Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems

Rebecca Marvin becky@jhu.edu
Philipp Koehn phi@jhu.edu
Department of Computer Science, Johns Hopkins University, Baltimore, 21218, USA

Abstract

Neural machine translation systems have been shown to achieve state-of-the-art translation performance for many language pairs. In order to produce a correct translation, MT systems must learn how to disambiguate words with multiple senses and pick the correct translation. We explore the extent to which the word embeddings for ambiguous words are able to disambiguate senses at deeper layers of the NMT encoder, which are thought to represent words with surrounding context. Consistent with previous research, we find that the NMT system fails to translate many ambiguous words correctly. We provide an evaluation framework to use for proposed improvements to word sense disambiguation abilities of NMT systems.

1 Introduction

Neural machine translation systems have to be able to perform many different linguistic tasks successfully in order to obtain good translations. For example, MT systems have to be able to deal with syntactic reordering, semantic relationships, co-reference, and discourse roles, among other phenomena. The obvious question that arises is: how well are state-of-the-art NMT systems doing at detecting linguistic features?

This question is not new. Statistical machine translation (SMT) systems have achieved consistently high BLEU scores because they explicitly try to model features such as word or phrase alignments. For lower-resource languages, SMT systems have been shown to outperform NMT systems, but NMT systems overtake SMT once there is enough training data (Koehn and Knowles, 2017). Recent work has looked at the ability of neural systems to learn syntactic and morphological features. Specifically, Belinkov et al. (2017) showed that recurrent neural networks are able to achieve high accuracy on tasks such as predicting morphological or part of speech tags and Linzen et al. (2016) showed that RNNs follow similar patterns as humans with respect to sentences that are grammatical or ungrammatical in agreement structure. Additionally, specific RNN cells can be shown to have high correlation with features such as sentence length (Karpathy et al., 2016), part of speech (Ding et al., 2017), or whether or not the RNN has finished a relative clause (Linzen et al., 2016).

Another linguistic issue NMT systems have to deal with is translating words in the source language that might have multiple translations in the target language. When these words don't differ orthographically, this task is known as *word sense disambiguation*. Typically, humans can successfully translate these kinds of words by looking at the contexts in which they appear. If NMT systems are able to successfully translate these words, it seems likely that they would have had to learn something about word sense disambiguation.

There has been much research on improving machine translation performance by simultaneously improving word sense disambiguation (Vickrey et al., 2005; Chan et al., 2007; Carpuat and Wu, 2007) for SMT systems, showing that adding word sense disambiguation to a baseline SMT system greatly improves translation performance. For NMT, recent work points out that NMT systems are not very reliable at translating rare word senses, but that disambiguation performance can be improved by using sense embeddings either as additional input to the encoder or to extract more structured lexical chains from the training data (Rios et al., 2017), or by using context-aware embeddings (Liu et al., 2017).

To the best of our knowledge, no work has yet attempted to examine the hidden activations of an NMT system to see whether it is able to disambiguate word senses. In this paper, we present means for evaluating the word sense disambiguation performance of NMT systems. Specifically, we visualize the hidden activations of an NMT encoder to see whether it is able to disambiguate word senses at deeper layers. We also present metrics that represent how well-disambiguated the senses are, with the hope that these metrics can be used to evaluate the word sense disambiguation performance of NMT systems in the future.

Word Sense Disambiguation

Word sense disambiguation (WSD) is the task of figuring out what a word with multiple potential senses means in context. For example, in the sentences below, the word *like* has four different meanings, or senses.

1. **similar:** Her English, *like* that of most people here, is flawless.
2. **speech:** We were *like*, what do we do?
3. **enjoy:** Of the youngers, I really *like* the work of Leo Arill.
4. **request:** I would *like* to be a part of them, but I cannot.

It is crucial for NMT systems to excel at this task in order to produce fluent translations. If the NMT systems do not correctly translate ambiguous words, the resulting translations could be incomprehensible or misleading.

Evaluation metrics have been proposed for assessing word sense disambiguation performance in the past. Lexical choice in MT systems has been evaluated using WSD tasks (Carpuat, 2013) or fill-in-the-blank tasks where the blank represents an ambiguous word (Vickrey et al., 2005), to name a couple methods. These are based on the idea that the entire sentential context should disambiguate the intended word sense. If MT systems are paying attention to the full context, they should be able to succeed at this task.

2 Methodology

We present experiments for examining the word sense disambiguation abilities of the attention-based encoder-decoder model (Bahdanau et al., 2015). In this model, since the encoder computes both forward and backward hidden states after reading the input sequence, each encoder hidden state h_i can be thought of as containing the entire context for the input word i. The idea continues as we add more layers to the encoder: each hidden state h_i should be learning more contextual information about the words surrounding word i. Intuitively, it seems that if the hidden states represent the context for a particular word, then these hidden states would be able to separate words with different senses based on the contexts in which they appear.

In order to formally examine the extent to which the hidden states of the encoder layer(s) of an NMT system disambiguate word senses, we look at the following metrics:

Distinctness. We will extract the hidden states from the last layer of the encoder and compute a principle component analysis (PCA) for these contextualized "embeddings." We can then plot the embeddings for an ambiguous word with different true senses. We also compute two metrics for how well-clustered our embeddings are.

Depth of encoder. We will look at these PCA embedding plots for NMT systems with different numbers of layers in the encoder. Since we are always extracting from the last layer of the encoder, we can get a sense of what the deeper layers in NMT systems are doing with respect to word sense disambiguation.

Correlation with translation performance. It might be that the NMT system only produces well-clustered embeddings for words that it correctly translates. We would like to look at the PCA embedding plots and internal cluster evaluation scores for all four layers when we only include the embeddings for correctly-translated words.

Cluster Measures

We use two intrinsic cluster evaluation metrics to score how well-clustered our resulting embeddings are. These are the *Dunn Index* and the *Davies-Bouldin Index*. We would like our plots to have reasonably distinct clusters which could indicate that word senses are being disambiguated in the encoder. Thus, the purpose of both of these metrics is to identify clusters that are compact and well-separated from other clusters.

The *Dunn Index* is defined as:

$D = \frac{min_{1 \leq i < j \leq n} d(i,j)}{max_{1 \leq k \leq n} d'(k)}$

where $d(i,j)$ represents the distance between cluster medians i and j and $d'(k)$ represents the maximal distance between any two points in cluster k. A higher Dunn Index corresponds to clusters that are dense and well-separated.

The *Davies-Bouldin (DB)* Index is defined as:

$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$

where n is the number of clusters, c_x is the median of cluster x, σ_x is the average distance of all points in cluster x to c_x, and $d(c_i, c_j)$ is the distance between the medians of clusters i and j. A lower DB Index corresponds to clusters that are dense and well-separated.

We hope to find that the Dunn Index increases and the DB Index decreases as we compute these scores for deeper layers of the NMT encoder. This would signify that our word senses were becoming more separated, which would likely correlate with disambiguation performance.

3 Experimental Design

We trained all of our NMT systems using the OpenNMT-py toolkit (Klein et al., 2017), which trains an attentional encoder-decoder model with the attention from Luong et al. (2015). We tokenized, cleaned, and truecased our data using the standard tools from the Moses toolkit (Koehn et al., 2007). We did not use byte-pair encoding in order to more easily do manual annotation of the data later. We used the default parameters of the OpenNMT-py toolkit for training, with the exception of the number of encoder layers, which we varied from 1 to 4.

For the current study, we extensively analyzed WSD performance on sentences containing four possible ambiguous words: *right*, *like*, *last*, or *case*. We manually annotated English sentences with their most appropriate sense (these were our "gold" sense labels), and fed the (un-annotated) sentences into our English-French NMT system. After feeding in the source sentence, we extracted the hidden activations of the NMT encoder and labeled them with their corresponding "gold" sense. We will refer to these hidden activations as the "extracted embeddings," since they are thought to represent a kind of word-and-context embedding.

We performed principle component analysis (PCA) on all of the extracted embeddings and plotted the first two components, where we marked these points based on their "gold" sense label. We then computed internal cluster evaluation scores for all of our embedding "clusters."

Data

The data we used to train our NMT systems comes from the Europarl Corpus (Koehn,

2005) and News Commentary corpus available through the WMT 2014 website. After removing sentences with more than 80 words, this amounted to slightly more than 2.1 million sentences of training data.[1] We used the 2013 news test dataset and the 2014 news test dataset from the WMT 2014 website to validate and to test our trained models, respectively. This amounted to 3000 validation sentences and 3003 test sentences. The 1 layer, 2 layer, 3 layer and 4 layer NMT systems achieved BLEU scores of 23.84, 23.71, 23.77, and 23.94 respectively when tested on the news test 2014 dataset from the WMT 2014 website.

For these initial experiments, we tested our systems on sentences containing one of four ambiguous words: *right*, *like*, *last*, or *case*. Test sentences containing any of these words were manually annotated with their associated sense, and labeled as "unclear" if the sense could not be easily determined from the sentential context. Some examples of sentences containing five different senses of the word *like* can be seen in the introduction.

There were 426 total test sentences that we examined. The number of sentences per each sense of a word is shown in Table 1.

Word	Sense 1	Sense 2	Sense 3	Sense 4	Sense 5	Sense 6	Unclear from context
Right	8	21	12	21	12	1	6
Like	130	1	21	16	n/a	n/a	6
Last	91	6	n/a	n/a	n/a	n/a	1
Case	46	4	16	3	1	n/a	3

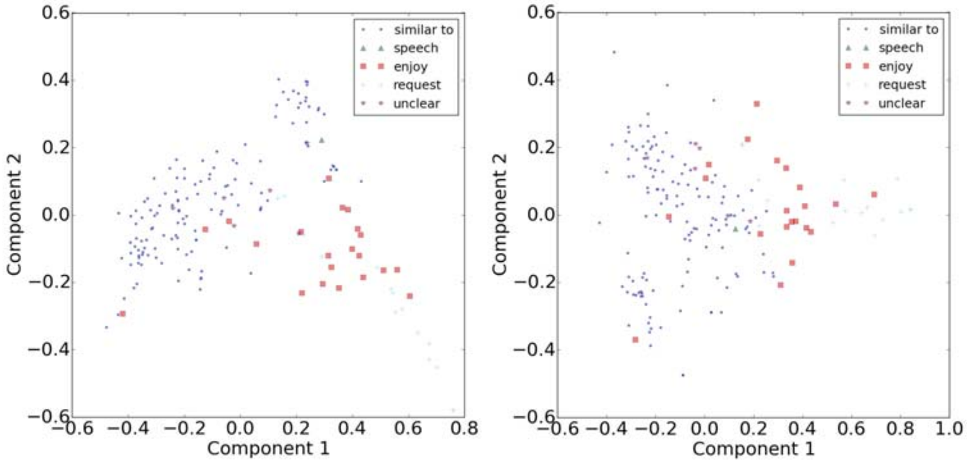
Table 1: Number of sentences for each sense of our ambiguous words. If "n/a" appears in a cell, the word did not have that many distinct senses.

Experiment 1 After removing sentences for which a sense label could not be easily determined from context, we used our manually annotated 410 sentences containing the word *right*, *like*, *last*, or *case* for our gold sense labels. Each sentence was translated by all four of our trained models, and we computed the first two principal components of the extracted embeddings, which were used to compute our internal cluster scores.

Experiment 2 In Experiment 1, we did cluster analysis on the extracted embeddings for *all* sentences containing different senses of our ambiguous words. However, we would expect that senses would be better clustered when the model correctly translates the word, since in that case the model would have had to first choose the correct meaning of the word in context and then translate it. In this experiment, we only looked at the extracted embeddings for sentences where the word *like* or *right* was correctly translated.

Experiment 3 It is possible that the first two principal components of the hidden activations of an NMT encoder might not best represent the amount of word sense information the NMT system is able to learn. That is, the first two components could represent information about the source sentence that has nothing to do with word senses. To examine the extent to which sense information was encoded in the full extracted embeddings, we trained a linear SVM to predict the sense of a word from hidden activations. We trained the SVM on 80% of the test extracted embeddings, and tested it on the remaining 20% of examples. We hope to achieve a high accuracy at this task if sense information was easily accessible from the hidden state vectors.

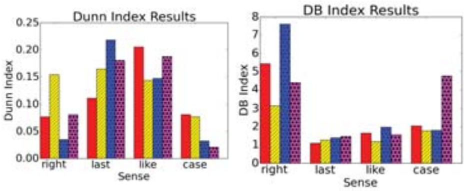
[1] 2 million sentences is enough data for an NMT system to come close to or even outperform an SMT system, according to Koehn and Knowles (2017).



(a) 1 layer embeddings.

(b) 4 layer embeddings.

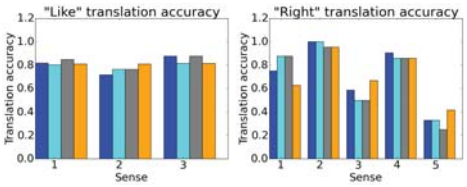
Figure 1: The embeddings of different senses of the word *like*, extracted from the 1 layer and 4 layer models.



(a) Dunn Index results.

(b) DB Index results.

Figure 2: The cluster metrics as we look at different numbers of encoder layers.



(a) *Like* accuracies.

(b) *Right* accuracies.

Figure 3: The translation accuracies for distinct senses for all four models.

4 Results

The plots of the extracted embeddings of different senses of *like* from two of our models can be seen in Figure 1.[2] Visually, the plots seem to show some separation between different senses.

The Dunn and DB Index scores for all four models in Experiment 1 are shown in Figure 2. The different colors represent the numbers of layers, with 1 layer being the leftmost bar and 4 layers being the rightmost within each bar cluster. There does not seem to be a general trend in either index as we look at deeper models.

For Experiment 2, we looked at how the translation accuracy for sentences containing instances of a particular sense varies with the number of layers in the NMT encoder. Figure 3 examines this for the words *right* and *like*.[3] Here again, we would hope to see a general increase in translation accuracy as we increased the number of encoder layers. However, these results and the lack of a general trend in either the Dunn or Davies-Bouldin Index suggest that standard NMT systems still struggle with the issue of word sense disambiguation.

The results for our classifier in Experiment 3 are shown in Table 2. The SVM gets above

[2] The plots for the 2 and 3 layer models looked very similar to these two plots.

[3] We excluded the senses which only had one training example. For *right*, all four models were unable to correctly translate the *90 degrees* sense. For *like*, all four models were able to correctly translate the *speech* sense.

word	1layer	2layer	3layer	4layer
right	0.44	0.63	0.56	0.44
last	0.84	1	0.95	0.84
like	0.88	0.91	0.94	0.94
case	0.79	0.86	0.79	0.57
average	0.76	0.88	0.85	0.74

Table 2: The SVM classifier accuracy at predicting sense from hidden activations.

84% accuracy for the extracted embeddings from all four models for both *last* and *like*, both of which had one sense which was significantly more dominant than the others. The accuracy of the SVM is much lower on *right* and *case*, which have slightly more equal sense distributions.

5 Limitations

Our results hint that standard NMT encoder layers are not encoding enough sentential context to do well at word sense disambiguation. However, we would like to treat these results as a starting point for future evaluations. In particular, we discuss a few limitations of this work:

Manual annotation. It is well-known that obtaining manually annotated data is expensive, sometimes prohibitively so. In this study, we hand-annotated 426 sentences for just four ambiguous words. In the future, we would like to get much more sense-labeled data, either through crowdsourcing to obtain more hand-labeled data, or by using other annotation strategies.

Small test data size. We presented a preliminary study using the ambiguous words *right*, *like*, *last*, and *case*. Perhaps the mixed results could be explained though some particular feature of *right*, and including other words in an evaluation could cancel out that noise. Future work should use more words with multiple senses and more sentences per sense of each word, in order to draw stronger conclusions about word sense disambiguation.

Encoder states. It could be that the NMT system learns how to disambiguate word senses at a different point in the architecture than the encoder. For example, perhaps the NMT system performs the disambiguation step during decoding, thus removing some of the burden of capturing sense information from the encoder. While we believe the NMT encoder should have access to enough sentence context to be able to disambiguate sense, future work could explore whether different components of the NMT architecture more efficiently store sense information.

6 Conclusion & Future Work

Despite these limitations, our preliminary results do suggest that NMT systems still need much improvement in the area of word sense disambiguation. The PCA embedding plots of extracted embeddings at varying levels of the encoder showed some evidence of distinct clusters, but the internal cluster scores varied when we looked at deeper layers of the encoder or considered only sentences that produced correct translations of *right* or *like*.

The results we see are limited by the small sample size we use in our experiments, but we have presented a methodology for examining the word sense disambiguation abilities of NMT systems. These kinds of visualizations and internal cluster evaluation metrics can be used in future research on improving word sense disambiguation in neural machine translation.

Acknowledgements This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via the MATERIAL research program.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *ICLR*.

Belinkov, Y., Durrani, N., Dalai, F., Sajjad, H., and Glass, J. (2017). What do neural machine translation models learn about morphology? *ACL*.

Carpuat, M. (2013). A semantic evaluation of machine translation lexical choice. *ACL*.

Carpuat, M. and Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. *EMNLP*.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. *ACL*.

Ding, Y., Liu, Y., Luan, H., and Sun, M. (2017). Visualizing and understanding neural machine translation. *ACL*.

Karpathy, A., Johnson, J., and Li, F.-F. (2016). Visualizing and understanding recurrent networks. *ICLR*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation.

Koehn, P., Hoang, H., Birch, A., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *ACL*.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *ACL*.

Liu, F., Lu, H., and Neubig, G. (2017). Handling homographs in neural machine translation. *arXiv preprint*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *EMNLP*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *ACL*.

Rios, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. *WMT*.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. *EMNLP*.