Soft Gazetteers for Low-Resource Named Entity Recognition

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
{srijhwan,shuyanzh,gneubig,jgc}@cs.cmu.edu

Abstract

Traditional named entity recognition models use gazetteers (lists of entities) as features to improve performance. Although modern neural network models do not require such hand-crafted features for strong performance, recent work (Wu et al., 2018) has demonstrated their utility for named entity recognition on English data. However, designing such features for low-resource languages is challenging, because exhaustive entity gazetteers do not exist in these languages. To address this problem, we propose a method of "soft gazetteers" that incorporates ubiquitously available information from English knowledge bases, such as Wikipedia, into neural named entity recognition models through cross-lingual entity linking. Our experiments on four low-resource languages show an average improvement of 4 points in F1 score.[1]

1 Introduction

Before the widespread adoption of neural networks for natural language processing tasks, named entity recognition (NER) systems used linguistic features based on lexical and syntactic knowledge to improve performance (Ratinov and Roth, 2009). With the introduction of the neural LSTM-CRF model (Huang et al., 2015; Lample et al., 2016), the need to develop hand-crafted features to train strong NER models diminished. However, Wu et al. (2018) have recently demonstrated that integrating linguistic features based on part-of-speech tags, word shapes, and manually created lists of entities called *gazetteers* into neural models leads to better NER on English data. Of particular interest to this paper are the gazetteer-based features – binary-valued features determined by whether or not an entity is present in the gazetteer.

Although neural NER models have been applied to low-resource settings (Cotterell and Duh, 2017; Huang et al., 2019), directly integrating gazetteer features into these models is difficult because gazetteers in these languages are either limited in coverage or completely absent. Expanding them is time-consuming and expensive, due to the lack of available annotators for low-resource languages (Strassel and Tracey, 2016).

As an alternative, we introduce "soft gazetteers", a method to create continuous-valued gazetteer features based on readily available data from high-resource languages and large English knowledge bases (e.g., Wikipedia). More specifically, we use entity linking methods to extract information from these resources and integrate it into the commonly-used CNN-LSTM-CRF NER model (Ma and Hovy, 2016) using a carefully designed feature set. We use entity linking methods designed for low-resource languages, which require far fewer resources than traditional gazetteer features (Upadhyay et al., 2018; Zhou et al., 2020).

Our experiments demonstrate the effectiveness of our proposed soft gazetteer features, with an average improvement of 4 F1 points over the baseline, across four low-resource languages: Kinyarwanda, Oromo, Sinhala, and Tigrinya.

2 Background

Named Entity Recognition NER identifies named entity spans in an input sentence, and classifies them into predefined types (e.g., location, person, organization). A commonly used method for doing so is the **BIO** tagging scheme, representing the **B**eginning, the **I**nside and the **O**utside of a text segment (Ratinov and Roth, 2009). The first word of a named entity is tagged with a "B-", subsequent words in the entity are "I-", and non-entity words are "O". For example:

[Mark]$_{\text{B-PER}}$ [Watney]$_{\text{I-PER}}$ [visited]$_{\text{O}}$ [Mars]$_{\text{B-LOC}}$

[1] Code and data are available at https://github.com/neulab/soft-gazetteers.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8118–8123
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

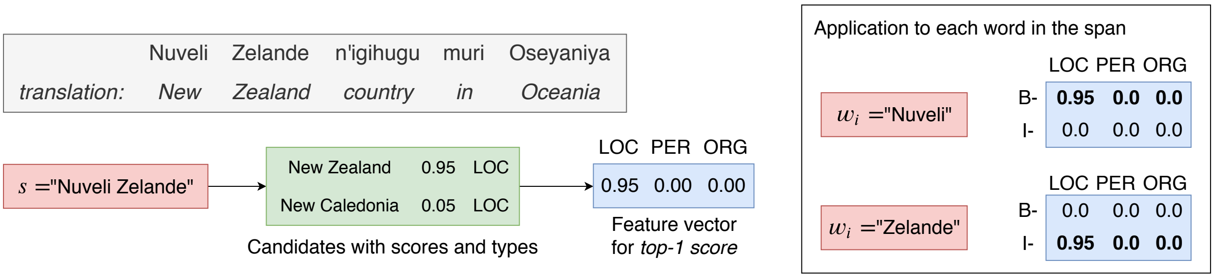

Figure 1: An example in Kinyarwanda to demonstrate soft gazetteer feature creation for each span $\boldsymbol{s}$ using candidate lists. The feature vector is applied to each word $\boldsymbol{w}_i$ in the span, depending on the position (“B-” or “I-”).

Binary Gazetteer Features Gazetteers are lists of named entities collected from various sources (e.g., nation-wide census, GeoNames, etc.). They have been used to create features for NER models, typically binary features indicating whether the corresponding n-gram is present in the gazetteer.

Entity Linking Entity linking (EL) is the task of associating a named entity mention with its corresponding entry in a structured knowledge base (KB) (Hachey et al., 2013). For example, linking the entity mention “Mars” with its Wikipedia entry.

In most entity linking systems (Hachey et al., 2013; Sil et al., 2018), the first step is shortlisting candidate KB entries, which are further processed by an entity disambiguation algorithm. Candidate retrieval methods, in general, also score each candidate with respect to the input mention.

3 Soft Gazetteer Features

As briefly alluded to in the introduction, creating binary gazetteer features is challenging for low-resource languages. The soft gazetteer features we propose instead take advantage of existing limited gazetteers and English knowledge bases using low-resource EL methods. In contrast to typical binary gazetteer features, the soft gazetteer feature values are continuous, lying between 0 and 1.

Given an input sentence, we calculate the soft gazetteer features for each span of n words, $\boldsymbol{s} = \boldsymbol{w}_i, \ldots, \boldsymbol{w}_{i+n-1}$, and then apply the features to each word in the span. We assume that we have an EL candidate retrieval method that returns candidate KB entries $\mathcal{C} = (c_1, c_2 \ldots)$ for the input span. c_1 is the highest scoring candidate.

As a concrete example, consider a feature that represents the *score of the top-1 candidate*. Figure 1 shows an example of calculating this feature on a sentence in Kinyarwanda, one of the languages used in our experiments. The feature vector $\boldsymbol{f}$ has an element corresponding to each named entity type in the KB (e.g., LOC, PER, and ORG).

For this feature, the element corresponding to the entity type of the highest scoring candidate c_1 is updated with the score of the candidate. That is,

$$\boldsymbol{f}_{\texttt{type}(c_1)} = \texttt{score}(c_1).$$

This feature vector is applied to each word in the span, considering the position of the specific word in the span according to the BIO scheme; we use the “B-” vector elements for the first word in the span, “I-” otherwise.

For a word $\boldsymbol{w}_i$, we combine features from different spans by performing an element-wise addition over vectors of all spans of length n that contain $\boldsymbol{w}_i$. The cumulative vector is then normalized by the number of spans of length n that contain $\boldsymbol{w}_i$, so that all values lie between 0 and 1. Finally, we concatenate the normalized vectors for each span length n from 1 to N ($N = 3$ in this paper).

We experiment with different ways in which the candidate list can be used to produce feature vectors. The complete feature set is:

1. **top-1 score**: This feature takes the score of the highest scoring candidate c_1 into account.

$$\boldsymbol{f}_{\texttt{type}(c_1)} = \texttt{score}(c_1)$$

2. **top-3 score**: Like the top-1 feature, we additionally create feature vectors for the second and third highest scoring candidates.

3. **top-3 count**: These features are type-wise counts of the top-3 candidates. Instead of adding the score to the appropriate feature element, we add 1.0 to the current value. For a candidate type t, such as LOC, PER or ORG,

$$\boldsymbol{f}_t = \sum_{c \in \{c_1, c_2, c_3\}} 1.0 \times \mathbf{1}_{\texttt{type}(c)=t}$$

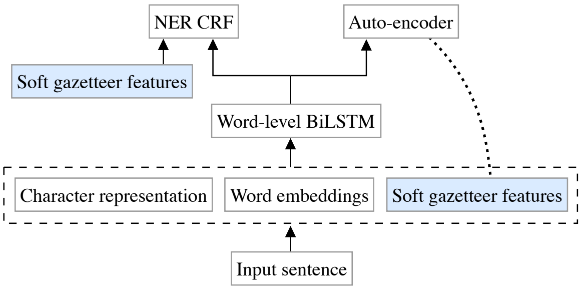

Figure 2: NER Model Architecture. The proposed soft gazetteer features are highlighted and the autoencoder reconstructs these features, indicated by a dotted line.

$\mathbf{1}_{\texttt{type}(c)=t}$ is an indicator function that returns 1.0 when the candidate type is the same as the feature element being updated, 0.0 otherwise.

4. **top-30 count**: This feature computes type-wise counts for the top-30 candidates.

5. **margin**: The margin between the scores of consecutive candidates within the top-4. These features are not computed type-wise. For example the feature value for the margin between the top-2 candidates is,

$$\boldsymbol{f}_{c_1,c_2} = \texttt{score}(c_1) - \texttt{score}(c_2)$$

We experiment with different combinations of these features by concatenating their respective vectors. The concatenated vector is passed through a fully connected neural network layer with a *tanh* non-linearity and then used in the NER model.

4 Named Entity Recognition Model

As our base model, we use the neural CRF model of Ma and Hovy (2016). We adopt the method from Wu et al. (2018) to incorporate linguistic features, which uses an autoencoder loss to help retain information from the hand-crafted features throughout the model (shown in Figure 2). We briefly discuss the model in this section, but encourage readers to refer to the original papers for a more detailed description.

NER objective Given an input sequence, we first calculate a vector representation for each word by concatenating the character representation from a CNN, the word embedding, and the soft gazetteer features. The word representations are then used as input to a bidirectional LSTM (BiLSTM). The hidden states from the BiLSTM and the soft gazetteer features are input to a Conditional Random Field (CRF), which predicts a sequence of NER labels. The training objective, $\mathcal{L}_{CRF}$, is the negative log-likelihood of the gold label sequence.

Lang.	Dataset size	Frac. of NIL	Gaz. size
kin	951	0.41	912
orm	2958	0.36	313
sin	1068	0.29	2738
tir	2202	0.28	92

Table 1: NER dataset and Wikipedia gazetteer sizes.

Autoencoder objective Wu et al. (2018) demonstrate that adding an autoencoder to reconstruct the hand-crafted features leads to improvement in NER performance. The autoencoder takes the hidden states of the BiLSTM as input to a fully connected layer with a sigmoid activation function and reconstructs the features. This forces the BiLSTM to retain information from the features. The cross-entropy loss of the soft gazetteer feature reconstruction is the autoencoder objective, $\mathcal{L}_{AE}$.

Training and inference The training objective is the joint loss: $\mathcal{L}_{CRF} + \mathcal{L}_{AE}$. The losses are given equal weight, as recommended in Wu et al. (2018). During inference, we use Viterbi decoding to obtain the most likely label sequence.

5 Experiments

In this section, we discuss our experiments on four low-resource languages and attempt to answer the following research questions: 1) "Although gazetteer-based features have been proven useful for neural NER on English, is the same true in the low-resource setting?" 2) "Do the proposed soft-gazetteer features outperform the baseline?" 3) "What types of entity mentions benefit from soft gazetteers?" and 4) "Does the knowledge base coverage affect performance?".

5.1 Experimental setup

NER Dataset We experiment on four low-resource languages: Kinyarwanda (kin), Oromo (orm), Sinhala (sin), and Tigrinya (tir). We use the LORELEI dataset (Strassel and Tracey, 2016), which has text from various domains, including news and social media, annotated for the NER task.

Table 1 shows the number of sentences annotated. The data is annotated with four named entity types: locations (LOC), persons (PER), organizations (ORG), and geopolitical entities (GPE). Following the CoNLL-2003 annotation standard, we merge the LOC and GPE types (Tjong Kim Sang and De Meulder, 2003). Note that these datasets are very low-resource, merely 4% to 13% the size of the CoNLL-2003 English dataset.

These sentences are also annotated with entity links to a knowledge base of 11 million entries, which we use *only* to aid our analysis. Of particular interest are "NIL" entity mentions that do not have a corresponding entry in the knowledge base (Blissett and Ji, 2019). The fraction of mentions that are NIL is shown in Table 1.

Gazetteer Data We also compare our method with binary gazetteer features, using entity lists from Wikipedia, the sizes of which are in Table 1.

Implementation Our model is implemented using the DyNet toolkit (Neubig et al., 2017), and we use the same hyperparameters as Ma and Hovy (2016). We use randomly initialized word embeddings since we do not have pretrained vectors for low-resource languages.[2]

Evaluation We perform 10-fold cross-validation for all experiments because of the small size of our datasets. Our primary evaluation metric is span-level named entity F1 score.

5.2 Methods

Baselines We compare with two baselines:

- NoFeat: The CNN-LSTM-CRF model (section 4) without any features.
- BinaryGaz: We use Wikipedia entity lists (Table 1) to create binary gazetteer features.

Soft gazetteer methods We experiment with different candidate retrieval methods designed for low-resource languages. These are trained *only* with small bilingual lexicons from Wikipedia, of similar size as the gazetteers (Table 1).

- WikiMen: The WikiMention method is used in several state-of-the-art EL systems (Sil et al., 2018; Upadhyay et al., 2018), where bilingual Wikipedia links are used to retrieve the appropriate English KB candidates.
- Pivot-based-entity-linking (Zhou et al., 2020): This method encodes entity mentions on the character level using n-gram neural embeddings (Wieting et al., 2016) and computes their similarity with KB entries. We experiment with two variants and follow Zhou et al. (2020) for hyperparameter selection:

  1) PbelSupervised: trained on the small number of bilingual Wikipedia links available in the target low-resource language.

  2) PbelZero: trained on some high-resource language ("the pivot") and transferred to the target language in a zero-shot manner. The transfer languages we use are Swahili for Kinyarwanda, Indonesian for Oromo, Hindi for Sinhala, and Amharic for Tigrinya.

Oracles As an upper-bound on the accuracy, we compare to two artificially strong systems:

- OracleEL: For soft gazetteers, we assume perfect candidate retrieval that always returns the correct KB entry as the top candidate if the mention is non-NIL.
- OracleGaz: We artificially inflate BinaryGaz by augmenting the gazetteer with all the named entities in our dataset.

5.3 Results and Analysis

Results are shown in Table 2. First, comparing BinaryGaz to NoFeat shows that traditional gazetteer features help somewhat, but gains are minimal on languages with fewer available resources.[3] Further, we can see that the proposed soft gazetteer method is effective, some variant thereof achieving the best accuracy on all languages.

For the soft gazetteer method, Table 2 shows the performance with the best performing features (which were determined on a validation set): **top-1** features for Kinyarwanda, Sinhala and Tigrinya,

[2] A note on efficiency: our method involves computing entity linking candidates for each n-gram span in the dataset. The most computationally intensive candidate retrieval method (Pbel, discussed in subsection 5.2) takes ≈1.5 hours to process all spans on a single 1080Ti GPU. Note that this is a preprocessing step and once completed, it does not add any extra computational cost to the NER training process.

[3] We note that binary gazetteer features usually refer to simply using the gazetteer as a lookup (Ratinov and Roth, 2009). However, we also attempt to use WikiMen and Pbel for retrieval, with scores converted to binary values at a threshold of 0.5. BinaryGaz in Table 2 is the best F1 score among these methods–this turns out to be the string lookup for all four languages. This is expected because, for low-resource languages, the other candidate retrieval methods are less precise than their high-resource counterparts. Binary-valued features are not fine-grained enough to be robust to this.

Model	kin	orm	sin	tir
NoFeat	67.16	71.07	49.68	75.44
BinaryGaz	69.05	71.24	54.08	75.84
WikiMen	68.36	71.58	51.34	75.69
PbelSuper.	68.94	71.61	**60.95**	76.49
PbelZero	**69.92**	**71.75**	51.69	**76.99**
OracleEL	82.89	87.69	81.98	89.85
OracleGaz	93.38	94.71	94.00	94.43

Table 2: 10-fold cross-validation NER F1 score. The best performing feature combination is shown here. Bold indicates the best non-oracle system.

and **top-30** features for Oromo. Although Sinhala (sin) has a relatively large gazetteer (Table 1), we observe that directly using the gazetteer as recommended in previous work with BinaryGaz, does not demonstrate strong performance. On the other hand, with the soft gazetteer method and our carefully designed features, PbelSupervised works well for Sinhala (sin) and improves the NER performance. PbelZero is the best method for the other three languages, illustrating how our proposed features can be used to benefit NER by leveraging information from languages closely related to the target. The improvement for Oromo (orm) is minor, likely because of the limited cross-lingual links available for training PbelSupervised and the lack of suitable transfer languages for PbelZero (Rijhwani et al., 2019).

Finally, we find that both OracleGaz and OracleEL improve by a large margin over all non-oracle methods, indicating that there is substantial headroom to improve low-resource NER through either the development of gazetteer resources or the creation of more sophisticated EL methods.

How do soft-gazetteers help? We look at two types of named entity mentions in our dataset that we expect to benefit from the soft gazetteer features: 1) non-NIL mentions with entity links in the KB that can use EL candidate information, and 2) mentions unseen in the training data that have additional information from the features as compared to the baseline. Table 3 shows that the soft gazetteer features increase the recall for both types of mentions by several points.

Knowledge base coverage Table 3 indicates that the soft gazetteer features benefit those entity mentions that are present in the KB. However, our dataset has a significant number of NIL-clustered mentions (Table 1). The ability of our features to add information to NIL mentions is diminished because they do not have a correct candidate in the KB. To measure the effect of KB coverage, we augment the soft gazetteer features with OracleGaz features, applied *only* to the NIL mentions. Large F1 increases in Table 4 indicate that higher KB coverage will likely make the soft gazetteer features more useful, and stresses the importance of developing KBs that cover all entities in the document.

	Non-NIL Recall		Unseen Recall	
Lang.	Baseline	SoftGaz	Baseline	SoftGaz
kin	66.5	73.3	35.4	43.9
orm	72.0	72.8	49.5	51.9
sin	57.3	69.8	20.3	35.3
tir	79.2	80.9	38.9	41.5
Avg.	68.7	**74.2**	36.0	**43.1**

Table 3: Recall for non-NIL mentions and mentions unseen in the training data. SoftGaz represents the best soft gazetteer model as seen in Table 2.

	kin	orm	sin	tir
Orig. KB	69.92	71.71	60.95	76.58
NIL augment	76.28	76.50	70.87	83.07

Table 4: NER F1 score of the best performing soft gazetteer model with the original KB and with augmenting NIL-clustered entity mentions.

6 Conclusion

We present a method to create features for low-resource NER and show its effectiveness on four low-resource languages. Possible future directions include using more sophisticated feature design and combinations of candidate retrieval methods.

Acknowledgements

Shruti Rijhwani is supported by a Bloomberg Data Science Ph.D. Fellowship. Shuyan Zhou is supported by the DARPA Information Innovation Office (I2O) Low Resource Languages for Emergent Incidents (LORELEI) program under Contract No. HR0011-15-C0114. We also thank Samridhi Choudhary for help with the model implementation and Deepak Gopinath for feedback on the paper.

References

Kevin Blissett and Heng Ji. 2019. Cross-lingual NIL entity clustering for low-resource languages. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 20–25, Minneapolis, USA. Association for Computational Linguistics.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Lifu Huang, Heng Ji, and Jonathan May. 2019. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3823–3833, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

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*, pages 142–147.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515.

Minghao Wu, Fei Liu, and Trevor Cohn. 2018. Evaluating the utility of hand-crafted features in sequence labelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2850–2856, Brussels, Belgium. Association for Computational Linguistics.

Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association of Computational Linguistics*.