Evaluating Syntactic Properties of Seq2seq Output with a Broad Coverage HPSG: A Case Study on Machine Translation

Johnny Tian-Zheng Wei
College of Natural Sciences
University of Massachusetts Amherst
jwei@umass.edu

Khiem Pham
Department of Computer Science
San Jose State University
khiem.pham@sjsu.edu

Brian Dillon
Department of Linguistics
University of Massachusetts Amherst
brian@linguist.umass.edu

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
brenocon@cs.umass.edu

Abstract

Sequence to sequence (seq2seq) models are often employed in settings where the target output is natural language. However, the syntactic properties of the language generated from these models are not well understood. We explore whether such output belongs to a formal and realistic grammar, by employing the English Resource Grammar (ERG), a broad coverage, linguistically precise HPSG-based grammar of English. From a French to English parallel corpus, we analyze the parseability and grammatical constructions occurring in output from a seq2seq translation model. Over 93% of the model translations are parseable, suggesting that it learns to generate conforming to a grammar. The model has trouble learning the distribution of rarer syntactic rules, and we pinpoint several constructions that differentiate translations between the references and our model.

French	Une situation grotesque.
Reference	It is a grotesque situation.
NMT Output	A generic_adj situation.

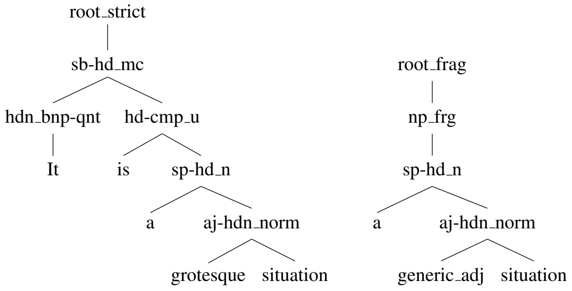

Figure 1: A test set source-reference pair and the NMT translation. Below are parser derivations in the ERG of both the reference and NMT translation. The ERG is described in §2. Non-syntactic rules have been omitted. The NMT model is trained and tested only on sentence pairs where the reference is parseable by the ERG. The NMT translation may not always be parseable. Analysis on model output parseability in §4.1.

1 Introduction

Sequence to sequence models (seq2seq; Sutskever et al., 2014; Bahdanau et al., 2014) have found use cases in tasks such as machine translation (Wu et al., 2016), dialogue agents (Vinyals and Le, 2015), and summarization (Rush et al., 2015), where the target output is natural language. However, the decoder side in these models is usually parameterized by gated variants of recurrent neural networks (Hochreiter and Schmidhuber, 1997), and are general models of sequential data not explicitly designed to generate conforming to the grammar of natural language.

The syntactic properties of seq2seq output is our central interest. We focus on machine translation as a case study, and situate our work among those of artificial language learning, where we train our translation model exclusively on sentence pairs where the target-side output is in our grammar, and test our models by evaluating the output with respect to a grammar. We attempt to understand seq2seq output with the English Resource Grammar (Flickinger, 2000), a broad coverage, linguistically precise HPSG-based grammar of English, and explore the advantages and potential of using such an approach.

This approach has three appealing properties in evaluating seq2seq output. First, the language of the ERG is a departure from studies on unrealistic artificial languages with regular or context-free grammars, which give exact analyses on grammars that bear little relation to human language

(Weiss et al., 2018; Gers and Schmidhuber, 2001). In fact, about 85% of the sentences found in Wikipedia are parseable by the ERG (Flickinger et al., 2010). Second, our methodology directly evaluates sequences the model outputs in practice with greedy or beam search, in contrast to methods rescoring pre-generated contrastive pairs to test implicit model knowledge (Linzen et al., 2016; Sennrich, 2016). Third, the linguistically precise nature of the ERG gives us detailed analyses of the linguistic constructions exhibited by reference translations and parseable seq2seq translations for comparison.

Figure 1 shows an example from our analysis. Each testing example records the reference derivation, the model translation, and the derivation of that translation, if applicable. The derivations richly annotate the rule types and the linguistic constructions present in the translations.

Our analysis in §4.1 presents results on parseability by the ERG and summarizes its relation to surface level statistics using Pearson correlation. In §4.2 we manually annotate a small sample of NMT output without ERG derivations for grammaticality. We find that 60% of exhaustively unparseable NMT translations are ungrammatical by humans. We also identify that 18.3% of the ungrammatical sentences could be corrected by fixing agreement attachment errors. We conduct a discriminatory analysis in §4.4 on reference and NMT rule usage to guide a qualitative analysis on our NMT output. In analyzing specific samples, we find a general trend that our NMT model prefers to translate literally.

2 Head-phrase Structure Grammars

A head-phrase structure grammar (HPSG; Pollard and Sag, 1994) is a highly lexicalized constraint based linguistic formalism. Unlike statistical parsers, these grammars are hand-built from lexical entries and syntactic rules. The English Resource Grammar (Flickinger, 2000) is an HPSG-based grammar of English, with broad coverage of linguistic phenomena, around 35K unique lexical entries, and handling of unknown words with both generic part-of-speech conditioned lexical types (Adolphs et al., 2008) and a comprehensive set of class based generic lexical entries captured by regular expressions. The syntactic rules give fine-grained labels to the linguistic constructions present.[1] While the ERG produces both syntactic and semantic annotations, we focus only on syntactic derivations in this study.

Suitable to our task, the ERG was engineered to capture as many grammatical strings as possible, while correctly rejecting ungrammatical strings. Parseability under the ERG should have linguistic reality in grammaticality. Ideally, there will be no parses for any ungrammatical string, and at least one parse for all grammatical strings, which can be unpacked in order of scores assigned by the included maximum entropy model. We make a distinction between parseability and grammaticality. For our purposes of evaluating with a specified grammar, we consider the parseability of sentences under the ERG in §4.1, regardless of human grammaticality judgments. In §4.2, we manually annotate unparseable sentences for English grammaticality.

All experiments are conducted with the 1214 version of the ERG, and the LKB/PET was used for all parsing (Copestake and Flickinger, 2000). We use the default parsing configuration (command line option "`--erg+tnt`"), which uses a parsing timeout of 60 seconds. A sentence is labeled unparseable either if the search space contains no derivations or if not a single derivation is found within the search space before the timeout. Figure 1 shows a simplified derivation tree.

3 Experimental Setup

This section details our setup of a French to English (FR → EN) neural machine translation system which we now refer to as NMT. Our goal was to test a baseline system for comparable results to machine translation and seq2seq models.

Dataset. From 2M French to English sentence pairs in the Europarl v7 parallel corpora (Koehn, 2005), we subset 1.6M where the English/reference sentence was parseable by the ERG. For these 1.6M sentence pairs, we record the best tree of the English sentence as determined by the maximum entropy model included in the ERG. All sentence pairs we now consider have at least one English translation within our grammar, and we make no constraint on French. About 1.4M pairs were used for training, 5K for validation, and the remaining 200K reserved for analysis.

Out of vocabulary tokens. On the source-side

[1] A list of rules types and their descriptions can be found at `http://moin.delph-in.net/ErgRules`.

	Strict		Informal		Unpar-
Source	Full	Frag	Full	Frag	seable
Ref	64.7	2.4	31.5	1.4	0.0
NMT	60.5	3.0	28.1	1.6	6.8
Δ	-4.2	+0.6	-3.4	+0.2	+6.8

Table 1: The distribution of root node conditions for the reference and NMT translations on the 200K analysis sentence pairs. Root node conditions are taken from the recorded best derivation. The best derivation is chosen by the maximum entropy model included in the ERG.

French sentences, simple rare word handling was applied, where all tokens with a frequency rank over 40K were replaced with an "UNK" token. However, when handling rare words in the target-side English sentences, "UNK" will significantly degrade ERG parsing performance on model output. We replace our output tokens based on the lexical entries recognized by the ERG in our best parses (as in Figure 1's NMT output). This form of rare word handling is similar to the 10K PTB dataset (Mikolov et al., 2011), but with more detailed part-of-speech and regular expression conditioned "UNK" tokens. After preprocessing, we had a source vocabulary size of 40000, and a target vocabulary size of 36292.

Model. Our translation model is a word-level neural machine translation system with an attention mechanism (Bahdanau et al., 2014). We used an encoder and decoder with 512 dimensions and 2 layers each, and word embeddings of size 1024. Dropout rates of 0.3 on the source, target, and hidden layers were applied. A dropout of 0.4 was applied to the word embedding, which was tied for both input and output. The model was trained for about 20 hours with early stopping on validation perplexity with patience 10 on a single Nvidia GPU Titan X (Maxwell). We used the NEMATUS (Sennrich et al., 2017) implementation, a highly ranked system in WMT16.

Translations. After training convergence on the 1M sentence pairs, the saved model is used for translation on the 200K sentences pairs left for analysis. A beam size of 5 is used to search for the best translation under our NMT model. We parse these translations with the ERG and record the best tree under the maximum entropy model. We have parallel data of the French sentence, the human/reference English translation, the NMT English translation, the parse of the reference translation, and the parse of NMT translation (if it was grammatical). Note that the NMT translation may have no parse.

Feature	Equation	r
LP NMT	$\log P_m(S_o)$	0.313
LP Unigr. (src-fr)	$\log P_u(S_i)$	0.289
LP Unigr. (ref-en)	$\log P_u(S_r)$	0.273
LP Unigr. (out-en)	$\log P_u(S_o)$	0.304
Length Output	$\|S_o\|$	-0.320
Mean LP	$\frac{\log P_m(S_o)}{\|S_o\|}$	0.093
Norm LP	$-\frac{\log P_m(S_o)}{\log P_u(S_o)}$	0.057

Table 2: Pearson's r of surface statistics against the binary parseability variable. Parseable is denoted with +1. S_i, S_r, S_o are the input, reference, and NMT output sentences, respectively. We abbreviate log probability as "LP." $P_m(S)$ is the probability of S occurring under the NMT model, and $P_u(S)$ is the probability of S occurring under a unigram model.

4 Results

4.1 Parseability

The NMT translations for the 200K test split were parsed. Parsing a sentence with the ERG yields one of four cases:

- Parseable. A derivation is found and recorded by the parser before the timeout. The best derivation is chosen by the included maximum entropy in the ERG. About 93.2% of the sentences were parseable.

- Unparseable due to resource limitations. The parser reached its limit of either memory or time before finding a derivation. This constitutes about 3.2% of all cases, and 47% of unparsable cases.

- Unparseable due to parser error. The parser encountered an error in retrieving lexical entries or instantiating the parsing chart. This constitutes about 0.5% of all cases, and 8% of unparsable cases.

- Unparseable due to exhaustion of search space. The parser exhausted the entire search space of derivations for a sentence, and concludes that it does not have a derivation in

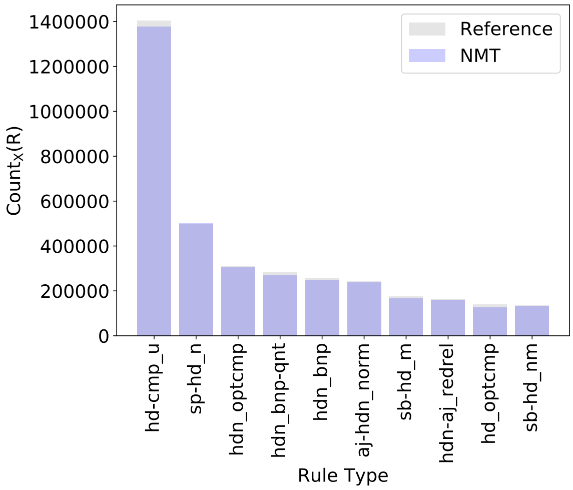

Figure 2: Count of rule usage for the 10 most frequent rules in the derivations of the reference and grammatical NMT translations.

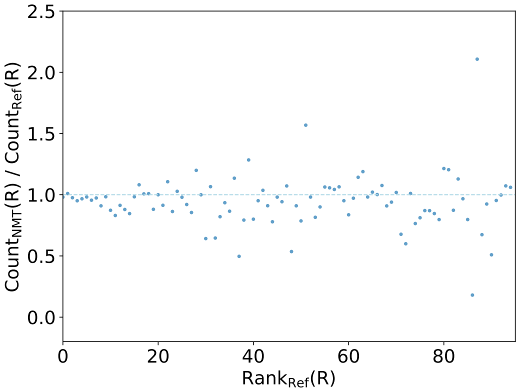

Figure 3: The ratio of each rule's count in grammatical NMT translations over count in reference translations, ordered by the rule's frequency rank in reference derivations. Only rules with over 1000 usages in the set of reference derivations are shown.

the ERG. This constitutes about 3.1% of all cases, and 45% of unparsable cases.

The distribution of the root node conditions for the reference and NMT translation derivations are listed in table 1, along with the parseability of the NMT translations. Root node conditions are used by the ERG to denote whether the parser had to relax punctuation and capitalization rules, with "strict" and "informal", and whether the derivation is of a full sentence or a fragment, with "full" and "frag". Fragments can be isolated noun, verb, or prepositional phrases. Both full sentence root node conditions saw a decrease in usage, with the strict full root condition having the largest drop out of all conditions. Both fragments have a small increase in usage.

We summarize the parseability of NMT translations with a few surface level statistics. In addition to log probabilities from our translation model, we provide several transformations of these scores, which were inspired by work in unsupervised acceptability judgments (Lau et al., 2015). In table 2, we calculate Pearson's r for each statistic and the binary parseability variable. The r coefficient is effectively a normalized difference in means.

From the correlation coefficients, we see that the probabilities from the NMT and unigram models are all indicative of parseability. The higher the probabilities, the more likely the translation is to be grammatical. Length is the only exception with a negative coefficient, where the longer a sentence is, the less likely a translation is grammatical. Length has the strongest correlation of all our features, but this correlation may be due to limitations in the ERG's ability to parse longer sentences, instead of the NMT model's to generate longer grammatical sentences. We see that the LP NMT has a higher correlation with grammaticality than the unigram models, but not by a large amount. Coefficients for length and LP NMT have the two greatest magnitudes.

4.2 Grammaticality

Out of the 14K unparseable NMT translations, there are 6.2K translations where the parser concluded unparseability after exhausting the search space for derivations. We will refer to these examples as "exhaustively unparseable." To understand the relation between English grammaticality and exhaustive unparseability under the ERG, two linguistics undergraduates (including the first author) labeled a random sample of 100 NMT translations from this subset. We sampled only those translations with less than 10 words to limit annotator confusion. Annotators were instructed to assign a binary grammatical judgment to each sentence, ignoring the coherence and meaning of the translation, to the best of their abilities. Punctuation was ignored in all annotations, although the ERG is sensitive to punctuation. When the sentence was ungrammatical, subject-verb agreement and noun phrase agreement errors were annotated.

Within our random sample, 60 sentences were labeled as ungrammatical. Of these ungrammatical sentences, 5 could be made grammatical if a

Reference		NMT	
Rule Type	**Annotations**	**Rule Type**	**Annotations**
xp_brck-pr	Paired bracketed phrase	j_sbrd-pre	Pred.subord phr fr.adj, prehead
cl-cl_runon	Run-on sentence w/two clauses	n-j_j-cpd	Compound from noun+adj
np-hdn_cpd	Compound proper-name+noun	j_n-ed	Adj-phr from adj + noun+ed
vp_sbrd-prd-prp	Pred.subord phr from prp-VP	aj-np_int-frg	Fragment intersctv modif + NP
hd-aj_int-sl	Hd+foll.int.adjct, gap in adj	vp_sbrd-prd-aj	Pred.subord phr from adjctv phr
hd-aj_vmod	Hd+foll.int.adjct, prec. NP cmp	np_frg	Fragment NP
vp_np-ger	NP from verbal gerund	flr-hd_nwh	Filler-head, non-wh filler
mrk-nh_atom	Paired marker + phrase	hdn-aj_rc-pr	NomHd+foll.rel.cl, paired pnct
vp_sbrd-pre	Pred.subord phr fr.VP, prehead	sb-hd_mc	Head+subject, main clause
num_prt-det-nc	Partitive NP fr.number, no cmp	num-n_mnp	Measure NP from number+noun

Table 3: The most discriminatory features of both the reference and NMT translations. Features are ranked by a logistic regression without an intercept and an L1 penalty $C = 0.01$, trained with LIBLINEAR within scikit-learn. Description of rule types are taken from the annotations in the ErgRules website.

subject-verb agreement error was corrected, and 5 other translations could be made grammatical by correcting an article or determiner attachment to a noun. One translation exhibited both forms of agreement attachment errors. Agreement attachment errors are better studied phenomenon (Linzen et al., 2016; Sennrich, 2016). However, correcting these errors only fixes 18.3% of ungrammaticality that we observed in our sample.

Out of the 100 sampled NMT translations that have no ERG derivations, we found 35 to be grammatical. 5 test examples were excluded. These include two cases where the source sentences were empty, and three cases where the sentence was parliament session information. Both annotators found annotating to be challenging, and possibly better annotated on an ordinal scale. Out of the exhaustively unparseable random sample, 37% was found to be grammatical. The ERG may have grammar gaps for near grammatical sentences.

4.3 Rule Counts

This section and those following will analyze the rules present in the derivations of the reference and the grammatical NMT translations. We consider only the appearance of the rule, disregarding the context it appears in, and define $\mathrm{Count}_{\mathrm{X}}(R)$ as the number of times rule R appears in the set $X \in \{\mathrm{Ref}, \mathrm{NMT}\}$ of derivations. In figure 2, we plot the counts of the 10 most frequent rule types in the reference and NMT translations. The rules were taken from the best derivations as determined by the included maximum entropy classifier in the ERG. Note that we have about 200K reference derivations and 189K NMT derivations we aggregate statistics from, as about 7% of the NMT translations are unparseable. We see that both distributions seem to be Zipfian, and that the rule counts in the NMT translations match the reference closely.

In figure 3, for each rule R, we plot the ratio $\mathrm{Count}_{\mathrm{NMT}}(R)/\mathrm{Count}_{\mathrm{Ref}}(R)$ of derivations against the rank of the rule type. The rank is computed from the set of reference derivations. The variance of the ratio seems to increase as the rank of the rule increases. While the occurrences of rarer constructions is low in the NMT translations, it seems not to match the usage in the reference translation dataset. This suggests that NMT has trouble learning the usage of rarer syntactic constructions.

4.4 Discriminative Rules

This section aims to understand which usage of rules distinguish the reference from the NMT translations. The analysis in this section is largely inspired by work in syntactic stylometrics (Feng et al., 2012; Ashok et al., 2013), where we vectorize each derivation as a bag of rules, and fit a logistic regression without an intercept to predict whether a derivation was from the set of reference or NMT translations. In total, there are 392K examples and we prepare an 80/20 training validation split. The model is fit with an L1 sparsity penalty of $C = 0.01$ with the LIBLINEAR solver in scikit learn (Pedregosa et al., 2011). On the validation set, the logistic regression achieves an accuracy of about 59.0% on the validation set up from the 51.9% majority class baseline. Of the 204 rules used as features, only 71 were non-zero. There are 47 rules that are discriminatory towards reference translations (positive weights), and 24 rules that are discriminatory towards NMT translations (negative weights). Table 3 shows the 10 most discriminative rules for each set.

4.5 Qualitative Analysis

We provide qualitative analysis for a few of the most discriminative rules for both the reference and NMT translations. When exploring discriminatory rules in the reference, we sampled for sentence pairs where the reference translation that contained the rule of interest, and the NMT translation did not. We only sampled within sentences with a length of less than 12. Our qualitative analysis is written after we looked through many samples, and we attempted to list a few of our general observations for each rule.

The "cl-cl_runon" rule type indicates a runon sentence with two conjoined clauses. This rule has a positive coefficient, and discriminates towards reference translations. An example is given below:

French	je le répète , vous avez raison .
Reference	i repeat ; you are quite right .
NMT Output	i repeat , you are right .

In this case, the NMT used a comma to conjoin two clauses instead of using a semi-colon, which is more similar in punctuation to the source sentence. In every case we saw, the NMT model seems to follow the French style of conjunction more closely, mirroring the punctuation of the source sentence. Reference translations seem to be more spurious in the usage of semicolons or periods. In more concerning cases, short conjoined clauses were dropped by the NMT translations; e.g. "thank you .".

We now analyze "np_frg" which denotes a noun phrase fragment. This rule that has a negative coefficient, and discriminates towards NMT translations. We give an example below:

French	quel paradoxe !
Reference	what a paradox this is !
NMT Output	what a paradox !

When looking through samples, we saw many examples where the expletive is dropped. This is similar to the case for the previous rule as it is a literal translation of the French source. In NMT translations we observed increases in the formal and strict fragment root conditions, and we believe these translations are a factor.

5 Related Work

Previous work in recurrent neural network based recognizers on artificial languages has studied the performance on context-free and limited context-sensitive languages (Gers and Schmidhuber, 2001). More recent research in this setting provide methods to extract the exact deterministic finite automaton represented by the RNN based recognizers of regular languages (Weiss et al., 2018). These studies give exact analyses of RNN recognizers for simple artificial languages.

In the evaluation of language models in natural language settings, recent work analyzes the rescoring of grammatical and ungrammatical sentence pairs based on specific linguistic phenomenon such as agreement attraction (Linzen et al., 2016). These contrastive pairs have also found use in evaluating seq2seq models through rescoring with the decoder side of neural machine translation systems (Sennrich, 2016). Both studies on contrastive pairs evaluate implicit grammatical knowledge of a language model.

HPSG-based grammars have found use in evaluating human produced language. To determine the degree of syntactic noisiness in social media text, parseability under the ERG was examined for newspaper and Twitter texts (Baldwin et al., 2013). In predicting grammaticality of L2 language learners with linear models, the parseability of sentences with the ERG was found to be a useful feature (Heilman et al., 2014). These studies suggest parseability in the ERG has some degree of linguistic reality.

Our work combines analysis of neural seq2seq models with an HPSG-based grammar, which begins to let us understand the syntactic properties in the model output. Recent work most similar to ours is in evaluating multimodel deep learning models with the ERG (Kuhnle and Copestake, 2017). While their work uses the ERG for language generation to test language understanding, we evaluate language generation with the parsing capabilities of the ERG, and study the syntactic properties.

6 Conclusion

Neural sequence to sequence models do not have any explicit biases towards inducing underlying grammars, yet was able to generate sentences conforming to an English-like grammar at a high rate. We investigated parseability and differences in syntactic rule usage for this neural seq2seq model, and these two analyses were made possible by the English Resource Grammar. Future work will involve using human ratings and machine translation quality estimation datasets to understand which syntactic biases are preferable for machine translation systems. The ERG also produces Minimal Recursion Semantics (MRS; Copestake et al., 2005), a semantic representation which our work does not yet explore. By matching the semantic forms produced, we can make evaluations of language generation systems on a semantic level as well. In using these deep resources for evaluation, there is a shortcoming in the biased coverage of the grammar. Future work will also study how to evaluate our models despite these limitations. We hope this paper spurs others' interest in HPSG-based or language-like grammar evaluations of neural networks.

Acknowledgments

This work was facilitated by the Center for Study of Language and Information REU site at Stanford University, where the first and second author were roommates. Andrew McCallum and IESL provided computing resources. Dan Flickinger, Chris Potts, and the anonymous reviewers provided helpful comments. Nicholas Thomlin co-annotated the grammaticality judgments. Technical support on DELPH-IN was provided by Michael Goodman and Stephen Oepen. The first author received mentorship from Omer Levy, Roy Schwarz, Chenhao Tan, Nelson Liu, and Noah Smith in the summer of 2017, and from Ari Kobren, Nicholas Monath, and Haw-Shiuan Chang in the years prior. The second author received mentorship from Guangliang Chen.

References

Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger, and Bernd Kiefer. 2008. Some fine points of hybrid natural language parsing. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1753–1764.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 356–364.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2):281–332.

Ann A. Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 1522–1533.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. 2010. Wikiwoods: Syntacto-semantic annotation for English wikipedia. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Felix A. Gers and Jürgen Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Networks*, 12(6):1333–1340.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel R. Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 174–180.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Alexander Kuhnle and Ann A. Copestake. 2017. Deep learning evaluation using deep linguistic processing. *CoRR*, abs/1706.01322.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1618–1628.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *TACL*, 4:521–535.

Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukás Burget, and Jan Cernocký. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 605–608.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. *CoRR*, abs/1612.04629.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. Extracting automata from recurrent neural networks using queries and counterexamples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5244–5253.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.