On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference

Adam Poliak[1] **Yonatan Belinkov**[2] **James Glass**[2] **Benjamin Van Durme**[1]
[1]Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218
[2]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, MA 02139
`{azpoliak,vandurme}@cs.jhu.edu`, `{belinkov,glass}@mit.edu`

Abstract

We propose a process for investigating the extent to which sentence representations arising from neural machine translation (NMT) systems encode distinct semantic phenomena. We use these representations as features to train a natural language inference (NLI) classifier based on datasets recast from existing semantic annotations. In applying this process to a representative NMT system, we find its encoder appears most suited to supporting inferences at the syntax-semantics interface, as compared to anaphora resolution requiring world-knowledge. We conclude with a discussion on the merits and potential deficiencies of the existing process, and how it may be improved and extended as a broader framework for evaluating semantic coverage.[1]

1. Introduction

What do neural machine translation (NMT) models learn about semantics? Many researchers suggest that state-of-the-art NMT models learn representations that capture the meaning of sentences (Gu et al., 2016; Johnson et al., 2017; Zhou et al., 2017; Andreas and Klein, 2017; Neubig, 2017; Koehn, 2017). However, there is limited understanding of how specific semantic phenomena are captured in NMT representations beyond this broad notion. For instance, how well do these representations capture Dowty (1991)'s thematic proto-roles? Are these representations sufficient for understanding paraphrastic inference? Do the sentence representations encompass complex anaphora resolution? We argue that existing semantic annotations recast as Natural Language Inference (NLI) can be leveraged to investigate whether sentence representations encoded by NMT models capture these semantic phenomena.

DPR	Sara adopted Jill, *she* wanted a child	✗
	Sara adopted Jill, *Jill* wanted a child	
FN+	Iran *possesses* five research reactors	✓
	Iran *has* five research reactors	
SPR	Berry Rejoins WPP Group	✓
	Berry was *sentient*	

Figure 1: Example sentence pairs for the different semantic phenomena. DPR deals with complex anaphora resolution, FN+ is concerned with paraphrastic inference, and SPR covers Reisinger et al. (2015)'s semantic proto-roles. ✓ / ✗ indicates that the first sentence entails / does not entail the second.

We use sentence representations from pretrained NMT encoders as features to train classifiers for NLI, the task of determining if one sentence (a *hypothesis*) is supported by another (a *context*).[2] If the sentence representations learned by NMT models capture distinct semantic phenomena, we hypothesize that those representations should be sufficient to perform well on NLI datasets that test a model's ability to capture these phenomena. Figure 1 shows example NLI sentence pairs with their respective labels and semantic phenomena.

We evaluate NMT sentence representations of 4 NMT models from 2 domains on 4 different NLI datasets to investigate how well they capture different semantic phenomena. We use White et al. (2017)'s *Unified Semantic Evaluation Framework* (USEF) that recasts three semantic phenomena NLI: 1) semantic proto-roles, 2) paraphrastic inference, 3) and complex anaphora resolution. Additionally, we evaluate the NMT sentence representations on 4) Multi-NLI, a recent extension of the Stanford Natural Language Inference dataset (SNLI) (Bowman et al., 2015) that includes multiple genres and domains (Williams et al.,

[1] Code developed and data used are available at `https://github.com/boknilev/nmt-repr-analysis`.

[2] Sometimes referred to as recognizing textual entailment (Dagan et al., 2006, 2013).

Proceedings of NAACL-HLT 2018, pages 513–523
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

2017). We contextualize our results with a standard neural encoder described in Bowman et al. (2015) and used in White et al. (2017).

Based on the recast NLI datasets, our investigation suggests that NMT encoders might learn more about semantic proto-roles than anaphora resolution or paraphrastic inference. We note that the target-side language affects how an NMT source-side encoder captures these semantic phenomena.

2. Motivation

Why use recast NLI? We focus on NLI, as opposed to a wide range of NLP taks, as a unified framework that can capture a variety of semantic phenomena based on arguments by White et al. (2017). Their recast dataset enables us to study whether NMT encoders capture "distinct types of semantic reasoning" under just one task. We choose these specific semantic phenomena for two reasons. First, a long term goal is to understand how combinations of different corpora and neural architectures can contribute to a system's ability to perform general language understanding. As humans can understand (annotate consistently) the sentence pairs used in our experiments, we would similarly like our final system to have this same capability. We posit that it is necessary but not necessarily sufficient for a language understanding system to be able to capture the semantic phenomena considered here. Second, we believe these semantic phenomena might be relevant for translation. We demonstrate this with a few examples.

Anaphora Anaphora resolution connects tokens, typically pronouns, to their referents. Anaphora resolution should occur when translating from morphologically poor languages into some morphologically rich languages. For example, when translating "The parent fed the child because she was hungry," a Spanish translation should describe *the child* as *la niña (fem.)* and not *el niño (masc.)* since *she* refers to *the child*. Because world knowledge is often required to perform anaphora resolution (Rahman and Ng, 2012; Javadpour, 2013), this may enable evaluating whether an NMT encoder learns world knowledge. In this example, *she* refers to *the child* and not *the parent* since world knowledge dictates that parents often feed children when children are hungry.

Proto-roles Dowty (1991)'s proto-roles may be expressed differently in different languages, and so correctly identifying them can be important for translation. For example, English does not usually explicitly mark *volition*, a proto-role, except by using adverbs like *intentionally* or *accidentally*. Other languages mark volitionality by using special affixes (e.g., Tibetan and Sesotho, a Bantu language), case marking (Hindi, Sinhalese), or auxiliaries (Japanese).[3] Correctly generating these markers may require the MT system to encode volitionality on the source side.

Paraphrases Callison-Burch (2007) discusses how paraphrases help statistical MT (SMT) when alignments from source words to target-language words are unknown. If the alignment model can map a paraphrase of the source word to a word in the target language, then the SMT model can translate the original word based on its paraphrase.[4] Paraphrases are also used by professional translators to deal with non-equivalence of words in the source and target languages (Baker, 2018).

3. Methodology

We use NMT models based on bidirectional long short-term memory (Bi-LSTM) encoder-decoders with attention (Sutskever et al., 2014; Bahdanau et al., 2015), trained on a parallel corpus. Given an NLI context-hypothesis pair, we pass each sentence independently through a trained NMT encoder to extract their respective vector representations. We represent each sentence by concatenating the last hidden state from the forward and backward encoders, resulting in $\mathbf{v}$ and $\mathbf{u}$ (in $\mathbb{R}^{2d}$) for the context and hypothesis.[5] We follow the common practice of feeding the concatenation $(\mathbf{v}, \mathbf{u}) \in \mathbb{R}^{4d}$ to a classifier (Rocktäschel et al., 2016; Bowman et al., 2015; Mou et al., 2016; Liu et al., 2016; Cheng et al., 2016; Munkhdalai and Yu, 2017).

Sentence pair representations are fed into a classifier with a softmax layer that maps onto the number of labels. Experiments with both linear and non-linear classifiers have not shown major differences, so we report results with the linear classifier unless noted otherwise. We report implementation details in Appendix B.

[3] For references and examples, see: `en.wikipedia.org/wiki/Volition_(linguistics)`.

[4] Using paraphrases can help NMT models generate text in the target language in some settings (Sekizawa et al., 2017).

[5] We experimented with other sentence representations and their combinations, and did not see differences in overall conclusions. See Appendix A for these experiments.

Test Train	DPR: 50.0					SPR: 65.4					FN+: 57.5				
	ar	es	zh	de	USEF	ar	es	zh	de	USEF	ar	es	zh	de	USEF
DPR	49.8	**50.0**	**50.0**	**50.0**	49.5	45.4	57.1	47.0	43.9	**65.2**	48.0	**55.9**	51.0	46.8	19.2
SPR	50.1	50.3	50.1	49.9	**50.7**	72.1	74.2	73.6	73.1	**80.6**	56.3	57.0	56.9	56.1	**65.8**
FN+	50.0	50.0	**50.4**	50.0	49.5	57.3	**63.6**	54.5	60.7	60.0	56.2	56.1	54.3	55.5	**80.5**

Table 1: Accuracy on NLI with representations generated by encoders of English→{ar,es,zh,de} NMT models. Rows correspond to the training and validation sets and major columns correspond to the test set. The column labeled "USEF" refers to the test accuracies reported in White et al. (2017). The numbers on the top row represents each dataset's majority baseline. Bold numbers indicate the highest performing model for the given dataset.

4. Data

MT data We train NMT models on four language pairs: English → {Arabic (ar), Spanish (es), Chinese (zh), and German (de)}. See Appendix B for training details. The first three pairs use the United Nations parallel corpus (Ziemski et al., 2016) and for English-German, we use the WMT dataset (Bojar et al., 2014). Although the entailment classifier only uses representations extracted from the English-side encoders as features, using multiple language pairs allows us to explore whether different target languages affect what semantic phenomena are captured by an NMT encoder.

Natural Language Inference data We use four distinct datasets to train classifiers: MultiNLI (Williams et al., 2017), a recent expansion of SNLI containing a broad array of domains that was used in the 2017 RepEval shared task (Nangia et al., 2017), and three recast NLI datasets from The JHU Decompositional Semantics Initiative (Decomp)[6] released by White et al. (2017). Sentence-pairs and labels were recast, i.e. automatically converted, from existing semantic annotations: FrameNet Plus (FN+) (Pavlick et al., 2015), Definite Pronoun Resolution (DPR) (Rahman and Ng, 2012), and Semantic Proto-Roles (SPR) (Reisinger et al., 2015). The FN+ portion contains sentence pairs based on paraphrastic inference, DPR's sentence pairs focus on identifying the correct antecedent for a definite pronoun, and SPR's sentence pairs test whether the semantic proto-roles from Reisinger et al. (2015) apply based on a given sentence.[7] Recasting makes it easy to determine how well an NLI method captures the fine-grained semantics inspired by Dowty (1991)'s thematic proto-roles, paraphrastic inference, and complex anaphora resolutions. Table 2 includes the datasets' statistics.

	DPR	SPR	FN+	MNLI
Train	2K	123K	124K	393K
Dev	.4K	15K	15K	9K
Test	1K	15K	14K	9K

Table 2: Number of sentences in NLI datasets.

5. Results

Table 1 shows results of NLI classifiers trained on representations from different NMT encoders. We also report the majority baseline and the results of Bowman et al.'s 3-layer deep 200 dimensional neural network used by White et al. ("USEF").

Paraphrastic entailment (FN+) Our classifiers predict FN+ entailment worse than the majority baseline, and drastically worse than USEF when trained on FN+'s training set. Since FN+ tests paraphrastic inference and NMT models have been shown to be useful to generate sentential paraphrase pairs (Wieting and Gimpel, 2017; Wieting et al., 2017), it is surprising that our classifiers using the representations from the NMT encoder perform poorly. Although the sentences in FN+ are much longer than in the other datasets, sentence length does not seem to be responsible for the poor FN+ results. The classifiers do not noticeably perform better on shorter sentences than longer ones, as noted in Appendix C.

Upon manual inspection, we noticed that in many *not-entailed* examples, swapped paraphrases had different part-of-speech (POS) tags. This begs the question of whether different POS tags for swapped paraphrases affects the accuracies. Using Stanford CoreNLP (Manning et al., 2014), we partition our validation set based on whether the paraphrases share the same POS tag. Table 3 reports dev set accuracies using classifiers trained on FN+. Classifiers using features from NMT encoders trained on the three languages from the UN corpus noticeably perform better on cases where paraphrases have different POS tags compared to paraphrases with the same POS tags. These dif-

[6] decomp.net

[7] We refer the reader to White et al. (2017) for detailed discussion on how the existing datasets were recast as NLI.

	ar	es	zh	de
Same Tag	52.9	52.6	52.6	50.2
Different Tag	55.8	59.1	53.4	46.0

Table 3: Accuracies on FN+'s dev set based on whether the swapped paraphrases share the same POS tag.

ferences might suggest that the recast FN+ might not be an ideal dataset to test how well NMT encoders capture paraphrastic inference. The sentence representations may be impacted more by ungrammaticality caused by different POS tags as opposed to poor paraphrases.

Anaphora entailment (DPR) The low accuracies for predicting NLI targeting anaphora resolution are similar to White et al. (2017)'s findings. They suggest that the model has difficulty in capturing complex anaphora resolution. By using contrastive evaluation pairs, Bawden et al. (2017) recently suggested as well that NMT models are poorly suited for co-reference resolution. Our results are not surprising given that DPR tests whether a model contains common sense knowledge (Rahman and Ng, 2012). In DPR, syntactic cues for co-reference are purposefully balanced out as each pair of pro-nouns appears in at least two context-hypothesis pairs (Table 9). This forces the model's decision to be informed by semantics and world knowledge – a model cannot use syntactic cues to help perform anaphora resolution.[8] Although the poor performance of NMT representations may be explained by a variety of reasons, e.g. training data, architectures, etc., we would still like ideal MT systems to capture the semantics of co-reference, as evidenced in the example in §2.

Even though the classifiers perform poorly when predicting paraphrastic entailment, they surprisingly outperform USEF by a large margin (around 25–30 %) when using a model trained on DPR.[9] This might suggest that an NMT encoder can pick up on how pronouns may be used as a type of lexical paraphrase (Bhagat and Hovy, 2013).

Proto-role entailment (SPR) When predicting SPR entailments using a classifier trained on SPR data, we noticeably outperform the majority baseline but are below USEF. Both ours and USEF's accuracies are lower than Teichert et al. (2017)'s best reported numbers. This is not surprising as Teichert et al. condition on observed semantic role labels when predicting proto-role labels.

Proto-Role	ar	es	zh	de	avg	MAJ
physically existed	70.6	70.8	**77.2**	70.8	72.4†	65.9
sentient	78.5	**82.2**	80.5	81.7	80.7†	75.5
aware	75.9	**77.0**	76.6	76.7	76.6†	60.9
volitional	74.3	**76.8**	74.7	73.7	74.9†	64.5
existed before	68.4	**70.5**	66.5	68.4	68.5†	64.8
caused	69.4	**74.1**	72.2	72.7	72.1†	63.4
changed	64.2	62.4	63.8	62.0	63.1	**65.1**
location	91.1	90.1	90.4	90.2	90.4	**91.7**
moved	90.6	88.8	90.1	90.3	89.9	**93.3**
used in	34.9	38.1	31.8	34.2	34.7	**55.2**
existed after	62.7	69.0	65.6	65.2	65.7	**69.7**
chang. state	61.8	60.7	60.9	60.7	61.0	**65.2**
chang. possession	89.6	88.6	89.9	88.3	89.1	**93.9**
stationary during	86.3	84.4	90.5	86.0	86.8	**96.3**
physical contact	85.0	82.0	84.5	84.4	84.0	**85.8**
existed during	59.3	71.8	60.8	64.4	64.1	**84.7**

Table 4: Accuracies on the SPR test set broken down by each proto-role. "avg" represents the score for the proto-role averaged across target languages. Bold and † respectively indicate the best results for each proto-role and whether all of our classifiers outperformed the proto-role's majority baseline.

Table 4 reports accuracies for each proto-role. Whenever one of the classifiers outperforms the baseline for a proto-role, all the other classifiers do as well. The classifiers outperform the majority baseline for 6 of the reported 16 proto-roles. We observe these 6 properties are more associated with proto-agents than proto-patients.

The larger improvements over the majority baseline for SPR compared to FN+ and DPR is not surprising. Dowty (1991) posited that proto-agent, and -patient should correlate with English syntactic subject, and object, respectively, and empirically the *necessity of [syntactic] parsing for predicate argument recognition* has been observed in practice (Gildea and Palmer, 2002; Punyakanok et al., 2008). Further, recent work is suggestive that LSTM-based frameworks implicitly may encode syntax based on certain learning objectives (Linzen et al., 2016; Shi et al., 2016; Belinkov et al., 2017b). It is unclear whether NMT encoders capture semantic proto-roles specifically or just underlying syntax that affects the proto-roles.

NMT target language Our experiments show differences based on which target language was used to train the NMT encoder, in capturing semantic proto-roles and paraphrastic inference. In Table 1, we notice a large improvement using sentence representations from an NMT encoder that was trained on en-es parallel text. The improvements are most profound when a classifier trained on DPR data predicts entailment focused on se-

[8] Appendix D includes some illustrative examples.

[9] This is seen in the last columns of the top row in Table 1.

	ar	es	zh	de	MAJ
MNLI-1	45.9	45.7	46.6	48.0	35.6
MNLI-2	46.6	46.7	48.2	48.9	36.5

Table 5: Accuracies for MNLI test sets. MNLI-1 refers to the matched case and MNLI-2 is the mismatched.

mantic proto-roles or paraphrastic inference. We also note that using the NMT encoder trained on en-es parallel text results in the highest results in 5 of the 6 proto-roles in the top portion of Table 4. When using other sentence representations (Appendix A), we notice that using representations from English-German encoders consistently outperforms using the other encoders (Tables 6 and 7). This prevents us from making generalizations regarding specific target side languages.

NLI across multiple domains Though our main focus is exploring what NMT encoders learn about distinct semantic phenomena, we would like to know how useful NMT models are for general NLI across multiple domains. Therefore, we also evaluate the sentence representations with Multi-NLI. As indicated by Table 5, the representations perform noticeably better than a majority baseline. However, our results are not competitive with state-of-the-art systems trained specifically for Multi-NLI (Nangia et al., 2017).

6. Related Work

In concurrent work, Poliak et al. (2018) explore whether NLI datasets contain statistical irregularities by training a model with access to only hypotheses. Their model significantly outperforms the majority baseline and our results on Multi-NLI, SPR, and FN+. They suggest that these, among other NLI datasets, contain statistical irregularities. Their findings illuminate issues with the recast datasets we consider, but do not invalidate our approach of using recast NLI to determine whether NMT encoders capture distinct semantic phenomena. Instead, they force us to re-evaluate the majority baseline as an indicator of whether encoders learn distinct semantics and to what extent we can make conclusions based on these recast datasets.

Prior work has focused on the relationship between semantics and machine translation. MEANT and its extension XMEANT evaluate MT systems based on semantics (Lo and Wu, 2011; Lo et al., 2014). Others have focused on incorporating semantics directly in MT. Chan et al. (2007) use word sense disambiguation to help statistical MT, Gao and Vogel (2011) add semantic-roles to improve phrase-based MT, and Carpuat et al. (2017) demonstrate how filtering parallel sentences that are not parallel in meaning improves translation. Recent work explores how representations learned by NMT systems can improve semantic tasks. McCann et al. (2017) show improvements in many tasks by using contextualized word vectors extracted from a LSTM encoder trained for MT. Their goal is to use NMT to improve other tasks while we focus on using NLI to determine what NMT models learn about different semantic phenomena.

Researchers have explored what NMT models learn about other linguistic phenomena, such as morphology (Dalvi et al., 2017; Belinkov et al., 2017a), syntax (Shi et al., 2016), and lexical semantics (Belinkov et al., 2017b), including word senses (Marvin and Koehn, 2018; Liu et al., 2018)

7. Conclusion and Future Work

Researchers suggest that NMT models learn sentence representations that capture meaning. We inspected whether distinct types of semantics are captured by NMT encoders. Our experiments suggest that NMT encoders might learn the most about semantic proto-roles, do not focus on anaphora resolution, and may poorly capture paraphrastic inference. We conclude by suggesting that target-side language affects how well an NMT encoder captures these semantic phenomena.

In future work, we would like to study how well NMT encoders capture other semantic phenomena, possibly by recasting other datasets. Comparing how semantic phenomena are represented in different NMT architectures, e.g. purely convolutional (Gehring et al., 2017) or attention-based (Vaswani et al., 2017), may shed light on whether different architectures may better capture semantic phenomena. Finally, investigating how multilingual systems learn semantics can bring a new perspective to questions of universality of representation (Schwenk and Douze, 2017).

Acknowledgements

This work was supported by JHU-HLTCOE, DARPA LORELEI, and Qatar Computing Research Institute. We thank anonymous reviewers. Views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Jacob Andreas and Dan Klein. 2017. Analogs of linguistic structure in deep representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2893–2897.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513* .

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. 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)*. Association for Computational Linguistics, pages 861–872. https://doi.org/10.18653/v1/P17-1080.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 1–10.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics* 39(3):463–472.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 12–58.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland. http://cis.upenn.edu/~ccb/publications/callison-burch-thesis.pdf.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, pages 69–79. http://aclweb.org/anthology/W17-3209.

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*. Association for Computational Linguistics, pages 33–40.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 551–561.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 103–111. http://www.aclweb.org/anthology/W14-4012.

Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS workshop*. EPFL-CONF-192376.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 670–680. http://aclweb.org/anthology/D17-1070.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, Springer, pages 177–190.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies* 6(4):1–220.

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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 142–151.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language* pages 547–619.

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*. Association for Computational Linguistics, pages 107–115.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*. PMLR, International Convention Centre, Sydney, Australia, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 239–246.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1631–1640.

Seyedeh Leili Javadpour. 2013. *Resolving pronominal anaphora using commonsense knowledge*. Ph.D. thesis, Louisiana State University.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a ViÃ©gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351.

Philipp Koehn. 2017. Neural machine translation. *arXiv preprint arXiv:1709.07809* .

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association of Computational Linguistics* 4(1):521–535.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling Homographs in Neural Machine Translation. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090* .

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. Xmeant: Better semantic mt evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 765–771.

Chi-kiu Lo and Dekai Wu. 2011. Meant: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 220–229.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Rebecca Marvin and Philipp Koehn. 2018. Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems. In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas (Volume 1: Research Track*. pages 125–131.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 6297–6308.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 130–136. `http://anthology.aclweb.org/P16-2022`.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural tree indexers for text understanding. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 11–21.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. pages 1–10.

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619* .

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 408–413.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines for natural language inference. In *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 34(2):257–287.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 777–789. http://www.aclweb.org/anthology/D12-1071.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics* 3:475–488.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. pages 157–167.

Yuuki Sekizawa, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Improving japanese-to-english neural machine translation by paraphrasing the target language. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 64–69.

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 1526–1534. https://aclweb.org/anthology/D16-1159.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew R Gormley. 2017. Semantic proto-role labeling. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 5998–6008.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 996–1005.

John Wieting and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732* .

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 274–285. https://www.aclweb.org/anthology/D17-1026.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* .

Hao Zhou, Zhaopeng Tu, Shujian Huang, Xiaohua Liu, Hang Li, and Jiajun Chen. 2017. Chunk-based bi-scale decoder for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 580–586. https://doi.org/10.18653/v1/P17-2092.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

A. Sentence Representations

In the experiments reported in the main paper, we used a simple sentence representation, the first and last hidden states of the forward and backward encoders. We concatenated them for both the context and the hypothesis and fed to a linear classifier. Here we compare the results of InferSent (Conneau et al., 2017), a more involved representation that was found to provide a good sentence representation based on NLI data. Specifically, we concatenate the forward and backward encodings for each sentence, and max-pool over the length of the sentence, resulting in $\mathbf{v}$ and $\mathbf{u}$ (in $\mathbb{R}^{2d}$) for the context and hypothesis. The InferSent representation is defined by

$$(\mathbf{u}, \mathbf{v}, |\mathbf{u} - \mathbf{v}|, \mathbf{u} * \mathbf{v}) \in \mathbb{R}^{8d}$$

where the product and subtraction are carried element-wise and commas denote vector-concatenation.

The pair representation is fed into a multi-layered perceptron (MLP) with one hidden layer and a ReLU non-linearity. We set the hidden layer size to 500 dimensions, similarly to Conneau et al. (2017). The softmax layer maps onto the number of labels, which is either 2 or 3 depending on the dataset.

InferSent results Table 6 shows the results of the classifier trained on NMT representations with the InferSent architecture. Here, the representations from NMT encoders trained on the English-German parallel corpus slightly outperforms the others. Since this data used a different corpus compared to the other language pairs, we cannot determine whether the improved results are due to the different target side language or corpus. The main difference with respects to the simpler sentence representation (Concat) is improved results on FN+. Table 7 shows the results on Multi-NLI. It is interesting to note that, when using the sentence representations from NMT encoders, concatenating the sentence vectors outperformed the InferSent method on Multi-NLI.

B. Implementation & Experimental Details

We use 4-layer NMT systems with 500-dimensional word embeddings and LSTM states (i.e., $d = 500$). The vocabulary size is 75K words. We train NMT models until convergence and take the models that performed best on the development set for generating representations to feed into the entailment classifier. We use the hidden states from the top encoding layer for obtaining sentence representations since it has been hypothesized that higher layers focus on word meaning, as opposed to syntax (Belinkov et al., 2017a,b). We remove long sentences (> 50 words) when training both the classifier and the NMT model, as is common NMT practice (Cho et al., 2014). During testing, we use all test sentences regardless of sentence length. Our implementation extends Belinkov et al. (2017a)'s implementation in Torch (Collobert et al., 2011).

		FN+	DPR	SPRL
NMT Concat	en-ar	56.2	49.8	72.1
	en-es	56.1	50.0	74.2
	en-zh	54.3	50.0	73.1
	en-de	55.5	50.0	73.1
NMT InferSent	en-ar	57.9	50.0	73.6
	en-es	58.0	50.0	72.7
	en-zh	57.8	49.8	72.4
	en-de	58.3	50.1	73.7
Majority		57.5	50.0	65.4
(White et al., 2017)		80.5	49.5	80.6

Table 6: NLI results on fine-grained semantic phenomena. FN+ = paraphrases; DPR = pronoun resolution; SPRL = proto-roles. NMT representations are combined with either a simple concatenation (results copied from Table 2) or the InferSent representation. State-of-the-art (SOTA) is from White et al. (2017).

		MNLI-1	MNLI-2
NMT Concat	en-ar	45.9	46.6
	en-es	45.7	46.7
	en-zh	46.6	48.2
	en-de	48.0	48.9
NMT InferSent	en-ar	40.1	41.8
	en-es	44.9	40.8
	en-zh	43.7	42.1
	en-de	41.3	41.1
Majority		35.6	36.5
SOTA		81.10	83.21

Table 7: Results on language inference on MultiNLI (Williams et al., 2017), matched/mismatched scenario (MNLI1/2).

We train English→Arabic/Spanish/Chinese NMT models on the first 2 million sentences of the United Nations parallel corpus training set (Ziemski et al., 2016), and the English→German model on the WMT dataset (Bojar et al., 2014). We use the official training/development/test splits.

In our NLI experiments, we do not train on Multi-NLI and test on the recast datasets, or vice-versa, since Multi-NLI since Multi-NLI uses a 3-way classification (*entailment*, *neutral*, and *contradictions*) while the recast datasets use just two labels (*entailed* and *not-entailed*). In preliminary experiments, we also used a 3-layered MLP. Although the results slightly improved, we noted similar trends to the linear classifier.

C. Sentence length

The average sentence in the FN+ test dataset is 31 words and almost 10 % of the test sentences are longer than 50 words. In SPR and DPR, each premise sentence has on average 21 and 15 words respectively and only 1 % of sentences in SPR have more than 50 words. No DPR sentences have > 50 words.

Table 8 reports accuracies for ranges of sentence lengths in FN+'s development set. When trained on sentence representations form an English→Chinese,German NMT encoder, the NLI accuracies steadily decrease. When using English→Arabic, the accuracies stay consistent until sentences have between 70–80 tokens while the results from English→Spanish quickly drops from 0–10 to 10–20 but then stays relatively consistent.

Sentence length	ar	es	zh	de	total
0-10	46.8	63.7	66.0	65.4	526
10-20	49.0	53.3	57.4	56.5	2739
20-30	48.4	54.0	53.2	54.9	4889
30-40	48.4	54.1	51.2	53.9	4057
40-50	47.7	59.0	55.0	58.7	2064
50-60	49.1	56.1	54.5	57.5	877
60-70	46.4	53.6	43.9	44.1	444
70-80	59.9	51.6	43.3	43.3	252

Table 8: Accuracies on FN+'s dev set based on sentence length. The first column represents the range of sentences length: first number is inclusive and second is exclusive. The last column represents how many context sentences have lengths that are in the given row's range.

D. World Knowledge in DPR

When released, Rahman and Ng (2012)'s DPR dataset confounded the best co-reference models because "its difficulty stems in part from its reliance on sophisticated knowledge sources." Table 9 includes examples that demonstrate how world knowledge is needed to accurately predict these recast NLI sentence-pairs.

Chris was running after John, because he stole his watch	
▶ Chris was running after John, because John stole his watch	✓
▶ Chris was running after John, because Chris stole his watch	✗
Chris was running after John, because he wanted to talk to him	
▶ Chris was running after John, because Chris wanted to talk to him	✓
▶ Chris was running after John, because John wanted to talk to him	✗
The plane shot the rocket at the target, then it hit the target	
▶ The plane shot the rocket at the target, then the rocket hit the target	✓
▶ The plane shot the rocket at the target, then the target hit the target	✗
Professors do a lot for students, but they are rarely thankful	
▶ Professors do a lot for students, but students are rarely thankful	✓
▶ Professors do a lot for students, but Professors are rarely thankful	✗
MIT accepted the students, because they had good grades	
▶ MIT accepted the students, because the students had good grades	✓
▶ MIT accepted the students, because MIT had good grades	✗
Obama beat John McCain, because he was the better candidate	
▶ Obama beat John McCain, because Obama was the better candidate	✓
▶ Obama beat John McCain, because John McCain was the better candidate	✗
Obama beat John McCain, because he failed to win the majority of the electoral votes	
▶ Obama beat John McCain, because John McCain failed to win the majority of the electoral votes	✓
▶ Obama beat John McCain, because Obama failed to win the majority of the electoral vote	✗

Table 9: Examples from DPR's dev set. The first line in each section is a context and lines with ▶ are corresponding hypotheses. ✓ (✗) in the last column indicates whether the hypothesis is entailed (or not) by the context.