Neural Text Generation from Rich Semantic Representations.

Valerie Hajdik, Jan Buys,
Michael Wayne Goodman and Emily M. Bender

DELPH-IN summary presentation, 16 July 2019
Presented originally at NAACL 2019

Table of contents

Overview

- Generation of English with ACE on Redwoods (1214)
  - High BLEU scores: ∼60–70
  - Coverage is low (78%) compared to parsing
- (Aside) The LKB does slightly better due to:
  - More support for generating from unknowns
  - Better realization ranking
- Neural generation gets us 100% coverage AND significantly higher BLEU (∼70–80)
- But error analysis reveals that BLEU is not adequately estimating quality (to no one's surprise)

Preprocessing

- MRS $\rightarrow$ DMRS $\rightarrow$ DMRS-PENMAN $\rightarrow$ Modified PENMAN
- Regarding PENMAN:
  - \+ Tree-like arrangement reduces leakage of word-order information
  - \+ Less markup than SimpleDMRS, DMRX
  - \+ Edges are local to their start nodes
  - – Loses information with disconnected graphs
  - – No overt distinction between arguments and properties

PENMAN Simplification

```
(10002 / _see_v_1
  :tense PRES
  :sf PROP
  :perf -
  :mood INDICATIVE
  :ARG1-NEQ (10001 / named
    :carg "Kim"
    :pers 3
    :num SG
    :ind +)
  :ARG2-NEQ (10004 / _boy_n_1
    :pers 3
    :num SG
    :ind +
    :RSTR-H-of (10003 / _a_q)))
```

- no node identifiers (thus no reentrancies)
- properties are consolidated
- named entities are anonymized

```
( _see_v_1 mood=INDICATIVE|perf=
    ↪ -|sf=PROP|tense=PRES
    ↪ ARG1-NEQ (named0 ind=+|
    ↪ num=SG|pers=3 ) ARG2-NEQ
    ↪  ( _boy_n_1 ind=+|num=SG
    ↪ |pers=3 RSTR-H-of ( _a_q
    ↪  ) ) )
```

Model

- Encoder-decoder
- Encoder is 2-layer LSTM
- Decoder uses global soft attention for alignment and pointer attention for copying unknowns to output
- Implemented with OpenNMT-py

```
python OpenNMT-py/train.py -data data/opennmt \
    -layers 2 -dropout 0.5 \
    -word_vec_size 500 -batch_type sents \
    -max_grad_norm 5 -param_init_glorot \
    -encoder_type brnn -decoder_type rnn \
    -rnn_type LSTM -rnn_size 800 \
    -save_model $MODEL_PREFIX \
    -learning_rate 0.001 -start_decay_at 25 \
    -opt adam -epochs 40 -gpuid $GPU_ID \
    > "logs/train_$MODEL_VERSION.log"
```

Semi-supervised Training

- The gold training data is augmented with *silver* data
- Produced by parsing 1M sentences from Gigaword with the ERG and ACE ($\sim$90% parse coverage)

Results

Model	**BLEU (All)**	**BLEU (WSJ)**	**BLEU (overlap)**	**Exact Match%**
Neural MRS (gold)	66.11	73.12	69.27	24.09
Neural MRS (silver)	75.43	81.76	77.13	25.82
Neural MRS (gold + silver)	77.17	83.37	79.15	32.07
ACE (ERG)	–	–	62.05	15.08
DAG transducer (Ye et al 2018)	–	68.07	–	–

Table: BLEU and exact-match scores over held-out test set

Out-of-domain evaluation

	Training Data	
Test domain	**WSJ**	**WSJ + Giga**
WSJ	65.78	83.42
Brown	45.00	76.99
Wikipedia	35.90	62.26

Table: BLEU scores for domain match experiments

Attribute ablation

Ablation	**BLEU**
All attributes	72.06
No node attributes	59.37
No node attr except num, tense	67.34
No edge features	71.27

Table: Results of semantic feature ablation, model trained with gold data only

Error analysis

Type	**B80-89**	**B60-69**	**B40-49**	**All**
Unproblematic	56.4	39.55	48.8	47.1
Slightly problematic	18.0	9.2	3.3	7.6
Moderately problematic	12.8	25.0	18.7	19.8
Ungrammatical	5.1	7.9	8.1	7.6
Other serious error	7.7	18.4	21.1	18.1
Number of errors	39	76	123	238
Errors per item	1.18	2.30	3.73	7.21

Table: Percentage of errors of each type, across 99 sampled items, grouped by BLEU score

Unproblematic

- Capitalization
- Non-meaning changing differences in punctuation
- Spelling variants
- Extraposition/intraposition
- What/which in determiner position
- Optional that & similar
- Contractions
- Very close synonyms
- Free word order choices/swapped word dependent word order variation
- Meaning preserving reduplication

Slightly problematic

- Differences in formatting/markup
- Spelled out numbers where numerals are preferred
- Close synonyms
- Spurious whitespace

Moderately problematic

- Meaning-changing differences in punctuation
- Meaning-changing difference in tense/mood/aspect
- Animacy error on relative pronoun
- A/an

Ungrammatical

- Non-replaced CARG
- Spurious additional token resulting in ungrammaticality
- UNK in output
- Ungrammatical difference in TMA
- Word order change resulting in ungrammaticality
- Ungrammatical contraction
- Ungrammatical inflection change

Serious error resulting in a grammatical string

- Dropped token
- Meaning-altering swapped word

- Subcase: wrong number
- Spurious additional token, still grammatical
- Meaning changing word order difference
- Pragmatics changing word order difference

Conclusion

- BLEU scores underestimate output quality
- Meaning signified by punctuation not fully captured in ERSs
- Neural seq2seq models are effective for generation from MRS
- MRS is an effective source for neural seq2seq generation