What do Neural Machine Translation Models Infer about Multiple Meanings of Homonyms?

Fransisca Susan (fsusan@mit.edu)
James Glass (glass@mit.edu)
Yonatan Belinkov (belinkov@mit.edu)

***Abstract*—Neural machine translation (NMT) has shown promising results in translating text or speech from one language to another and achieved state-of-the-art performances for various language pairs. However, little is known about what these models learn about language. Recent work has started evaluating the quality of vector representations learned by NMT models on morphological, syntactic, and part-of-speech tasks. In this paper, we empirically evaluate the quality of NMT representations for learning the different meanings of homonyms through an extrinsic meaning classification task. We conduct a thorough investigation for both unidirectional and bidirectional NMT models trained with various target languages. We analyze how various target languages used in training NMT models affect how well NMT representations convey the different meanings of homonyms. We observe that the effect of the target languages on source-side representations is more substantial for lower quality NMT models than for higher quality NMT models. Our data-driven, quantitative evaluation serves as the necessary first step to incorporate the best meaning disambiguation system to the models, which would improve NMT model's ability to translate sentences containing homonyms.**[1]

I. Introduction

In many major languages around the world, some words, called homonyms, represent multiple meanings/senses. When a homonym appears in a sentence, a machine translation system needs to know which meaning it represents to obtain the accurate translation. The inability to differentiate homonyms hinders the use of machine translation in many fields such as the legal and medical fields where getting the accurate translation is crucial. For example, the word loop can mean repeatedly doing something or a circular ring in the medical field. When a doctor informs a foreign patient that he needs to get a loop recorder implant, which records the electrical activity of the heart, it is crucial that the translation implies that loop is a circular device instead of a repeated procedure. Machine translation models need to be able to differentiate the meanings of homonyms to produce an adequately accurate translation in many fields.

Studies consistently show that even the state-of-the-art method of machine translation, neural machine translation (NMT), struggles with translating sentences containing words with multiple meanings ([6], [**?**]). Furthermore, word sense (meaning) disambiguation has already improved non-neural machine translation models performance on translating sentences containing homonyms [6]. We are trying to improve the NMT models performance on sentences containing homonyms by allowing them to do word sense disambiguation. We believe that combining a good word sense disambiguation system with NMT models would result in a higher translation accuracy, enabling the models to be more applicable in many fields.

In this paper, we are taking the first step towards improving NMT models in translating homonyms by investigating how well the NMT models differentiate the various meanings of homonyms. Although NMT models are quickly becoming the predominant approach to machine translation, little is known about what and how much NMT models learn about each language and its features. Recent work has started exploring what kind of linguistic information NMT models learn on morphological [10] and syntactic levels ([9], [7]), but research studies are yet to answer important questions in the semantic levels. In this paper, we investigate how well the NMT models differentiate the various meanings of homonyms by providing quantitative, data-driven answers to the following specific questions:

- Do NMT models representations convey any information about the different meanings of homonyms?
- What is the effect of one-directional and bidirectional training on the NMT models representations' ability in differentiating the different meanings of homonyms?
- What impact do different target languages used in training NMT models have on how well the corresponding representations convey the different meanings of homonyms?

[1]Our code is available at https://github.com/fsusan/homonym-repr-analysis.

To achieve this, we follow a simple but effective procedure with three steps: (i) train NMT models on a parallel corpus; (ii) use the trained models to extract feature representations for homonyms in a language of interest; and (iii) train a classifier using extracted features as input to predict the chosen meaning among multiple meanings of a homonym. We then evaluate the quality of trained classifier on the meaning classification task as a proxy to the quality of the extracted representations. We trained both one-directional and bidirectional NMT models using various target languages: French, German, and Finnish. Our analysis reveals interesting insights reported in section V.

II. Related Work

Neural network models are quickly becoming a popular approach to machine translation. NMT systems have become competitive with, or better than, the previous state-of-the-art statistical machine translation system, especially since the introduction of sequence-to-sequence models and the attention mechanism [4]. These sequence-to-sequence NMT models have been implemented across various programming languages, including OpenNMT [5].

Previous work has explored the interpretability of NMT models feature representations in three linguistic areas: syntax, morphology, and semantics. Shi et al. investigate what hidden representations convey about syntax and morphology [9]; Belinkov et al. explore what hidden representations convey about the morphology and POS (part of speech) tagging properties ([10], [11]); while Sennrich et al. has previously analyzed the NMT models capability to perform word sense disambiguation [7].

Our end goal is most similar with Sennrich [7] since we focus on evaluating the NMT models ability on differentiating multiple semantics of homonyms, given a particular context. However, our methodology is most similar to Shi and Belinkov ([9], [10], [11]), who use hidden vectors from an NMT encoder to predict linguistic properties of the source side; Shi and Belinkov used this technique with great success, so we employ a similar one.

In particular, Sennrich creates a word sense disambiguation system by incorporating sense (meaning) embeddings as inputs or parameters to the NMT model. In contrast with Sennrichs work [7], our work trains the NMT model independently and fixes the resulting weights to extract representations, which we later feed to the meaning disambiguation classifier, mimicking the technique used by Shi and Belinkov ([9], [10], [11]).

Besides, Carpuat et al. show that even though machine translation models still struggle with translating sentences containing homonyms, incorporating word sense (meaning) disambiguation has already improved statistical machine translation models ability in translating homonyms [6]. Similarly, we hope to incorporate the best homonyms meaning disambiguation system to the neural machine translation models to improve their ability in translating homonyms, and learning what NMT models infer about the different semantics of homonyms is the necessary first step.

III. Methodology

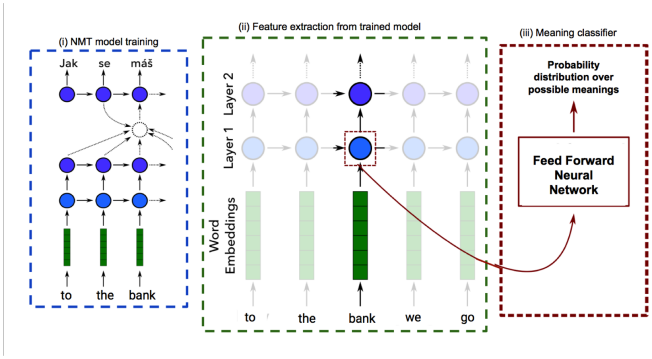


Fig. 1: An overview of our approach: (i) NMT (neural machine translation) model is trained on a parallel corpus, (ii) Features are extracted from the trained models, (iii) Classifiers are trained with extracted features, homonyms, and its context as inputs, producing meaning of the homonym in the context as output. Adapted from Belinkov et al. [10].

An overview of the main methodology is shown in figure 1, which includes three main steps: 1) training the NMT model on a parallel corpus (pairs of sentences in source and target languages), 2) extracting the representations for the homonym from the trained model, and 3) building a classifier to produce a meaning given a homonym and its context.

First, given a parallel corpus of source and target sentence pairs, we train an NMT system with a standard sequence-to-sequence model with attention implemented by Klein et al. as mentioned before ([4], [5]). To extract the vector representations, we freeze/fix the parameters (weights and functions between layers/components of the model) of the NMT model that performs best in the development set. Given a sentence containing a homonym, we use the encoder part of the model to obtain a vector representation of the homonym. If the homonym is the j-th word in the sentence, for layer k, of the model,

we look at h_j^k, which is the output of the k-th layer for the j-th word.

We then build a classifier for each homonym to evaluate the quality of the encoding vector extracted from our trained NMT model. This classifier takes a vector representation of a homonym with its context sentence as input and produces the probability distribution over the possible meanings of the homonym, as seen in figure 2. The final output of this classifier is a meaning of the input homonym. This classifier can be modeled in different ways. However, since we are interested in assessing the quality of the representations learned by the MT system, we choose to model the classifier as a simple feed-forward neural network with one hidden layer and a ReLU non-linearity.

To evaluate how well the representations generated by the NMT model disambiguate word meanings, we measure the performance of the homonym classifiers. We do the same experiments on different target and source languages and compare the accuracy of the homonym classifier we get from each model. We also extract the vector representations of a homonym from the different layers of the model to determine the layer whose corresponding classifier has the highest accuracy (in determining a meaning of a homonym).

IV. Data and Experimental Setup

A. Data

*1) **Machine translation training data**:* We use the EuroSense corpus [1] for training NMT models, which includes 2 million parallel sentences (sentences with the same meanings) in 21 languages. EuroSense is sense-annotated, automatically built via the joint disambiguation of the Europarl corpus with almost 123 million sense annotations for over 155 thousand distinct concept and entities, drawn from the multilingual sense inventory of BabelNet. We train EN-to-* (translation from English to another language specified by *) models on the first 1 million sentences of the train set, splitting them into train/dev/test split. We choose French (FR), German (DE), and Finnish (FI) as our target languages because they have the most number of sense-annotations in the corpus. Our choice of target languages covers an adequately wide range of languages – both languages that are structurally similar with English (German, French, and English are members of Indo-European language family) and language that is structurally different with English (Finnish is a member of Uralic language family). The statistics of the EuroSense corpus on the target languages chosen can be found in Table 1.

*2) **Homonym meaning disambiguation data**:* The homonym meanings classifier is trained on the supervised data derived from EuroSense corpus. We choose fifty most sense-annotated words with multiple meanings in the corpus as our homonym dataset. We use the refined sense-annotations in EuroSense. For each chosen homonym, we extract all sentences containing the homonym, denote the sense of the homonym corresponding to each sentence where it appears, and combine them as our classifier dataset. We split this into train/dev/test set.

B. Experimental Setup

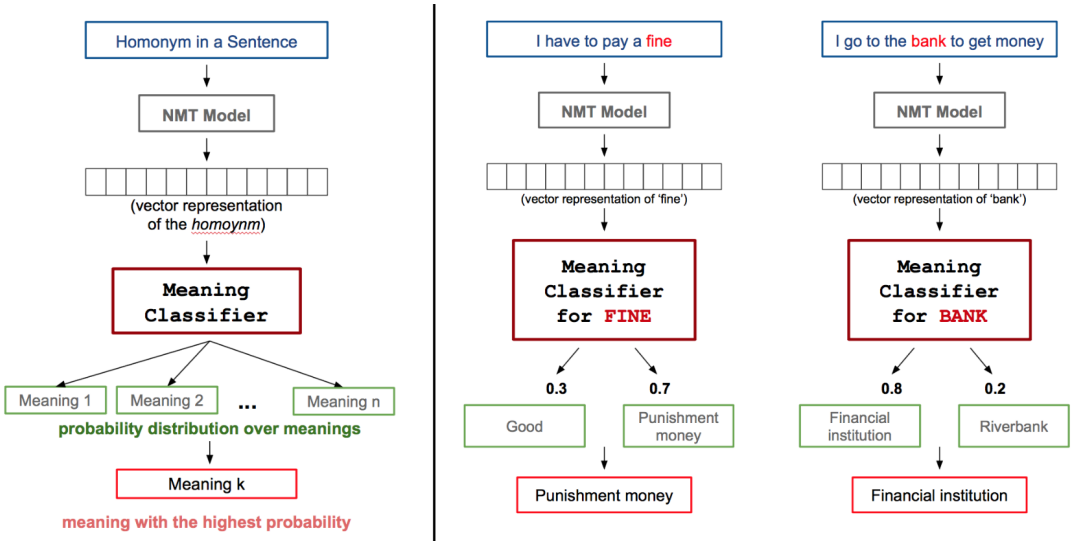
*1) **Neural Machine Translation**:* We use the `seq2seq-attn` toolkit [5] to train 2-layered long short term memory (LSTM) [8] attentional encoder-decoder NMT systems with 500 dimensions for both word embeddings and LSTM states. We compare both unidirectional and bidirectional encoders and experiment with different target languages used in training. Each system is trained with SGD for 20 epochs and the model with the best loss on the development set is used for generating features for the classifier.

*2) **Homonym Meaning Classifier**:* We build a distinct classifier for each chosen homonym. The homonym meaning classifier is modeled as a feed-forward neural network with one hidden layer, dropout ratio of 0.5, a ReLU activation function, and a softmax layer onto the set of the different meanings of the homonym. The hidden layer is of the same size as the input coming from the NMT system (500 dimensions). The classifier has no explicit access to context other than the hidden representation generated by the NMT system, allowing us to focus on the quality of the representation. We chose a simple architecture for the classifier as our main goal is to analyze the quality of the NMT vector representations in differentiating the meanings of homonyms. We train the classifier for 50 epochs each by minimizing the cross-entropy loss using Adam optimizer [3] with default settings. Again, we use the model with the best loss on the development set for evaluation.

*3) **Baseline**:* We use the simplest baselines, the majority baseline: assigning to each homonym in a sentence the meaning that appears the most often for the corresponding homonym according to the training set.

V. Results

Recall that after training the NMT model, we freeze its parameters and use the model only to generate features for the different meaning classifiers. Given a trained



(i) The general architecture for meaning classifier

(ii) Example of Meaning Classifiers

Fig. 2: An overview of the homonym classifier that is distinct for every word. Each takes a homonym and a sentence as input and produces a probability distribution over the possible meanings of the homonym.

TABLE I
THE STATISTICS ON EUROSENSE'S MEANING ANNOTATION

Eurosense Meaning Annotations Data				
	EN	**FR**	**DE**	**FI**
Annotations	15,441,667	12,955,469	9,165,112	8,819,231
Lemmas	42,947	23,603	50,681	31,980
Meanings	86,881	49,189	52,425	52,859
Score	0.28	0.25	0.28	0.26

Table 1. The statistics on Eurosense [1] corpus for the target languages we chose. Distinct lemma denotes the number of different words (in base form) annotated in each language, distinct meaning denotes the number of different meanings found in the annotated words for each language, while score measures how good the meaning annotation is for each language.

encoder ENC in the NMT model, we generate word features ENC_i^k(s) for the homonym in the sentence, given that the homonym is the i-th word in the sentence. Here, k denotes the layer we extract our vector representation from. We then train the meaning classifiers that use the features ENC_i^k(s) as input to predict the corresponding most probable meanings.

A. *Effect of bidirectional model*

Table 2 summarizes the results of training meaning classifiers using representations extracted from trained unidirectional and bidirectional NMT models compared to the majority baseline for several homonyms. In this case, we use English homonyms and train the NMT model with the EN-to-FR parallel corpus, which has the highest BLEU score (quality translation) among others.

TABLE II
AVERAGE ACCURACY OF THE MEANING CLASSIFIERS

	Majority baseline	Unidirectional NMT	Bidirectional NMT
Training	0.59	0.70	**0.74**
Validation	0.58	0.69	**0.71**
Testing	0.55	0.67	**0.71**

Table 2. The comparison of the average accuracy of the homonyms' meaning classifiers on the training, validation, and testing dataset using feature representations from trained unidirectional and bidirectional NMT models (as input), compared to the majority baseline.

The BLEU scores of our unidirectional and bidirectional models are 35.96 and 37.25 respectively. There are two main results we can observe from the table. First, both NMT models produce vector representations for homonyms that are abe to differentiate between the different meanings of homonyms, as shown by roughly 20% increase in average accuracy from the accuracy

TABLE III
THE ACCURACY OF MEANING CLASSIFIER CORRESPONDING TO VARIOUS TARGET LANGUAGES USED IN TRAINING NMT MODELS

Homonym	Majority baseline		EN-FR model		EN-DE model		EN-FI model	
	small	big	small	big	small	big	small	big
apply	0.67	0.68	0.70	0.72	0.75	0.73	**0.79**	**0.77**
new	0.60	0.58	**0.67**	0.70	0.70	0.70	0.73	**0.73**
situation	0.39	0.38	0.52	0.50	0.49	0.49	**0.55**	**0.52**
good	0.72	0.73	0.75	0.75	0.78	0.76	**0.80**	**0.78**
force	0.44	0.44	0.57	**0.57**	0.53	0.55	**0.58**	0.56
principle	0.50	0.50	0.58	0.68	**0.69**	**0.70**	0.68	**0.70**
aid	0.40	0.39	0.52	0.54	0.56	**0.55**	**0.58**	**0.55**
subject	0.51	0.50	0.54	0.54	0.54	0.66	**0.69**	**0.67**
account	0.65	0.65	0.65	0.65	0.65	**0.69**	**0.72**	0.68
freedom	0.65	0.66	0.66	0.67	0.66	0.71	**0.71**	**0.72**

Table 3. The effect of training the NMT models (bidirectional) with various target languages: French (FR), German (DE), Finnish (FI) with big and small dataset on the meaning classifier accuracy. Bold numbers represent the highest accuracy among the three models (blue for the small dataset and magenta for the large dataset.)

we get using majority baseline method. This shows that the in-between vector representations of trained NMT models actually contain some information about the meaning of a homonym in a context (sentence).

Second, bidirectional NMT models produce vector representations for homonyms that better convey the meaning of homonyms than the vector representations produced by unidirectional NMT models, as shown by the higher average accuracy of the meaning classifiers in all training, validation, and testing dataset. One main reason might be because bidirectional NMT models take into account the context from the whole sentence (including the sequence of words after the homonym) to encode an information (vector representation) of a homonym, while unidirectional NMT models' representation of a homonym in a sentence only has the information of all words up to the homonym (in a sentence). This means that bidirectional NMT models' vector representation for a homonym has more context, which is better for meaning interpretation. This is consistent with quality of the translation, as depicted from the BLEU score (bidirectional NMT has higher BLEU score than unidirectional NMT).

These two trends are still true for the accuracy of each homonym's meaning classifier. Thus, we do not present the accuracy data of each meaning classifier.

B. *Effect of target language*

Does translating into different languages make the NMT system learn different source-side representations? In previous work ([10], [11]), the effect of the target language on the quality of encoder representations for POS and morphological tagging tasks is fairly consistent, with differences of 2-3% in accuracy. Here we examine if such an effect exists in differentiating homonym senses.

In generating the parallel corpus used to train the NMT system, we ensure that the homonyms show up the same amount in the training sets for each language. This prevents any bias caused by data skewness, for example if a homonym appears way more in EN-FR corpus but never appears in EN-FI corpus, then the NMT model trained with EN-FI corpus would never learn about the particular homonym.

To amplify the effect of the target language used in training the NMT model on the quality of the model's vector representation of a homonym, in training the NMT model, we use two kinds of dataset (parallel corpus): small and big dataset. For the big dataset, we use all

sentences in the corpus to train the NMT model while for the small dataset, we limit the training set to 200,000 parallel sentences, chosen at random.

Table 3 shows the results of using features obtained by training NMT systems on different target languages (the English source remains fixed) using big and small dataset. We can see small differences with different target languages ~1% for the big dataset. We can see bigger differences with different target languages ~3% for small dataset. While the differences are small, they are mostly statistically significant.

In general, NMT models trained with English-Finnish parallel corpus produce NMT representations that better convey the different meanings of homonyms than NMT models trained with English-German and English-French parallel corpus. This is because a homonym in English is translated into different words in Finnish depending on the senses. When a same word in English is translated into different words in another language (Finnish in this case), the vector representation of the word produced by the model learns to differentiate the meaning based on the context. In contrast, French is more similar to English, so most of the time, a homonym in English is also translated into a homonym in French, and the word vector representations do not learn about differentiating the meanings in this case.

VI. Conclusion

Although a neural network model provides an elegant architecture and a good performance for machine translation, it is difficult to interpret what it learns about linguistic properties. In this work, we analyze whether neural network models learn something about the different semantics in homonyms and investigate the effect of target languages used in training the NMT models on how well the corresponding models infer the different meanings of homonym. In the future, we hope to incorporate a good word sense disambiguation system with NMT models to produce a translation model that can accurately translate sentences containing homonyms. We would also like to extend this work to other semantics tasks that require building relations among word such as evaluating semantic representations on multi-word expressions/phrases. We believe that understanding how linguistic properties are learned in NMT is a crucial step for creating better machine translation systems that can be reliably applied in the real world.

References

[1] Claudio Delli Bovi, Jos Camacho Collados, Alessandro Raganato and Roberto Navigli. EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. Proceedings of 55th annual meeting of the Association for Computational Linguistics (ACL 2017), pages 594600, Vancouver, Canada, 30 July-4 August 2017.

[2] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optiimzation. arXiv preprint arXiv:1412.6980.

[3] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optiimzation. arXiv preprint arXiv:1412.6980.

[4] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate.

[5] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation.

[6] Marine Carpuat, Dekai Wu. Improving Statistical Machine Translation using Word Sense Disambiguation.

[7] Rico Sennrich. 2017. How Grammatical is Characterlevel Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, pages 376382. http://aclweb.org/anthology/E17-2060.

[8] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735-1780.

[9] Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pages 15261534. https://aclweb.org/anthology/D16-1159.

[10] Yonatan Belinkov, Nadin Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume I: Long Papers). Association for Computational Linguistics, pages 861-872.

[11] Yonatan Belinkov, Nadin Durrani, Lluis Marquez, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks.