

# Named Entity Disambiguation for Noisy Text

Anonymous ACL submission

## Abstract

We address the task of Named Entity Disambiguation (NED) for noisy text. We present WikilinksNED, a large-scale NED dataset of short web-page fragments. In contrast to existing news-based datasets WikilinksNED is significantly noisier and less coherent, making it more difficult to disambiguate. We propose a model based on Attention-RNNs to both exploit the sequential nature of text and attend to the most important signals in it. We evaluate both on WikilinksNED and on a standard, smaller, news-based dataset and find our model significantly outperforms existing state-of-the-art methods on WikilinksNED while achieving reasonable performance on the smaller dataset.

## 1 Introduction

General comment about citations – take whatever bibs you can from the ACL anthology: <http://aclweb.org/anthology/> The bibs from Google scholar lack a lot of information.

Named Entity Disambiguation (NED) is the task of linking mentions of entities in text to a given knowledge base, such as Freebase or Wikipedia. NED is a key component in Entity Linking (EL) systems, focusing on the disambiguation task itself, independently from the tasks of Named Entity Recognition (detecting mention bounds) and Candidate Generation (retrieving the set of potential candidate entities). NED has been recognized as an important component in semantic parsing (Berant and Liang, 2014), as well as other NLP tasks.

NED algorithms can broadly be divided into local and global approaches. Local algorithms dis-

ambiguate each mention independently using local context (e.g. the sentence in which the mention appeared), whereas global approaches assume coherence among mentions within a single document, and try to disambiguate all mentions simultaneously. Global algorithms have significantly outperformed the local approach on standard datasets (Guo and Barbosa, 2014; Pershina et al., 2015; Globerson et al., 2016). However, most of these datasets are based on news corpora and Wikipedia, which are naturally coherent, well-structured, and rich in context. Other domains, such as web page fragments, social media, or questions, lack the sufficient coherence and context for global models to pay off. Take for example this fragment taken from the web:

“I had no choice but to experiment with other indoor games. I was born in Atlantic City so the obvious next choice was **Monopoly**. I played until I became a successful Captain of Industry.”

This fragment is considerably less structured and with a more personal tone than news reports. It clearly references the entity *Monopoly\_(Game)*, however expressions such as ‘experiment’, and ‘Industry’ can generate a lot of noise when disambiguating *Monopoly\_(Game)* from the much more common entity *Monopoly* (economics term). Some sense of local semantics must be considered in order to separate the useful signals (e.g. indoor games, played) from the noisy ones.

In this work, we investigate the task of NED in a setting where only local and noisy context is available. In particular, we create a dataset of 3.2M short text fragments extracted from web pages, each containing a mention of a named entity. Our dataset contains 18K unique mentions linking to over 100K unique entities. This dataset is two orders of magnitude larger than previously

collected ones such as CoNLL-YAGO (Hoffart et al., 2011), TAC KBP (Ji et al., 2010) and ACE 2010 (Bentivogli et al., 2010). We have empirically found our dataset to be significantly noisier and more challenging than standard datasets, suggesting new approaches that can better model semantics are required.

We propose a novel neural network architecture based on recurrent neural networks (RNNs) with an attention mechanism. Our model differs from non-neural approaches by automatically learning feature representations for entity and context, allowing it to extract features from noisy and unexpected context patterns where it can be hard to manually design useful features. We differ from existing neural-based approaches by accounting for the sequential nature of textual context using RNNs and by adapting an attention model to reduce the impact of noise by assigning weights to different contextual signals based on the specific candidate entity being evaluated.

We also describe a novel method for initializing word and entity embeddings, and demonstrate its importance for model performance and training efficiency.

We demonstrate our model significantly outperforms existing state-of-the-art NED algorithms on the noisy web-based dataset, showing that RNNs with attention are adapt to modeling short and noisy context. In addition, we evaluate our algorithm on the CoNLL-YAGO dataset (Hoffart et al., 2011), a dataset of annotated newswire articles. We use a simple domain adaptation technique since CoNLL-YAGO lacks a large enough training set for our model, and achieve comparable results to other state-of-the-art methods. We conclude that RNNs with attention are well-suited for disambiguation in real-world scenarios where only a short, noisy and less coherent context is given.

## 2 Related Work

Early work on Named Entity Disambiguation, such as Bunescu and Paşca (2006) and Mihalcea and Csomai (2007) have focused on local approaches where a mention is disambiguated using hand-crafted statistical and contextual features. While providing a hard-to-beat baseline (Ratinov et al., 2011), increasing attention was recently given to global approaches, which add a layer of sophistication on top of local approaches by considering the coherency of entities assign-

ment within a document. For example the local component of the GLOW algorithm (Ratinov et al., 2011) was exploited as part of the Relational inference system suggested by Cheng and Roth (2013). Similarly, Globerson et al. (2016) achieved state-of-the-art results by extending the local-based selective-context model of Lazic et al. (2015) with an attention-like coherence mechanism. However, global approaches are inherently restricted to documents of sufficient length and coherency.

The first published attempt of using DNNs for NED was by He et al. (2013), which used stacked auto-encoders to learn a similarity measure between mention-context structures and entity candidates. Recently the increasing popularity of DNNs has inspired a number of works that used Convolutional Neural Nets (CNN) for learning semantic similarity between context, mention and candidate inputs (Sun et al., 2015; Francis-Landau et al., 2016). Neural Embedding techniques have also inspired a number of works that measure entity-context relatedness using entity and context embeddings (Yamada et al., 2016; Hu et al., 2015).

In this paper, we train a Recurrent Neural Network (RNN) model, which unlike CNNs and most methods purely based on embeddings, are naturally adapt to exploit the sequential structure of text. Moreover, Lazic et al. (2015) have used a probabilistic attention-like model and have shown only few context words have value in disambiguating a mention. We therefore experimented with a neural version of an attention mechanism.

Chisholm and Hachey (2015) experimented with using web-link data from the Wikilinks corpus (Singh et al., 2012) for training a disambiguation algorithm. They have shown that despite the noisy nature of web data, augmenting Wikipedia derived data with web-links can lead to improved performance on standard datasets. We make use of a subset of Wikilinks as a noisy test case in itself, rather than using it for training alone. Moreover, we argue that manually designing features for noisy text can be difficult and therefore use DNNs to automatically capture useful features.

Commonly used benchmarks for NED systems have mostly focused on news-based corpora. CoNLL-YAGO is a dataset based on Reuters newswire articles that was created by Hoffart et al. (2011) by hand-annotating the CoNLL 2003 Named Entity Recognition task dataset with

YAGO entities. It contains 1393 documents split into train, development and test sets. TAC KBP 2010 (Ji et al., 2010) is another, smaller, dataset for NED based on news articles. ACE 2005 corpus is another news based dataset annotated by Bentivogli et al. (2010). Ratnikov et al. (2011) have used a random sample of paragraphs from Wikipedia for evaluation, however they did not make the precise sample they used publicly available.

Our WikilinksNED dataset is substantially different from currently available datasets since these are all based on high quality content from either news-articles or Wikipedia, while WikilinksNED is a test-case for generally noisier, less coherent and lower quality data. The annotation process is significantly different as well, as our dataset reflects the annotation preferences of real-world website authors and is not annotated by either experts, or a high-quality community effort in the case of Wikipedia. It is also significantly larger in size, being two orders of magnitude larger than existing news-based datasets.

Recently, a number of Tweet based datasets have been composed as well (Fromreide et al., 2014). These represent a much more extreme case than our dataset in terms of noise, shortness and spelling variations, and are much smaller in size. Due to the unique nature of Tweet data, proposed algorithms tend to be substantially different from algorithms used for other NED tasks.

### 3 WikilinksNED Dataset: Entity Mentions in the Web

We introduce a new large-scale NED dataset based on text fragments from the web. Our dataset is derived from the Wikilinks corpus (Singh et al., 2012), which was constructed by crawling the web and collecting hyperlinks (mentions) linking to Wikipedia concepts (entities) and their surrounding text (context). Wikilinks contains 40 million mentions covering 3 million entities, collected from over 10 million web pages.

Wikilinks can be seen as a large-scale, naturally-occurring, crowd-sourced dataset where thousands of human annotators provide ground truths for mentions of interest. This means that the dataset also contains various kinds of noise, including erroneous ground-truth labels, malformed mentions, and incoherent contexts. The contextual noise in particular presents an interesting test-case

that supplements existing datasets that are sourced from mostly coherent and well-formed text (news and Wikipedia).

To get a sense of textual noise and entity-context coherence we have set up a small experiment where we measured the similarity between entities mentioned in WikilinksNED and their surrounding context, and compared the results to CoNLL-YAGO. We used state-of-the-art word and entity embeddings obtained from Yamada et al. (2016) and computed cosine similarity between an entity embedding and the mean of context words embeddings. We compared results from all mentions in CoNLL-YAGO to a sample of 50000 web fragments taken from WikilinksNED, using a window of words of size 40 around entity mentions. We have found the similarity between context and the correct entity is indeed lower for web mentions, and found this result to be statistically significant with very high probability ( $p < 10^{-5}$ ). We believe this result indicates that web fragments in WikilinksNED are indeed less coherent and noisier compared to CoNLL-YAGO documents.

We prepared our dataset from the local-context version of Wikilinks<sup>1</sup>, and resolved ground-truth links using a Wikipedia dump from April 2016<sup>2</sup>. We used the *page* and *redirect* tables for resolution, and kept the database *pageid* column as a unique identifier for Wikipedia entities. We discarded mentions where the ground-truth could not be resolved (only 3% of mentions).

We collected all pairs of mention  $m$  and entity  $e$  appearing in the dataset, and computed the number of times  $m$  refers to  $e$  ( $\#(m, e)$ ), as well as the conditional probability of  $e$  given  $m$ :  $P(e|m) = \#(m, e) / \sum_{e'} \#(m, e')$ . Examining these distributions revealed many mentions belong to two extremes – either they had very little ambiguity, or they appeared in the dataset only a handful of times and referred to different entities only a couple of times each. We deemed the former to be less interesting for the purpose of NED, and suspected the latter to be noise with high probability. To filter these cases, we kept only mentions for which at least two different entities have 10 mentions each ( $\#(m, e) \geq 10$ ) and consist of at least 10% of occurrences ( $P(e|m) \geq 0.1$ ). This procedure aggressively filtered our dataset and we were left with 3.2M mentions.

<sup>1</sup><http://www.iesl.cs.umass.edu/data/wiki-links>

<sup>2</sup><https://dumps.wikimedia.org/>

Not sure I understand the second case. Example? YOTAM: tried to improve explanation.

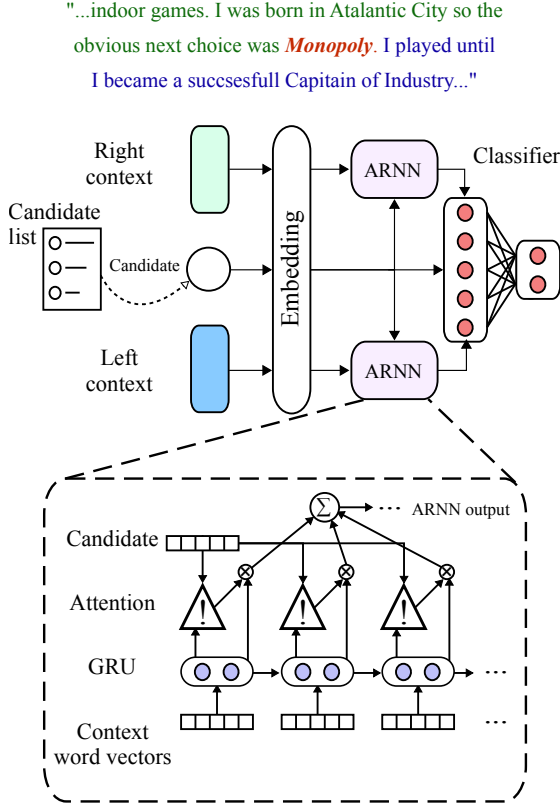


Figure 1: Architecture of our Neural Network model. A close up of the Attention-RNN component appears in the dashed box.

Finally, we randomly split the data into train (90%), validation (10%), and test (10%), according to website domains in order to minimize lexical memorization (see (Levy et al., 2015)).

## 4 Algorithm

Our DNN model is a discriminative model which takes a pair of local context and candidate entity, and outputs a likelihood for the candidate entity being correct. Both words and entities are represented using embedding dictionaries and we interpret local context as a window-of-words to the left and right of a mention. The left and right contexts are fed into a duo of Attention-RNN (ARNN) components which process each side and produce a fixed length vector representation. The resulting vectors along with the entity embedding are then fed into a classifier network with two output units that are trained to emit the likelihood of the candidate being a correct or corrupt assignment.

### 4.1 Model Architecture

Our architecture has two main components: a duo of ARNNs, each processing one side of the con-

text (left and right)<sup>3</sup>, and a classifier. The classifier network consists of a hidden layer<sup>4</sup> and an output layer with two output units in a softmax. The output units are trained to emit the likelihood of the candidate being a correct or corrupt assignment by optimizing a cross-entropy loss function.

Equation 1 represents the general semantics of an RNN unit. An RNN reads a sequence of vectors  $\{v_t\}$  and maintains a hidden state vector  $\{h_t\}$ . At each step a new hidden state is computed based on the previous hidden state and the next input vector using some function  $f$ , and an output is computed using  $g$ . This allows the RNN to 'remember' important signals while scanning the context and to recognize signals spanning multiple words.

$$\begin{aligned} h_t &= f_{\Theta_1}(h_{t-1}, v_t) \\ o_t &= g_{\Theta_2}(h_t) \end{aligned} \quad (1)$$

In our implementation we have used a standard GRU unit (Cho et al., 2014), but any RNN can be a drop-in replacement. We fit the RNN unit with an additional attention mechanism, commonly used with state-of-the-art encoder-decoder models (Bahdanau et al., 2014; Xu et al., 2015). Since our model lacks a decoder, we use the entity embedding as a control signal for the attention mechanism.

Equation 2 details the equations governing the attention model.

$$\begin{aligned} a_t &\in \mathbb{R}; a_t = r_{\Theta_3}(o_t, v_{candidate}) \\ a'_t &= \frac{1}{\sum_{i=1}^t \exp\{a_i\}} \exp\{a_t\} \\ o_{attn} &= \sum_{i=1}^t a'_i o_i \end{aligned} \quad (2)$$

The function  $r$  computes an attention value at each step using  $o_t$ , the RNN output and  $v_{candidate}$ , the candidate entity. The final output vector  $o_{attn}$  is a fixed-size vector, which is the sum of all the output vectors of the RNN weighted according to the attention values. This allows the attention mechanism to decide on the importance of different context parts when examining a specific candidate. We follow Bahdanau et al. (2014) and parametrize the attention function  $r$  as a single layer NN as shown in equation 3.

$$r_{\Theta_3}(o_t, v_{candidate}) = A o_t + B v_{candidate} + b \quad (3)$$

<sup>3</sup>Right context is fed into the ARNN in reverse order

<sup>4</sup>300 dimensions with ReLU, and  $p = 0.5$  dropout.

Omer, check this again. Is it a factor or a probability?



## 4.2 Training

We assume our model is only given examples of correct entity assignments during training and therefore use *corrupt-sampling*, where we automatically generate examples of corrupt assignments. For each context-entity pair  $(c, e)$ , where  $e$  is the correct assignment for  $c$ , we produce  $k$  corrupt examples with the same context  $c$  but with a different, corrupt entity  $e'$ . We have considered two alternatives for corrupt sampling:

**Near-Misses:** Sampling out of the candidate set of each mention. We have found this to be more effective where the training data reliably reflects the test-set distribution.

**All-Entity:** Sampling from the entire dictionary of entities. Better suited to cases where the training data or candidate generation does not reflect the test-set well. Has an added benefit of allowing us to utilize unambiguous training examples where only a single candidate is found.

In our evaluation we specify exactly which approach was used for each experiment and provide an empirical comparison of the two approaches.

We sample corrupt examples uniformly in both alternatives. With uniform sampling the ratio between the number of positive and negative examples of an entity is higher for popular entities, thus biasing the network towards popular entities. In the All-Entity case, this ratio is approximately proportional to the prior probability of the entity.

We note that preliminary experiments revealed that corrupt-sampling according to the distribution of entities in the dataset (as is done by Mikolov et al. (2013)), rather than uniform sampling, produces an interesting entity-context similarity measure. However, it does not perform well in our settings due to the lack of biasing toward popular entities.

Model optimization was carried out using standard backpropagation and an AdaGrad optimizer (Duchi et al., 2011). We allowed the error to propagate through all parts of the network and fine tune all trainable parameters, including the word and entity embeddings themselves. We found the performance of our model substantially improves for the first few epochs and then continues to slowly converge with marginal gains, and therefore trained all models for 8 epochs with  $k = 5$  for corrupt example generation.

## 4.3 Embedding Initialization

Training our model implicitly embeds the vocabulary of words and collection of entities in a common space. However, we find that explicitly initializing these embeddings with vectors pre-trained over a large collection of unlabeled data significantly improved both performance and training speed (see Section 5). To this end, we implemented an SGNS-based approach (Mikolov et al., 2013) that simultaneously trains both word and entity vectors.

We used `word2vecf`<sup>5</sup> (Levy and Goldberg, 2014a), which allows one to train word and context embeddings using arbitrary definitions of "word" and "context" by providing a dataset of word-context pairs  $(w, c)$ , rather than a textual corpus. In our usage, we define a context as an entity  $e$ . To compile a dataset of  $(w, e)$  pairs, we consider every word  $w$  that appeared in the Wikipedia article describing entity  $e$ . We limit our vocabularies to words that appeared at least 20 times in the corpus and entities that contain at least 20 words in their articles. We ran the process for 10 epochs and produced vectors of 300 dimensions; other hyperparameters were set to their defaults.

Levy and Goldberg (2014b) showed that SGNS implicitly factorizes the word-context PMI matrix. Our approach is doing the same for the word-entity PMI matrix, which is highly related to the word-entity TFIDF matrix used in Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007).

## 5 Evaluation

In this section, we describe our experimental setup and compare our model to the state of the art on two datasets: our new WikilinksNED dataset, as well as the commonly-used CoNLL-YAGO dataset (Hoffart et al., 2011). We also examine the effect of different corrupt example generation schemes, and of initializing our model with pre-trained word and entity embeddings.

In all experiments, our model was trained with fixed-size left and right contexts (20 words in each direction). We used a special padding symbol when the actual context was shorter than the window. Further, we filtered stopwords using NLTK's stop-word list prior to selecting the window in order to focus on more informative words.

<sup>5</sup><http://bitbucket.org/yoavgo/word2vecf>

Refer to the exact subsection where you discuss this result

Wikilinks Test-Set Evaluation		
Model	Sampled Test Set (10K)	Full Test Set (300K)
Baseline (MPS)	60	59.6
Cheng et al.	50.7	-
Yamada et al.	67.6	66.9
Our Attention-RNN	73.2	73
Our RNN, w/o Attention	72.1	72.2

Table 1: Evaluation on Web-Fragment data (Wikilinks)

## 5.1 WikilinksNED

**Training** we use Near-Misses corrupt-sampling which was found to perform well due to a large training data that represents the test-data well.

**Candidate Generation** To isolate the effect of candidate generation algorithms, we used the following simple method for all systems: given a mention  $m$ , consider all candidate entities  $e$  that appeared as the ground-truth entity for  $m$  at least once in the training corpus. This simple method yields 97% ground-truth recall on the test set.

**Baselines** Since we are the first to evaluate NED algorithms on WikilinksNED, we ran a selection of existing local NED systems and compared their performance to our algorithm's. **Yamada et al.** (2016) created a state-of-the-art NED system that models entity-context similarity with word and entity embeddings trained using the skip-gram model. We obtained the original embeddings from the authors, and trained the statistical features and ranking model on the WikilinksNED training set. Our configuration of Yamada et al.'s model used only their local features.

**Cheng et al.** (2013) have made their global NED system publicly available<sup>6</sup>. This algorithm uses GLOW (Ratinov et al., 2011) for local disambiguation. We compare our results to the ranking step of the algorithm, without the global component. Due to the long running time of this system, we only evaluated their method on the smaller test set, which contains 10,000 randomly sampled instances from the full 320,000-example test set.

Finally, we include the **Most Probable Sense (MPS)** baseline, which selects the entity that was seen most with the given mention during training.

**Results** We used standard micro P@1 accuracy for evaluation on WikilinksNED. Experimental re-

sults are reported in Table 1. Our algorithm significantly outperforms Yamada et al. on this data by over 5 points. This result indicates that the skip-gram model used by Yamada et al. which averages the embedding vectors of all context words is non-optimal compared to our more sophisticated context modeling on this dataset. Our method outperforms the Baseline as well by a very large margin indicating our RNN model is indeed able to capture meaningful contextual features despite the noisy and short context.

When running Cheng et al (2013) we have used a pre-trained model supplied by Cheng et al which, similarly to the setting used for evaluating the GLOW algorithm by Ratinov et al (Ratinov and Roth, 2011), was not directly trained on our training set. This has resulted in poor performance, emphasizing the greater importance of training a model directly on the training set compared to existing datasets based on news corpora and annotated by experts.

We find that the attention mechanism significantly improves results by attending to the most discriminative parts of the context given a specific candidate mention.

## 5.2 CoNLL-YAGO

**Training** CoNLL-YAGO has a training set with 18505 non-NIL mentions, which preliminary experiments showed is not sufficient to train our model on. We therefore resorted to a more complex training method. We first trained our model on a large corpus of Wikipedia derived data, and fine-tuned on CoNLL-YAGO training set. We then used the model in a similar setting to Yamada et al. (2016) where a GBRT was trained with our model as a feature along with the statistical and string based features defined by Yamada.

To derive the Wikipedia training corpus we have extracted all cross-reference links from Wikipedia along with their context, resulting in over 80 mil-

<sup>6</sup>[https://cogcomp.cs.illinois.edu/page/software/\\_view/Wikifier](https://cogcomp.cs.illinois.edu/page/software/_view/Wikifier)

lion training examples. We set  $k = 5$  for corrupt example generation and trained for 1 epoch with All-Entity corrupt-sampling. The resulting model was then fine-tuned on CoNLL-YAGO training set, where corrupt examples were produced by considering all possible candidates for each mention.

**Candidate Generation** For comparability with existing methods we used two publicly available candidates datasets:

- PPRforNED - Pershina et al. (2015)
- YAGO - Hoffart et al. (2011)

**Baselines** As a baseline we took the standard Most Probable Sense (MPS) prediction, which corresponds to the  $\arg \max_{e \in E} P(e|m)$ , where  $E$  is the group of all candidate entities. We also compare to the following papers - Francis-Landau et al. (2016), Yamada et al. (2016), and Chisholm et al. (2015), as they are all strong local approaches and a good source for comparison.

**Results** The micro and macro P@1 scores on CoNLL-YAGO test-b are displayed in table 2. On this dataset our model achieves reasonable results, however it cannot beat state-of-the-art results since it requires large quantities of training data to properly train the large number of parameters in the model. Further, the relative cleanliness of the data allows existing approaches to perform very well and marginalizes the effect of utilizing a deep and powerful neural model.

CoNLL-YAGO test-b (Local methods)		
Model	Micro P@1	Macro P@1
PPRforNED		
Our ARNN + GBRT	87.3	88.6
Yamada et al. local	90.9	92.4
Yago		
Our ARNN + GBRT	83.3	86.3
Yamada et al. local	87.2	89.6
Francis-Landau et al.	85.5	-
Chisholm et al. local	86.1	-

Table 2: Evaluation on CoNLL-YAGO.

### 5.3 Effects of initialized embeddings and corrupt-sampling schemes

We performed a study of the effects of using pre-initialized embeddings for our model, and of

using either All-Entity or Near-Misses corrupt-sampling. The evaluation was done on a 10% sample of the evaluation set of the WikilinksNED corpus and can be seen in Table 4. We have found that using pre-initialized embeddings gives a much better starting point for training compared to random initialization. This results in faster convergence and a significant performance gain.

Comparison of All-Entity and Near-Misses corrupt-sampling reveals that when using Near-Misses, our model achieves significantly improved performance. We attribute this difference both to the more efficient nature of training with near misses, and to Near-Misses preserving the conditional-prior of entities which is a stronger signal than the prior probability preserved by All-Entity corrupt-sampling.

YOTAM: \*\* = Missing one more epoch for Near-misses, random init. - tomorrow

Wikilinks Evaluation-Set	
Model	Micro accuracy
All-Entity, with init.	70
All-Entity, random init.	67.1
Near-misses, with init.	72.5
Near-misses, random init.	67.2 * *

Table 3: Corrupt-sampling and Initialization

## 6 Error analysis

Omer: Error Analysis needs a table with an overview of how the errors are distributed. Don't worry about the space.

We randomly sampled and manually analyzed 200 individual cases of false disambiguations that our neural model generated. This error subset was obtained from a Wikilinks-based validation set that was not used for training.

Working with crowd-labeled data, we expected some of the mentions to be associated with wrong Wikipedia titles. Accordingly, we found out that 19% (38/200) of the error-labeled predictions were not false, where 66% of the mistakes originated in wrong annotations of the author and 34% were predictions with an equivalent meaning as the correct entity.

Interestingly, in 92% (23/25) of the annotation error events, the model suggested a more convinc-

ing solution than the original author by using specific hints from the context. In this manner, the mention 'Supreme leader', which was contextually associated to the Iranian leader Ali Khamenei, was linked by our model with 'supreme leader of Iran' while the "correct" tag was the general 'supreme leader' entity. From the remaining 162 error-entries, 19% (31/162) were cases where a Wikipedia disambiguation-page was chosen as either the correct or predicted entity (3% and 16%, respectively). We decided to discount this error type from further analysis, as it is logically equivalent to remain unresolved. Eventually we ended up with 131 error cases to analyze.

Error type	Fraction
Annotation errors	18
Disambiguation	
specific general	
cross-ref	
Low count	

Table 4: Error distribution in 200 samples

First, we noticed that in 32% of the errors (42/131) the model selected an entity that can be understood as a specific (7%) or general (25%) realization of the correct solution. For example, instead of predicting 'Fu Manchu mustache' for a facial hair related text, the model addressed the mention with the fictional character 'Fu Manchu', with whom the beard style originated. We further investigated and saw that in almost half of those cases, either the predicted or the correct web-page contained a non-trivial link to the other page, thus implying a strong correlation between both entities. Overall, we observed this type of cross-reference relation in 26% of the error cases (35/131). A closer look discovered two prominent types of cross-reference errors that occurred repeatedly in the data. The first category was of a film/book/theater type of error (9%), where the actual and the predicted entities were a different display of the same narrative. Even though having a different jargon and producers, those fields share extremely similar content, which may explain why they tend to be frequently confused by the algorithm. The second issue of the model was in differentiating between adjectives that are used to describe properties of a nation. Particularity, mentions such as 'Germanic', 'Chinese' and 'Dutch' were falsely assigned to entities that describe lan-

guage instead of people, and vice versa. We observed this type of mistake in 8.4% of the errors (12/131).

A further interesting trend was the prediction policy of the model in cases where the entity had low-count. We defined low-count entities as entities which appeared less than 10 times in the data, and hence were probably not seen in the model's training period. We saw that the model followed the MPS in 75% of the low-count events. This shows that entity-mention pairs, which are alien to the model, tend to bias its prediction towards the baseline. Further, the amount of generalization error in low-count conditions was also significant (35.7%), as the uncertainty of the prediction was captured by a similar but less specific entity suggestion.

Omer: For examples like movie/book - how many of these can be disambiguated consistently by humans? You could argue that some cases are too ambiguous for humans to solve (very common in semantic tasks).

## 7 Conclusions

We believe these results demonstrate that our ARNN model is better than existing state-of-the-art techniques at modeling noisy context when sufficient amounts of training examples are available. This allows us to significantly outperform state-of-the-art algorithms which are found to be suboptimal in a noisy environment such as web content.

However the gap between results of all systems tested on both CoNLL-YAGO and WikilinksNED indicates that web mentions are indeed a more challenging test case than news-based corpora. We believe this to be an important and challenging real-world scenario where noise in the form of mis-structured text, incoherent topics and grammatical/spelling mistakes are present. This scenario represents a distinct test-case that lays in between the standard news-based datasets and the much noisier Tweeter data that has been receiving increasing attention lately. We believe recursive neural models are a promising direction for this task, while there is still much room for improvement.

Omer: Disambiguation pages should be filtered out of the candidate generation process. Mention this as something to fix in future work.



## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko, 2010. *Proceedings of the 2nd Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*, chapter Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia, pages 19–27. Coling 2010 Organizing Committee.
- Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425. Association for Computational Linguistics.
- Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796. Association for Computational Linguistics.
- Andrew Chisholm and Ben Hachey. 2015. Entity disambiguation with web links. *Transactions of the Association of Computational Linguistics*, 3:145–156.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.
- John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.
- Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261. Association for Computational Linguistics.
- Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating ner for twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 1606–1611.
- Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631. Association for Computational Linguistics.
- Zhaochen Guo and Denilson Barbosa. 2014. Entity linking with a unified semantic representation. In *Proceedings of the 23rd International Conference on World Wide Web, WWW ’14 Companion*, pages 1305–1310, New York, NY, USA. ACM.
- Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–34. Association for Computational Linguistics.
- Johannes Hoffart, Amir Mohamed Yosef, Ilaria Bordino, Hagen Fürstenu, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.
- Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric Xing. 2015. Entity hierarchy embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1292–1300. Association for Computational Linguistics.
- Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffith, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, volume 3, pages 3–3.
- Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. *Transactions of the Association of Computational Linguistics*, 3:503–515.
- Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308. Association for Computational Linguistics.

900	Omer Levy and Yoav Goldberg. 2014b. Neural word	<i>Intelligence, IJCAI 2015, Buenos Aires, Argentina,</i>	950
901	embedding as implicit matrix factorization. In <i>Ad-</i>	<i>July 25-31, 2015</i> , pages 1333–1339.	951
902	<i>vances in Neural Information Processing Systems</i>		952
903	27: <i>Annual Conference on Neural Information Pro-</i>	Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun	953
904	<i>cessing Systems 2014, December 8-13 2014, Mon-</i>	Cho, Aaron C. Courville, Ruslan Salakhutdinov,	954
905	<i>trealm, Quebec, Canada</i> , pages 2177–2185.	Richard S. Zemel, and Yoshua Bengio. 2015. Show,	955
906	Omer Levy, Steffen Remus, Chris Biemann, and Ido	attend and tell: Neural image caption generation	956
907	Dagan. 2015. Do supervised distributional meth-	with visual attention. In <i>Proceedings of the 32nd In-</i>	957
908	ods really learn lexical inference relations? In <i>Pro-</i>	<i>ternational Conference on Machine Learning, ICML</i>	958
909	<i>ceedings of the 2015 Conference of the North Amer-</i>	2015, Lille, France, 6-11 July 2015, pages 2048–	959
910	<i>ican Chapter of the Association for Computational</i>	2057.	960
911	<i>Linguistics: Human Language Technologies</i> , pages		961
912	970–976. Association for Computational Linguis-	Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and	962
913	tics.	Yoshiyasu Takefuji. 2016. Joint learning of the em-	963
914	Rada Mihalcea and Andras Csomai. 2007. Wikify!:	bedding of words and entities for named entity dis-	964
915	Linking documents to encyclopedic knowledge. In	ambiguation. In <i>Proceedings of The 20th SIGNLL</i>	965
916	<i>Proceedings of the Sixteenth ACM Conference on</i>	<i>Conference on Computational Natural Language</i>	966
917	<i>Conference on Information and Knowledge Man-</i>	<i>Learning</i> , pages 250–259. Association for Compu-	967
918	<i>agement, CIKM '07</i> , pages 233–242, New York,	tational Linguistics.	968
919	NY, USA. ACM.		969
920	Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S.		970
921	Corrado, and Jeffrey Dean. 2013. Distributed rep-		971
922	resentations of words and phrases and their com-		972
923	positionality. In <i>Advances in Neural Information</i>		973
924	<i>Processing Systems 26: 27th Annual Conference on</i>		974
925	<i>Neural Information Processing Systems 2013. Pro-</i>		975
926	<i>ceedings of a meeting held December 5-8, 2013,</i>		976
927	<i>Lake Tahoe, Nevada, United States.</i> , pages 3111–		977
928	3119.		978
929	Maria Pershina, Yifan He, and Ralph Grishman. 2015.		979
930	Personalized page rank for named entity disam-		980
931	biguation. In <i>Proceedings of the 2015 Conference of</i>		981
932	<i>the North American Chapter of the Association for</i>		982
933	<i>Computational Linguistics: Human Language Tech-</i>		983
934	<i>nologies</i> , pages 238–243. Association for Computa-		984
935	tional Linguistics.		985
936	Lev-Arie Ratinov and Dan Roth. 2011. GLOW		986
937	TAC-KBP2011 entity linking system. In <i>Proceed-</i>		987
938	<i>ings of the Fourth Text Analysis Conference, TAC</i>		988
939	<i>2011, Gaithersburg, Maryland, USA, November 14-</i>		989
940	<i>15, 2011.</i>		990
941	Lev Ratinov, Dan Roth, Doug Downey, and Mike An-		991
942	derson. 2011. Local and global algorithms for dis-		992
943	ambiguation to wikipedia. In <i>Proceedings of the</i>		993
944	<i>49th Annual Meeting of the Association for Com-</i>		994
945	<i>putational Linguistics: Human Language Technolo-</i>		995
946	<i>gies</i> , pages 1375–1384. Association for Computa-		996
947	tional Linguistics.		997
948	Sameer Singh, Amarnag Subramanya, Fernando		998
949	Pereira, and Andrew McCallum. 2012. Wikilinks:		999
	A large-scale cross-document coreference corpus la-		
	beled via links to wikipedia. <i>University of Mas-</i>		
	<i>sachusetts, Amherst, Tech. Rep. UM-CS-2012-015.</i>		
	Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhen-		
	zhou Ji, and Xiaolong Wang. 2015. Modeling men-		
	tion, context and entity with neural networks for en-		
	tity disambiguation. In <i>Proceedings of the Twenty-</i>		
	<i>Fourth International Joint Conference on Artificial</i>		