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

University of Massachusetts Amherst

Johnny Wei[1], Khiem Pham[2], Brian Dillon[1], Brendan O'Connor[1]
[1]University of Massachusetts, Amherst
[2]San Jose State University
jwei@umass.edu

College of Information and Computer Sciences

Summary:

- **Goal**: Evaluate grammaticality and syntactic properties of seq2seq output.
- **Key idea:** Train a seq2seq model on examples where the output sequence is in an HPSG. Observe output with respect to the same grammar.
- **Advantages:** (1) The HPSG grammar is language-like. (2) Directly evaluates sequences produced in practice. (3) HPSG gives detailed analyses of syntactic constructions.

The English Resource Grammar (ERG)

- The ERG is an HPSG, a highly lexicalized constraint based linguistic formalism.
- It is a rule-based grammar with 35K lexical entries and 250 syntactic rules. Parses 85% of Wikipedia.
- Train a vanilla seq2seq to translate FR→EN, where the reference sentence is in the ERG.

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

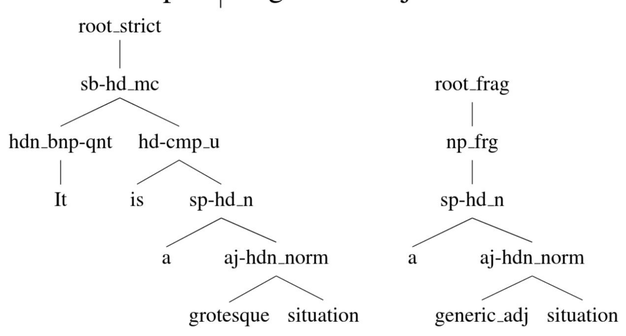

Figure 1. A test set source reference pair and the seq2seq NMT translation.

Parse NMT output with ERG. Record parseability. If parseable, record best ERG derivation.

Parseability

Source	Strict Full	Strict Frag	Informal Full	Informal Frag	Unparseable
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. Parseability by root condition.

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	$\lvert S_o \rvert$	-0.320
Mean LP	$\frac{\log P_m(S_o)}{\lvert S_o \rvert}$	0.093
Norm LP	$-\frac{\log P_m(S_o)}{\log P_u(S_o)}$	0.057

Table 2. Correlation of parseability with various statistics.

- 93.2% is ERG-parseable.
- Among the unparseable 7%, only 45% cases have search space exhausted.
- Correlations for the binary parseability variable (+1 parseable).
- NMT LP scores have highest correlation.
- NMT LP correlation only slightly higher than unigram.

Grammaticality

100 samples
- **60** ungrammatical
  - **5** ungrammatical - subject verb agreement
  - **5** ungrammatical - determiner noun agrm
  - **1** ungrammatical - both agreement errors
- **30** grammatical
- **5** excluded

- Human grammaticality judgments of 100 cases that are exhaustively unparseable (3.2% of total test set).
- Restricted to length < 10.

Rule Statistics & Discriminative Analysis

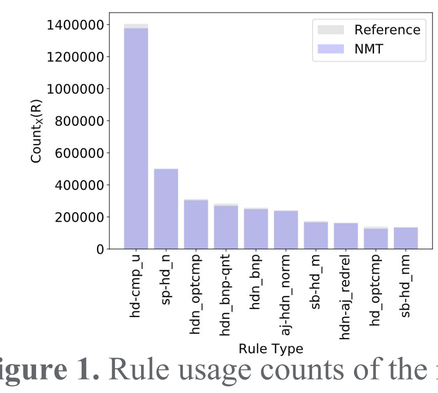

Figure 1. Rule usage counts of the reference and grammatical NMT translations.

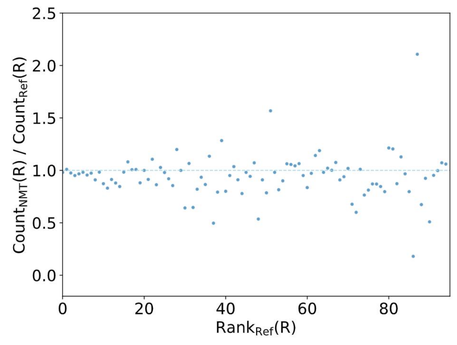

Figure 2. Ratio of each rule count in grammatical NMT translations to reference by rank.

Reference Rule Type	Reference Annotations	NMT Rule Type	NMT 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 syntactic rule usages between reference and NMT derivations, ranked by a logistic regression with sparsity penalty.

Qualitative Analysis

- Sample those sentences where reference or NMT translation uses a rule but the other translation does not.

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

- cl-cl_runon
- Discriminates towards reference translations.

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

- np_frg
- Discriminates towards NMT translations.

- NMT translates more literally of French source; human translations of syntactic rules are not as faithful.

Future Directions

- Ablating training data to observe whether some syntactic constructions can be learned without supervision.
- Correlating differences in syntactic and semantic representation with human judgments.
- Presenter is currently applying to PhD programs with an interest in parsing + language generation.