Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve

Oshin Agarwal
University of Pennsylvania
Department of Computer and
Information Science
oagarwal@seas.upenn.edu

Yinfei Yang
Google Research
yinfeiy@google.com

Byron C. Wallace
Northeastern University
Khoury College of Computer Sciences
b.wallace@northeastern.edu

Ani Nenkova
University of Pennsylvania
Department of Computer and
Information Science
nenkova@seas.upenn.edu

Named entity recognition systems achieve remarkable performance on domains such as English news. It is natural to ask: What are these models actually learning to achieve this? Are they merely memorizing the names themselves? Or are they capable of interpreting the text and inferring the correct entity type from the linguistic context? We examine these questions by contrasting the performance of several variants of architectures for named entity recognition, with some provided only representations of the context as features. We experiment with GloVe-based BiLSTM-CRF as well as BERT. We find that context does influence predictions, but the main factor driving high performance is learning the named tokens themselves. Furthermore, we find that BERT is not always better at recognizing predictive contexts compared to a BiLSTM-CRF model. We enlist human annotators to evaluate the feasibility of inferring entity types from context alone and find that humans are also mostly unable to infer entity types for the majority of examples on which the context-only system made errors. However, there is room for improvement: A system should be able to recognize any named entity in a predictive context correctly

Submission received: 28 April 2020; revised version received: 12 November 2020; accepted for publication: 8 December 2020.

https://doi.org/10.1162/COLI_a_00397

and our experiments indicate that current systems may be improved by such capability. Our human study also revealed that systems and humans do not always learn the same contextual clues, and context-only systems are sometimes correct even when humans fail to recognize the entity type from the context. Finally, we find that one issue contributing to model errors is the use of "entangled" representations that encode both contextual and local token information into a single vector, which can obscure clues. Our results suggest that designing models that explicitly operate over representations of local inputs and context, respectively, may in some cases improve performance. In light of these and related findings, we highlight directions for future work.

1. Introduction

Named Entity Recognition (NER) is the task of identifying words and phrases in text that refer to a person, location, or organization name, or some finer subcategory of these types. NER systems work well on domains such as English news, achieving high performance on standard data sets like MUC-6 (Grishman and Sundheim 1996), CoNLL 2003 (Tjong Kim Sang and De Meulder 2003), and OntoNotes (Pradhan and Xue 2009). However, prior work has shown that the performance deteriorates on entities unseen in the training data (Augenstein, Derczynski, and Bontcheva 2017; Fu et al. 2020) and when entities are switched with a diverse set of entities even within the same data set (Agarwal et al. 2020).

In this article, we examine the interpretability of models used for the task, focusing on the type of textual clues that lead systems to make predictions. Consider, for instance, the sentence "Nicholas Romanov abdicated the throne in 1917." The correct identification of "Nicholas Romanov" as a person may be due to (i) knowing that *Nicholas* is a fairly common name and that (ii) the capitalized word after that ending with "-ov" is likely a Slavic last name, too. Alternatively, (iii) a competent user of language would know the selectional restrictions (Framis 1994; Akbik et al. 2013; Chersoni et al. 2018) for the subject of the verb abdicate, namely, that only a person may abdicate the throne. The presence of two words indicates that it cannot be a pronoun, so *X* in the context "*X* abdicated the throne" can only be a person.

Such probing of the reasons behind a prediction is in line with early work on NER that emphasized the need to consider both internal (features of the name itself) and external (context features) evidence when determining the semantic types of named entities (McDonald 1993). We specifically focus on the interplay between learning names as in (i), and recognizing constraining contexts as in (iii), given that (ii) can be construed as a more general case of (i), in which word shape and morphological features may indicate that a word is a name even if the exact name is never explicitly seen by the system (Table 1 in Bikel, Schwartz, and Weischedel [1999]).

Below are some examples of constraining contexts for different entity types. The type of X in each of these contexts should always be the same, irrespective of the identity of X.

PER My name is *X*.
LOC The flight to *X* leaves in two hours.
ORG The CEO of *X* resigned.

As a foundation for our work, we conduct experiments with BiLSTM-CRF models using GloVe input representations (Huang, Xu, and Yu 2015) modified to use *only context* representations or *only word* identities to quantify the extent to which systems exploit word and context evidence, respectively (Section 3). We test these systems on

three different data sets to identify trends that generalize across corpora. We show that context does somewhat inform system predictions, but the major driver of performance is recognition of certain words as names of a particular type. We modify the model by introducing gates for word and context representations to determine what it focuses on. We find that on average, only the gate value of the word representation changes when there is a misprediction; the context gate value remains the same. We also examine the performance of a BERT-based NER model and find that it does not always incorporate context better than the BiLSTM-CRF models.

We then ask if systems should be expected to do better from the context text (Section 5). Specifically, we task human annotators with inferring entity types using only (sentential) context, for the instances on which a BiLSTM-CRF relying solely on context made a mistake. We find that in the majority of cases, annotators are (also) unable to choose the correct type. This suggests that it may be beneficial for systems to similarly recognize situations in which there is a lack of reliable semantic constraints for determining the entity type. Annotators sometimes make the same mistakes as the model does. This may hint at why conventional systems tend to ignore context features: The number of examples for which relying primarily on contextual features will result in an accurate prediction is almost the same as the number for which relying on the context will lead to an erroneous prediction. For some cases, however, annotators are able to correctly identify the entity type from the context alone when the system makes an error, indicating that systems do not identify all constraining contexts and that there is room for improvement.

We conclude by running "oracle" experiments in Section 3 that show that systems with access to different parts of the input can be better combined to improve results. These experiments also show that the GloVe-based BiLSTM-CRF and BERT based only on the context representation are correct on very different examples; consequently, combining the two leads to performance gains. There is thus room for better exploiting the context to develop robust NER systems with improved generalizability.

Finally, we discuss the implications of our findings for the direction of future research, summarizing the answers to three questions we explore in this article: (1) What do systems learn—word identity vs. context? (2) Can context be utilized better so that generalization is improved? (3) If so, what are some possible directions for it?

2. Related Work

Most related prior work has focused on learning to recognize certain words as names, either using the training data, gazetteers or, most recently, pretrained word representations. Early work on NER did explicitly deal with the task of scoring contexts on their ability to predict the entity types in that context. More recent neural approaches have only indirectly incorporated the learning of context, namely, via contextualized word representations (Peters et al. 2018; Devlin et al. 2019).

2.1 Names Seen in Training

NER systems recognize entities seen in training more accurately than entities that were not present in the training set (Augenstein, Derczynski, and Bontcheva 2017; Fu et al. 2020). The original CoNLL NER task used as a baseline a name look-up table: Each word that was part of a name that appeared in the training data with a unique class was correspondingly classified in the test data as well. All other words were marked as

non-entities. Even the simplest learning systems outperform such a baseline (Tjong Kim Sang and De Meulder 2003), as it will clearly achieve poor recall. At the same time, overviews of NER systems indicate that the most successful systems, both old (Tjong Kim Sang and De Meulder 2003) and recent (Yadav and Bethard 2018), make use of gazetteers listing numerous names of a given type. Wikipedia in particular has been used extensively as a source for lists of names of given types (Kazama and Torisawa 2007; Ratinov and Roth 2009). The extent to which learning systems are effectively 'better' look-up models—or if they actually learn to recognize contexts that suggest specific entity types—is not obvious.

Even contemporary systems that do not use gazetteers expand their knowledge of names through the use of pretrained word representations. With distributed representations trained on large background corpora, a name is "seen in training" if its representation is similar to names that explicitly appeared in the training data for NER. Consider, for example, the commonly used Brown cluster features (Brown et al. 1992; Miller, Guinness, and Zamanian 2004). Both in the original paper and the reimplementation for NER, authors show examples of representations that would be the same for classes of words (John, George, James, Bob or John, Gerald, Phillip, Harold, respectively). In this case, if one of these names is seen in training, any of the other names are also treated as seen, because they have the exact same representation.

Similarly, using neural embeddings, words with representations similar to those seen explicitly in training would likely be treated as "seen" by the system as well. Tables 6 and 7 in Collobert et al. (2011) show the impact of word representations trained on small training data also annotated with entity types compared with those making use of large amounts of unlabeled text. When using only the limited data, the words with representations closest to *france* and *jesus*, respectively, are "persuade, faw, blackstock, giorgi" and "thickets, savary, sympathetic, jfk," which seem unlikely to be useful for the task of NER. For the word representations trained on Wikipedia and Reuters,[1] the corresponding most similar words are "austria, belgium, germany, italy" and "god, sati, christ, satan." These representations clearly have higher potential for improving NER.

Systems with character-level representations further expand their ability to recognize names via word shape (capitalization, dashes, apostrophes) and basic morphology (Lample et al. 2016).

We directly compare the lookup baseline with a system that uses only predictive contexts learned from the training data, and an expanded baseline drawing on pretrained word representations that cover many more names than the limited training data itself.

2.2 Unsupervised Name–Context Learning

Approaches for database completion and information extraction use free unannotated text to learn patterns predictive of entity types (Riloff and Jones 1999; Collins and Singer 1999; Agichtein and Gravano 2000; Etzioni et al. 2005; Banko et al. 2007) and then use these to find instances of new names. Given a set of known names, they rank all n-gram contexts for their ability to predict the type of entities, discovering for example that "the mayor of X" or "Mr. Y" or "permanent resident of Z" are predictive of city, person, and country, respectively.

1 CoNLL data, one of the standard data sets used to evaluate NER systems, is drawn from Reuters.

Early NER systems also attempted to use additional unannotated data, mostly to extract names not seen in training but also to identify predictive contexts. These, however, had little to no effect on system performance (Tjong Kim Sang and De Meulder 2003), with few exceptions where both names and contexts were bootstrapped to train a system (Cucerzan and Yarowsky 2011; Nadeau, Turney, and Matwin 2006; Talukdar et al. 2006).

Recent work in NLP relies on neural representations to expand the ability of the systems to learn context and names (Huang, Xu, and Yu 2015). In these approaches the learning of names is powered by the pretrained word representations, as described in the previous section, and the context is handled by an LSTM representation. So far, there has not been analysis of which parts of contexts are properly captured by the LSTM representations, especially what they do better than more local representations of just the preceding/following word.

Acknowledged state-of-the-art approaches have demonstrated the value of *contextualized* word embeddings, as in ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019); these are representations derived both from input tokens and the context in which they appear. They have the clear benefit of making use of large data sets for pretraining that can better capture a diversity of contexts. But at the same time these contextualized representations make it difficult to interpret the system prediction and which parts of the input led to a particular output. Contextualized representations can in principle disambiguate the meaning of words based on their context, for example, the canonical example of Washington being a person, a state, or a city, depending on the context. This disambiguation may improve the performance of NER systems. Furthermore, token representations in such models reflect their context by construction, so may specifically improve performance on entity tokens not seen during training but encountered in contexts that constrain the type of the entity.

To understand the performance of NER systems, we should be able to probe the justification for the predictions: Did they recognize a context that strongly indicates that whatever follows is a name of a given type (as in "Czech Vice-PM _"), or did they recognize a word that is typically a name of a given type ("Jane"), or a combination of the two? In this article, we present experiments designed to disentangle to the extent possible the contribution of the two sources of confidence in system predictions. We perform in-depth experiments on systems using non-contextualized word representations and a human/system comparison with systems that exploit contextualized representations.

3. Context-Only and Word-Only Systems

Here we perform experiments to disentangle the performance of systems based on the word identity and the context. We compare two look-up baselines and several systems that vary the representations fed into a sequential Conditional Random Field (CRF) (Lafferty, McCallum, and Pereira 2001), described below.

Lookup. Create a table of each word preserving its case, and its most frequent tag from the training data. In testing, lookup a word in this table and assign its most frequent tag. If the word does not appear in the training data or there is a tie in the tag frequency, mark as O (outside, not a named entity).

LogReg. Logistic Regression using the GloVe representation of the word only (no context of any kind). This system is equivalent to lookup in both the NER training data and GloVe representations as determined by the data they were trained on.

GloVe fixed + CRF. This system uses GloVe word representations as features in a linear chain CRF model. Any word in training or testing that does not have a GloVe representation is assigned representation equal to the average of all words represented in GloVe. The GloVe input vectors are fixed in this setting, that is, we do not backpropagate into these.

GloVe fine-tuned + CRF. The same as the preceding model, except that GloVe embedding parameters are updated during training. This method nudges word representations to become more similar depending on how they manifest in the NER training data, and generally performs better than relying on fixed representations.

FW context + CRF. This system uses LSTM (Hochreiter and Schmidhuber 1997) representations only for the text preceding the current word (i.e., run forward from the start to this word), with GloVe as inputs. Here we take the hidden state of the previous word as the representation of the current word. This incorporates non-local information not available to the two previous systems, from the part of the sentence before the word.

BW context + CRF. Backward context-only LSTM with GloVe as inputs. Here we reverse the sentence sequence and take the hidden state of the next word in the original sequence as the output representation of the current word.

BI context + CRF. Bidirectional context-only LSTM (Graves and Schmidhuber 2005) with GloVe as input. We concatenate the forward and backward context-only representations taking the hidden state as in the two systems above and not the hidden state of the current word.

BI context + word + CRF. Bidirectional LSTM as in Huang, Xu, and Yu (2015). The feature representing the word is the hidden state of the LSTM after incorporating the current word; the backward and forward representations are concatenated.

We use 300 dimensional cased GloVe (Pennington, Socher, and Manning 2014) vectors trained on Common Crawl.[2] The models are trained for 10 epochs using Stochastic Gradient Descent with a learning rate of 0.01 and weight decay of 1e-4. A dropout of 0.5 is applied on the embeddings and the hidden layer dimension used for the LSTM is 100. We use the IO labeling scheme and evaluate the systems via micro-F1, at the token level. We use the word-based model for all the above variations, but believe a character-level model would yield similar results: Such models would differ only in how they construct the independent context and word representations that we consider.

Although the above systems would show how the model behaves when it has access to only specific information—context or word—they do not capture what the model would focus on with access to both types of information. For this reason, we build a gated system as follows.

2 `http://nlp.stanford.edu/data/glove.840B.300d.zip`.

Gated BI + word + CRF. This system comprises a bidirectional LSTM that uses both context and the word, but the two representations are combined using gates. This provides a mechanism for revealing the degree to which the prediction was influenced by the word versus the context.

More concretely, let $X = (x_1, x_2, \ldots, x_n)$ be the input sequence where $x_i \in R^d$ is the pretrained word embedding of the i^{th} word in the sequence. The input representation of each word is projected to a lower dimensional space, yielding a word-only representation $x_i^w \in R^{d_w}$.

$$x_i^w = W_w x_i$$

For each word, we also induce a context-only representation using two LSTMs. One computes the representation of the forward context h_i^{fw}; the other reverses the sequence to derive a representation of the backward context h_i^{bw}. As noted earlier, these hidden states incorporate the representation of the i^{th} word. Because we want a context-only representation, we instead use the hidden state for the previous word in the fw LSTM, and hidden state for the following word yielded from the bw LSTM. We concatenate these to form a context-only representation $x_i^c \in R^{d_c}$.

$$x_i^c = \left[h_{i-1}^{fw}; h_{i+1}^{bw}\right]$$

Next we combine the word-only and context-only representations using parameterized gates:

$$g_w = \sigma\left(W^{gw}\left[x_i^w; x_i^c\right]\right)$$
$$g_c = \sigma\left(W^{gc}\left[x_i^w; x_i^c\right]\right)$$
$$x_i^{wc} = g_w x_i^w + g_c x_i^c$$

This is followed by a linear layer that consumes the resultant x_i^{wc} representations and produces logits for potential class assignments for tokens that are subsequently decoded using a linear chain CRF. This architecture is illustrated in Figure 1.

In addition to the above systems that are based on GloVe representations, we also perform experiments using the following systems that use contextual representations, for the sake of completeness.

Full BERT. We use the original public large model[3] and apply the default fine-tuning strategy. We use the NER fine-tuning implementation in Wolf et al. (2020) and train the model for three epochs with a learning rate of 5e-06 and maximum sequence length of 128. We use the default values of all other hyperparameters as in the implementation.

Word-only BERT. Fine-tuned using only the word as the input resulting in a non-contextual word-only representation. Each word token can appear multiple times with various entity types in the training data and the frequency for each is maintained in the training data.

3 `cased_L-24_H-1024_A-16`.

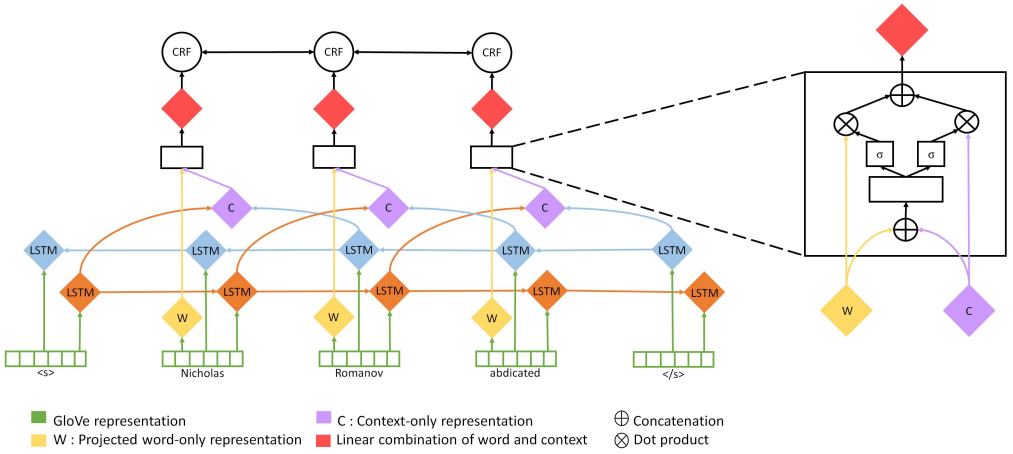

Figure 1
Architecture of the gated system. For each word, a token-only (yellow) and a context-only (purple) representation is learned. These are combined using gates, as illustrated on the right, and fed into a CRF.

Context-only BERT. Because decomposition of BERT representations into word-only and context-only is not straightforward,[4] we adopt an alternate strategy to test how BERT fares without seeing the word itself. We use a reserved token from the vocabulary "[unused0]" as a mask for the input token so that the system is forced to make decisions based on the context and does not have a prior entity type bias associated with the mask. We do this for the entire data set, masking one word at a time. It is important to note that the word is only masked during testing and not during fine-tuning.

4. Results

We evaluate these systems on the CoNLL 2003 and MUC-6 data. Our goal is to quantify how well the models can work if the identity of the word is not available, and to compare that to situations in which *only* the word identity is known. Additionally, we evaluate the systems trained on CoNLL data on the Wikipedia data set (Balasuriya et al. 2009), to assess how data set–dependent the performance of the system is. Table 1 reports our results. The first line in the table (BI context + word + CRF) corresponds to the system presented in Huang, Xu, and Yu (2015).

4.1 Word-Only Systems

The results in the *Word only* rows are as expected: We observe low recall and high precision. All systems that rely on the context alone, without taking the identity of the word into account, have worse F1 than the Lookup system. The results are consistent across CoNLL as well as MUC6 data sets. On the cross domain evaluation, however, when the system trained on CoNLL is evaluated on Wikigold, the recall drops considerably.

4 We tried a few techniques such as projecting the BERT representations into word and context spaces by learning to predict the word itself and the context words as two simultaneous tasks, but this did not work well.

Table 1
Performance of GloVe word-level BiLSTM-CRF and BERT. All rows are for the former and only the last two rows for BERT. Local context refers to high precision constraints due to sequential CRF. Non-local context refers to the entire sentence. No document level context is included. The first two panels were trained on the Original English CoNLL 03 training data and tested on the original English CoNLL 03 test data and the WikiGold data. The last panel was trained and tested on the respective splits of MUC-6. Highest F1 in each panel is **boldfaced**, excluding the full systems.

System	CoNLL			Wikipedia			MUC-6		
	P	R	F1	P	R	F1	P	R	F1
Full system									
BI context + word + CRF	90.7	91.3	91.0	66.6	60.8	63.6	90.1	91.8	90.9
Words only									
Lookup	84.1	56.6	67.7	66.3	28.5	39.8	81.4	54.4	65.2
LogReg	80.2	74.3	**77.2**	58.8	48.9	**53.4**	75.1	71.7	**73.4**
Words + local context									
GloVe fixed + CRF	67.9	63.4	65.6	53.7	37.6	44.2	74.1	68.1	70.9
GloVe fine-tuned + CRF	80.8	77.3	**79.0**	63.3	45.8	**53.1**	82.1	77.0	**79.5**
Non-local context-only									
FW context-only + CRF	71.3	39.4	50.8	53.3	19.3	28.4	71.9	58.9	**64.7**
BW context-only + CRF	69.5	47.7	56.6	46.6	21.7	29.6	74.0	49.4	59.2
BI context-only + CRF	70.1	52.1	**59.8**	51.2	21.4	**30.2**	66.4	56.5	61.1
BERT									
Full	91.9	93.1	92.5	75.4	75.1	75.2	96.1	97.2	96.7
Word-only	80.0	80.5	**80.3**	61.6	54.8	**58.0**	77.9	75.3	**76.5**
Context-only	43.1	64.1	51.6	39.7	76.2	52.2	75.6	71.6	73.5

This behavior may be attributed to the data set: Many of the entities in the CoNLL training data also appear in testing, a known undesirable fact (Augenstein, Derczynski, and Bontcheva 2017). We find that 64.60% and 64.08% of entity tokens in CoNLL and MUC6 test are seen in the respective training sets. However, only 41.55% of entity tokens in Wikigold are found in the CoNLL training set.

The use of word representations (*LogReg* row) contributes substantially to system performance, especially for the MUC-6 data set in which few names appear in both train and test. Given the impact of the word representations, it would seem important to track how the choice and size of data for training the word representations influences system performance.

4.2 Word and Local Context Combinations

Next, we consider the systems in the *Word + local context* rows. CRFs help to recognize high precision entity-type local contextual constraints, for example, force a LOC in the pattern "ORG ORG LOC" to be an ORG as well. Another type of high-precision constraining context is word-identity based, similar to the information extraction work discussed above, and constrains X in the pattern "X said" to be PER. Both of these

Table 2
Mean gate values in CoNLL when entities and non-entities are correct and incorrect. For entities, the average value of context gates remains the same irrespective of the predicted values. For both entities and non-entities, the word gate has a much higher value when the prediction is correct. The word identity itself is the major driver of performance.

	Context	**Word**
ENT correct	0.906	0.831
ENT incorrect	0.906	0.651
O correct	0.613	0.897
O incorrect	0.900	0.613

context types were used in Liao and Veeramachaneni (2009) for semi-supervised NER. The observed improved precision and recall of *GloVe fine-tuned + CRF* over *LogReg* indicates that the CRF layer modestly improves performance. However, fine-tuning the representations on the training set is far more important than including such constraints with CRFs as *fixed GloVe + CRF* performs consistently worse than *LogReg*.

4.3 Context Features Only

We compare *Context-only* systems with non-local context. In CoNLL data, the context after a word appears to be more predictive, whereas in MUC-6 the forward context is more predictive. In CoNLL, some of the right contexts are too corpus-specific, such as "X 0 Y 1" being predictive of X and Y as organizations with the example occurring in reports of sports games, such as "France 0 Italy 1." We find that 32.37% of the *ORG* entities in CoNLL test split occur in such sentences. MUC-6, on the other hand, contains many examples that include honorifics, such as "Mr. X." In the MUC-6 test split, 35.08% of PER entities are preceded by an honorific, whereas in CoNLL and MUC6, this is the case only for 2.5% and 4.6% entities, respectively. Statistics on these two patterns are shown in Table 3. We provide the list of honorifics and regular expressions for sports scores used to calculate this in the Appendix.

Finally, we note that combining the backward and forward contexts by concatenating their representations results in a better system for CoNLL but not for MUC-6.

Clearly, systems with access to only word identity perform better than those with access to only the context (drop of ∼20 F1 in all the three data sets). Next, we use the Gated BI + word + CRF system in Figure 1 to investigate what the system focuses on when it has access to both the word and to the context, as distinct input representations. We compare the average value of the word and context gates when the system is correct vs. incorrect in Table 2. For entities, while the context gate value is higher than the word gate value, its average remains the same irrespective of whether the prediction is correct or not. On the other hand, the word gate value drops considerably when the system makes an error. Similarly, the word gate value drops considerably for non-entities as well on error. Surprisingly, the context gate value increases for non-entities when an error is made. These results suggest that systems over-rely on word identity to make their predictions.

Moreover, whereas one would have expected that the context features have high precision and low recall, this is indeed not the case: The precision of the BI+CRF system is consistently lower than the precision for the full system and the logistic regression

Table 3
Repetitive context patterns in the data sets. In CoNLL, a large percentage of organizations occur in sports scores. In MUC-6, a large percentage of PER entities are preceded by an honorific.

	Honorifics % PER	**Sports Scores** % ORG
CoNLL train	2.74	25.89
CoNLL testa	2.19	22.13
CoNLL testb	2.59	32.37
Wikipedia	4.65	0.00
MUC-6 train	27.46	0.00
MUC-6 test	35.08	0.00

model. This means that a better system will not only learn to recognize more contexts but also would be able to override contextual predictions based on features of the word in that context.

4.4 Contextualized Word Representations

Finally, we experiment with BERT (Devlin et al. 2019). Word-only BERT is fine-tuned using only one word at a time without any context. Full BERT and context-only BERT are the same system fine-tuned on the original unchanged data set and differ only in inference. For context-only BERT, a reserved vocabulary token "[unused0]" is used to hide one word at a time to get a context-only prediction. We report these results in the last two rows of Table 1. Full BERT improves in F1 over the GloVe-based BiLSTM-CRF as reported in the original paper. Word-only BERT performs better than the context-only BERT but the difference in performance is not as pronounced as in the case of the GloVe-based BiLSTM-CRF, except on the CoNLL corpus due to the huge overlap of the training and testing set entities, as noted earlier. We also note the difference in performance of the context-only systems using GloVe and BERT. Context-only BERT performs better or worse than context-only BiLSTM, depending on the corpora. These results show that BERT is not always better at capturing contextual clues. Although it is better in certain cases, it also misses these clues in some instances for which the BiLSTM makes a correct prediction.

In Section 6, we provide a more detailed oracle experiment that demonstrates that the BiLSTM-CRF and BERT context-only systems capture different aspects of context and make errors on different examples. Here we preview this difference in strengths by examining BERT performance on a sample of sentences for which BiLSTM context representations are sufficient to make a correct prediction, and a sample of sentences where it is not. We randomly sampled 200 examples from CoNLL 03 where the context-only LSTM was correct (Sample-C) and another 200 where it was incorrect (Sample-I). Context-only BERT is correct on 71.5% examples in Sample-C but fails to make the correct prediction on the remaining 28.5% that are easy for the BiLSTM context representation. In contrast, it is also able to correctly recognize the entity type in 53.22% of the cases in Sample-I, where BiLSTM context is insufficient for prediction.

Similar to the gated system, we looked at the attention values in the case of BERT. However, these are much harder to interpret as compared to the gate values in the LSTM. The gated LSTM had only one value for the entire context and one value for

the entire word. In contrast, BERT has an attention value for each subword in the context and each subword in the word. Moreover, there are multiple attention heads. It is unclear how to interpret the individual values and combine them into a single number for the context and the word. We looked into this by taking the maximum (or mean) value for each subword across all heads and maximum (or mean) value of all subwords in the context and the word but found similar final values for both in all cases.

5. Human Evaluation

In this section, we describe a study with humans, assessing if they tend to be more successful at using contextual cues to infer entity types. For comparison, we also conduct another study to assess the success of humans in using just the word to infer entity type.

5.1 Context-Only

We performed a human study to determine if humans can infer entity types solely from the context in which they appear. Only sentence-level context is used as all systems operate at the sentence-level. For each instance with a target word in a sentence context, we show three annotators the sentence with the target word masked and ask them to determine its type as PER, ORG, LOC, MISC, or O (not a named entity). We allow them to select multiple options. We divide all examples into batches of 10. We ensure the quality of the annotation by repeating one example in every batch and removing annotators that are not consistent on this repeated example. Furthermore, we include an example either directly from the instructions or very similar to an example in the instructions. If an annotator does not select the type from the example at a minimum, we assume that they either did not read the instructions carefully, or that they did not understand the task, and we remove them as well. Since humans are allowed to select multiple options, we do not expect them to fully agree on all of the options. The goal is to select the most likely option and so we take as the human answer the option with the most votes: We will refer to this as the majority label.

Similar to the comparison between BiLSTM and BERT, we now compare BiLSTM behavior and that of humans. For the study, we select 20 instances on which the context-only BiLSTM model was correct and 200 instances for which the context-only BiLSTM made errors. The sample from correct prediction is smaller because we see overwhelming evidence that whenever the BiLSTM prediction is correct, humans can also easily infer the correct entity type from contextual features alone. For 85% of these correctly labeled instances, the majority label provided by the annotators was the same as the true (and predicted) label. Table 4 shows the three (out of 20) examples in which humans did not agree on the category or made a wrong guess. These 20 instances serve as a sanity check as well as a check for annotator quality for the examples where the system made incorrect predictions.

We received a variety of responses for the 200 instances in sentences where the context-only BiLSTM-CRF model made an incorrect prediction. Below we describe the results from the study. We break down the human predictions in two classes. Examples of each are shown in Table 5.

Error class 1. Humans correct: The human annotators were able to correctly determine the label for 23.5% (47) of the sentences where the context-only BiLSTM-CRF made errors, indicating some room for improvement in a context-only system.

Table 4
Examples of human evaluation where the context-only system was correct but humans incorrect.

Sentence	Word	Label	Human
Lang said he ___ conditions proposed by Britain's Office of Fair Trading, which was asked to examine the case last month.	supported	O	-
___ Vigo 15 5 5 5 17 17 20	Celta	ORG	O
The years I spent as manager of the Republic of ___ were the best years of my life.	Ireland	LOC	-

Error class 2. Human incorrect or no majority: For the remaining 76.5% instances, humans could not predict the entity type from only the context.

For 55.5% of the cases, there was a human majority label but it was not the same as the true label. In these cases, the identity of the word would have provided clues conflicting those in the context alone. The large number of such cases—where context and word cues need to be intelligently combined—may provide a clue as to why modern NER systems largely ignore context: They do not have the comprehension and reasoning abilities to combine the semantic evidence, and instead resort to over-reliance on the word identity, which in such cases would override the contextual cues in human comprehension.

For 21%, there was no majority label in the human study, suggesting that the context did not sufficiently constrain the entity type. In such cases, clearly the identity of the words would dominate the interpretation.

In sum, humans could correctly guess the type without seeing the target word for less than a quarter of the errors made by the BiLSTM-CRF. Remarkably, BERT has correct as well as incorrect predictions on the examples from both BiLSTM-CRF error classes. It was correctly able to determine the entity type for 65.9% of cases in error class 1 and 49.3% of cases in error class 2. These results show that neither system is learning the same contextual clues as humans. Humans find the context insufficient in error class 2 but BERT is able to capture something predictive in the context. Future work could collect more human annotations with humans specifying the reason for selecting an answer. A carefully designed framework would collect human reasoning for their answers and incorporate this information while building a system.

5.2 Word-Only

For completeness, we also performed human evaluation to determine whether humans can infer the type solely from the word identity. In this case, we do not show the annotators any context. We follow the same annotation and quality control process as above. Because words are ambiguous (Washington can be PER, LOC, or ORG), and data sets have different priors for words being of each type, we do not expect the annotators to get all of the answers correct. However, we do expect them to be correct more often than they were in the context-only evaluation.

We select the same 200 instances for the evaluation. We consider the GloVe-based LogReg system for comparison as it does not include any context at all. Out of these 200, the LogReg system was correct on 146 (73%) instances and incorrect for the remaining 27% instances. Of the 146 cases in LogReg-correct, humans are correct for 116 (79.4%) instances. For the remaining instances in LogReg-correct, humans are incorrect due to

Table 5
Examples of human evaluation.

Sentence	Word	Label	Human	GloVe	BERT
Error Class 1					
Analysts said the government, while anxious about ___ 's debts, is highly unlikely to bring the nickel, copper, cobalt, platinum and platinum group metals producer to its knees or take measures that could significantly affect output.	Norilisk	ORG	✓	O	O
6. Germany III (Dirk Wiese, Jakobs ___) 1:46.02	Marco	PER	✓	O	O
- Gulf ___ Mexico:	of	LOC	✓	MISC	O
About 200 Burmese students marched briefly from troubled Yangon ___ of Technology in northern Rangoon on Friday towards the University of Yangon six km (four miles) away, and returned to their campus, witnesses said.	Institute	ORG	✓	O	LOC
NOTE - Sangetsu Co ___ is a trader specialising in interiors.	Ltd	ORG	✓	O	✓
Russ Berrie and Co Inc said on Friday that A. ___ Cooke will retire as president and chief operating officer effective July 1, 1997 .	Curts	PER	✓	ORG	✓
ASEAN groups Brunei, Indonesia, Malaysia, the ___, Singapore, Thailand and Vietnam .	Philippines	LOC	✓	O	✓
Error Class 2					
Their other marksmen were Brazilian defender Vampeta ___ Belgian striker Luc Nilis, his 14th of the season.	and	O	PER	PER	✓
On Monday and Tuesday, students from the YIT and the university launched street protests against what they called unfair handling by police of a brawl between some of their colleagues and restaurant owners in ___ .	October	O	LOC	LOC	LOC
Alessandra Mussolini, the granddaughter of ___ 's Fascist dictator Benito Mussolini, said on Friday she had rejoined the far-right National Alliance (AN) party she quit over policy differences last month.	Italy	LOC	PER	O	MISC
Public Service Minister David Dofara, who is the head of the national Red Cross, told Reuters he had seen the bodies of former interior minister ___ Grelombe and his son, who was not named.	Christophe	PER	ORG	O	✓
The longest wait to load on the West ___ was 13 days.	Coast	O	MISC	LOC	LOC
, 41, was put to death in ___ 's electric chair Friday.	Florida	LOC	-	O	✓
Wall Street, since the bid, has speculated that any deal between Newmont and ___ Fe would be a "bear hug, " or a reluctantly negotiated agreement where the buyer is not necessarily a friendly suitor.	Santa	ORG	-	LOC	LOC

domain/data set priors. For example, in CoNLL, cities such as Philadelphia are sports teams and hence ORG (not LOC), but without that knowledge, humans categorize them as LOC. Out of all 54 instances in LogReg-incorrect, humans were correct on 16. While

Table 6
Performance (F1) of GloVe word-based BiLSTM-CRF and BERT. System 1 denotes the oracle combination of separate systems which access to specific input representation only. System 2 refers to a single system with access to all the input of the various systems in system 1. The first two panels were trained on the Original English CoNLL 03 training data and tested on the original English CoNLL 03 test data and the WikiGold data. The last panel was trained and tested on the respective splits of MUC-6.

System 1 (Oracle)	System 2	CoNLL		Wikipedia		MUC-6	
		sys 1	sys 2	sys 1	sys 2	sys 1	sys 2
FW context – BW context LSTM-CRF	Bi context LSTM-CRF	75.3	59.8	49.3	30.2	85.3	61.1
FW context – BW context – GloVe fine-tuned LSTM-CRF	Full LSTM-CRF	92.2	91.0	72.4	63.6	94.9	90.9
Full system – FW context – BW context – GloVe fine-tuned LSTM-CRF	Full LSTM-CRF	95.1	91.0	76.8	63.6	96.1	90.9
word-only – context-only BERT	Full BERT	92.5	92.5	86.8	75.2	93.7	96.7
Full – word-only – context-only BERT	Full BERT	96.5	92.5	90.1	75.2	98.5	96.7

this may seem to suggest room for system improvement, it is due to the same ambiguity and data set bias. For example, CoNLL has a bias for cities like the Philadelphia example above to be sports teams and hence ORGs but there are a few instances where the city name is actually LOC. In these cases, the LogReg system is incorrect but humans are correct.

To summarize, human evaluation using only the word identity is consistent with predictions of the word-only system and observed differences are likely due to data set biases and so expected.

6. Oracle Experiments

In the human evaluation we saw some mixed results, with some room for improvement on 23.5% of the errors on one side and some errors that seem to be due to over-reliance on context on the other. This motivates investigating whether a more sophisticated approach that decides how to combine cues would be a better approach.

6.1 Combining All Systems

We perform an oracle experiment where the oracle knows which of the systems is correct. If neither is correct, it defaults to one of the systems. We report results in Table 6. The default system in each case is the one listed first. Row 1 in the table shows that an oracle combination of the forward context-only and backward context-only does much better than the system that simply concatenates both context representations to make the prediction. The gains are about 15, 20, and 24 points F1 on CoNLL, Wiki, and MUC-6, respectively. This improvement captures 22 of the 47 examples (46.8%) that human annotators got right but not the context-only system in Section 5.

Table 7
Oracle combination of the context-only systems using GloVe and BERT. *min* and *max* denote the minimum and maximum F1 of the systems combined and *comb* denotes the F1 of the combined system. The first two panels were trained on the Original English CoNLL 03 training data and tested on the original English CoNLL 03 test data and the WikiGold data. The last panel was trained and tested on the respective splits of MUC-6.

System	CoNLL			Wikipedia			MUC-6		
	min	max	comb	min	max	comb	min	max	comb
context-only LSTM – context-only BERT	51.6	59.8	83.7	30.2	52.2	79.0	61.1	73.5	84.6

We performed more such experiments with the full system and the word-only and context-only systems. These are shown in rows 2 and 3. In each case, there are gains over the full BiLSTM-CRF. An oracle with the four systems (row 3) shows the highest gains with ∼4 points F1 on CoNLL and 6 points F1 on MUC-6. The gains are especially pronounced in the case of cross-domain evaluation, namely, the system trained on CoNLL when evaluated on Wikipedia has an increase of 13 points F1.

These results indicate that when given access to different components—word, forward context, backward context—systems recognize different entities correctly, as they should. However, when all of these components are thrown at the system at once, they are not able to combine these in the best possible way.

6.2 Combining BERT Variants

We performed similar experiments with the full, word-only, and context-only BERT as well (rows 4 and 5). Remember that the context-only BERT is the same trained system as the full BERT and the difference comes from masking the word during evaluation. The word-only and context-only combination shows no improvement over full BERT on CoNLL but does so on the two other data sets. However, even for CoNLL, the oracle of these two systems is correct on somewhat different examples than the full BERT as evident in the last row where combining all three systems gives improvement on all three data sets. Again, the improvement is highest on cross-domain evaluation on the Wikipedia data set as with the combination of the GloVe-based systems. We hypothesize that the attention on the current word likely dominates in the output contextual representation, because of which the relevant context words do not end up contributing much. However, when the word is masked in the context-only variant, the attention weight on it will likely be small because we use one of the tokens unused in the vocabulary and the relevant context words end up contributing to the final representation. Future work could involve some regularization to reduce the impact of the current word in a few random training examples so that systems do not overfit to the word identity and focus more on the context.

6.3 Combining Context-Only LSTM and BERT

Lastly, we performed an oracle experiment with the context-only LSTM-CRF and the context-only BERT (see Table 7). We see the biggest jumps in performance with this particular experiment: 24, 27, and 11 points F1 on CoNLL, Wiki, and MUC-6, respectively. Both the systems use the context differently and are correct on a large number

of different examples, again showing that better utilization of context is possible. Note that the human study showed that the room for improvement from context is small. The study was based on 200 random examples, where BERT was able to get some correct even when humans could not. We hypothesize that while the context was possibly not sufficient to identify the type, a similar context had been seen in the training data that was learned by BERT .

All the oracle experiments show room for improvement and future work would involve looking into strategies/systems to combine these components better. The progress toward this can be measured by breaking down the systems and conducting oracle experiments as done here.

7. Discussion and Future Work

We started the article with three questions about the workings of NER systems. Here, we return to these questions and synthesize the insights related to each gained from the experiments presented in the article. We zeroed in on the question of interpretability of NER systems, specifically examining the performance of systems that represent differently the current word, the context, and their combination. We tested the systems on two corpora and one tested across domains and show that some of the answers to these questions are at times corpus dependent.

7.1 What Do Systems Learn?

Word types, mostly.

We find that current systems, including those build on top of contextualized word representations, pay more attention to the current word than to the contextual features. Partly this is due to the fact that contextual features do not have high precision and in many cases need to be overridden by evidence from the current word. Moreover, we find that contextual representations, namely, BERT, are not always better at capturing context as compared to systems such as GloVe-based BiLSTM-CRFs. Their higher performance could be the results of better pretraining data and learning the subword structure better. We leave this analysis for future work and instead focus on the extent of word versus context utilization with more focus on context utilization for better generalization.

7.2 Can Context Be Utilized Better?

Only a bit better, if we want to emulate humans. But a lot better if willing to incorporate the superhuman abilities of transformer models.

We carry out a study to test the ability of human annotators to predict the type of an entity without seeing the entity word. Humans seem to easily do the task on examples where the context-only system predicts the entity type correctly. The examples on which the context-only system makes a mistake are difficult for humans as well. Humans can guess the correct label only in about a quarter of all such examples. The opportunity for improvement from better contextual representations that recognize more constraining contexts exists but is relatively small.

7.3 How Can We Utilize Context Better?

By developing reasoning, more flexible combination of evidence, and informed regularization. Based on our experiments, several avenues for better context utilization emerge.

Human-like reasoning. Our human study shows that systems are not capturing the same information from textual context as humans do. BERT is able in many cases to correctly recognize the type from the context even when humans fail to do so. A direction for future work would involve collecting human reasoning behind their answers and incorporating this information in building the systems. This means learning to identify when the context is constraining and possibly what makes it constraining.

Avoiding concatenation. Another promising direction for the overall improvement of the GloVe-based BiLSTM-CRF system appears to be the better combination of features representing different types of context and the word. Oracle experiments show that different parts of the sentence—word, forward context, backward context—can help recognize entities correctly when used standalone but not when used together. A simple concatenation of features is not as meaningful and a smarter combination of several types of features can lead to better performance.

Attention regularization. Oracle experiments involving BERT show that hiding the word itself can sometimes correctly identify the word type even when seeing the word leads to an incorrect prediction. BERT uses attention weights to combine different parts of the input instead of concatenation as in the just discussed BiLSTM approaches. We hypothesize that the attention on the word is likely much larger than on the rest of the input because of which the relevant context words do not end up contributing much. Future work could involve some regularization to reduce the impact of the current word in a few random training examples so that it does not overfit to the word identity and can focus more on the context.

Lastly, another direction for future work could expand the vocabulary of entities instead of trying to learn context better so that more entities are seen either directly in the training data or have similar representation in the embedding space by virtue of being seen in the pretraining data. This could be done by having an even larger pretraining data from diverse sources for better coverage or by incorporating resources such as knowledge bases and gazetteers in the contextual systems.

Appendix

Entity Recognition vs. Typing

Entity recognition results are shown in Table 8. Here, we check if a word is correctly recognized as an entity, even if the type is wrong. The results are better than Recognition + Typing results in Table 1 for all systems and data sets but still not very high. Therefore, the errors made are in both recognition and typing. The same performance pattern is observed with word-only systems better than context-only systems even in entity recognition. The word-only systems have an F1 much higher than context-only systems, meaning that looking at a word, it is much easier to say whether it is an entity or not. But looking at context, it is still hard to say whether something is an entity or not. The breakdown by type is also shown in Table 9 for all three data sets for Full BERT.

Table 8
Performance of GloVe word-level BiLSTM-CRF and BERT on entity recognition (no typing). All rows are for the former and only the last two rows for BERT. Local context refers to high precision constraints due to sequential CRF. Non-local context refers to the entire sentence. No document level context is included. The first two panels were trained on the Original English CoNLL 03 training data and tested on the original English CoNLL 03 test data and the WikiGold data. The last panel was trained and tested on the respective splits of MUC-6.

System	CoNLL			Wikipedia			MUC-6		
	P	R	F1	P	R	F1	P	R	F1
Full system									
BI context + word + CRF	96.2	96.8	96.5	86.7	79.2	82.8	94.5	96.3	95.4
Words only									
Lookup	96.2	64.8	77.4	93.4	40.1	56.1	93.6	62.6	75.0
LogReg	93.8	86.9	90.2	89.9	74.8	81.6	87.5	83.6	85.5
Words + local context									
GloVe fixed + CRF	83.2	77.6	80.3	81.6	57.1	67.2	83.5	76.7	80.0
GloVe fine-tuned + CRF	91.9	88.0	89.9	88.6	64.1	74.4	91.3	85.6	88.4
Non-local context-only									
FW context-only + CRF	78.8	43.6	56.1	65.8	23.9	35.0	76.6	62.8	69.0
BW context-only + CRF	76.6	52.6	62.3	62.0	28.9	39.5	76.8	51.2	61.4
BI context-only + CRF	78.0	58.0	66.5	64.8	27.1	38.2	72.1	61.4	66.3
BERT									
Full	97.0	98.3	97.7	88.3	87.9	88.1	97.2	98.3	97.7
Word-only	94.8	95.4	95.1	93.0	82.8	87.6	93.1	89.9	91.5
Context-only	50.3	74.7	60.1	47.7	91.5	62.7	77.7	73.5	75.5

Honorifics

Data sets have different common context patterns. For example, forward context is highly predictive in MUC-6. In MUC-6 test, 35.08% of PER entities are preceded by an honorific whereas in CoNLL and MUC-6, this number is only 2.5% and 4.6%, respectively. Following is the list of honorifics (mostly English, owing to the nature of the data sets) used to calculate these numbers — Dr, Mr, Ms, Mrs, Mstr, Miss, Dr., Mr., Ms., Mrs., Mx., Mstr., Mister, Professor, Doctor, President, Senator, Judge, Governor, Officer, General, Nurse, Captain, Coach, Reverend, Rabbi, Ma'am, Sir, Father, Maestro, Madam, Colonel, Gentleman, Sire, Mistress, Lord, Lady, Esq, Excellency, Chancellor, Warden, Principal, Provost, Headmaster, Headmistress, Director, Regent, Dean, Chairman, Chairwoman, Chairperson, Pastor.

Sports Scores

We use the following three regular expressions to determine if a sentence contains sports scores -

([0-9]+.)?([A-Za-z]+){1,3}([0-9]+){0,6}(([0-9]+)(?!))(12)?(-)?

and

([A-Za-z]+){1,3}([0-9]+){1,3}([A-Za-z]+){1,3}
([0-9]+){0,2}[0-9]

and

([A-Z]+){1,3}AT ([A-Z]+){1,2}[A-Z]+

Human Labels Breakdown

Here we provide the breakdown between the different labels as marked by the human annotators. Note that the annotators could select multiple options and the majority label was selected as the final label. The results are shown in Table 10.

Table 9
Confusion matrices for BERT-Full. Rows represent the true class and columns represent the predicted class.

	PER	ORG	LOC	MISC	O
CoNLL					
PER	2,709	33	15	5	11
ORG	8	2,320	86	47	35
LOC	6	72	1,789	39	19
MISC	10	77	25	736	70
O	18	81	22	126	38,076
Wikipedia					
PER	1,509	14	13	6	92
ORG	50	1,362	200	192	154
LOC	14	141	1,013	39	240
MISC	22	74	61	944	291
O	51	154	230	312	31,827
MUC-6					
PER	584	4	1	0	1
ORG	6	830	5	0	22
LOC	0	0	105	0	4
MISC	0	0	0	0	0
O	3	41	1	0	12,498

Table 10
Breakdown of human labels. The majority label is selected as the final label.

Sentence	Word	True Label	Human label PER	ORG	LOC	MISC	O
Error Class 1							
Analysts said the government, while anxious about ___ 's debts, is highly unlikely to bring the nickel, copper, cobalt, platinum and platinum group metals producer to its knees or take measures that could significantly affect output.	Norilisk	ORG	0	2	1	0	0
6. Germany III (Dirk Wiese, Jakobs ___) 1:46.02	Marco	PER	2	1	1	1	0
- Gulf ___ Mexico:	of	LOC	1	0	2	0	0
About 200 Burmese students marched briefly from troubled Yangon ___ of Technology in northern Rangoon on Friday towards the University of Yangon six km (four miles) away, and returned to their campus, witnesses said.	Institute	ORG	0	2	1	1	0
NOTE - Sangetsu Co ___ is a trader specialising in interiors.	Ltd	ORG	0	3	0	0	0
Russ Berrie and Co Inc said on Friday that A. ___ Cooke will retire as president and chief operating officer effective July 1, 1997.	Curts	PER	3	0	0	0	0
ASEAN groups Brunei, Indonesia, Malaysia, the ___, Singapore, Thailand and Vietnam.	Philippines	LOC	1	0	2	0	0
Error Class 2							
Their other marksmen were Brazilian defender Vampeta ___ Belgian striker Luc Nilis, his 14th of the season.	and	O	2	0	1	0	0
On Monday and Tuesday, students from the YIT and the university launched street protests against what they called unfair handling by police of a brawl between some of their colleagues and restaurant owners in ___ .	October	O	1	0	3	0	0
Alessandra Mussolini, the granddaughter of ___ 's Fascist dictator Benito Mussolini, said on Friday she had rejoined the far-right National Alliance (AN) party she quit over policy differences last month.	Italy	LOC	3	0	1	1	0
Public Service Minister David Dofara, who is the head of the national Red Cross, told Reuters he had seen the bodies of former interior minister ___ Grelombe and his son, who was not named.	Christophe	PER	1	2	0	0	0
The longest wait to load on the West ___ was 13 days .	Coast	O	1	0	0	2	0
, 41, was put to death in ___ 's electric chair Friday .	Florida	LOC	0	2	2	0	0
Wall Street, since the bid, has speculated that any deal between Newmont and ___ Fe would be a "bear hug, " or a reluctantly negotiated agreement where the buyer is not necessarily a friendly suitor.	Santa	ORG	2	2	1	0	0
Sanity check errors							
Lang said he ___ conditions proposed by Britain's Office of Fair Trading, which was asked to examine the case last month.	supported	O	0	0	0	2	2
___ Vigo 15 5 5 5 17 17 20	Celta	ORG	0	1	0	0	2
The years I spent as manager of the Republic of ___ were the best years of my life.	Ireland	LOC	1	2	2	0	0

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under grant no. NSF 1901117.

References

Agarwal, Oshin, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2020. Entity-switched data sets: An approach to auditing the in-domain robustness of named entity recognition models.

Agichtein, Eugene and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM Conference on Digital Libraries*, pages 85–94, San Antonio, TX. DOI: https://doi.org/10.1145/336597.336644

Akbik, Alan, Larysa Visengeriyeva, Johannes Kirschnick, and Alexander Löser. 2013. Effective selectional restrictions for unsupervised relation extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1312–1320, Nagoya.

Augenstein, Isabelle, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83. DOI: https://doi.org/10.1016/j.csl.2017.01.012

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, Suntec. DOI: https://doi.org/10.3115/1699765.1699767

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 7:2670–2676, New York, NY.

Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231. DOI: https://doi.org/10.1023/A:1007558221122

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Chersoni, Emmanuele, Adrià Torrens Urrutia, Philippe Blache, and Alessandro Lenci. 2018. Modeling violations of selectional restrictions with distributional semantics. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 20–29, Santa Fe, NM.

Collins, Michael and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Cucerzan, Silviu and David Yarowsky, 2002. Language independent NER using a unified model of internal and contextual evidence. *In COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://www.aclweb.org/anthology/W02-2007

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, MN.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134. DOI: https://doi.org/10.1016/j.artint.2005.03.001

Framis, Francesc Ribas. 1994. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*, pages 769–774, Kyoto. DOI: https://doi.org/10.3115/991250.991270

Fu, Jinlan, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethinking generalization of neural models: A named entity recognition case study. DOI: https://doi.org/10.1609/aaai.v34i05.6276

Graves, Alex and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint*

Conference on Neural Networks, 2005, 4:2047–2052.

Grishman, Ralph and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 466–471, Copenhagen. DOI: https://doi.org/10.1162/neco.1997.9.8.1735, PMID: 9377276

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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

Kazama, Jun'ichi and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, Prague.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA.

Lample, Guillaume, 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, CA. DOI: https://doi.org/10.18653/v1/N16-1030

Liao, Wenhui and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 58–65, Stroudsburg, PA. DOI: https://doi.org/10.3115/1621829.1621837, PMID: 18992300

McDonald, David. 1993. Internal and external evidence in the identification and semantic categorization of proper names. In *Acquisition of Lexical Knowledge from Text*. MIT Press, pages 32–43.

Miller, Scott, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 337–342, Boston, MA.

Nadeau, David, Peter D. Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 266–277, Berlin. DOI: https://doi.org/10.1007/11766247_23

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha. DOI: https://doi.org/10.3115/v1/D14-1162

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, LA. DOI: https://doi.org/10.18653/v1/N18-1202

Pradhan, Sameer S. and Nianwen Xue. 2009. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, CO. DOI: https://doi.org/10.3115/1620950.1620956

Ratinov, Lev 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, CO. DOI: https://doi.org/10.3115/1596374.1596399

Riloff, Ellen and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 474–479, USA.

Talukdar, Partha Pratim, Thorsten Brants, Mark Liberman, and Fernando Pereira. 2006. A context pattern induction method

for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 141–148, New York City.

Tjong Kim Sang, Erik F. 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.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, and others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. DOI: https://doi.org/10.18653/v1/2020.emnlp-demos.6, PMCID: PMC7365998

Yadav, Vikas and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, NM.