Global Attention for Name Tagging

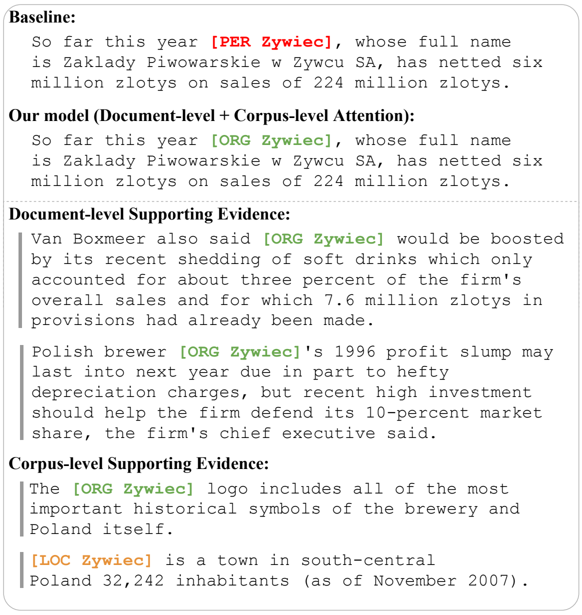
Boliang Zhang, Spencer Whitehead, Lifu Huang and Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
{zhangb8,whites5,huangl7,jih}@rpi.edu

Abstract

Many name tagging approaches use local contextual information with much success, but fail when the local context is ambiguous or limited. We present a new framework to improve name tagging by utilizing local, document-level, and corpus-level contextual information. We retrieve document-level context from other sentences within the same document and corpus-level context from sentences in other topically related documents. We propose a model that learns to incorporate document-level and corpus-level contextual information alongside local contextual information via global attentions, which dynamically weight their respective contextual information, and gating mechanisms, which determine the influence of this information. Extensive experiments on benchmark datasets show the effectiveness of our approach, which achieves state-of-the-art results for Dutch, German, and Spanish on the CoNLL-2002 and CoNLL-2003 datasets.[1].

1 Introduction

Name tagging, the task of automatically identifying and classifying named entities in text, is often posed as a sentence-level sequence labeling problem where each token is labeled as being part of a name of a certain type (*e.g.*, location) or not (Chinchor and Robinson, 1997; Tjong Kim Sang and De Meulder, 2003). When labeling a token, local context (*i.e.*, surrounding tokens) is crucial because the context gives insight to the semantic meaning of the token. However, there are many instances in which the local context is ambiguous or lacks sufficient content. For example, in Figure 1, the query sentence discusses "Zywiec" selling a product and profiting from these sales, but the local contextual information is ambiguous as more than one entity type could be involved in a sale. As a result, the baseline model mistakenly tags "Zywiec" as a person (PER) instead of the correct tag, which is organization (ORG). If the model has access to supporting evidence that provides additional, clearer contextual information, then the model may use this information to correct the mistake given the ambiguous local context.



Baseline:

So far this year [PER Zywiec], whose full name is Zaklady Piwowarskie w Zywcu SA, has netted six million zlotys on sales of 224 million zlotys.

Our model (Document-level + Corpus-level Attention):

So far this year [ORG Zywiec], whose full name is Zaklady Piwowarskie w Zywcu SA, has netted six million zlotys on sales of 224 million zlotys.

Document-level Supporting Evidence:

Van Boxmeer also said [ORG Zywiec] would be boosted by its recent shedding of soft drinks which only accounted for about three percent of the firm's overall sales and for which 7.6 million zlotys in provisions had already been made.

Polish brewer [ORG Zywiec]'s 1996 profit slump may last into next year due in part to hefty depreciation charges, but recent high investment should help the firm defend its 10-percent market share, the firm's chief executive said.

Corpus-level Supporting Evidence:

The [ORG Zywiec] logo includes all of the most important historical symbols of the brewery and Poland itself.

[LOC Zywiec] is a town in south-central Poland 32,242 inhabitants (as of November 2007).

Figure 1: Example from the baseline and our model with some supporting evidence.

Additional context may be found from other sentences in the same document as the query sentence (**document-level**). In Figure 1, the sentences in the document-level supporting evidence provide clearer clues to tag "Zywiec" as ORG, such as the references to "Zywiec" as a "firm". A concern of leveraging this information is the amount of noise that is introduced. However, across all the

[1] The programs are publicly available for research purpose: https://github.com/boliangz/global_attention_ner

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 86–96
Brussels, Belgium, October 31 - November 1, 2018.

data in our experiments (Section 3.1), we find that an average of 35.43% of named entity mentions in each document are repeats and, when a mention appears more than once in a document, an average of 98.78% of these mentions have the same type. Consequently, one may use the document-level context to overcome the ambiguities of the local context while introducing little noise.

Although a significant amount of named entity mentions are repeated, 64.57% of the mentions are unique. In such cases, the sentences at the document-level cannot serve as a source of additional context. Nevertheless, one may find additional context from sentences in other documents in the corpus (**corpus-level**). Figure 1 shows some of the corpus-level supporting evidence for "`Zywiec`". In this example, similar to the document-level supporting evidence, the first sentence in this corpus-level evidence discusses the branding of "`Zywiec`", corroborating the ORG tag. Whereas the second sentence introduces noise because it has a different topic than the current sentence and discusses the Polish town named "`Zywiec`", one may filter these noisy contexts, especially when the noisy contexts are accompanied by clear contexts like the first sentence.

We propose to utilize local, document-level, and corpus-level contextual information to improve name tagging. Generally, we follow the *one sense per discourse* hypothesis introduced by Yarowsky (2003). Some previous name tagging efforts apply this hypothesis to conduct majority voting for multiple mentions with the same name string in a discourse through a cache model (Florian et al., 2004) or post-processing (Hermjakob et al., 2017). However, these rule-based methods require manual tuning of thresholds. Moreover, it's challenging to explicitly define the scope of discourse. We propose a new neural network framework with global attention to tackle these challenges. Specifically, for each token in a query sentence, we propose to retrieve sentences that contain the same token from the document-level and corpus-level contexts (*e.g.*, document-level and corpus-level supporting evidence for "`Zywiec`" in Figure 1). To utilize this additional information, we propose a model that, first, produces representations for each token that encode the local context from the query sentence as well as the document-level and corpus-level contexts from the retrieved sentences. Our model uses a *document-level attention* and *corpus-level attention* to dynamically weight the document-level and corpus-level contextual representations, emphasizing the contextual information from each level that is most relevant to the local context and filtering noise such as the irrelevant information from the mention "`[LOC Zywiec]`" in Figure 1. The model learns to balance the influence of the local, document-level, and corpus-level contextual representations via gating mechanisms. Our model predicts a tag using the local, gated-attentive document-level, and gated-attentive corpus-level contextual representations, which allows our model to predict the correct tag, ORG, for "`Zywiec`" in Figure 1.

The major contributions of this paper are: First, we propose to use multiple levels of contextual information (local, document-level, and corpus-level) to improve name tagging. Second, we present two new attentions, document-level and corpus-level, which prove to be effective at exploiting extra contextual information and achieve the state-of-the-art.

2 Model

We first introduce our baseline model. Then, we enhance this baseline model by adding document-level and corpus-level contextual information to the prediction process via our document-level and corpus-level attention mechanisms, respectively.

2.1 Baseline

We consider name tagging as a sequence labeling problem, where each token in a sequence is tagged as the beginning (B), inside (I) or outside (O) of a name mention. The tagged names are then classified into predefined entity types. In this paper, we only use the person (PER), organization (ORG), location (LOC), and miscellaneous (MISC) types, which are the predefined types in CoNLL-02 and CoNLL-03 name tagging dataset (Tjong Kim Sang and De Meulder, 2003).

Our baseline model has two parts: 1) Encoding the sequence of tokens by incorporating the preceding and following contexts using a bi-directional long short-term memory (Bi-LSTM) (Graves et al., 2013), so each token is assigned a local contextual embedding. Here, following Ma and Hovy (2016a), we use the concatenation of pre-trained word embeddings and character-level word representations composed by a convolutional neural network (CNN) as input to the Bi-LSTM. 2) Using a Conditional Random Fields (CRFs) output layer to render predictions for each token, which can efficiently capture dependencies among name tags (*e.g.*, "I-LOC" cannot follow "B-ORG").

The Bi-LSTM CRF network is a strong baseline due to its remarkable capability of modeling contextual information and label dependencies. Many recent efforts combine the Bi-LSTM CRF network with language modeling (Liu et al., 2017; Peters et al., 2017, 2018) to boost the name tagging performance. However, they still suffer from the limited contexts within individual sequences. To overcome this limitation, we introduce two attention mechanisms to incorporate document-level and corpus-level supporting evidence.

2.2 Document-level Attention

Many entity mentions are tagged as multiple types by the baseline approach within the same document due to ambiguous contexts (14.43% of the errors in English, 18.55% in Dutch, and 17.81% in German). This type of error is challenging to address as most of the current neural network based approaches focus on evidence within the sentence when making decisions. In cases where a sentence is short or highly ambiguous, the model may either fail to identify names due to insufficient information or make wrong decisions by using noisy context. In contrast, a human in this situation may seek additional evidence from other sentences within the same document to improve judgments.

In Figure 1, the baseline model mistakenly tags "`Zywiec`" as PER due to the ambiguous context "`whose full name is...`", which frequently appears around a person's name. However, contexts from other sentences in the same document containing "`Zywiec`" (*e.g.*, s_q and s_r in Figure 2), such as "`'s 1996 profit...`" and "`would be boosted by its recent shedding...`", indicate that "`Zywiec`" ought to be tagged as ORG. Thus, we incorporate the document-level supporting evidence with the following attention mechanism (Bahdanau et al., 2015).

Formally, given a document $D = \{s_1, s_2, ...\}$, where $s_i = \{w_{i1}, w_{i2}, ...\}$ is a sequence of words, we apply a Bi-LSTM to each word in s_i, generating local contextual representations $h_i = \{\mathbf{h}_{i1}, \mathbf{h}_{i2}, ...\}$. Next, for each w_{ij}, we retrieve the sentences in the document that contain w_{ij} (*e.g.*, s_q and s_r in Figure 2) and select the local contextual representations of w_{ij} from these sentences as supporting evidence, $\tilde{h}_{ij} = \{\tilde{\mathbf{h}}_{ij}^1, \tilde{\mathbf{h}}_{ij}^2, ...\}$ (*e.g.*, $\tilde{\mathbf{h}}_{qj}$ and $\tilde{\mathbf{h}}_{rk}$ in Figure 2), where h_{ij} and $\tilde{h}_{ij}$ are obtained with the same Bi-LSTM. Since each representation in the supporting evidence is not equally valuable to the final prediction, we apply an attention mechanism to weight the contextual representations of the supporting evidence:

$$e_{ij}^k = \mathbf{v}^\top \tanh\left(W_h \mathbf{h}_{ij} + W_{\tilde{h}} \tilde{\mathbf{h}}_{ij}^k + \mathbf{b}_e\right) ,$$

$$\alpha_{ij}^k = \text{Softmax}\left(e_{ij}^k\right) ,$$

where $\mathbf{h}_{ij}$ is the local contextual representation of word j in sentence s_i and $\tilde{\mathbf{h}}_{ij}^k$ is the k-th supporting contextual representation. W_h, $W_{\tilde{h}}$ and $\mathbf{b}_e$ are learned parameters. We compute the weighted average of the supporting representations by

$$\tilde{\mathbf{H}}_{ij} = \sum_{k=1} \alpha_{ij}^k \tilde{\mathbf{h}}_{ij}^k ,$$

where $\tilde{\mathbf{H}}_{ij}$ denotes the contextual representation of the supporting evidence for w_{ij}.

For each word w_{ij}, its supporting evidence representation, $\tilde{\mathbf{H}}_{ij}$, provides a summary of the other contexts where the word appears. Though this evidence is valuable to the prediction process, we must mitigate the influence of the supporting evidence since the prediction should still be made primarily based on the query context. Therefore, we apply a gating mechanism to constrain this influence and enable the model to decide the amount of the supporting evidence that should be incorporated in the prediction process, which is given by

$$\mathbf{r}_{ij} = \sigma(W_{\tilde{H},r} \tilde{\mathbf{H}}_{ij} + W_{h,r} \mathbf{h}_{ij} + \mathbf{b}_r) ,$$

$$\mathbf{z}_{ij} = \sigma(W_{\tilde{H},z} \tilde{\mathbf{H}}_{ij} + W_{h,z} \mathbf{h}_{ij} + \mathbf{b}_z) ,$$

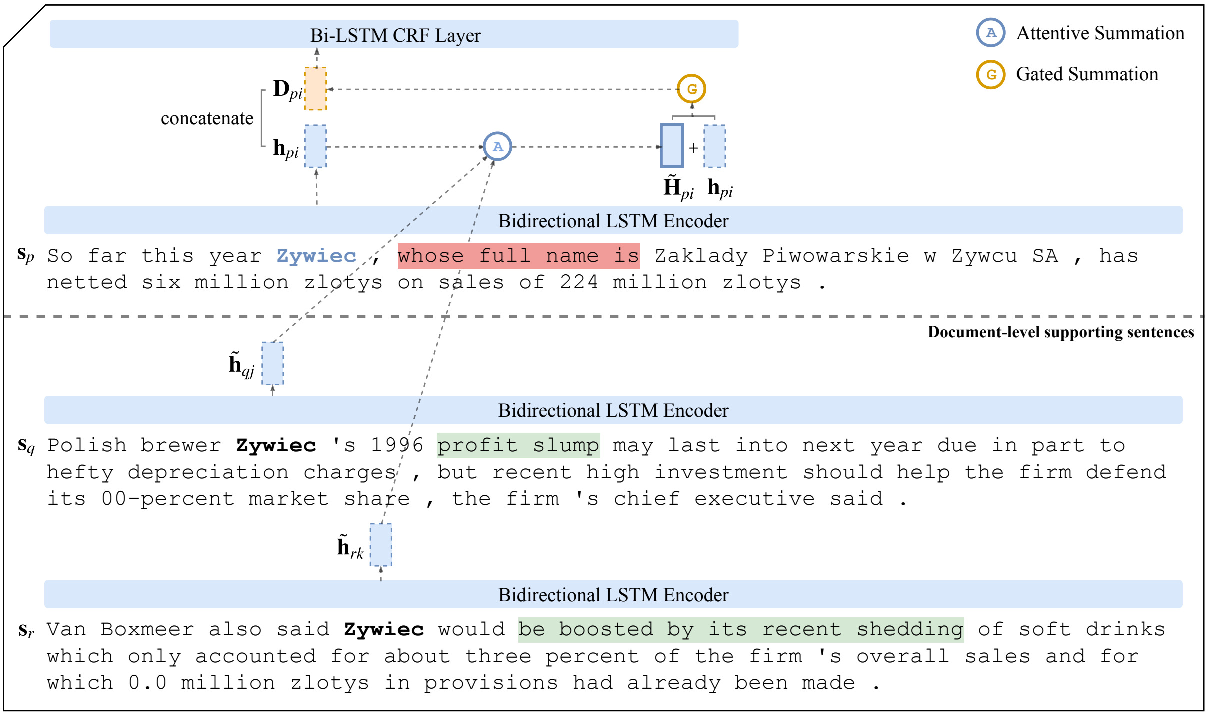
$$\mathbf{g}_{ij} = \tanh(W_{h,g} \mathbf{h}_{ij} + \mathbf{z}_{ij} \odot (W_{\tilde{H},g} \tilde{\mathbf{H}}_{ij} + \mathbf{b}_g)) ,$$

$$\mathbf{D}_{ij} = \mathbf{r}_{ij} \odot \mathbf{h}_{ij} + (1 - \mathbf{r}_{ij}) \odot \mathbf{g}_{ij} ,$$

where all W, $\mathbf{b}$ are learned parameters and $\mathbf{D}_{ij}$ is the gated supporting evidence representation for w_{ij}.

2.3 Topic-aware Corpus-level Attention

The document-level attention fails to generate supporting evidence when the name appears only once in a single document. In such situations, we analogously select supporting sentences from the entire corpus. Unfortunately, different from


Bi-LSTM CRF Layer
Attentive Summation
Gated Summation
concatenate
$\mathbf{D}_{pi}$
$\mathbf{h}_{pi}$
A
G
+
$\tilde{\mathbf{H}}_{pi}$
$\mathbf{h}_{pi}$
Bidirectional LSTM Encoder
$\mathbf{s}_p$
Document-level supporting sentences
$\tilde{\mathbf{h}}_{qj}$
Bidirectional LSTM Encoder
$\mathbf{s}_q$
$\tilde{\mathbf{h}}_{rk}$
Bidirectional LSTM Encoder
$\mathbf{s}_r$

$\mathbf{s}_p$ So far this year **Zywiec** , whose full name is Zaklady Piwowarskie w Zywcu SA , has netted six million zlotys on sales of 224 million zlotys .

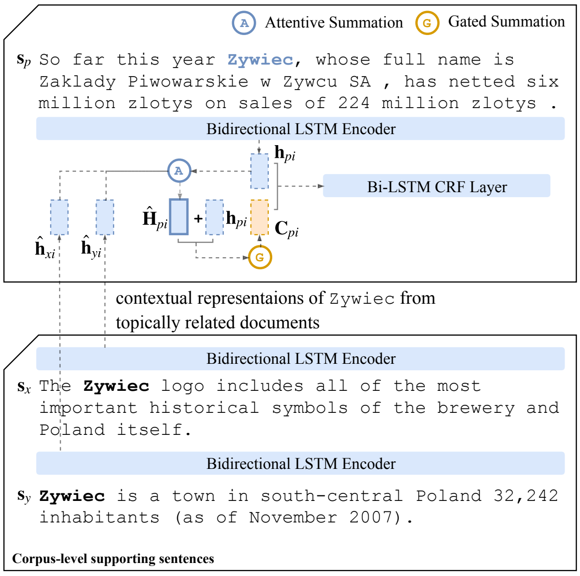
$\mathbf{s}_q$ Polish brewer **Zywiec** 's 1996 profit slump may last into next year due in part to hefty depreciation charges , but recent high investment should help the firm defend its 00-percent market share , the firm 's chief executive said .

$\mathbf{s}_r$ Van Boxmeer also said **Zywiec** would be boosted by its recent shedding of soft drinks which only accounted for about three percent of the firm 's overall sales and for which 0.0 million zlotys in provisions had already been made .

Figure 2: Document-level Attention Architecture. (Within-sequence context in red incorrectly indicates the name as PER, and document-level context in green correctly indicates the name as ORG.)

the sentences that are naturally topically relevant within the same documents, the supporting sentences from the other documents may be about distinct topics or scenarios, and identical phrases may refer to various entities with different types, as in the example in Figure 1. To narrow down the search scope from the entire corpus and avoid unnecessary noise, we introduce a topic-aware corpus-level attention which clusters the documents by topic and carefully selects topically related sentences to use as supporting evidence.

We first apply Latent Dirichlet allocation (LDA) (Blei et al., 2003) to model the topic distribution of each document and separate the documents into N clusters based on their topic distributions.[2] As in Figure 3, we retrieve supporting sentences for each word, such as "Zywiec", from the topically related documents and employ another attention mechanism (Bahdanau et al., 2015) to the supporting contextual representations, $\hat{h}_{ij} = \{\hat{\mathbf{h}}^1_{ij}, \hat{\mathbf{h}}^2_{ij}, ...\}$ (e.g., $\tilde{\mathbf{h}}_{xi}$ and $\tilde{\mathbf{h}}_{yi}$ in Figure 3). This yields a weighted contextual representation of the corpus-level supporting evidence, $\hat{\mathbf{H}}_{ij}$, for each w_{ij}, which is similar to the document-level supporting evidence representation, $\tilde{\mathbf{H}}_{ij}$, described in section 2.2. We use another gating mechanism to combine $\hat{\mathbf{H}}_{ij}$ and the local contextual representation, $\mathbf{h}_{ij}$, to obtain the corpus-level gated supporting evidence representation, $\mathbf{C}_{ij}$, for each w_{ij}.



$\mathbf{s}_p$ So far this year **Zywiec**, whose full name is Zaklady Piwowarskie w Zywcu SA , has netted six million zlotys on sales of 224 million zlotys .

$\mathbf{s}_x$ The **Zywiec** logo includes all of the most important historical symbols of the brewery and Poland itself.

$\mathbf{s}_y$ **Zywiec** is a town in south-central Poland 32,242 inhabitants (as of November 2007).

Figure 3: Corpus-level Attention Architecture.

[2] $N = 20$ in our experiments.

2.4 Tag Prediction

For each word w_{ij} of sentence s_i, we concatenate its local contextual representation $\mathbf{h}_{ij}$, document-level gated supporting evidence representation $\mathbf{D}_{ij}$, and corpus-level gated supporting evidence representation $\mathbf{C}_{ij}$ to obtain its final representation. This representation is fed to another Bi-LSTM to further encode the supporting evidence and local contextual features into an unified representation, which is given as input to an affine-CRF layer for label prediction.

3 Experiments

3.1 Dataset

We evaluate our methods on the CoNLL-2002 and CoNLL-2003 name tagging datasets (Tjong Kim Sang and De Meulder, 2003). The CoNLL-2002 dataset contains name tagging annotations for Dutch (NLD) and Spanish (ESP), while the CoNLL-2003 dataset contains annotations for English (ENG) and German (DEU). Both datasets have four pre-defined name types: person (PER), organization (ORG), location (LOC) and miscellaneous (MISC).[3]

Code	Train	Dev.	Test
NLD	202,931 (13,344)	37,761 (2,616)	68,994 (3,941)
ESP	264,715 (18,797)	52,923 (4,351)	51,533 (3,558)
ENG	204,567 (23,499)	51,578 (5,942)	46,666 (5,648)
DEU	207,484 (11,651)	51,645 (4,669)	52,098 (3,602)

Table 1: # of tokens in name tagging datasets statistics. # of names is given in parentheses.

We select at most four document-level supporting sentences and five corpus-level supporting sentences.[4] Since the document-level attention method requires input from each individual document, we do not evaluate it on the CoNLL-2002 Spanish dataset which lacks document delimiters. We still evaluate the corpus-level attention on the Spanish dataset by randomly splitting the dataset into documents (30 sentences per document). Although randomly splitting the sentences does not yield perfect topic modeling clusters, experiments show the corpus-level attention still outperforms the baseline (Section 3.3).

Hyper-parameter	Value
CharCNN Filter Number	25
CharCNN Filter Widths	[2, 3, 4]
Lower Bi-LSTM Hidden Size	100
Lower Bi-LSTM Dropout Rate	0.5
Upper Bi-LSTM Hidden Size	100
Learning Rate	0.005
Batch Size	N/A*
Optimizer	SGD (Bottou, 2010)

* Each batch is a document. The batch size varies as the different document length.

Table 2: Hyper-parameters.

3.2 Experimental Setup

For word representations, we use 100-dimensional pre-trained word embeddings and 25-dimensional randomly initialized character embeddings. We train word embeddings using the word2vec package.[5] English embeddings are trained on the English Giga-word version 4, which is the same corpus used in (Lample et al., 2016). Dutch, Spanish, and German embeddings are trained on corresponding Wikipedia articles (2017-12-20 dumps). Word embeddings are fine-tuned during training.

Table 2 shows our hyper-parameters. For each model with an attention, since the Bi-LSTM encoder must encode the local, document-level, and/or corpus-level contexts, we pre-train a Bi-LSTM CRF model for 50 epochs, add our document-level attention and/or corpus-level attention, and then fine-tune the augmented model. Additionally, Reimers and Gurevych (2017) report that neural models produce different results even with same hyper-parameters due to the variances in parameter initialization. Therefore, we run each model ten times and report the mean as well as the maximum F1 scores.

3.3 Performance Comparison

We compare our methods to three categories of baseline name tagging methods:

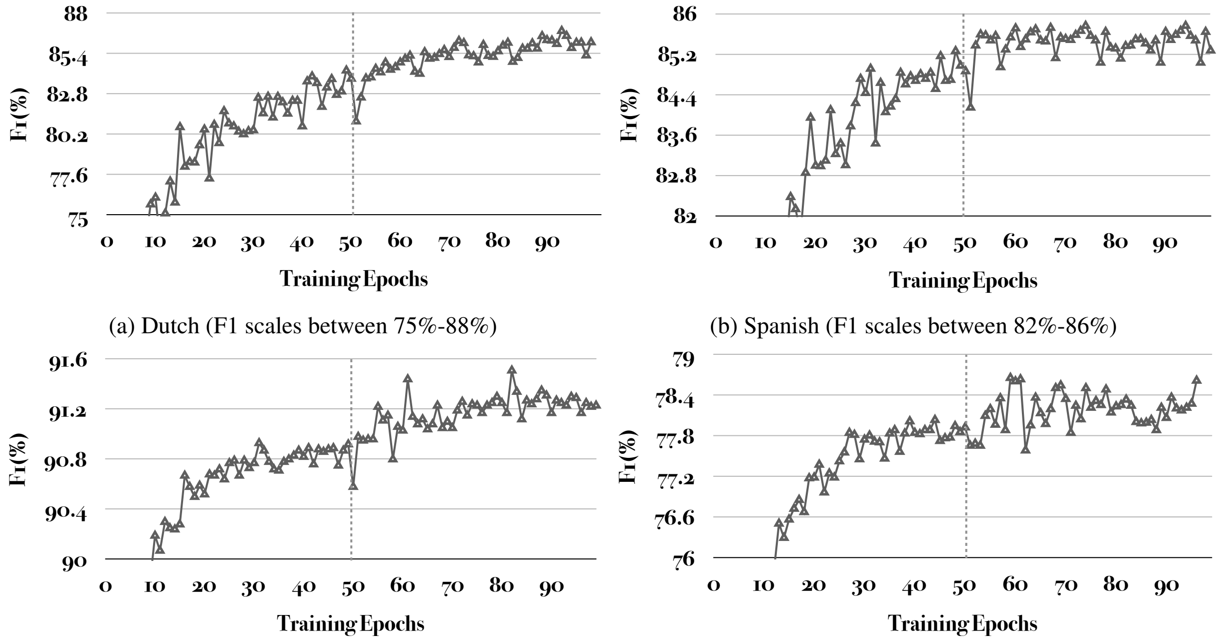
- **Vanilla Name Tagging** Without any additional resources and supervision, the current state-of-the-art name tagging model is the Bi-LSTM-CRF network reported by Lample et al. (2016) and Ma and Hovy (2016b), whose difference lies in using a LSTM or CNN to encode characters. Our methods fall in this category.

- **Multi-task Learning** Luo et al. (2015); Yang et al. (2017) apply multi-task learning to boost

[3] The miscellaneous category consists of names that do not belong to the other three categories.

[4] Both numbers are tuned from 1 to 10 and selected when the model performs best on the development set.

[5] https://github.com/tmikolov/word2vec



(a) Dutch (F1 scales between 75%-88%)

(b) Spanish (F1 scales between 82%-86%)

(c) English (F1 scales between 90%-91.6%)

(d) German (F1 scales between 76%-79%)

Figure 4: Average F1 score for each epoch of the ten runs of our model with both document-level and corpus-level attentions. Epochs 1-50 are the pre-training phase and 51-100 are the fine-tuning phase.

name tagging performance by introducing additional annotations from related tasks such as entity linking and part-of-speech tagging.

- **Join-learning with Language Model** Peters et al. (2017); Liu et al. (2017); Peters et al. (2018) leverage a pre-trained language model on a large external corpus to enhance the semantic representations of words in the local corpus. Peters et al. (2018) achieve a high score on the CoNLL-2003 English dataset using a giant language model pre-trained on a 1 Billion Word Benchmark (Chelba et al., 2013).

Table 3 presents the performance comparison among the baselines, the aforementioned state-of-the-art methods, and our proposed methods. Adding only the document-level attention offers a F1 gain of between 0.37% and 1.25% on Dutch, English, and German. Similarly, the addition of the corpus-level attention yields a F1 gain between 0.46% to 1.08% across all four languages. The model with both attentions outperforms our baseline method by 1.60%, 0.56%, and 0.79% on Dutch, English, and German, respectively. Using a paired t-test between our proposed model and the baselines on 10 randomly sampled subsets, we find that the improvements are statistically significant ($p \leq 0.015$) for all settings and all languages.

By incorporating the document-level and corpus-level attentions, we achieve state-of-the-art performance on the Dutch (NLD), Spanish (ESP) and German (DEU) datasets. For English, our methods outperform the state-of-the-art methods in the "Vanilla Name Tagging" category. Since the document-level and corpus-level attentions introduce redundant and topically related information, our models are compatible with the language model enhanced approaches. It is interesting to explore the integration of these two methods, but we leave this to future explorations.

Figure 4 presents, for each language, the learning curves of the full models (*i.e.*, with both document-level and corpus-level attentions). The learning curve is computed by averaging the F1 scores of the ten runs at each epoch. We first pre-train a baseline Bi-LSTM CRF model from epoch 1 to 50. Then, starting at epoch 51, we incorporate the document-level and corpus-level attentions to fine-tune the entire model. As shown in Figure 4, when adding the attentions at epoch 51, the F1 score drops significantly as new parameters are introduced to the model. The model gradually adapts to the new information, the F1 score rises, and the full model eventually outperforms the pre-trained model. The learning curves strongly prove the effectiveness of our proposed methods.

Code	Model	F1 (%)	
NLD	(Gillick et al., 2015)	reported	82.84
	(Lample et al., 2016)	reported	81.74
	(Yang et al., 2017)	reported	85.19
	Our Baseline	mean	85.43
		max	85.80
	Doc-lvl Attention	mean	86.82
		max	87.05
	Corpus-lvl Attention	mean	86.41
		max	86.88
	Both	mean	87.14
		max	**87.40**
		Δ	**+1.60**
ESP	(Gillick et al., 2015)	reported	82.95
	(Lample et al., 2016)	reported	85.75
	(Yang et al., 2017)	reported	85.77
	Our Baseline	mean	85.33
		max	85.51
	Corpus-lvl Attention	mean	85.77
		max	**86.01**
		Δ	**+0.50**
ENG	(Luo et al., 2015)	reported	91.20
	(Lample et al., 2016)	reported	90.94
	(Ma and Hovy, 2016b)	reported	91.21
	(Liu et al., 2017)	reported	91.35
	(Peters et al., 2017)	reported	91.93
	(Peters et al., 2018)	reported	**92.22**
	Our Baseline	mean	90.97
		max	91.23
	Doc-lvl Attention	mean	91.43
		max	91.60
	Corpus-lvl Attention	mean	91.41
		max	91.71
	Both	mean	91.64
		max	**91.81**
		Δ	**+0.58**
DEU	(Gillick et al., 2015)	reported	76.22
	(Lample et al., 2016)	reported	78.76
	Our Baseline	mean	78.15
		max	78.42
	Doc-lvl Attention	mean	78.90
		max	79.19
	Corpus-lvl Attention	mean	78.53
		max	78.88
	Both	mean	78.83
		max	**79.21**
		Δ	**+0.79**

Table 3: Performance of our methods versus the baseline and state-of-the-art models.

We also compare our approach with a simple rule-based propagation method, where we use token-level majority voting to make labels consistent on document-level and corpus-level. The score of document-level propagation on English is 90.21% (F1), and the corpus-level propagation is 89.02% which are both lower than the BiLSTM-CRF baseline 90.97%.

3.4 Qualitative Analysis

Table 5 compares the name tagging results from the baseline model and our best models. All examples are selected from the development set.

In the Dutch example, “`Granada`” is the name of a city in Spain, but also the short name of “`Granada Media`”. Without ORG related context, “`Granada`” is mistakenly tagged as LOC by the baseline model. However, the document-level and corpus-level supporting evidence retrieved by our method contains the ORG name “`Granada Media`”, which strongly indicates “`Granada`” to be an ORG in the query sentence. By adding the document-level and corpus-level attentions, our model successfully tags “`Granada`” as ORG.

In example 2, the OOV word “`Kaczmarek`” is tagged as ORG in the baseline output. In the retrieved document-level supporting sentences, PER related contextual information, such as the pronoun “`he`”, indicates “`Kaczmarek`” to be a PER. Our model correctly tags “`Kaczmarek`” as PER with the document-level attention.

In the German example, “`Grünen`” (Greens) is an OOV word in the training set. The character embedding captures the semantic meaning of the stem “`Grün`” (Green) which is a common non-name word, so the baseline model tags “`Grünen`” as O (outside of a name). In contrast, our model makes the correct prediction by incorporating the corpus-level attention because in the related sentence from the corpus “`Bundesvorstandes der Grünen`” (Federal Executive of the Greens) indicates “`Grünen`” to be a company name.

3.5 Remaining Challenges

By investigating the remaining errors, most of the named entity type inconsistency errors are eliminated, however, a few new errors are introduced due to the model propagating labels from negative instances to positive ones. Figure 5 presents a negative example, where our model, being influenced by the prediction “`[B-ORG Indianapolis]`” in the supporting sentence, incorrectly predicts “`Indianapolis`” as ORG in the query sentence. A potential solution is to apply sentence classification (Kim, 2014; Ji and Smith, 2017) to the documents, divide the document into fine-grained clusters of sentences, and select supporting sentences within the same cluster.

In morphologically rich languages, words may have many variants. When retrieving supporting evidence, our exact query word match criterion misses potentially useful supporting sentences that contain variants of the word. Normalization and

#1 Dutch	
Baseline	[B-LOC Granada] overwoog vervolgens een bod op Carlton uit te brengen, maar daar ziet het concern nu van af. *Granada then considered issuing a bid for Carlton, but the concern now sees it.*
Our model	[B-ORG Granada] overwoog vervolgens een bod op Carlton uit te brengen, maar daar ziet het concern nu van af.
D-lvl sentences	[B-ORG Granada] [I-ORG Media] neemt belangen in United News. *Granada Media takes interests in United News.*
C-lvl sentences	Het Britse concern [B-ORG Granada] [I-ORG Media] heeft voor 1,75 miljard pond sterling (111 miljard Belgische frank) aandelen gekocht van United News Media. *The British group Granada Media has bought shares of GBP 1.75 trillion (111 billion Belgian francs) from United News Media.*
#2 English	
Baseline	Initially Poland offered up to 75 percent of Ruch but in March [ORG Kaczmarek] cancelled the tender and offered a minority stake with an option to increase the equity.
Our model	Initially Poland offered up to 75 percent of Ruch but in March [PER Kaczmarek] cancelled the tender and offered a minority stake with an option to increase the equity.
D-lvl sentences	[PER Kaczmarek] said in May he was unhappy that only one investor ended up bidding for Ruch.
#3 German	
Baseline	Diese Diskussion werde ausschlaggebend sein für die Stellungnahme der Grünen in dieser Frage. *This discussion will be decisive for the opinion of the Greens on this question.*
Our model	Diese Diskussion werde ausschlaggebend sein für die Stellungnahme der [B-ORG Grünen] in dieser Frage.
C-lvl sentences	Auch das Mitglied des Bundesvorstandes der [B-ORG Grünen], Helmut Lippelt, sprach sich für ein Berufsheer au. *Helmut Lippelt, a member of the Federal Executive of the Greens, also called for a professional army.*
#4 Negative Example	
Reference	[B-LOC Indianapolis] 1996-12-06
Our model	[B-ORG Indianapolis] 1996-12-06
D-lvl sentence	The injury-plagued [B-ORG Indianapolis] [I-ORG Colts] lost another quarterback on Thursday but last year's AFC finalists rallied together to shoot down the Philadelphia Eagles 37-10 in a showdown of playoff contenders.

* D-lvl sentences: document-level supporting sentences.

* C-lvl sentences: corpus-level supporting sentences.

Figure 5: Comparison of name tagging results between the baseline and our methods.

morphological analysis can be applied in this case to help fetch supporting sentences.

4 Related Work

Name tagging methods based on sequence labeling have been extensively studied recently. Huang et al. (2015) and Lample et al. (2016) proposed a neural architecture consisting of a bi-directional long short-term memory network (Bi-LSTM) encoder and a conditional random field (CRF) output layer (Bi-LSTM CRF). This architecture has been widely explored and demonstrated to be effective for sequence labeling tasks. Efforts incorporated character level compositional word embeddings, language modeling, and CRF re-ranking into the Bi-LSTM CRF architecture which improved the performance (Ma and Hovy, 2016a; Liu et al., 2017; Sato et al., 2017; Peters et al., 2017, 2018). Similar to these studies, our approach is also based on a Bi-LSTM CRF architecture. However, considering the limited contexts within each individual sequence, we design two attention mechanisms to further incorporate topically related contextual information on both the document-level and corpus-level.

There have been efforts in other areas of information extraction to exploit features beyond individual sequences. Early attempts (Mikheev et al., 1998; Mikheev, 2000) on MUC-7 name tagging dataset used document centered approaches. A number of approaches explored document-level features (*e.g.*, temporal and co-occurrence patterns) for event extraction (Chambers and Jurafsky, 2008; Ji and Grishman, 2008; Liao and Grishman, 2010; Do et al., 2012; McClosky and Manning, 2012; Berant et al., 2014; Yang and Mitchell, 2016). Other approaches leveraged features from external resources (*e.g.*, Wiktionary or FrameNet) for low resource name tagging and event extraction (Li et al., 2013; Huang et al., 2016; Liu et al., 2016; Zhang et al., 2016; Cotterell and Duh, 2017; Zhang et al., 2017; Huang et al., 2018). Yaghoobzadeh and Schütze (2016) aggregated corpus-level contextual information of each entity to predict its type and Narasimhan et al. (2016) incorporated contexts from external information sources (*e.g.*, the documents that contain the desired information) to resolve ambiguities. Compared with these studies, our work incorporates both document-level and corpus-level contextual information with attention mechanisms, which is a more advanced and efficient way to capture meaningful additional features. Additionally, our model is able to learn how to regulate the influence of the information outside the local context using gating mechanisms.

5 Conclusions and Future Work

We propose document-level and corpus-level attentions for name tagging. The document-level attention retrieves additional supporting evidence from other sentences within the document to enhance the local contextual information of the query word. When the query word is unique in the document, the corpus-level attention searches for topically related sentences in the corpus. Both attentions dynamically weight the retrieved contextual information and emphasize the information most relevant to the query context. We present gating mechanisms that allow the model to regulate the influence of the supporting evidence on the predictions. Experiments demonstrate the effectiveness of our approach, which achieves state-of-the-art results on benchmark datasets.

We plan to apply our method to other tasks, such as event extraction, and explore integrating language modeling into this architecture to further boost name tagging performance.

Acknowledgments

This work was supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014, LORELEI Program No. HR0011-15-C-0115, Air Force No. FA8650-17-C-7715, NSF IIS-1523198 and U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 2010 International Conference on Computational Statistics*.

Nathanael Chambers and Dan Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.

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*.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2015. Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*.

Alan Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*.

Ulf Hermjakob, Qiang Li, Daniel Marcu, Jonathan May, Sebastian J. Mielke, Nima Pourdamghani, Michael Pust, Xing Shi, Kevin Knight, Tomer Levinboim, Kenton Murray, David Chiang, Boliang Zhang, Xiaoman Pan, Di Lu, Ying Lin, and Heng Ji. 2017. Incident-driven machine translation and name tagging for low-resource languages. *Machine Translation*, pages 1–31.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. *arXiv preprint arXiv:1804.07875*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. *Proceedings of the 2008 Annual Meeting of the Association for Computational Linguistics*.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. *Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2017. Empower sequence labeling with task-aware neural language model.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging framenet to improve automatic event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Xuezhe Ma and Eduard Hovy. 2016a. End-to-end sequence labeling via bi-directional lstm-cnns-crf.

Xuezhe Ma and Eduard Hovy. 2016b. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

David McClosky and Christopher D Manning. 2012. Learning constraints for consistent timeline extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Andrei Mikheev. 2000. Document centered approach to text normalization. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 136–143. ACM.

Andrei Mikheev, Claire Grover, and Marc Moens. 1998. Description of the ltg system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. *arXiv preprint arXiv:1603.07954*.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Motoki Sato, Hiroyuki Shindo, Ikuya Yamada, and Yuji Matsumoto. 2017. Segment-level neural conditional random fields for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

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 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Corpus-level fine-grained entity typing using contextual information. *arXiv preprint arXiv:1606.07901*.

Bishan Yang and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. *arXiv preprint arXiv:1609.03632*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

David Yarowsky. 2003. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. ACL1995*.

Boliang Zhang, Di Lu, Xiaoman Pan, Ying Lin, Halidanmu Abudukelimu, Heng Ji, and Kevin Knight. 2017. Embracing non-traditional linguistic resources for low-resource language name tagging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.