# ParaAMR: A Large-Scale Syntactically Diverse Paraphrase Dataset by AMR Back-Translation

**Kuan-Hao Huang**[1] Varun Iyer[2] I-Hung Hsu[3] Anoop Kumar[4]
Kai-Wei Chang[1,4] Aram Galstyan[3,4]

[1] University of California, Los Angeles [2] University of Illinois Chicago
[2] Information Science Institute, University of Southern California [4] Amazon Alexa AI

ACL 2023




# Paraphrase Generation

- Paraphrase generation benefits many NLP applications
  - Question answering
  - Chatbots
  - Creative generation
  - Data augmentation
  - Robustness

*We will go hiking if tomorrow is a sunny day.*  *If it is sunny tomorrow, we will go hiking.*

# Challenge: Large-Scale High Quality Paraphrase Data

- Human-annotated dataset
  - MRPC [1], PAN [2], Quora [3]
  - High quality but limited scale
- Automatically generated dataset
  - Back-translation [4,5,6]
  - Large scale but lack of syntactic diversity

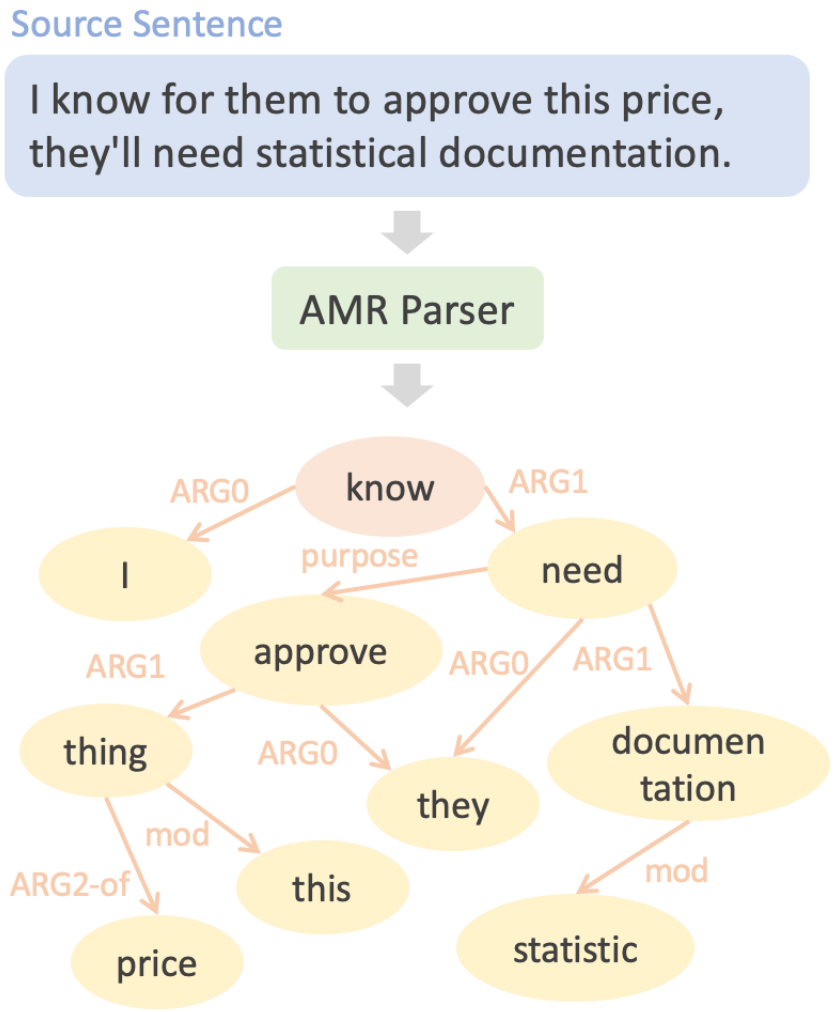
*I am pretty interested in this research direction.* ⟺ *I am very interested in this research direction.*

**Our Goal**

Construct a large-scale syntactically diverse paraphrase dataset

[1] Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources, 2004.
[2] Re-examining machine translation metrics for paraphrase identification, 2012.
[3] First Quora dataset release: Question pairs, 2017.
[4] Paranmt-50m: Pushing the limits of paraphrastic sentence embed- dings with millions of machine translation, 2018.
[5] PARABANK: monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation, 2019.
[6] Large-scale, diverse, paraphrastic bitexts via sampling and clustering, 2019.

# Key Idea: Leveraging AMR Graphs




**Abstract Meaning Representations (AMR)**

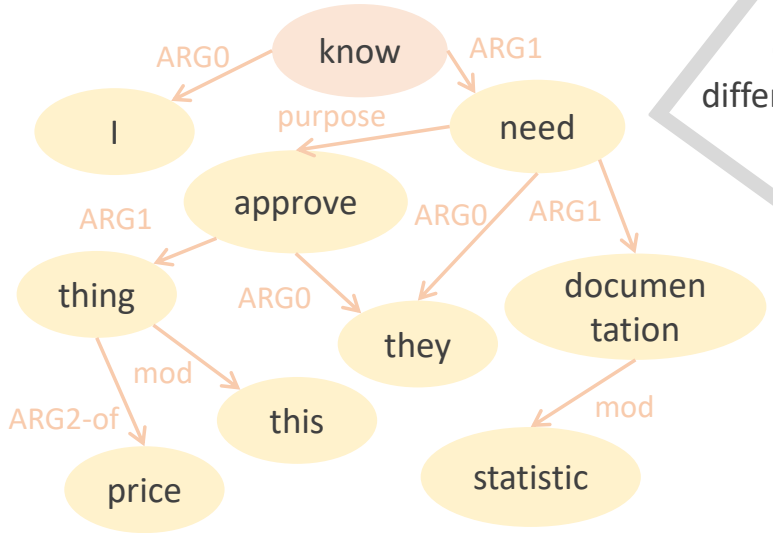
- A directed graph capturing the abstract meaning of a sentence
- Nodes represent semantic concepts
- Edges represent semantic relations
- Focus (root node) represents main assertion of the sentence

# Generating Paraphrases by AMR Back-Translation

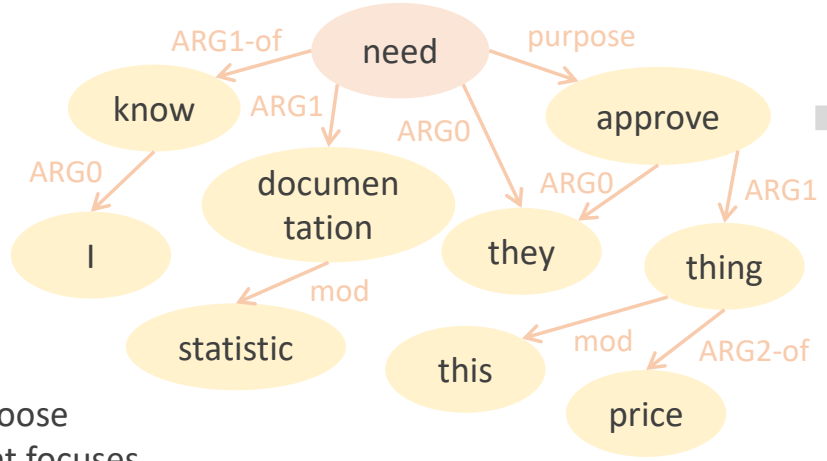
Source Sentence

I know for them to approve this price, they'll need statistical documentation.

AMR Parser



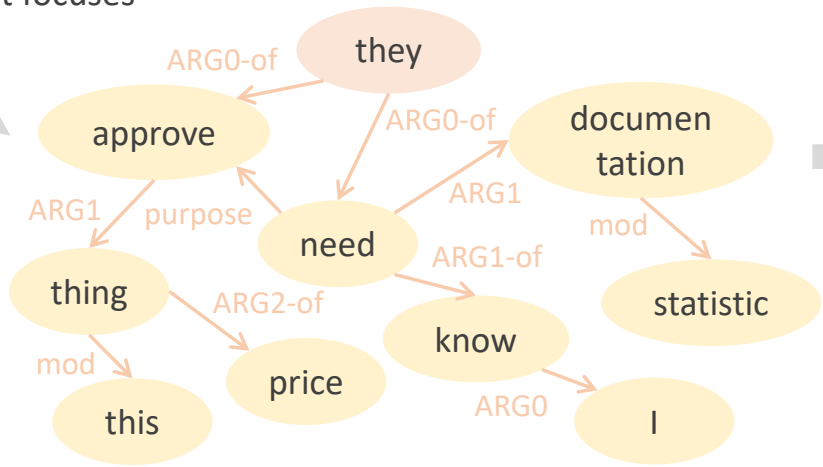

Choose different focuses




Linearized AMR Graph → AMR-to-text Generator

I know they need statistical documentation to approve this price.

Syntactically Diverse Paraphrase




Linearized AMR Graph → AMR-to-text Generator

They need statistical documentation to approve these prices, I know.

Syntactically Diverse Paraphrase

# Proposed Dataset: ParaAMR

- Around 15.5 million source sentences, 6.92 paraphrase per sentence

| | |
|---|---|
| Source Sentence | I know for them to approve this price, they'll need statistical documentation. |
| ParaNMT | I know that in order to accept this award, they'll need a statistical analysis. |
| ParaBank1 | I know that to accept this prize, they're going to need statistical analysis.<br>I know that in order to accept this prize, they're going to need a statistic analysis.<br>I know that if they accept this prize, they're gonna need a statistical analysis. |
| ParaBank2 | I know that to accept that prize, they're going to need a statistical analysis.<br>I know that in order to accept this prize, they will require a statistical analysis.<br>I know they'll require statistical analysis to accept that prize. |
| ParaAMR | I know they need statistical documentation to approve this price.<br>There is statistic documentation I know they need to approve these prices.<br>They need statistical documentation to approve these prices, I know. |

# Quantitative Analysis for ParaAMR

Automatic Scores

| Dataset | Semantic Similarity (↑) | Syntactic Diversity TED-3 (↑) | Syntactic Diversity TED-F (↑) |
|---|---|---|---|
| ParaNMT (Wieting and Gimpel, 2018) | **84.28** | 3.28 | 13.94 |
| ParaBank1 (Hu et al., 2019a) | 81.77 | 3.59 | 14.53 |
| ParaBank2 (Hu et al., 2019b) | 82.50 | 4.04 | 17.41 |
| ParaAMR (Ours) | 82.05 | **5.86** | **22.07** |

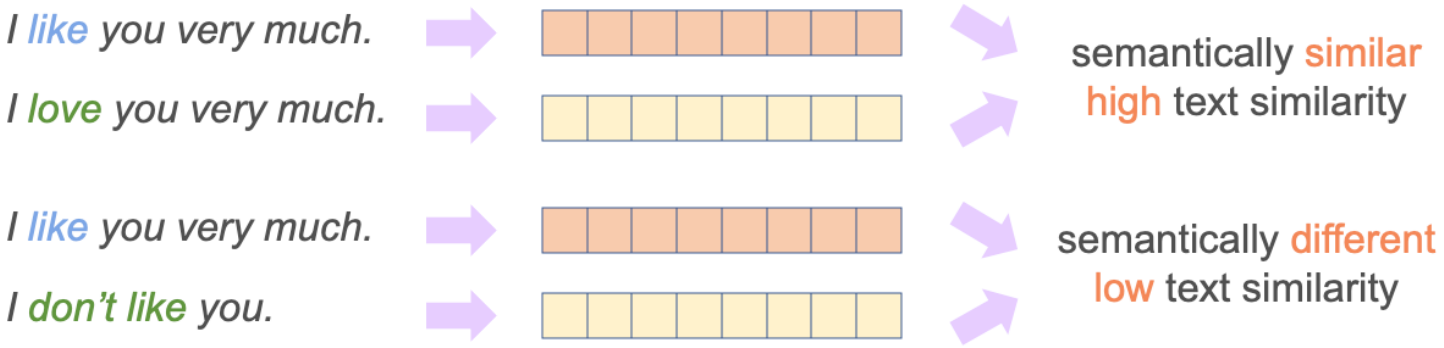
Human Evaluation Scores

| Datasets | Semantic Similarity 3(%) | 2(%) | 1(%) | Average | Syntactic Diversity 3(%) | 2(%) | 1(%) | Average |
|---|---|---|---|---|---|---|---|---|
| ParaNMT (Wieting and Gimpel, 2018) | 28.7 | 46.7 | 24.6 | **2.04** | 16.7 | 45.0 | 38.3 | 1.78 |
| ParaBank1 (Hu et al., 2019a) | 26.8 | 49.0 | 24.2 | 2.03 | 15.1 | 47.8 | 37.1 | 1.78 |
| ParaBank2 (Hu et al., 2019b) | 26.8 | 50.3 | 22.9 | **2.04** | 14.2 | 51.8 | 34.0 | 1.80 |
| ParaAMR (Ours) | 26.5 | 47.2 | 26.3 | 2.00 | 18.2 | 53.8 | 28.0 | **1.90** |

ParaAMR is syntactically more diverse compared to existing datasets while preserving good semantic similarity.

# Application 1: Learning Sentence Embeddings

Semantic Textual Similarity (STS)

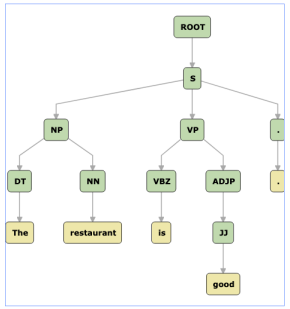



| Dataset | Pearson's r | Spearman's r |
|---|---|---|
| ParaNMT | 74.38 ± 0.70 | 73.80 ± 0.42 |
| ParaBank1 | 74.80 ± 1.33 | 74.56 ± 1.02 |
| ParaBank2 | 75.39 ± 0.29 | 75.17 ± 0.25 |
| ParaAMR (ours) | **77.70** ± 0.40 | **75.72** ± 0.43 |

# Application 2: Syntactically Controlled Paraphrase Generation

## Syntactically Controlled Paraphrase Generation

*This is a good restaurant.*




*The restaurant is good.*

| Dataset | Quora | MRPC | PAN |
|---|---|---|---|
| PARANMT | 47.38 ± 0.39 | 45.24 ± 0.61 | 39.45 ± 0.50 |
| PARABANK1 | 46.21 ± 0.26 | 44.52 ± 0.18 | 39.85 ± 0.11 |
| PARABANK2 | 46.86 ± 0.45 | 45.17 ± 0.39 | 40.20 ± 0.56 |
| PARAAMR (ours) | **48.50** ± 0.11 | **47.38** ± 0.19 | **40.30** ± 0.10 |

# Application 3: Data Augmentation for Few-Shot Learning

| Dataset | MRPC | QQP | RTE |
|---|---|---|---|
| 15-Shot Learning | | | |
| 15-Shot Baseline | 59.93 | 63.18 | 54.05 |
| PARANMT | 49.26 | 63.54 | **55.68** |
| PARABANK1 | 59.56 | 63.72 | 54.59 |
| PARABANK2 | 58.46 | 63.54 | 54.05 |
| PARAAMR (ours) | **62.87** | **64.08** | 52.97 |
| 30-Shot Learning | | | |
| 30-Shot Baseline | 68.38 | 64.93 | 54.51 |
| PARANMT | 67.65 | 66.20 | 52.71 |
| PARABANK1 | 64.46 | 64.86 | 53.79 |
| PARABANK2 | 68.38 | 64.91 | 54.15 |
| PARAAMR (ours) | **69.36** | **67.03** | **55.60** |

# Conclusion

- We proposed ParaAMR
  - Constructed by AMR back-translation
  - Large scale and syntactically diverse
- ParaAMR benefits several NLP applications
  - Learning sentence embeddings
  - Syntactically controlled paraphrase generation
  - Data augmentation for few-shot learning


Dataset is available at https://github. com/uclanlp/ParaAMR