Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks

Yonatan Belinkov[1] **Lluís Màrquez**[2] **Hassan Sajjad**[2]
Nadir Durrani[2] **Fahim Dalvi**[2] **James Glass**[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
{belinkov, glass}@mit.edu
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
{lmarquez, hsajjad, ndurrani, faimaduddin}@qf.org.qa

Abstract

While neural machine translation (NMT) models provide improved translation quality in an elegant framework, it is less clear what they learn about language. Recent work has started evaluating the quality of vector representations learned by NMT models on morphological and syntactic tasks. In this paper, we investigate the representations learned at different layers of NMT encoders. We train NMT systems on parallel data and use the models to extract features for training a classifier on two tasks: part-of-speech and semantic tagging. We then measure the performance of the classifier as a proxy to the quality of the original NMT model for the given task. Our quantitative analysis yields interesting insights regarding representation learning in NMT models. For instance, we find that higher layers are better at learning semantics while lower layers tend to be better for part-of-speech tagging. We also observe little effect of the target language on source-side representations, especially in higher quality models.[1]

1 Introduction

Neural machine translation (NMT) offers an elegant end-to-end architecture, while at the same time improving translation quality. However, little is known about the inner workings of these models and their interpretability is limited. Recent work has started exploring what kind of linguistic information such models learn on morphological (Vylomova et al., 2016; Belinkov et al., 2017; Dalvi et al., 2017) and syntactic levels (Shi et al., 2016; Sennrich, 2017).

[1]Our code is available at http://github.com/boknilev/nmt-repr-analysis.

One observation that has been made is that lower layers in the neural MT network learn different kinds of information than higher layers. For example, Shi et al. (2016) and Belinkov et al. (2017) found that representations from lower layers of the NMT encoder are more predictive of word-level linguistic properties like part-of-speech (POS) and morphological tags, whereas higher layer representations are more predictive of more global syntactic information. In this work, we take a first step towards understanding what NMT models learn about semantics. We evaluate NMT representations from different layers on a semantic tagging task and compare to the results on a POS tagging task. We believe that understanding the semantics learned in NMT can facilitate using semantic information for improving NMT systems, as previously shown for non-neural MT (Chan et al., 2007; Liu and Gildea, 2010; Gao and Vogel, 2011; Wu et al., 2011; Jones et al., 2012; Bazrafshan and Gildea, 2013, 2014).

For the semantic (SEM) tagging task, we use the dataset recently introduced by Bjerva et al. (2016). This is a lexical semantics task: given a sentence, the goal is to assign to each word a tag representing a semantic class. The classes capture nuanced meanings that are ignored in most POS tag schemes. For instance, proximal and distal demonstratives (e.g., *this* and *that*) are typically assigned the same POS tag (DT) but receive different SEM tags (PRX and DST, respectively), and proper nouns are assigned different SEM tags depending on their type (e.g., geopolitical entity, organization, person, and location). As another example, consider pronouns like *myself*, *yourself*, and *herself*. They may have reflexive or emphasizing functions, as in (1) and (2), respectively:

(1) Sarah bought herself a book

(2) Sarah herself bought a book

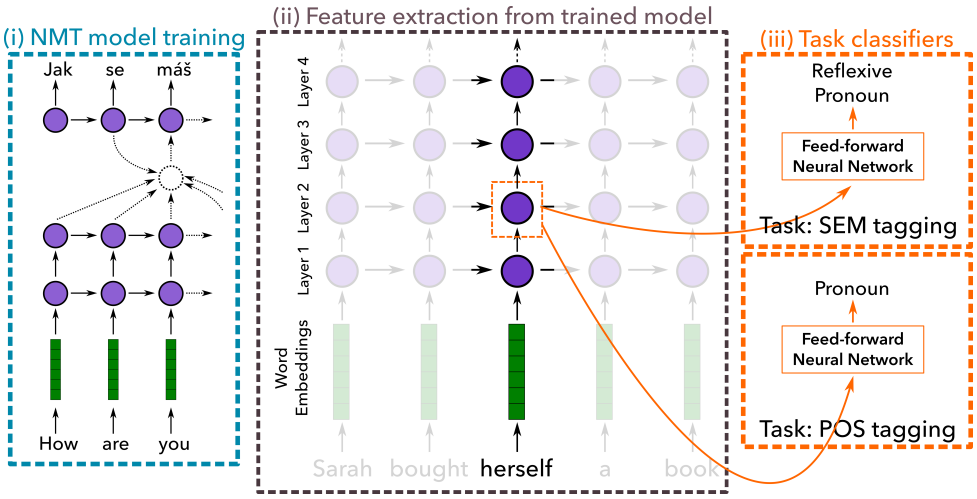

Figure 1: Illustration of our approach, after (Belinkov et al., 2017): (i) NMT system trained on parallel data; (ii) features extracted from pre-trained model; (iii) classifier trained using the extracted features. We train classifiers on either SEM or POS tagging using features from different layers (here: layer 2).

In these examples, *herself* has the same POS tag (PRP) but different SEM tags: REF for a reflexive function and EMP for an emphasizing function.

Capturing semantic distinctions of this sort can be important for producing accurate translations. For instance, example (1) would be translated to Spanish with the reflexive pronoun *se*, whereas (2) would be translated with the intensifier *misma*. Thus, a translation system needs to learn different representations of *herself* in the two sentences.

In order to assess the quality of the representations learned by NMT models, we adopt the following methodology from Shi et al. (2016) and Belinkov et al. (2017). We first train an NMT system on parallel data. Given a sentence, we extract representations from the pre-trained NMT model and train a word-level classifier to predict a tag for each word. Our assumption is that the performance of the classifier reflects the quality of the representation for the given task.

We compare POS and SEM tagging quality with representations from different layers or from models trained on different target languages, while keeping the English source fixed. Our results yield useful insights on representation learning in NMT:

- Consistent with previous work, we find that lower layer representations are usually better for POS tagging. However, we also find that representations from higher layers are better at capturing semantics, even though these are word-level labels. This is especially true with tags that are more semantic in nature such as discourse functions or noun concepts.
- In contrast to previous work, we observe little effect of the target language on source-side representation. We find that the effect of target language diminishes as the size of data used to train the NMT model increases.

2 Methodology

Given a parallel corpus of source and target sentence pairs, we train an NMT system with a standard sequence-to-sequence model with attention (Bahdanau et al., 2014; Sutskever et al., 2014). After training the NMT system, we fix its parameters and treat it as a feature generator for our classification task. Let $\mathbf{h}_j^k$ denote the output of the k-th layer of the encoder at the j-th word. Given another corpus of sentences, where each word is annotated with a label, we train a classifier that takes $\mathbf{h}_j^k$ as input features and maps words to labels. We then measure the performance of the classifier as a way to evaluate the quality of the representations generated by the NMT system. By extracting different NMT features we can obtain a quantitative comparison of representation learning quality in the NMT model for the given task. For instance, we may vary k in order to evaluate representations learned at different encoding layers.

In our case, we first train NMT systems on parallel corpora of an English source and several target languages. Then we train separate classifiers for predicting POS and SEM tags using the features $\mathbf{h}_j^k$ that are obtained from the English encoder and evaluate their accuracies. Figure 1 illustrates the process.

3 Data and Experimental Setup

3.1 Data

MT We use the fully-aligned United Nations corpus (Ziemski et al., 2016) for training NMT models, which includes 11 million multi-parallel sentences in six languages: Arabic (Ar), Chinese (Zh), English (En), French (Fr), Spanish (Es), and Russian (Ru). We train En-to-* models on the first 2 million sentences of the train set, using the official train/dev/test split. This dataset has the benefit of multiple alignment of the six languages, which allows for comparable cross-linguistic analysis.

Note that the parallel dataset is only used for training the NMT model. The classifier is then trained on the supervised data (described next) and all accuracies are reported on the English test sets.

Semantic tagging Bjerva et al. (2016) introduced a new sequence labeling task, for tagging words with semantic (SEM) tags in context. This is a good task to use as a starting point for investigating semantics because: *i)* tagging words with semantic labels is very simple, compared to building complex relational semantic structures; *ii)* it provides a large supervised dataset to train on, in contrast to most available datasets on word sense disambiguation, lexical substitution, and lexical similarity; and *iii)* the proposed SEM tagging task is an abstraction over POS tagging aimed at being language-neutral, and oriented to multi-lingual semantic parsing, all relevant aspects to MT. We provide here a brief overview of the task and its associated dataset, and refer to (Bjerva et al., 2016; Abzianidze et al., 2017) for more details.

The semantic classes abstract over redundant POS distinctions and disambiguate useful cases inside a given POS tag. Examples (1-2) above illustrate how fine-grained semantic distinctions may be important for generating accurate translations. Other examples of SEM tag distinctions include determiners like *every*, *no*, and *some* that are typically assigned a single POS tag (e.g., `DT` in the Penn Treebank), but have different SEM tags, reflecting universal quantification (`AND`), negation (`NOT`), and existential quantification (`DIS`), respectively. The comma, whose POS tag is a punctuation mark, is assigned different SEM tags representing conjunction, disjunction, or apposition, according to its discourse function. Proximal and distant demonstratives (*this* vs. *that*) have different SEM tags but the same POS tag. Named-entities, whose POS tag is usually a single tag for proper nouns, are disambiguated into several classes such as geo-political entity, location, organization, person, and artifact. Other nouns are divided into "role" entities (e.g., *boxer*) and "concepts" (e.g., *wheel*), a distinction reflecting existential consistency: an entity can have multiple roles but cannot be two different concepts.

The dataset annotation scheme includes 66 fine-grained tags grouped in 13 coarse categories. We use the silver part of the dataset; see Table 1 for some statistics.

		Train	Dev	Test
POS	Sentences	38K	1.7K	2.3K
	Tokens	908K	40K	54K
SEM	Sentences	42.5K	6.1K	12.2K
	Tokens	937K	132K	266K

Table 1: Statistics of the part-of-speech and semantic tagging datasets.

Part-of-speech tagging For POS tagging, we simply use the Penn Treebank with the standard split (parts 2-21/22/23 for train/dev/test); see Table 1 for statistics. There are 34 POS tags.

3.2 Experimental Setup

Neural MT We use the `seq2seq-attn` toolkit (Kim, 2016) to train 4-layered long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) 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 numbers of layers. 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.

Classifier The 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 label set size.[2] The hidden layer is of the same size as the input coming from the NMT system (i.e., 500 dimensions). The classifier has no explicit access to context other than the hidden representation gen-

[2]We use a non-linear classifier because previous work found that it outperforms a linear classifier, while showing very similar trends (Qian et al., 2016b; Belinkov et al., 2017).

	MFT	UnsupEmb	Word2Tag
POS	91.95	87.06	95.55
SEM	82.00	81.11	91.41

Table 2: POS and SEM tagging accuracy with baselines and an upper bound. MFT: most frequent tag; UnsupEmb: classifier using unsupervised word embeddings; Word2Tag: upper bound encoder-decoder.

erated by the NMT system, which allows us to focus on the quality of the representation. We chose this simple formulation as our goal is not to improve the state-of-the-art on the supervised task, but rather to analyze the quality of the NMT representation for the task. We train the classifier for 30 epochs by minimizing the cross-entropy loss using Adam (Kingma and Ba, 2014) with default settings. Again, we use the model with the best loss on the development set for evaluation.

Baselines and an upper bound we consider two baselines: assigning to each word the most frequent tag (MFT) according to the training set (with the global majority tag for unseen words); and training with unsupervised word embeddings (UnsupEmb) as features for the classifier, which shows what a simple task-independent distributed representation can achieve. For the unsupervised word embeddings, we train a Skip-gram negative sampling model (Mikolov et al., 2013) with 500 dimensional vectors on the English side of the parallel data, to mirror the NMT word embedding size. We also report an upper bound of directly training an encoder-decoder on word-tag sequences (Word2Tag), simulating what an NMT-style model can achieve by directly optimizing for the tagging tasks.

4 Results

Table 2 shows baseline and upper bound results. The UnsupEmb baseline performs rather poorly on both POS and SEM tagging. In comparison, NMT word embeddings (Table 3, rows with $k = 0$) perform slightly better, suggesting that word embeddings learned as part of the NMT model are better syntactic and semantic representations. However, the results are still below the most frequent tag baseline (MFT), indicating that non-contextual word embeddings are poor representations for POS and SEM tags.

k	Ar	Es	Fr	Ru	Zh	En
POS Tagging Accuracy						
0	88.0*	87.9*	87.9*	87.8*	87.7*	87.4*
1	92.4	91.9	92.1	92.1	91.5	89.4
2	91.9*	91.8	91.8	91.8*	91.3	88.3
3	92.0*	92.3*	92.1	91.6**	91.2*	87.9*
4	92.1*	92.4*	92.5*	92.0	90.5*	86.9*
SEM Tagging Accuracy						
0	81.9*	81.9*	81.8*	81.8*	81.8*	81.2*
1	87.9	87.7	87.8	87.9	87.7	84.5
2	87.4*	87.5*	87.4*	87.3*	87.2*	83.2*
3	87.8	87.9*	87.9**	87.3*	87.3*	82.9*
4	88.3*	88.6*	88.4*	88.1*	87.7*	82.1*
BLEU						
	32.7	49.1	38.5	34.2	32.1	96.6

Table 3: SEM and POS tagging accuracy using features from the k-th encoding layer of 4-layered NMT models trained with different target languages. "En" column is an English autoencoder. BLEU scores are given for reference. Statistically significant differences from layer 1 are shown at $p < 0.001^{(*)}$ and $p < 0.01^{(**)}$. See text for details.

4.1 Effect of network depth

Table 3 summarizes the results of training classifiers to predict POS and SEM tags using features extracted from different encoding layers of 4-layered NMT systems.[3] In the POS tagging results (first block), as the representations move above layer 0, performance jumps to around 91–92%. This is above the UnsupEmb baseline but only on par with the MFT baseline (Table 2). We note that previous work reported performance above a majority baseline for POS tagging (Shi et al., 2016; Belinkov et al., 2017), but used a weak global majority baseline (all words are assigned a single tag) whereas here we compare with a stronger baseline that assigns to each word the most frequent tag according to the training data. The results are also far below the Word2Tag upper bound (Table 2).

Comparing layers 1 through 4, we see that in 3/5 target languages (Ar, Ru, Zh), POS tagging accuracy peaks at layer 1 and does not improve

[3]The results given are with a unidirectional encoder; in section 4.5 we compare with a bidirectional encoder and observe similar trends.

at higher layers, with some drops at layers 2 and 3. In 2/5 cases (Es, Fr) the performance is higher at layer 4. This result is partially consistent with previous findings regarding the quality of lower layer representations for the POS tagging task (Shi et al., 2016; Belinkov et al., 2017). One possible explanation for the discrepancy when using different target languages is that French and Spanish are typologically closer to English compared to the other languages. It is possible that when the source and target languages are more similar, they share similar POS characteristics, leading to more benefit in using upper layers for POS tagging.

Turning to SEM tagging (Table 3, second block), representations from layers 1 through 4 boost the performance to around 87-88%, far above the UnsupEmb and MFT baselines. While these results are below the Word2Tag upper bound (Table 2), they indicate that NMT representations contain useful information for SEM tagging.

Going beyond the 1st encoding layer, representations from the 2nd and 3rd layers do not consistently improve semantic tagging performance. However, representations from the 4th layer lead to significant improvement with all target languages except for Chinese. Note that there is a statistically significant difference ($p < 0.001$) between layers 0 and 1 for all target languages, and between layers 1 and 4 for all languages except for Chinese, according to the approximate randomization test (Padó, 2006).

Intuitively, higher layers have a more global perspective because they have access to higher representations of the word and its context, while lower layers have a more local perspective. Layer 1 has access to context but only through one hidden layer which may not be sufficient for capturing semantics. It appears that higher representations are necessary for learning even relatively simple lexical semantics.

Finally, we found that En-En encoder-decoders (that is, English autoencoders) produce poor representations for POS and SEM tagging (last column in Table 3). This is especially true with higher layer representations (e.g., around 5% below the MT models using representations from layer 4). In contrast, the autoencoder has excellent sentence recreation capabilities (96.6 BLEU). This indicates that learning to translate (to any foreign language) is important for obtaining useful representations for both tagging tasks.

	Ar	Es	Fr	Ru	Zh	En
POS	88.7	90.0	89.6	88.6	87.4	85.2
SEM	85.3	86.1	85.8	85.2	85.0	80.7

Table 4: SEM and POS tagging accuracy using features extracted from the 4th NMT encoding layer, trained with different target languages on a smaller parallel corpus (200K sentences).

4.2 Effect of target language

Does translating into different languages make the NMT system learn different source-side representations? In previous work (Belinkov et al., 2017), we found a fairly consistent effect of the target language on the quality of encoder representations for POS and morphological tagging, with differences of ∼2-3% in accuracy. Here we examine if such an effect exists in both POS and SEM tagging.

Table 3 also shows results using features obtained by training NMT systems on different target languages (the English source remains fixed). In both POS and SEM tagging, there are very small differences with different target languages (∼0.5%), except for Chinese which leads to slightly worse representations. While the differences are small, they are mostly statistically significant. For example, at layer 4, all the pairwise comparisons with different target languages are statistically significant ($p < 0.001$) in SEM tagging, and all except for two comparisons (Ar vs. Ru and Es vs. Fr) are significant in POS tagging.

The effect of target language is much smaller than that reported in (Belinkov et al., 2017) for POS and morphological tagging. This discrepancy can be attributed to the fact that our NMT systems in the present work are trained on much larger corpora (10x), so it is possible that some of the differences disappear when the NMT model is of better quality. To verify this, we trained systems using a smaller data size (200K sentences), comparable to the size used in (Belinkov et al., 2017). The results are shown in Table 4. In this case, we observe a variance in classifier accuracy of 1-2%, based on target language, which is consistent with our earlier findings. This is true for both POS and SEM tagging. The differences in POS tagging accuracy are statistically significant ($p < 0.001$) for all pairwise comparisons except for Ar vs. Ru; the differences in SEM tagging accuracy are significant for all comparisons except for Ru vs. Zh.

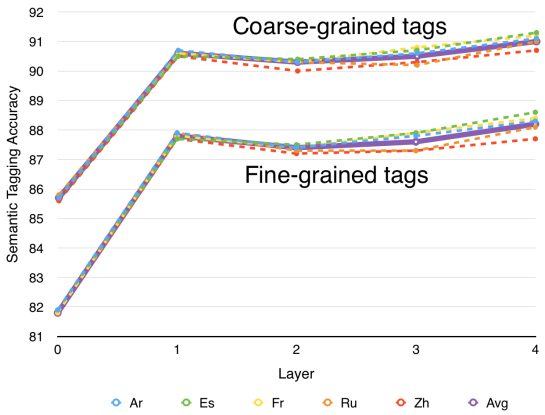

Figure 2: SEM tagging accuracy with fine/coarse-grained tags using features extracted from different encoding layers of 4-layered NMT models trained with different target languages.

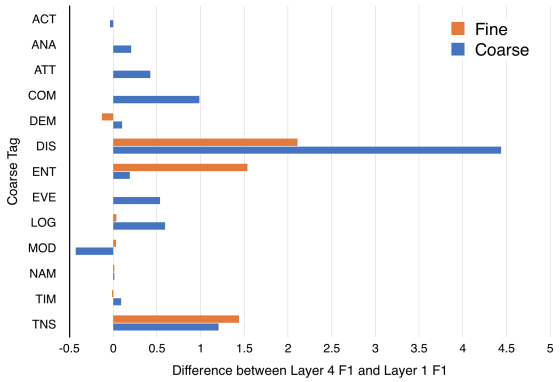

Figure 3: Difference in F_1 when using representations from layer 4 compared to layer 1, showing F_1 when directly predicting coarse tags (blue) and when predicting fine-grained tags and averaging inside each coarse tag (red).

Finally, we note that training an English autoencoder on the smaller dataset results in much worse representations compared to MT models, for both POS and SEM tagging (Table 4, last column), consistent with the behavior we observed on the larger data (Table 3, last column).

4.3 Analysis at the semantic tag level

The SEM tags are grouped in coarse-grained categories such as events, names, time, and logical expressions (Bjerva et al., 2016). In Figure 2 (top lines), we show the results of training and testing classifiers on coarse tags. Similar trends to the fine-grained case arise, with higher absolute scores: significant improvement using the 1st encoding layer and some additional improvement using the 4th layer, both statistically significant ($p < 0.001$). Again, there is a small effect of the target language.

Figure 3 shows the change in F_1 score (averaged over target languages) when moving from layer 1 to layer 4 representations. The blue bars describe the differences per coarse tag when directly predicting coarse tags. The red bars show the same differences when predicting fine-grained tags and micro-averaging inside each coarse tag. The former shows the differences between the two layers at distinguishing among coarse tags. The latter gives an idea of the differences when distinguishing between fine-grained tags within a coarse category. The first observation is that in the majority of cases there is an advantage for classifiers trained with layer 4 representations, i.e., higher layer representations are better suited for learning the SEM tags, at both coarse and fine-grained levels.

Considering specific tags, higher layers of the NMT model are especially better at capturing semantic information such as *discourse relations* (`DIS` tag: subordinate vs. coordinate vs. apposition relations), semantic properties of nouns (*roles* vs. *concepts*, within the `ENT` tag), *events* and *predicate tense* (`EVE` and `TNS` tags), *logic relations* and *quantifiers* (`LOG` tag: disjunction, conjunction, implication, existential, universal, etc.), and *comparative constructions* (`COM` tag: equatives, comparatives, and superlatives). These examples represent semantic concepts and relations that require a level of abstraction going beyond the lexeme or word form, and thus might be better represented in higher layers in the deep network.

One negative example that stands out in Figure 3 is the prediction of the `MOD` tag, corresponding to *modality* (necessity, possibility, and negation). It seems that such semantic concepts should be better represented in higher layers following our previous hypothesis. Still, layer 1 is better than layer 4 in this case. One possible explanation is that words tagged as `MOD` form a closed class, with only a few and mostly unambiguous words ("no", "not", "should", "must", "may", "can", "might", etc.). It is enough for the classifier to memorize these words in order to predict this class with high F_1, and this is something that occurs better in lower layers. One final case worth mentioning is the `NAM` category, which stands for different types of named entities (person, location, organization, artifact, etc.). In principle, this seems a clear case of semantic abstractions suited for higher layers,

	L1	L4	
1	REL	*SUB*	Zimbabwe 's President Robert Mugabe has freed three men who were jailed for murder and sabotage *as* they battled South Africa 's anti-apartheid African National Congress in 1988 .
2	REL	*SUB*	The military says the battle erupted *after* gunmen fired on U.S. troops and Afghan police investigating a reported beating of a villager .
3	IST	*SUB*	Election authorities had previously told Haitian-born Dumarsais Simeus that he was not eligible to run *because* he holds U.S. citizenship .
4	AND	*COO*	Fifty people representing 26 countries took the Oath of Allegiance this week (Thursday) *and* became U.S. citizens in a special ceremony at the Newseum in Washington , D.C.
5	AND	*COO*	But rebel groups said on Sunday they would not sign *and* insisted on changes .
6	AND	*COO*	A Fox asked him , " How can you pretend to prescribe for others , when you are unable to heal your own lame gait *and* wrinkled skin ? "
7	NIL	*APP*	But Syria 's president , Bashar al-Assad , has already rejected the commission 's request [...]
8	NIL	*APP*	Hassan Halemi , head of the pathology department at Kabul University where the autopsies were carried out , said hours of testing Saturday confirmed [...]
9	NIL	*APP*	Mr. Hu made the comments Tuesday during a meeting with Ichiro Ozawa , the leader of Japan 's main opposition party .
10	*AND*	COO	[...] abortion opponents will march past the U.S. Capitol *and* end outside the Supreme Court .
11	*AND*	COO	Van Schalkwyk said no new coal-fired power stations would be approved unless they use technology that captures *and* stores carbon emissions .
12	*AND*	COO	A MEMBER of the Kansas Legislature meeting a Cake of Soap was passing it by without recognition , but the Cake of Soap insisted on stopping *and* shaking hands .

Figure 4: Examples of cases of disagreement between layer 1 (L1) and layer 4 (L4) representations when predicting SEM tags. The correct tag is *italicized* and the relevant word is *underlined*.

but the results from layer 4 are not significantly better than those from layer 1. This might be signaling a limitation of the NMT system at learning this type of semantic classes. Another factor might be the fact that many named entities are out of vocabulary words for the NMT system.

4.4 Analyzing discourse relations

In this section, we analyze specific cases of disagreement between predictions using representations from layer 1 and layer 4. We focus on discourse relations, as they show the largest improvement when going from layer 1 to layer 4 representations (DIS category in Figure 3). Intuitively, identifying discourse relations requires a relatively large context so it is expected that higher layers would perform better in this case.

There are three discourse relations in the SEM tags annotation scheme: subordinate (SUB), coordinate (COO), and apposition (APP) relations. For each of those, Figure 4 (examples 1-9) shows the first three cases in the test set where layer 4 representations correctly predicted the tag but layer 1 representations were wrong. Examples 1-3 have subordinate conjunctions (*as*, *after*, *because*) connecting a main and an embedded clause, which layer 4 is able to correctly predict. Layer 1 mistakes these as attribute tags (REL, IST) that are usually used for prepositions. In examples 4-5, the coordinate conjunction *and* is used to connect sentences/clauses, which layer 4 correctly tags as COO. Layer 1 wrongly predicts the tag AND, which is used for conjunctions connecting shorter expressions like words (e.g., "murder *and* sabotage" in example 1). Example 6 is probably an annotation error, as *and* connects the phrases "lame gait" and "wrinkled skin" and should be tagged as AND. In this case, layer 1 is actually correct. In examples 7-9, layer 4 correctly identifies the comma as introducing an apposition, while layer 1 predicts NIL, a tag for punctuation marks without semantic content (e.g., end-of-sentence period). As expected, in most of these cases identifying the discourse function requires a fairly large context.

Finally, we show in examples 10-12 the first three occurrences of AND in the test set, where layer 1 was correct and layer 4 was wrong. Interestingly, two of these (10-11) are clear cases of *and* connecting clauses or sentences, which should have been annotated as COO, and the last (12) is a conjunction of two gerunds. The predictions from layer 4 in these cases thus appear justifiable.

4.5 Other architectural variants

Here we consider two architectural variants that have been shown to benefit NMT systems: bidirectional encoder and residual connections. We also experiment with NMT systems trained with

different depths. Our motivation in this section is to see if the patterns we observed thus far hold in different NMT architectures.

Bidirectional encoder Bidirectional LSTMs have become ubiquitous in NLP and also give some improvement as NMT encoders (Britz et al., 2017). We confirm these results and note improvements in both translation (+1-2 BLEU) and SEM tagging quality (+3-4% accuracy), across the board, when using a bidirectional encoder. Some of our bidirectional models obtain 92-93% accuracy, which is close to the state-of-the-art on this task (Bjerva et al., 2016). We observed similar improvements on POS tagging. Comparing POS and SEM tagging (Table 5), we note that higher layer representations improve SEM tagging, while POS tagging peaks at layer 1, in line with our previous observations.

Residual connections Deep networks can sometimes be trained better if residual connections are introduced between layers. Such connections were also found useful for SEM tagging (Bjerva et al., 2016). Indeed, we noticed small but consistent improvements in both translation (+0.9 BLEU) and POS and SEM tagging (up to +0.6% accuracy) when using features extracted from an NMT model trained with residual connections (Table 5). We also observe similar trends as before: POS tagging does not benefit from features from the upper layers, while SEM tagging improves with layer 4 representations.

Shallower MT models In comparing network depth in NMT, Britz et al. (2017) found that encoders with 2 to 4 layers performed the best. For completeness, we report here results using features extracted from models trained originally with 2 and 3 layers, in addition to our basic setting of 4 layers. Table 6 shows consistent trends with our previous observations: POS tagging does not benefit from upper layers, while SEM tagging does, although the improvement is rather small in the shallower models.

		0	1	2	3	4
Uni	POS	87.9	92.0	91.7	91.8	91.9
	SEM	81.8	87.8	87.4	87.6	88.2
Bi	POS	87.9	93.3	92.9	93.2	92.8
	SEM	81.9	91.3	90.8	91.9	91.9
Res	POS	87.9	92.5	91.9	92.0	92.4
	SEM	81.9	88.2	87.5	87.6	88.5

Table 5: POS and SEM tagging accuracy with features from different layers of 4-layer Uni/Bidirectional/Residual NMT encoders, averaged over all non-English target languages.

		0	1	2	3	4
4	POS	87.9	92.0	91.7	91.8	91.9
	SEM	81.8	87.8	87.4	87.6	88.2
3	POS	87.9	92.5	92.3	92.4	–
	SEM	81.9	88.2	88.0	88.4	–
2	POS	87.9	92.7	92.7	–	–
	SEM	82.0	88.5	88.7	–	–

Table 6: POS and SEM tagging accuracy with features from different layers of 2/3/4-layer encoders, averaged over all non-English target languages.

5 Related Work

Techniques for analyzing neural network models include visualization of hidden units (Elman, 1991; Karpathy et al., 2015; Kádár et al., 2016; Qian et al., 2016a), which provide illuminating, but often anecdotal information on how the network works. A number of studies aim to obtain quantitative correlations between parts of the neural network and linguistic properties, in both speech (Wu and King, 2016; Alishahi et al., 2017; Belinkov and Glass, 2017; Wang et al., 2017) and language processing models (Köhn, 2015; Qian et al., 2016a; Adi et al., 2016; Linzen et al., 2016; Qian et al., 2016b). Methodologically, our work is most similar to Shi et al. (2016) and Belinkov et al. (2017), who also used hidden vectors from neural MT models to predict linguistic properties. However, they focused on relatively low-level tasks (syntax and morphology, respectively), while we apply the approach to a semantic task and compare the results with a POS tagging task.

Our methodology is reminiscent of the approach taken by Pérez-Ortiz and Forcada (2001), who trained a recurrent neural network POS tagger in two steps. However, their goal was to improve POS tagging while we use it as a task to evaluate neural MT models.

6 Conclusion

While neural network models have improved the state-of-the-art in machine translation, it is difficult to interpret what they learn about language. In this work, we explore what kind of linguistic information such models learn at different layers. Our experimental evaluation leads to interesting insights about the hidden representations in NMT models such as the effect of layer depth and target language on part-of-speech and semantic tagging.

In the future, we would like to extend this work to other syntactic and semantic tasks that require building relations such as dependency relations and predicate-argument structure or to evaluate semantic representations of multi-word expressions. We believe that understanding how semantic properties are learned in NMT is a key step for creating better machine translation systems.

Acknowledgments

This research was carried out in collaboration between the HBKU Qatar Computing Research Institute (QCRI) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247. Association for Computational Linguistics.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *arXiv preprint arXiv:1608.04207*.

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, Vancouver, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic Roles for String to Tree Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–423, Sofia, Bulgaria. Association for Computational Linguistics.

Marzieh Bazrafshan and Daniel Gildea. 2014. Comparing Representations of Semantic Roles for String-To-Tree Decoding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1786–1791, Doha, Qatar. Association for Computational Linguistics.

Yonatan Belinkov, Nadir 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 1: Long Papers)*, pages 861–872. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2017. Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems. In *Advances in Neural Information Processing Systems (NIPS)*.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic Tagging with Deep Residual Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541, Osaka, Japan. The COLING 2016 Organizing Committee.

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*.

Seng Yee Chan, Tou Hwee Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan. Association for Computational Linguistics.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225.

Qin Gao and Stephan Vogel. 2011. Utilizing Target-Side Semantic Role Labels to Assist Hierarchical Phrase-based Machine Translation. In *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 107–115. Association for Computational Linguistics.

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

Bevan Jones, Jacob Andreas, Daniel Bauer, Moritz Karl Hermann, and Kevin Knight. 2012. Semantics-Based Machine Translation with Hyperedge Replacement Grammars. In *Proceedings of COLING 2012*, pages 1359–1376. The COLING 2012 Organizing Committee.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2016. Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and Understanding Recurrent Networks. *arXiv preprint arXiv:1506.02078*.

Yoon Kim. 2016. Seq2seq-attn. `https://github.com/harvardnlp/seq2seq-attn`.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Arne Köhn. 2015. What's in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Ding Liu and Daniel Gildea. 2010. Semantic Role Features for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 716–724. Coling 2010 Organizing Committee.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Sebastian Padó. 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*. `https://www.nlpado.de/~sebastian/software/sigf.shtml`.

Juan Antonio Pérez-Ortiz and Mikel L. Forcada. 2001. Part-of-Speech Tagging with Recurrent Neural Networks. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, volume 3, pages 1588–1592.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016a. Analyzing Linguistic Knowledge in Sequential Model of Sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, Austin, Texas. Association for Computational Linguistics.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016b. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich. 2017. How Grammatical is Character-level 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*, pages 376–382. Association for Computational Linguistics.

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*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2016. Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. *arXiv preprint arXiv:1606.04217*.

Yu-Hsuan Wang, Cheng-Tao Chung, and Hung-yi Lee. 2017. Gate Activation Signal Analysis for Gated Recurrent Neural Networks and Its Correlation with Phoneme Boundaries. *arXiv preprint arXiv:1703.07588*.

Dekai Wu, Pascale N Fung, Marine Carpuat, Chi-kiu Lo, Yongsheng Yang, and Zhaojun Wu. 2011. Lexical Semantics for Statistical Machine Translation. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

Zhizheng Wu and Simon King. 2016. Investigating Gated Recurrent Networks for Speech Synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5140–5144. IEEE.

Micha Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).