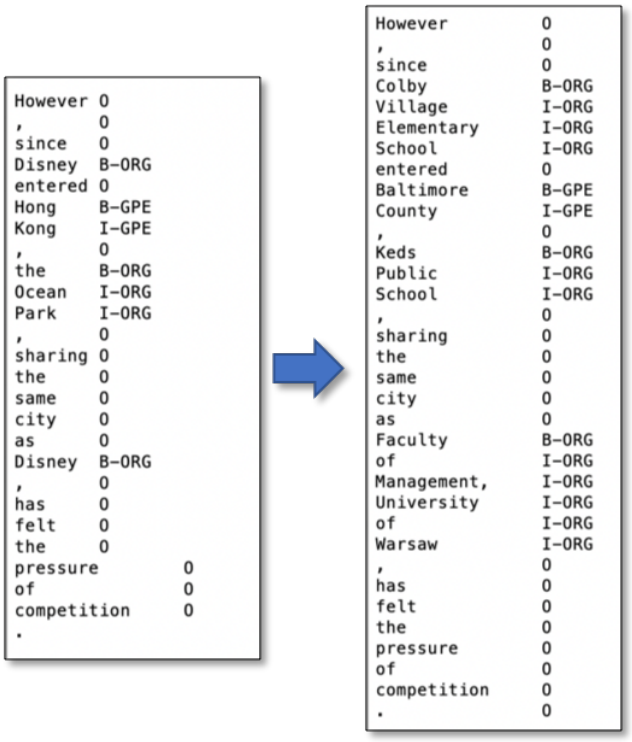

Figure 3: New training data is generated by replacing existing annotated names with gazetteer names of the same type.

one-hot vectors for each entity type. Those one-hot 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 one-hot 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.

We use the BIO (Beginning-Inside-Outside) tagging scheme where *B-<type>* tags the first token of an entity, *I-<type>* the subsequent ones, and *O* tagging non-entity tokens. For gazetteer matches, the gazetteer uses two matching schemes: full match and partial match.

- Full match: an n-gram in the dataset matches fully with a gazetteer entry. If there are multiple matches in same entity category, the longest match is preferred.
- Partial match: an n-gram in the dataset matches partially with the gazetteer. Only partial matches of length greater than one are accepted, except for the PER type, due to frequency of one-word person names.

As an example, consider a gazetteer that contains {Jack:PER, Lantau Island:LOC, Hong Kong Government:GPE, JFK International Airport:ORG}, Figure 4 shows how full matches and partial matches are handled. *Jack* is fully matched with PER, so it is tagged B-PER. *Lantau Island* is also fully matched and tagged LOC. *Hong Kong* partially matches *Hong Kong Government*, and since the match length is greater than one, it is considered a match and thus tagged as GPE. *Hong Kong International Airport* is also partially matched with gazetteer entry *JFK International Airport*, so *International Airport* is tagged with ORG. As seen in the example match, tags of the gazetteer entries are assigned during a partial match. For character-level tokenized text, like Chinese, we forgo partial matching because it produces too many false matches. However for all other language, we both utilize full match and partial match as shown in the experiments.

Hyperparameters	
BiLSTM layers	1
BiLSTM hidden size	256
BiLSTM dropout	.5
Optimizer	adafactor
Gradient clipping	1.0
Learning rate scheduler	cosine decay
BERT layers used	$-4, -3, -2, -1$
Weight decay	.005
Mini batch size	8

Table 2: Default hyperparameters used in the baseline model

After matching, matches are one-hot encoded with each tag type assigned a separate one-hot vector. Therefore, for each token in the text, it gets assigned a *number of tag types* of one-hot vectors. These one-hot vectors are concatenated to the other features, which are fed into the BiLSTM.

4.2 Generating Augmented Training Data

We experimented with a second application of gazetteers that uses them to generate additional training data. In this approach, we select sentences from our initial human-annotated training data, replace one or more of the annotated entities with a randomly selected gazetteer entity of the same type, and retrain the system. However, this approach did not produce statistically significant improvements.

We developed a script that takes as input a BIO-tagged file, a type, and gazetteer for that type, and produces a modified version of the file with entity instances of the type replace with a random entity selected from the gazetteer. Additional arguments control whether all instances of a given entity in the input BIO file are replaced with the same gazetteer entity and specify the random seed, to support repeatable experiments. Our current experiments were run by replacing entities for all types and to allowing a given input entity to be replaced with different gazetteer entities each time it appears. Figure 3.2 shows an example with an original annotated sentence from OntoNotes on the left, and a new, generated training instance on the right.

5 Architecture

We use a common baseline Bi-LSTM-CRF model like many sequence to sequence closed-world NER systems (Huang, Xu, and Yu 2015), which includes a stacked bi-directional recurrent neural network with long short-term memory units and a conditional random field decoder and is similar to

Text	Jack	is	on	Hong	Kong	International	Airport	in	Lantau	Island	,	Hong	Kong
PER	B-PER	O	O	O	O	O	O	O	O	O	O	O	O
LOC	O	O	O	O	O	O	O	O	B-LOC	I-LOC	O	O	O
GPE	O	O	O	B-GPE	I-GPE	O	O	O	O	O	O	B-GPE	I-GPE
ORG	O	O	O	O	O	I-ORG	I-ORG	O	O	O	O	O	O

Figure 4: This example shows gazetteer matches for the sentence "Jack is on Hong Kong International Airport in Lantau Island, Hong Kong" showing both full and partial matches.

Chiu and Nichols (2016) without the character-level CNN. We combine this system with BERT (Devlin et al. 2018), which is a stack of bi-directional transformer encoders. We keep the BERT 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 fed into the Bi-LSTM-CRF. Table 2 shows the hyperparameters used for our experiments. We did not perform a hyperparameter search.

6 Experimental Data Sets

To demonstrate the effectiveness of our new approaches to NER, datasets labeled with names are required. 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. Table 3 contains the statistics for these datasets. Russian, on the other hand, has little labeled data for NER. We chose to create our own dataset over Russian informal text. This section describes the construction of this collection, as well as its tag set and statistics.

The Russian Reddit collection comes from Russian comments collected over 433 threads on the Reddit online discussion forum (Reddit 2019). Reddit organizes threads around a submission that is posted to a channel. The first step in building the collection was to identify Russian threads. Annotators examined threads with at least ten comments in a majority of Cyrillic characters to determine whether the thread was written in Russian. We eliminated images and movies from the thread seeds, as well as seeds from sites primarily devoted to image content; such seeds typically contain few named entities in their comments. Over 30,000 threads met these criteria. Threads were prioritized based on the source of the material in the submission, where newswire and blogs were preferred.

Annotators examined around 800 of these threads and identified the language of the comments, 433 of which were in Russian. These comments were automatically sentence-segmented using CoreNLP (Manning et al. 2014), so that named entity tagging could be performed by annotators at the sentence level. The Dragonfly annotation tool (Lin et al. 2018) was used to record the entity tags through an in-house Mechanical Turk-like interface.

One goal of this collection was to have a wider variety of entity types so that future research could investigate types that have varying frequencies of attestation. Beyond the common core types, types were chosen that were sufficiently attested in the data. In addition, we desired to have a few subtypes of the common types to be able to experiment with this hierarchical relationship.

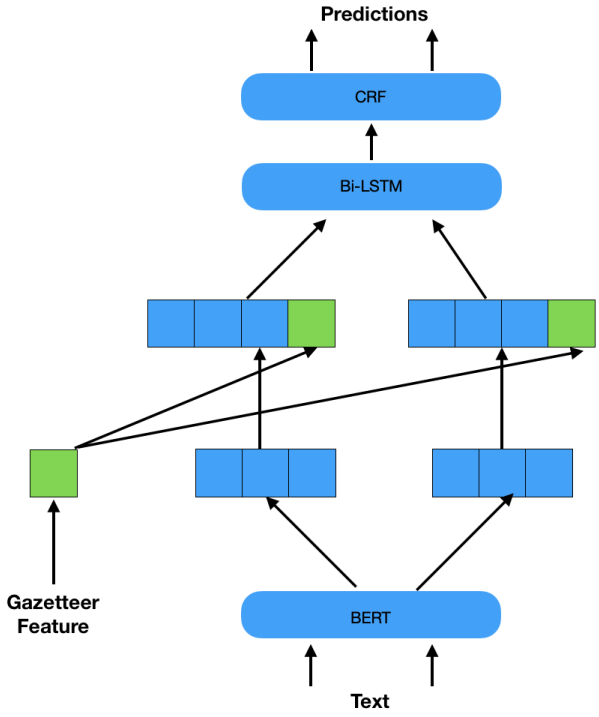

Figure 5: Architecture used in our models, with baseline components in blue and additional gazetteer features in green.

To assure quality annotations, either sentences were doubly-annotated and a third annotator reviewed disagreements or the sentence was singly-annotated and a second annotator reviewed the annotations. The inter-annotator agreement on the doubly annotated text was 53%. The annotators agreed on whether a token was part of a name 63% of the time. Since agreement was measured at the token level, both a name's tag and span had to match exactly. Finally the collection was split 80-10-10 into train, development, and test respectively. Table 3 shows the size of the collection, which was labeled with 16 types as shown in Table 1. The frequency of each type is shown in Table 4. The data set is available from https://github.com/hltcoe/rus-reddit-ner-dataset.

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 3: Statistics of dataset sizes

Tag Type	Frequency by Collection		
	English OntoNotes	Chinese OntoNotes	Russian Reddit
PER	27.4k	14.1k	3.3k
ORG	30.0k	10.1k	1.1k
COMM	-	-	409
POL	-	-	174
GPE	28.2k	20.2k	5.5k
LOC	2.7k	2.7k	451
FAC	-	-	50
GOVT	-	-	36
AIR	-	-	5
EVNT	-	-	152
VEH	-	-	63
COMP	-	-	273
MIL_G	-	-	260
MIL_N	-	-	139
CHEM	-	-	21
MISC	-	-	1.5k

Table 4: Statistics of datasets by tag type

7 Results using Gazetteer Features

We use our models for NER tasks on the English OntoNotes, Chinese OntoNotes Russian Reddit datasets. For each, we run our baseline models and the models with added gazetteer features at least ten times, depending on the size of the collection. Smaller collections were run a greater number of times because of the greater variability in the output. Performance is reported as precision (P), recall (R), and their harmonic mean (F1). Statistics of the dataset are shown in Tables 3 and 4. The statistics for gazetteer coverage for individual datasets are shown in Table 6. 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 experiment with both core and extended types for the Russian dataset. For each experiment, we train for fixed epochs and choose the model that shows the minimum loss on the development set.

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

7.1 English and Chinese OntoNotes

We use the English OntoNotes v5.0 dataset compiled for the CoNLL-2013 shared task (Pradhan and Ramshaw 2017) and follow the standard train/dev/test split as presented in (Pradhan et al. 2013). We use pre-trained Cased BERT-Base with 12-layer, 768-hidden, 12-heads, 110M parameters available on the Google Github. 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 5 shows our experiment results. We compute the p-value of the distribution using a t-test. We show that adding gazetteer features increases 0.52 F1 score, 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 6, as well as a high coverage (high 80s) for GPE entities, the entity type with the largest F1 gain.

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 use the pre-trained Chinese BERT-Base for simplified and traditional Chinese which has 12-layer, 768-hidden, 12-heads, 110M parameters. The experiment is run for 10 trials and trained for 30 epochs. The model with minimum loss on dev set is selected for testing. The gazetteer feature leads to a statically significant improvement ($p = 0.003$), which we attribute this 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 due to the high frequency of partial matched n-grams stemming from the language's logographic nature.

7.2 Russian Reddit Dataset

We use the Russian Reddit dataset to evaluate the performance of Russian NER. We use pre-trained Multilingual Cased BERT-Base with 12-layer, 768-hidden, 12-heads, 110M parameters. We use the same baseline BERT-BiLSTM-CRF model with gazetteer feature added. For the Russian Reddit 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 is due to the

Dataset	Type	Train	Test	Dev
English OntoNotes	PER	38.2%	44.3%	37.5%
	ORG	19.3%	17.2%	19.0%
	GPE	88.7%	86.8%	87.2%
	LOC	26.3%	23.7%	30.0%
Chinese OntoNotes	PER	24.0%	21.2%	21.6%
	ORG	18.0%	17.4%	23.4%
	GPE	76.2%	75.4%	77.2%
	LOC	18.1%	17.4%	14.1%
Russian Reddit	**PER**	12.5%	15.4%	11.4%
	ORG	16.9%	26.6%	9.8%
	COMM	31.4%	33.3%	5.3%
	GPE	23.4%	23.1%	20.2%
	LOC	7.4%	0%	5.8%
	FAC	4.3%	50.0%	0%
	GOVT	7.7%	0%	0%
	AIR	0%	0%	0%
	EVNT	3.4%	0%	0%
	VEH	6.1%	11.1%	20.0%
	COMP	24.0%	17.6%	0%
	MIL_G	0%	0%	0%
	MIL_N	0%	0%	0%
	CHEM	0%	0%	14.3%
	MISC	0%	0%	0%

Table 6: Statistics for the entity types and subtypes for each of the three collections. Our OntoNotes data only covered the core types while our Russian Reddit included additional types and subtypes.

smaller size of this dataset, as is shown in Table 3. We report experiments with both core types and extended types using the Russian Reddit dataset. Table 7 shows Russian dataset 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 is seen in Table 6. That table reports additional results from using inflected and familiar forms of entity canonical names and aliases, as described in Section 2. However, our takeaway here is that adding gazetteer features does not hurt the performance of the neural systems, but only improves it when the gazetteer has high coverage, as can be seen in the English and Chinese experiments.

8 Results for Training Data Augmentation

Using our gazetteers to produce additional training data produced mixed results and generally was inconsistent in improving performance for our models using BERT for either the four core types or the extended set. We ran early experiments using FastText embeddings (Bojanowski et al. 2017)

Tags-Model	P	R	F1
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 7: Performance of BERT-BiLSTM-CRF baseline and model with gazetteer features on Russian Reddit datasets for the Core (C) and Extended (E) tag sets

and found that using our gazetteers with Russian inflections for PER improved performance for most types. However, we did not see the same gains when using BERT.

This may be due to several reasons. First, both core and extended types are quite broad. Replacing the annotated ORG *Harvard University* with the gazetteer ORG *Disneyland* in the sentence "Professor Pinker teaches at Harvard University" seems anomalous to us and probably also to our model. Second, our experiments were done with relatively small amounts of annotated training data, especially for Russian. While drawing on gazetteer data may help introduce new patterns not present in the training data, such as ORGs beginning with "Association of", the chances of this helping when evaluated with the relatively small test partition is low. Third, entity names were extracted from Wikidata without regard to their utility, including both very prominent entities (e.g., the LOC *Atlantic Ocean*) and very obscure ones (e.g., *Avalonia*, a microcontinent in the Paleozoic era).

We plan to further explore this use case by replacing annotated entities with gazetteer entities that are in a same finer-grained Wikidata type. We can readily identify all of the Wikidata types to which a reasonably prominent entity (e.g., Harvard University) belongs. We will select several hundred of these types as targets for gazetteer entity replacement (e.g., Q2385804 – education institution) that are similar to an annotated entity. We can then associate gazetteer entities with these target types, replacing an entity such as "Harvard University" with an entity that is more similar, e.g., "Swarthmore College" or "Loyola Academy"). We also plan to limit obscure entities using a measure of prominence derived from Wikidata metrics, such as their number of incoming and outgoing links.

9 Conclusion and Future Work

We present a simple way to generate a gazetteer, and show how it can be used in a neural NER systems. We also present a new Russian NER corpus gathered from Reddit. We show that with enough coverage on the dataset, gazetteer features improve neural NER systems, even systems using deep pretrained models such as BERT. We hypothesize that paying attention to how tuned the signal is between the gazetteer and the training set greatly impacts ho much the neural sys-

tem learns to pay attention to the gazetteer. Modifying the gazetteer based on the training data is a path we plan to explored. In general, we believe gazetteer features should be a standard addition to any NER system and show that even with low coverage, the gazetteer features do not hurt the performance of neural NER systems.

While our gazetteer data augmentation did not show consistent improvement, we believe that future work in more sophisticated and contextualized replacement scheme will benefit low-resource languages such as Russian. In addition the noisiness of the gazetteers may have a great impact on performance since the NER system may learn not to trust a gazetteer that does not assist with tagging a sufficient number of time. Future work will identify techniques to produce gazetteers that are trustworthy relative to the training data to see if such gazetteers can be shown to be more helpful.

Gazetteers and associated software is available from https://github.com/hltcoe/gazetteer-collection.

Acknowledgments

We thank Johns Hopkins University, the Human Language Technology Center of Excellence and the 2019 SCALE workshop for their hospitality and for facilitating an excellent research environment.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Chiu, J. P., and Nichols, E. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Trans. of the 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.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. ACL*, 363–370.

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 *Proc. LREC*.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18(5-6):602–610.

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 *Proceedings of ACL 2018, System Demonstrations*, 1–6. Association for Computational Linguistics.

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

Ma, S., and Sun, X. 2016. A new recurrent neural CRF for learning non-linear edge features. arXiv:1611.04233.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL Demos*.

Mayfield, J.; McNamee, P.; and Piatko, C. 2003. Named entity recognition using hundreds of thousands of features. In *7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, 184–187. Association for Computational Linguistics.

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

Pellissier Tanon, T.; Vrandečić, D.; Schaffert, S.; Steiner, T.; and Pintscher, L. 2016. From freebase to wikidata: The great migration. In *Int. Conf. on world wide web*.

Pradhan, S., and Ramshaw, L. 2017. Ontonotes: Large scale multi-layer, multi-lingual, distributed annotation. In *Handbook of Linguistic Annotation*. Springer. 521–554.

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*.

Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *Conf. on computational natural language learning*, 147–155.

Reddit. 2019. Reddit web site. http://reddit.com/.

Song, C. H.; Lawrie, D.; Finin, T.; and Mayfield, J. 2020. Gazetteer generation for neural named entity recognition. In *Florida Artificial Intelligence Research Symposium*.

Sundheim, B. M. 1993. Tipster/muc-5: information extraction system evaluation. In *Proceedings of the 5th Conference on Message Understanding*, 27–44. Association for Computational Linguistics.

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