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

Oshin Agarwal
University of Pennsylvania
oagarwal@seas.upenn.edu

Yinfei Yang
Google AI
yinfeiy@google.com

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

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 LSTM-CRF architectures for named entity recognition, with some provided only representations of the context as features. We also perform similar experiments for BERT. We find that context representations do contribute to system performance, but that the main factor driving high performance is learning the name tokens themselves. We enlist human annotators to evaluate the feasibility of inferring entity types from the context alone and find that, while people are not able to infer the entity type either for the majority of the errors made by the *context-only* system, there is some room for improvement. A system should be able to recognize any name in a predictive context correctly and our experiments indicate that current systems may be further improved by such capability.

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 datasets 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 et al., 2017) and when entities are

switched with a diverse set of entities even within the same dataset (Agarwal et al., 2020).

In this paper, we examine the interpretability and explainability 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, i.e., that only a person may abdicate the throne, 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 (cf. Table 1 in (Bikel et al., 1999)).

As a foundation for our work, we conduct experiments with BiLSTM-CRF models (Huang et al., 2015) modified to use *only context* representations or *only word* identities to quantify the extent to which systems exploit word and context evidence (Section 3). We test these systems on three different datasets to identify trends that generalize across corpora. We show that context does somewhat inform system predictions, but the major driver of perfor-

mance is recognition of certain words as names of a particular type. We also modify the full system to use 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 briefly 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 people 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, people are not able 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. People sometimes make the same mistakes as the system, which may hint as to why conventional systems tend to ignore the context features: the number of examples where relying more on contextual features will lead to more accurate prediction are almost the same number as those where relying more on the context features will lead to an erroneous prediction.

We finish with some oracle experiments with systems in Section 3 and discuss the implications of our findings for the direction of future research.

2 Related Work

Most past effort has been spent on learning to recognize certain words as names, either from training data, using gazetteers and most recently from pretrained representations of words. 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 et al., 2017). 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 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). It is not obvious to what extent learning systems are effectively 'better' look-up models, or if they actually learn to recognize contexts that suggest specific entity types.

Even contemporary systems that do not use gazetteers expand their knowledge of names through the use of pre-trained 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 et al., 2004). Both in the original paper and the re-implementation 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 would also be 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. Table 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 to 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"

---

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

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 original CoNLL lookup baseline with a system that uses only predictive contexts learned from the training data, and an expanded baseline drawing on pre-trained word representations which 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 et al., 1999; Collins and Singer, 1999; Agichtein and Gravano, 2000; Etzioni et al., 2005; Banko et al., 2007) and finding 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.

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, 2002; Nadeau et al., 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 et al., 2015). In these approaches the learning of names is powered by the pre-trained 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. The 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 pre-training data that can better capture a diversity of contexts but at the same time make it difficult to interpret the system prediction and which parts of the input to the system led to a particular prediction. Contextualized representations can in principle disambiguate the meaning of words based on their context, e.g., 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 paper, we present experiments designed to disentangle to the extent possible the contribution of the two sources of certainty in system predictions. We perform in-depth experiments on systems using non-contextualized word representations and a human/system comparison with contextualized representation system.

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 which vary the representations fed into a sequential Conditional Random Field (CRF) (Lafferty et al., 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.

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 CRF. Any word in training or testing that does not have a
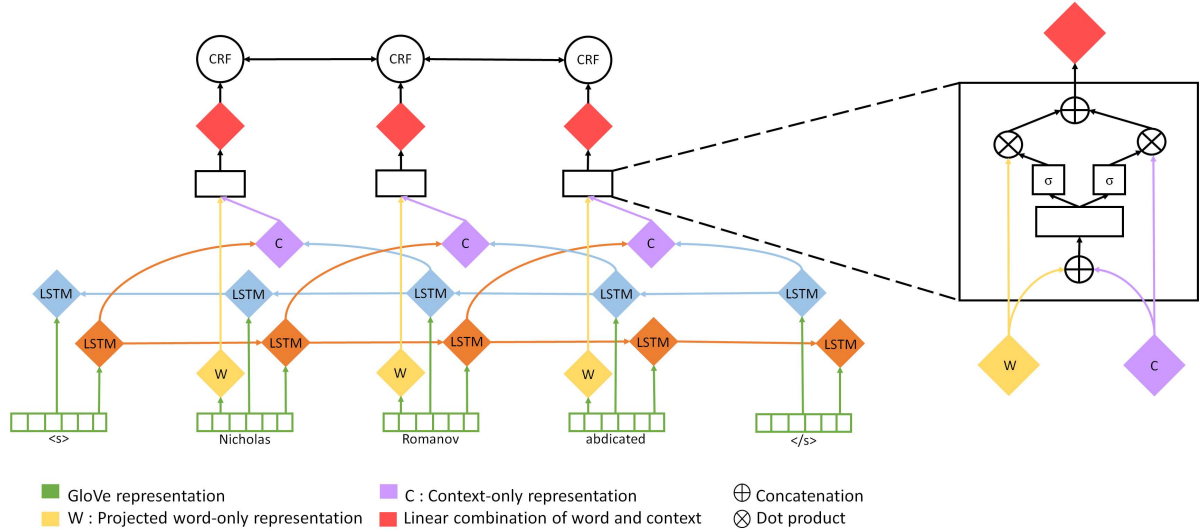
Figure 1: Architecture of the gated system. For each word, a token only (yellow) and a context only (purple) representation is learned. These are combined using gates, as illustrated on the right, and fed into a CRF.

GloVe representation is assigned representation equal to the average of all words represented in GloVe. The GloVe input vectors are fixed in this setting, i.e., 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 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 et al., 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 et al., 2014) vectors trained on Common Crawl.[2] 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.

While 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 -

Gated BI + word + CRF Bidirectional LSTM that uses both the context and the word, but the two representations are combined using gates. The gate values for both the word and the context are based on the concatenated word and context representation. The architecture is illustrated in Figure 1. This provides a mechanism for revealing the degree to which the prediction was influenced by the word versus the context.

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

---

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

Full BERT We use the original public large model[3] and apply the default fine-tuning strategy.

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

We do not build a word-only BERT because having a single word in an attention-based system where the pre-training objective involves predicting the word based on the context does not seem as meaningful.

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 dataset (Balasuriya et al., 2009), to assess how dataset-dependent the performance of the system is. Table 1 shows the results. The last line in the table, BI context + word corresponds to the system presented in Huang et al. (2015).

The results in the *Word only* rows reveal notable differences between the CoNLL and MUC datasets. The Lookup system achieves low recall, as expected, but is not the worst system when trained on CoNLL: in this setting all systems that rely on the context alone, without taking the identity of the word into account, have worse F1 than the Lookup system. This is not the case at all for the system trained on MUC-6. For this dataset, context only systems have double the F1 performance of the Lookup system. This behavior may be attributed to the dataset: many of the entities in the CoNLL training data also appear in testing, a known undesirable fact (Augenstein et al., 2017). Recall for the Lookup approach on CoNLL is 57%, whereas it is only 18% on MUC-6. Moreover, the names

in the CoNLL dataset appear to be less ambiguous in terms of their class than those in MUC-6. The Lookup method achieves 84% precision on CoNLL, and only 67% on MUC-6.

The use of word representations contributes substantially to system performance, especially for the MUC-6 dataset 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.

Next, we consider the systems in the *Word + local context* rows. CRFs help recognize high precision entity-type local contextual constraints, e.g., 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 context types were used in Liao and Veeramachaneni (2009) for semi-supervised NER. The observed improved precision and recall of GloVe finetuned + CRF over LogReg indicates that CRFs help modestly improve performance. However, finetuning the representations on the training set is far more important than including such constraining constraints with CRFs as fixed GloVe + CRF performs consistently worse than LogReg.

Finally, we compare *Context only* systems with non-local context. In CoNLL data, the context after a word appears to be more predictive, while 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 being locations, with the example occurring in reports of sports outcomes, such as 'France 0 Italy 1'. MUC-6, on the other hand, contains many examples that includes honorifics, such as 'Mr. X'. 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 datasets). Next, we use the Gated BI + word + CRF system in Figure 1 to investigate what it focuses on when it has access to both the word and the context. 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 av-

---

[3] cased_L-24_H-1024_A-16

[4] We tried a few techniques such as projections that did not work well

System	CoNLL			Wikipedia			MUC-6		
	P	R	F1	P	R	F1	P	R	F1
Words only									
Lookup	84.1	56.6	67.7	66.3	28.5	39.8	67.1	18.2	28.7
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 finetuned + 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
Full system									
BI context + word + CRF	90.7	91.3	91.0	66.6	60.8	63.6	90.1	91.8	90.9
BERT									
Full	91.9	93.1	92.5	75.4	75.1	75.2	96.1	97.2	96.7
Context-only	43.1	64.1	51.6	39.7	76.2	52.2	75.6	71.6	73.5

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.

	Context	Span
ENT correct	0.906	0.831
ENT incorrect	0.906	0.651
O correct	0.613	0.897
O incorrect	0.900	0.613

Table 2: Mean gate values 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/span gate has a much higher value when the prediction is correct. The word identity itself is the major driver of performance.

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	-

Table 3: Examples of human evaluation where the context-only system was correct but humans incorrect.

erage 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, while one would have expected that the context features have high precision but 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 system. 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.

Finally, we experiment with BERT. The results are in the last two rows of Table 1. Full BERT improves in F1 over the biLSTM as reported in the original paper. Context-only BERT does not perform as well and performs better or worse than context-only LSTM depending on the corpora. These results show BERT isn't always better at capturing contextual clues. While it is better in certain cases, it also misses these clues in certain cases where the biLSTM is correct. Its higher performance on the full dataset could be a result of having a better pretraining data or learning the subword structure. We leave this analysis for the future. In this work, we only focus on extent of context vs word utilization by the systems. Moreover, even when one system performs better than the other, both are correct on different examples. 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 only 71.5% examples in Sample-C. However, it is also able to

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 ...	Norilsk	ORG	✓	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 ...	Institute	ORG	✓	O	LOC
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	✓
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
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

Table 4: Examples of human evaluation.

correctly recognize the entity type in 53.22% of the cases in Sample-I.

Next, we describe an study with people, testing if they can be more successful in using contextual cues to figure out the type of entities.

5 Human Evaluation

We performed a human study to determine if people can infer entity types solely from the context in which they appear. 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 and/or O (not named entity). We allow them to select multiple options. The human answer is taken to be the majority label selected by them. We select 200 instances for which the context-only biL-STM made errors and 20 instances in which it was correct. These 20 instances serve as a sanity check as well as a check for annotator quality. For 85% of these correctly labeled instances, the majority label provided by the annotators was the same as the true (and predicted) label. Table 3 shows the three (out of 20) examples in which people did not agree on the category or made a wrong guess.

In contrast, we received a variety of responses for the 200 instances with errors. Below we describe the results from the study. We break down the human predictions in two classes. Examples of each are shown in Table 4.

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

Error class 2 Human incorrect or no majority - For 55.5% cases, there was a human majority label but it was not the same as the true label. For 21%, there was no majority label in the human study. In either same, humans could not predict the entity type from only the context.

In sum, a person could correctly guess the type without seeing the target word for less than a quarter of the errors made by the biLSTM model.

In contrast, BERT has correct answers in both the error classes. It was correctly able to determine the entity type for 65.9% cases in error class 1 and 49.3% of cases in error class 2. These results show that both the systems aren't 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.

6 Oracle Experiments

In the human evaluation we saw some mixed results, with some room for improvement on 23% of the errors on one side and some errors that seem to be due to over-reliance on context on the other. This leads us to wonder if a more sophisticated

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
Oracle systems									
FW context – BW context	94.3	62.6	75.3	87.2	34.4	49.3	95.5	77.1	85.3
FW context – BW context – Glove finetuned	97.8	87.2	92.2	95.5	58.2	72.4	98.4	91.7	94.9
Full system – Bi context only	93.6	92.2	92.9	77.7	63.7	70.0	94.1	94.6	94.3
Full system – FW context – BW context	95.2	92.7	93.9	81.9	66.9	73.6	96.6	94.9	95.8
Full system – FW context – BW context – Glove finetuned	96.4	93.9	95.1	85.3	69.8	76.8	97.0	95.2	96.1

Table 5: Performance of GloVe word-based BiLSTM-CRF. 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.

approach that decides how to combine cues would lead to a better system. We perform an oracle experiments where the oracle knows which of the systems is correct. If neither is correct, it defaults to one of the system. The results are shown in Table 5. 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 which looks at both these contexts together. The gains are about 15, 20 and 24 points F1 on CoNLL, Wiki and MUC-6 respectively. This improvement likely captures many of the examples that people got right but not the context-only system.

We performed more such experiments with the full system and the word-only and context-only systems. These are shown in row 2 onwards. In each case, there are gains over the full BiLSTM-CRF. An oracle with the four systems (last row) shows the highest gains with ~4 points F1 on CoNLL and 6 points F1 on MUC-6. The gains are especially pronounced in case of cross-domain evaluation i.e. 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. All the oracle experiments show room for improvement and future work would involve looking into strategies/systems to combine these components better. The progress towards this can be measured by breaking down the systems and conducting oracle experiments as here.

7 Discussion and Future Work

In this paper we zero in on the question of interpretability of named entity recognition systems, specifically examining the performance of systems that represent differently the current word, the context and their combination. We test the systems on two corpora and one tested across domains and show that some of the answers to these questions are times corpus dependent. 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 has to be overridden by evidence from the current word. Moreover, we find that contextual representations, namely BERT aren't always better at capturing context as compared to systems such as Glove-based biLSTMs. Their higher performance could be 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 context utilization.

Furthermore, we carry out a human study to test the ability of people to predict the type of an entity without seeing the entity word. People 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 people as well. People can guess the correct label only in about a quarter of all such examples. The opportunity for improvement from better contextual representations exists but is relatively small. Future work in NER would have to expand the vocabulary of entities instead 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 pre-training data from diverse sources for better coverage or by incorporating resources such as gazetteers in the contextual systems.

The human study also shows that the systems are not capturing the same information as humans. BERT is able to correctly recognize the type from the context even when human fails to do so in many cases. Another direction for future work would involve collecting human reasoning behind their answers and incorporating this information in building the systems.

Another promising direction for the overall improvement of NER systems appears to be the better combination of features representing different types of context and the word. Oracle experiments shows that different parts of the sentence – word, forward context, backward context – can help recognize entities correctly when used standalone but not when used as the input together. A simple concatenation of features is not as meaningful and that a smarter combination of several types of features can lead to better performance.

References

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

Eugene Agichtein 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. ACM.

Alan Akbik, 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, Japan.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Dominic Balasuriya, 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, Singapore. Association for Computational Linguistics.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007.

Open information extraction from the web. In *Ijcai*, volume 7, pages 2670–2676.

Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. 1999. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Emmanuele Chersoni, 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.

Michael Collins 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*.

Ronan Collobert, 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.

Silviu Cucerzan 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)*.

Jacob Devlin, 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, Minnesota. Association for Computational Linguistics.

Oren Etzioni, 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.

Francesc Ribas Framis. 1994. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

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

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

Junichi Kazama 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.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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

Wenhui Liao 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, USA. Association for Computational Linguistics.

David McDonald. 1993. Internal and external evidence in the identification and semantic categorization of proper names. In *Acquisition of Lexical Knowledge from Text*.

Scott Miller, 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.

David Nadeau, 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. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, 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, Louisiana. Association for Computational Linguistics.

Sameer S. Pradhan 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, Colorado. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.

Ellen Riloff, Rosie Jones, et al. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479.

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

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Vikas Yadav 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.