MRS-Based Generation Using Transformer

Gyu-min Lee
(Korea University, gyuminlee@korea.ac.kr)

DELPH-IN
July 21, 2021

Two Ways of Doing NLP

- Symbolic NLP
  - Rules and algorithm based
  - Engineered grammars
  - High precision
  - Easy to control
  - Low recall (i.e., narrow coverage)
- Stochastic NLP
  - Data and probability based
  - Neural language models
  - High empirical performance
  - Easy to build (conceptually)
  - Low controllability (the black box problem)

Integrating the Two Ways

- *Can we integrate the two methods?*
- The precision of symbolic NLP
- ...with the coverage of stochastic NLP!
- Hajdik et al. (2019)!
  - Integrates the neural generation with ERG
  - NLG as MT from MRS to natural language sentence
  - Sequence-to-sequence (BiLSTM + Attention) can faithfully generate natural language text from MRS representation!

Possible Issue with Hajdik et al. (2019)

- MRS representation is long!
- `(|_ unknown|mood=INDICATIVE|perf=-|sf=PROP-OR-QUES ARG-NEQ|_ (|_ _and_c|num=PL|pers=3 L-INDEX-NEQ|_ (|_ _cathedral_n_1|ind=+|num=SG|pers=3 RSTR-H-of|_ (|_ _the_q|_ )|_ )|_ R-INDEX-NEQ|_ (|_ _bazaar_n_1|ind =+|num=SG|pers=3 RSTR-H-of|_ (|_ _the_q|_ )|_ )|_ )|_ )|_`
- RNNs suffer from vanishing gradient problem
- i.e., RNNs inherently have difficulty processing long sequence of tokens
- . . . like MRS!

Let's Apply Transformer

- Transformer is not an RNN
- Transformer processes long sequences relatively better
- Transformer is generally a good choice for an MT task

Stochastic NLP I

- Machine learning techniques has brought massive improvement to NLP
- The “BERTology”
  - Study on BERT and other Transformer-based language models
  - Finds that neural language models learned some syntactic information
  - It appears that neural language models are “learning” hierarchical structure of language
- Some claims that symbolic, domain knowledge of linguistics is not necessary in NLP

Stochastic NLP II

- The Black Box problem
  - We can't have a clear look into the neural language models
  - Debugging and improving the model is extremely challenging
  - Making language models gigantic may work, but what about the environment?
  - Can those gigantic language models ever be able to acquire the language properly? (Bender et al., 2021)

NLG from Linguistic Meaning Representation

- Konstas et al. (2017)
    - NLG as MT from AMR
- Hajdik et al. (2019)
    - NLG as MT from MRS
    - ERG was used to get the MRS representation of the training data
    - MRS contains much richer information
    - Significant improvement from Konstas et al. (2017)

Data

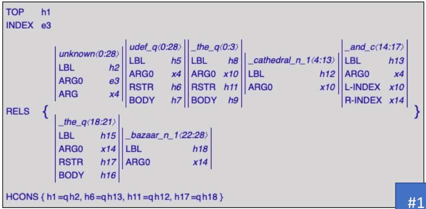
- Gold dataset
  - Redwoods Treebank (Oepen et al., 2004) release 1214
- Silver dataset
  - A million sentences from the Gigaword Corpus
  - MRS derived using ERG
  - Making use of silver dataset significantly improved the performance (Hajdik et al., 2019)
- Total data: 984,679 MRS-sentence pairs
- Anonymized according to ERG's NER to reduce data sparsity (Hajdik et al., 2019)

MRS Linearization I

- Not easy and practical to feed the multilinear MRS representation to a sequence-to-sequence model
- Konstas et al. (2017) used PENMAN format to express the directed graph of AMR
- DMRS is a directed graph representation of MRS which is interchangeable with it
- MRS $\rightarrow$ DMRS $\rightarrow$ PENMAN $\rightarrow$ Single Line String (Hajdik et al., 2019)

MRS Linearization II

Example Sentence:
“The Cathedral and the Bazaar”



#1

```
# ::id 1010
# ::snt The Cathedral and the Bazaar
(10000 / unknown
 :lnk "<0:28>"
 :sf PROP-OR-QUES
 :tense UNTENSED
 :mood INDICATIVE
 :perf -
 :ARG-NEQ (10004 / _and_c
  :lnk "<14:17>"
  :pers 3
  :num PL
  :L-INDEX-NEQ (10003 / _cathedral_n_1
   :lnk "<4:13>"
   :pers 3
   :num SG
   :ind +
   :RSTR-H-of (10002 / _the_q
    :lnk "<0:3>"))
  :R-INDEX-NEQ (10006 / _bazaar_n_1
   :lnk "<22:28>"
   :pers 3
   :num SG
   :ind +
   :RSTR-H-of (10005 / _the_q
    :lnk "<18:21>"))))
```

#2

```
(|_ unknown|mood=INDICATIVE|perf=-|sf=PROP-OR-QUES
ARG-NEQ|_ (|_ _and_c|num=PL|pers=3 L-INDEX-NEQ|_
(|_ _cathedral_n_1|ind=+|num=SG|pers=3 RSTR-H-of|_
(|_ _the_q|_ )|_ )|_ R-INDEX-NEQ|_ (|_
_bazaar_n_1|ind=+|num=SG|pers=3 RSTR-H-of|_ (|_
_the_q|_ )|_ )|_ )|_ )|_
```

#3

Model and Implementation

- Transformer
- Using OpenNMT-py (Klein et al., 2017)
- Hyperparameters from `https://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model`
- Mimics Vaswani et al. (2017) model
- Validation from every 10,000 steps to every 5,000 steps
- Using Google Colab
  - Decent cloud GPU environment
  - 12-hour limitation
  - Training up to 70,000 steps using `train_from`

Evaluation

- BLEU (Papineni et al., 2002)
  - Automatically evaluates machine translation
  - The task is NLG through MT
  - Therefore, BLEU can be used to measure how faithfully the model translated the meaning representation
- SacreBLEU (Post, 2018) for more comprehensive results and for comparison with Hajdik et al. (2019)

BLEU Score

Model	BLEU
Konstas et al. (2017)	33.8
Hajdik et al. (2019)	77.17
Ours	64.2

- BLEU measured for every 5,000 steps
- Score peaked at 30,000 steps with 64.2 BLEU
- Score decreased afterward with the accuracy, perplexity, and cross entropy plateauing

Translation Samples I

(1) a. **prediction**: If I am correct, they will help you understand exactly what it is saying the Linux community of good software - and perhaps they will help you become more productive yourself.

b. **answer**: If I'm correct, they'll help you understand exactly what it is that makes the Linux community such a fountain of good software—and, perhaps, they will help you become more productive yourself.

- PREDICTION is the detokenized and deanonymized prediction of the model
- ANSWER is the original text the model is suppsoed to translate to

Translation Samples II

(2) a. **prediction**: The myth and the sword.

b. **answer**: The Cathedral and the Bazaar

(3) a. **prediction**: = = = Objectives = = =

b. **answer**: Abstract

Translation Samples III

- The model struggled with seemingly easy task of lexical choice
- even when the lexical item is explicitly given in the MRS representation

```
(|_ unknown|mood=INDICATIVE|perf=-|sf=PROP-OR-QUES ARG-NEQ|
   _ (|_ _and_c|num=PL|pers=3 L-INDEX-NEQ|_ (|_
   _cathedral_n_1|ind=+|num=SG|pers=3 RSTR-H-of|_ (|_
   _the_q|_ )|_ )|_ R-INDEX-NEQ|_ (|_ _bazaar_n_1|ind=+|num
   =SG|pers=3 RSTR-H-of|_ (|_ _the_q|_ )|_ )|_ )|_ )|_
```

Error Analysis I

- Manual inspection over 100 randomly selected translation samples
- Tagged with: no error, lexical choice error, syntactic error, punctuation error, and missing elements error
- Some were not counted as errors:
    - Location of adverbial phrase that does not alter the meaning
    - Aspect (e.g., present on behalf of present progressive)
    - Use of clitcs (e.g., 'll)
    - Unreasonable punctuation

Error Analysis II

Error	Number	Sample Prediction
No Error	47	*Okay , we have card0 options .*
Lexical	31	*I assume there is a full* **salon** *on the shipping costs .*
Punctuation	8	**: *** *named0*
Lexical & Missing Argument	5	**Don 't Linger**
Lexical & Syntactic	4	*When ad dollars is tight , the high page cost is* **generally** *a major* **UNKcontributor0** *for UNKadvertisers0 who want to appear regularly in a publication or not at all .*
Missing Argument	3	*Requesting immediately .*
Syntactic	2	**polite0 refund .**
SUM	100	

Table: Number of errors from the 100 translation samples. The errors in the sample prediction are marked in bold face.

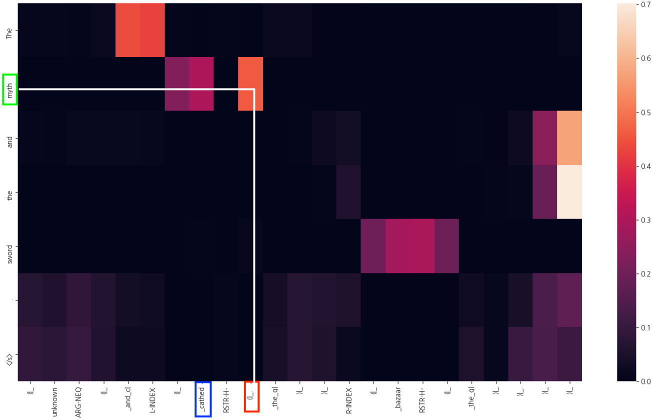
Error Analysis III

- 47 cases showed no error – coincides with Hajdik et al. (2019) that BLEU underestimate the model
- 40 cases out of 53 erroneous cases involved lexical choice problem
- Only 6 cases showed syntactic error

Attention Weight Distribution I

- Perhaps Attention weight distribution caused the poor performance
- Used OpenNMT-py's `attn_debug` to draw Attention weight for (2)
- Is this method safe?
  - Several concerns on the use of Attention weight as explanation
  - (e.g., Serrano and Smith, 2019; Brunner et al., 2019; Pruthi et al., 2019)
  - "Attention is Not Explanation" (Jain and Wallace, 2019)
  - But Wiegreffe and Pinter (2019) claims it depends on the definition of "explanation"
  - Explanation: transparency, explainability, interpretability
  - Attention weight does offer transparency
- The result as a piece of evidence
- pointing to the direction that Attention weight distribution caused the poor performance

Attention Weight Distribution II



Findings

- HPSG-based computational grammars like ERG do help neural NLG
- A neural model can faithfully generate sentences in terms of syntax from MRS
- Attention-based approach may be suboptimal for processing such rich linguistic representation

The Significance of the Study

- Why do we need this when ERG can already generate sentences?
- ERG is a robust system that can precisely and strictly generate English sentences
- ...but it lacks coverage (Bender and Emerson, 2021)
- Perhaps we can make a model with high precision *and* recall by joining ERG with deep learning!
- Plus, we can have more controllable neural model
- Despite the advancements of NLG, template-based models are still in use widely (Dale, 2019; Mahamood and Zembrzuski, 2019)
- Neural NLG models lack reliability
- ERG can provide that

Summary

- Reproduced Hajdik et al. (2019) with Transformer
- Transformer model struggles with the lexical choices
- MRS is a concise representation of syntactic and semantic information of a sentence
- Each item of linearized MRS contain essential information, unlike the natural language
- By paying attention to the certain part of MRS, it neglected the lexical items

Future Steps

- Further probe of the model
- Comparison with Hajdik et al. (2019) model to evaluate how beneficial Transformer model is in terms of syntax
- To advance the performance of the model,
  - Make use of RNN models that can handle longer sequence better
  - Adjust the Attention mechanism so that it can pay more attention to lexical items

References I

Bender, E. M. and Emerson, G. (2021). *Computational linguistics and grammar engineering*. Berlin: Language Science Press, prepublished version edition.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2019). On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.

Dale, R. (2019). Nlp commercialisation in the last 25 years. *Natural Language Engineering*, 25(3):419–426.

Hajdik, V., Buys, J., Goodman, M. W., and Bender, E. M. (2019). Neural text generation from rich semantic representations. *arXiv preprint arXiv:1904.11564*.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Konstas, I., Iyer, S., Yatskar, M., Choi, Y., and Zettlemoyer, L. (2017). Neural amr: Sequence-to-sequence models for parsing and generation. *arXiv preprint arXiv:1704.08381*.

Mahamood, S. and Zembrzuski, M. (2019). Hotel scribe: Generating high variation hotel descriptions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 391–396.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). Lingo redwoods. *Research on Language and Computation*, 2(4):575–596.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

References II

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., and Lipton, Z. C. (2019). Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.