Translation and Evaluation of AMRs

Daniel Gildea

December 13, 2020

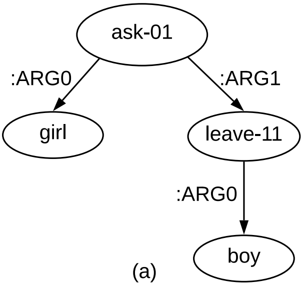
University of Rochester

Evaluation and Translation of AMRs

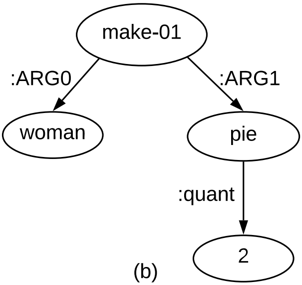
- SemBleu: A Robust Metric for AMR Parsing Evaluation (Song and Gildea, ACL 2019)
- Semantic Neural Machine Translation using AMR (Song, Gildea, Zhang, Wang, and Su, TACL 2019)

Evaluation for Semantic Parsing

"The girl asked the boy to leave."

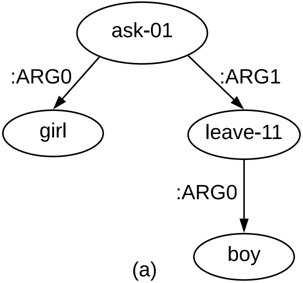


"The woman made two pies."

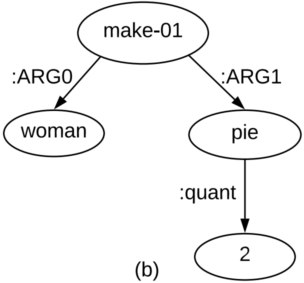


Both the system output and gold reference are graphs. The system's score is based on the similarity of the two graphs.

"The girl asked the boy to leave."

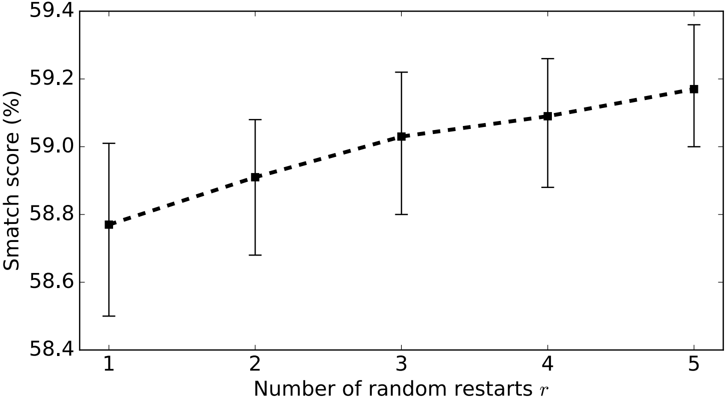


"The woman made two pies."



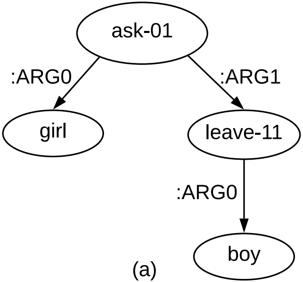
The widely used SMATCH score searches over mappings between the vertices of the two graphs, and measures the number of corresponding nodes and edges with the same label.

SMATCH is non-deterministic, and depends on the number of random restarts used in search:

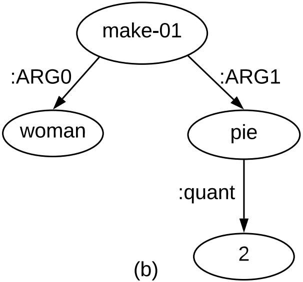


Average, minimal and maximal SMATCH scores over 100 runs on 100 sentences. The running time increases from 6.6 seconds (r=1) to 21.0 ($r = 4$).

“The girl asked the boy to leave.” “The woman made two pies.”



(a)



(b)

BLEU

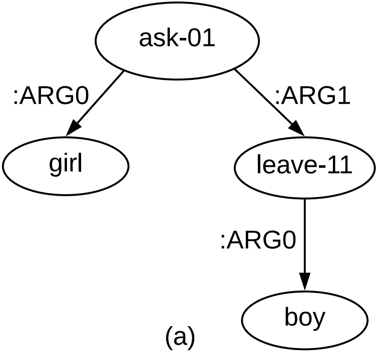
Measures overlap in n-grams between source and reference:
"The girl asked the boy to leave."
"The girl asked the boy to go."

$$\textsc{Bleu} = BP \cdot \exp\left(\sum_{k=1}^{4} \frac{1}{4} \log p_k\right)$$

$$p_k = \text{ k-gram precision } = \text{correct k-grams} \;/\; \text{predicted k-grams}$$

$$BP = \text{brevity penalty} = e^{\min\{1-\frac{|\boldsymbol{r}|}{|\boldsymbol{h}|},0\}}$$

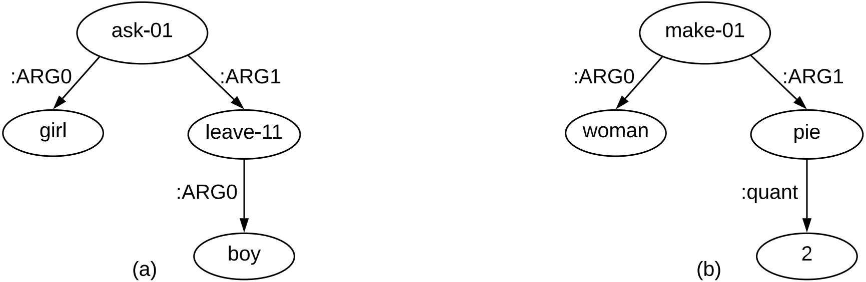
"The girl asked the boy to leave."



(a)

n	Extracted n-grams
1	ask-01; girl; leave-11; boy
2	ask-01 :ARG0 girl; ask-01 :ARG1 leave-11; leave-11 :ARG0 boy;
3	ask-01 :ARG1 leave-11 :ARG0 boy;

"The girl asked the boy to leave."   "The woman made two pies."



SemBleu considers higher order information through longer n-grams. These two graphs have a SemBleu of 0, because they have no matching n-grams.

Evaluation of Evaluation

Agreement with human judgments.
Three raters evaluated 100 pairs of outputs from four systems.

- sentence-level experiment
- corpus-level experiment

Sentence-level experiment

We measure how often the ordering of the score of two outputs is consistent with human judgments.

Metric	Percent (%)
SMATCH	76.5
SEMBLEU (n=1)	69.5
SEMBLEU (n=2)	78.0
SEMBLEU (n=3)	**81.5**
SEMBLEU (n=4)	80.0

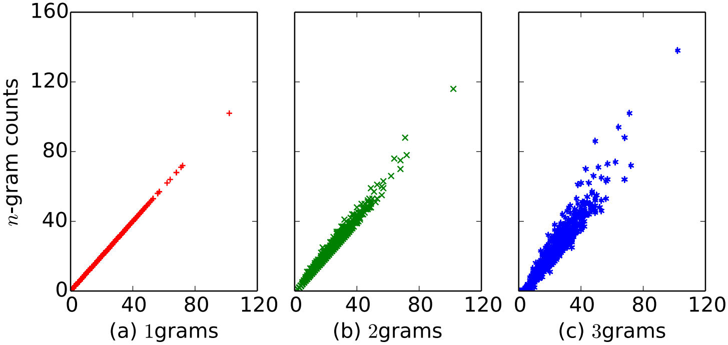
Corpus-level experiment

We measure how often the ordering of scores between two systems agrees with human judgments.
We use bootstrap resampling to measure the significance of each system pair.

Metric	CAMR vs JAMR	CAMR vs Gros	CAMR vs Lyu	JAMR vs Gros	JAMR vs Lyu	Gros vs Lyu
SMATCH	67.9	99.9	100.0	100.0	100.0	90.3
SEMBLEU	69.0	99.9	100.0	100.0	100.0	90.9

Efficiency

The number of n-grams extracted from a graph is potentially exponential in the graph size, but is roughly linear in linear in the graph size for AMRs.

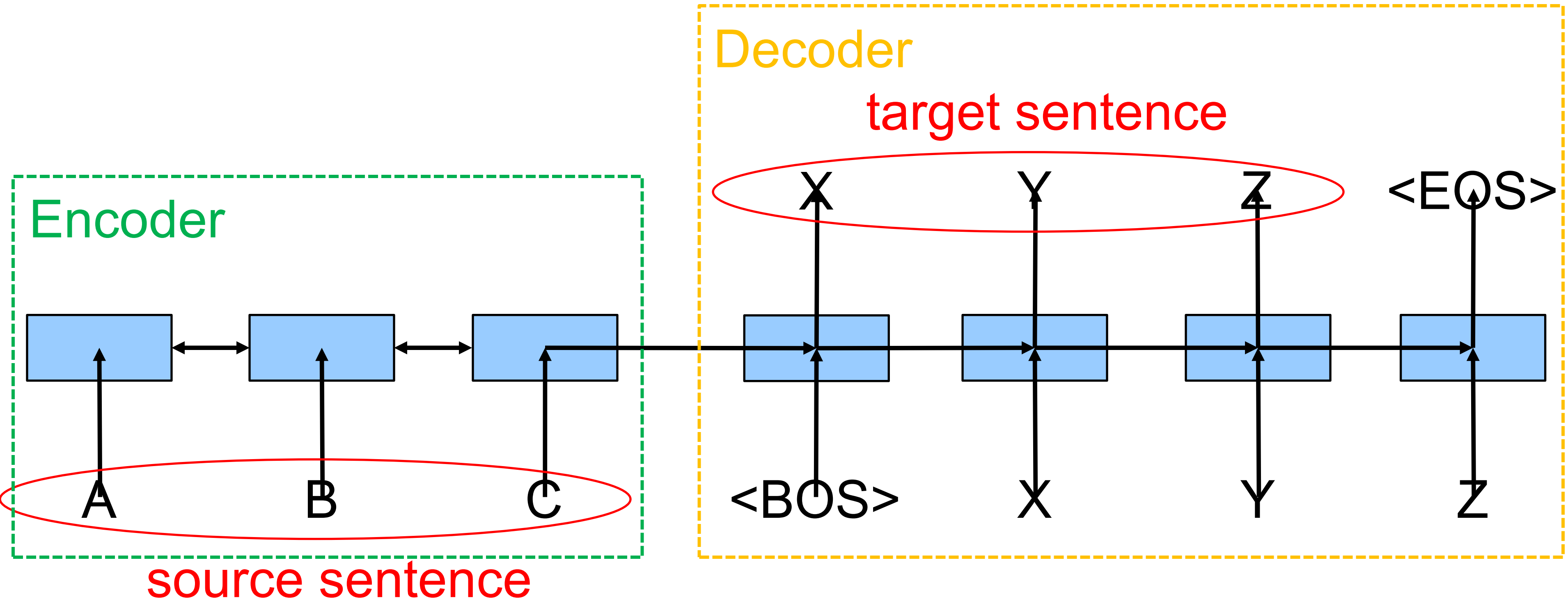


On a dataset of 1368 pairs of AMRs, SemBleu takes 0.5 seconds, while Smatch takes almost 2 minutes.

Conclusion

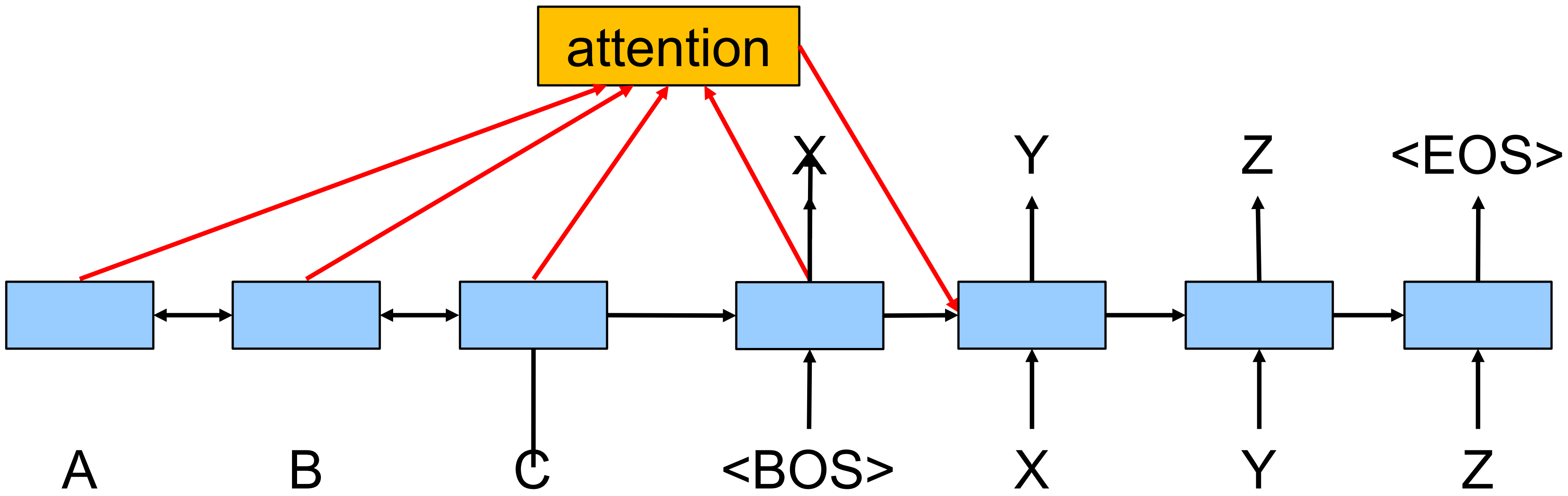
- SemBleu has the advantage of being deterministic, and fast to compute in practice.
- It correlates at least as well as Smatch with human judgments.

Sequence-to-sequence model for NMT



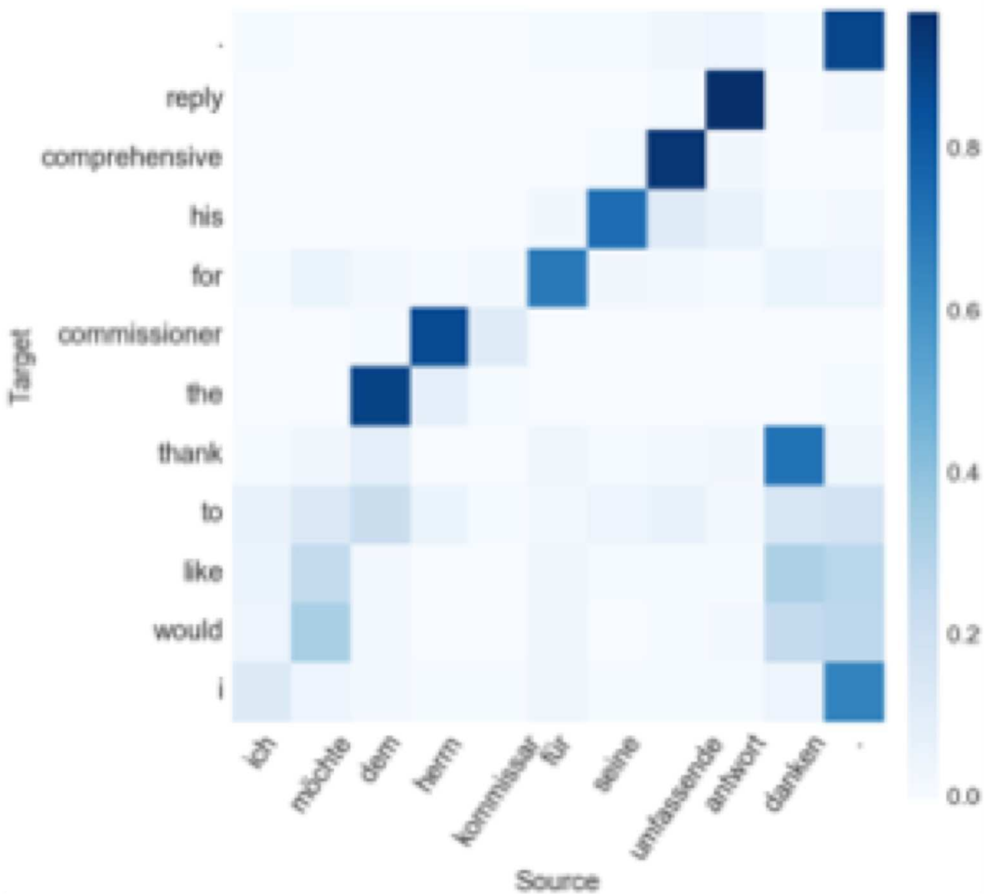
Bahdanau et al., ICLR 2015

Sequence-to-sequence model with attention mechanism

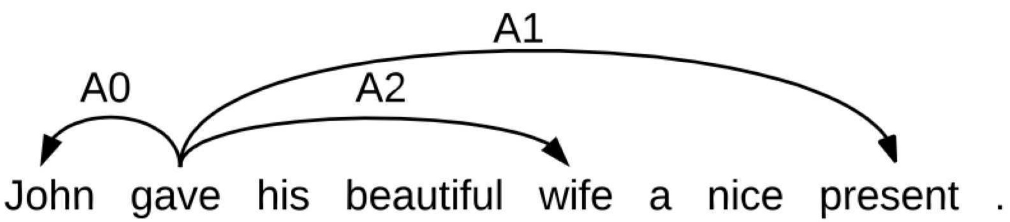


Bahdanau et al., ICLR 2015

Sequence-to-sequence model with attention mechanism

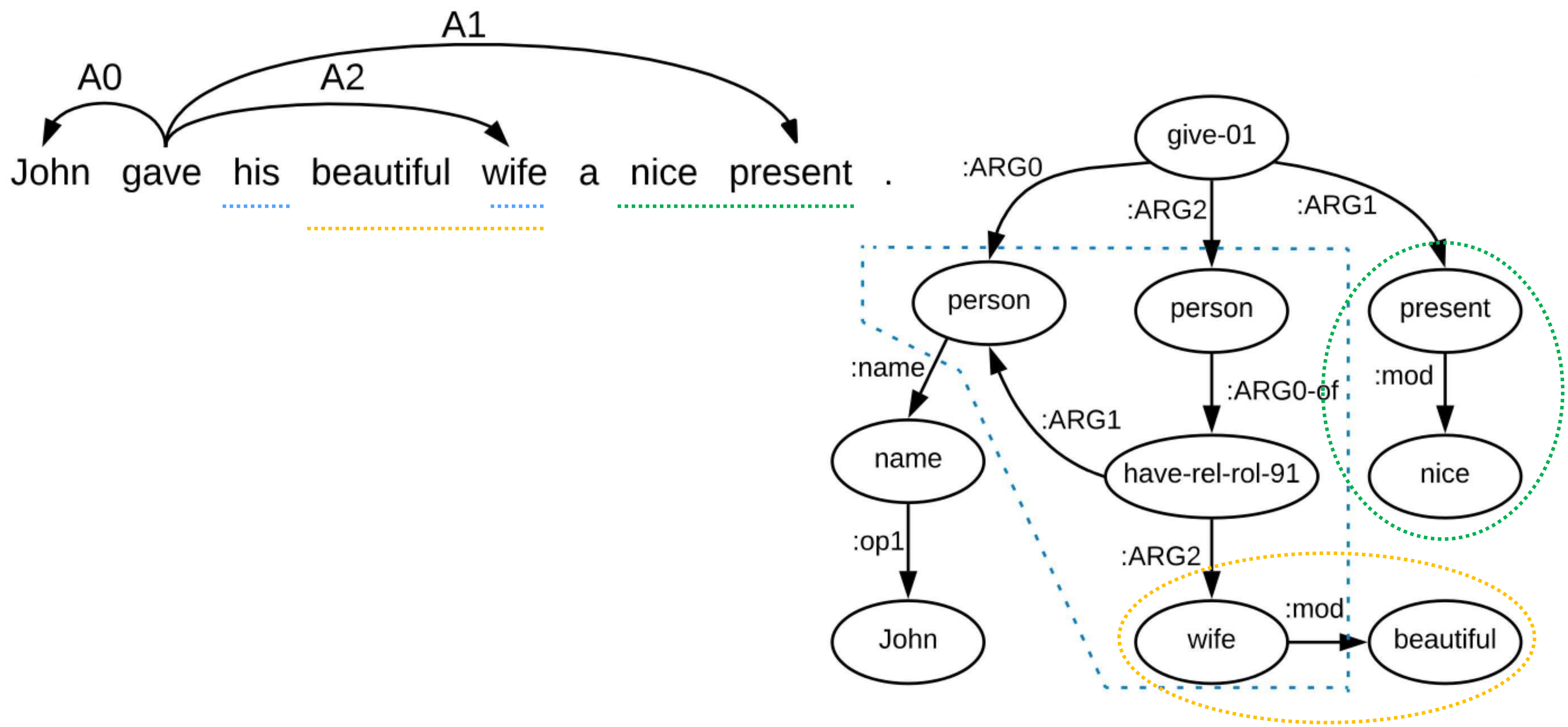


NMT with semantic roles



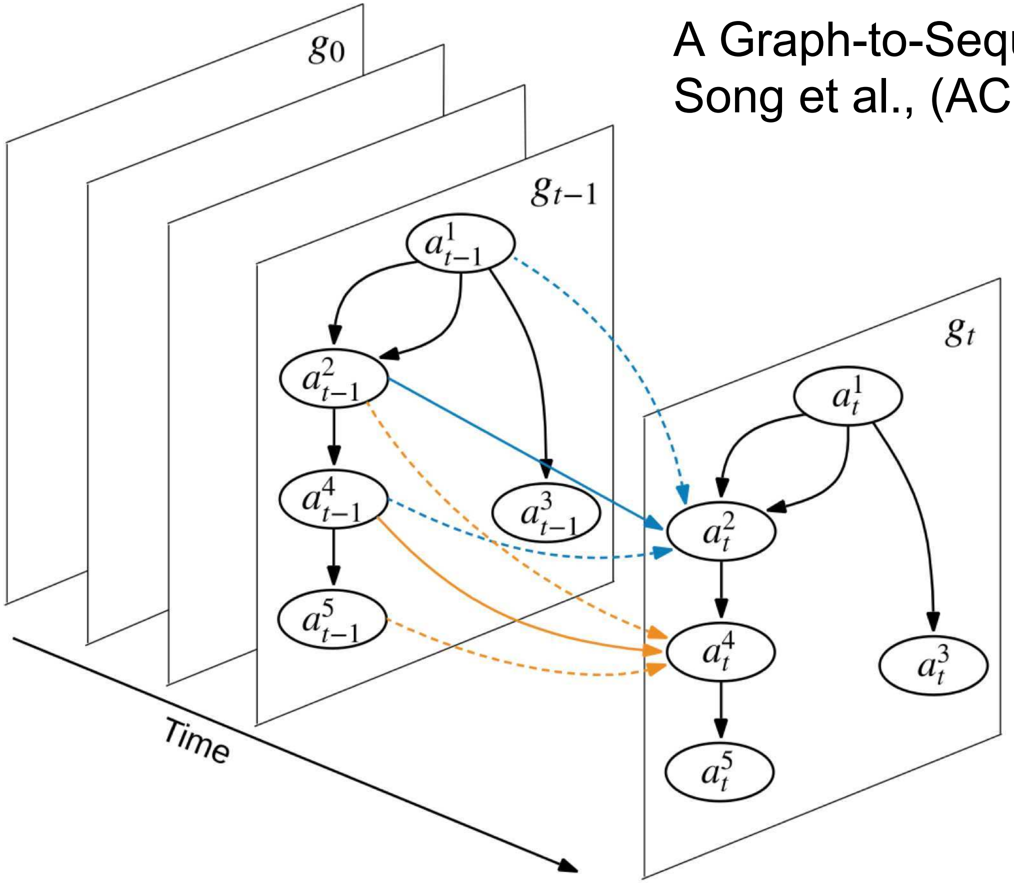
Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks. Marcheggiani et al., (NAACL 2018).

NMT with abstract meaning representation (AMR)

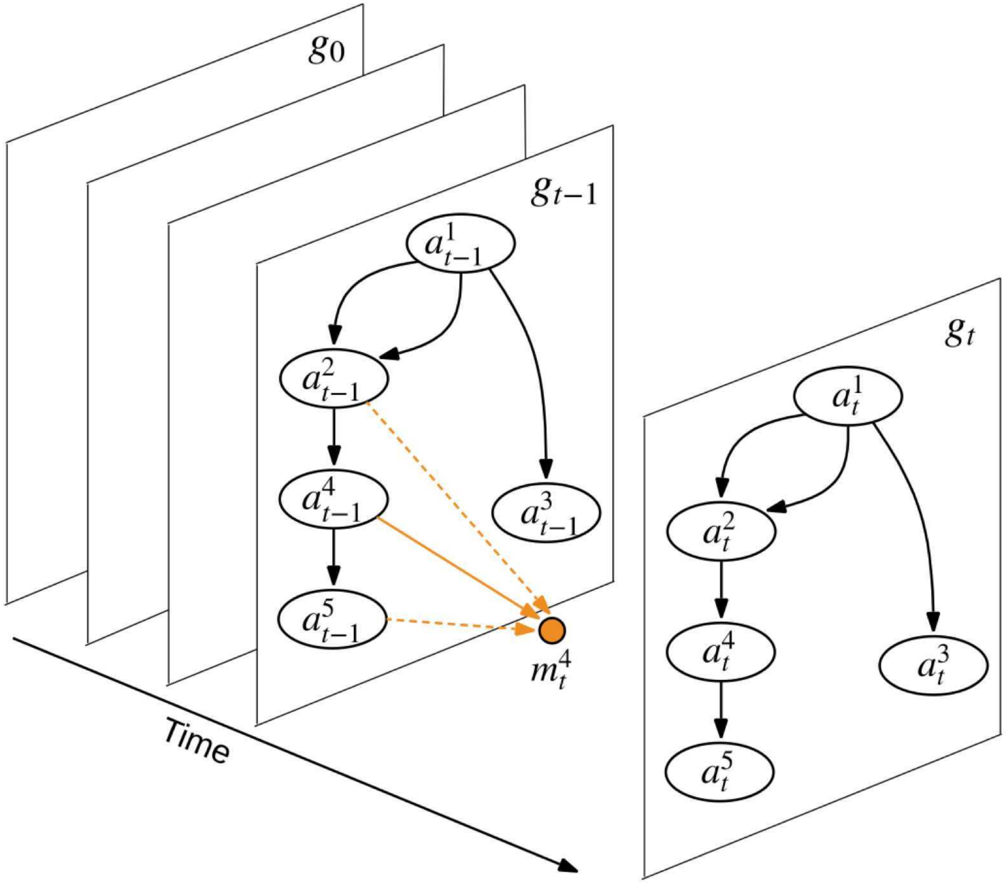


Encoding AMR with graph recurrent network (GRN)

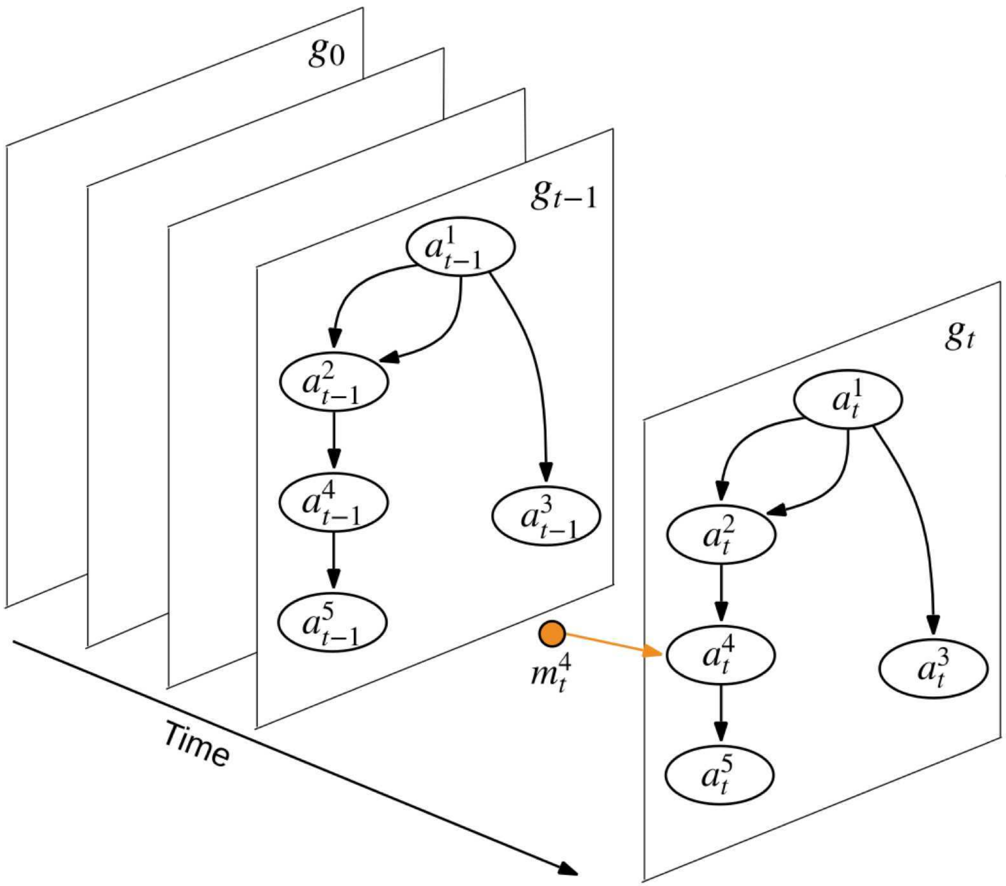
A Graph-to-Sequence Model for AMR-to-Text Generation
Song et al., (ACL 2018)



Encoding AMR with graph recurrent network (GRN)

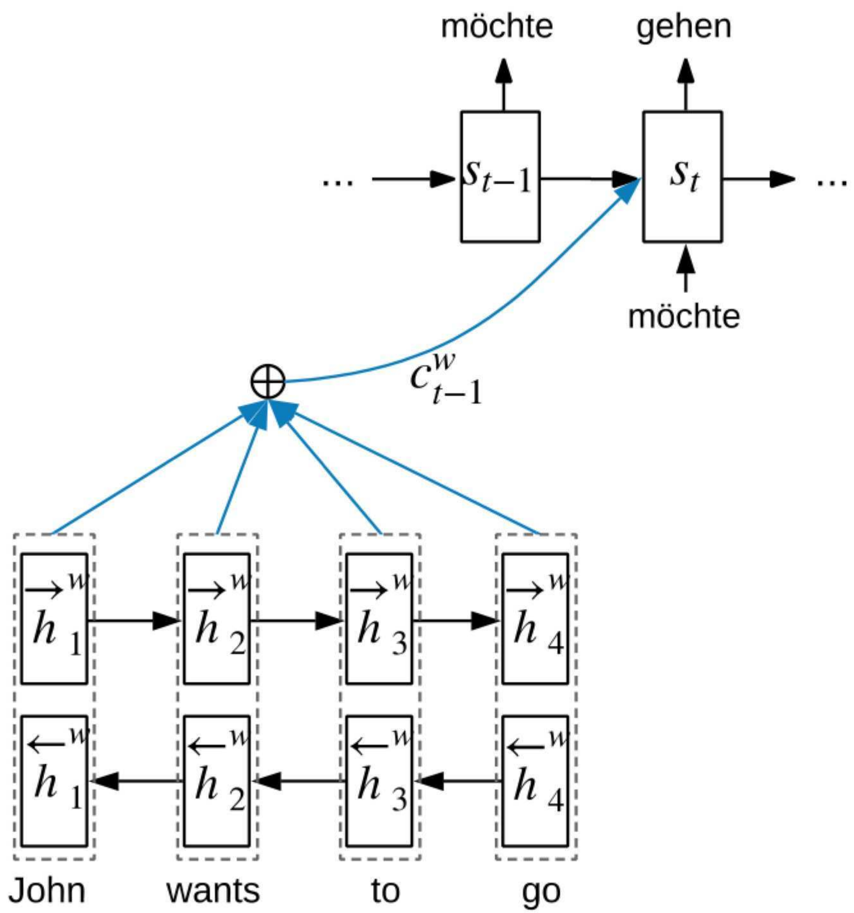


Encoding AMR with graph recurrent network (GRN)



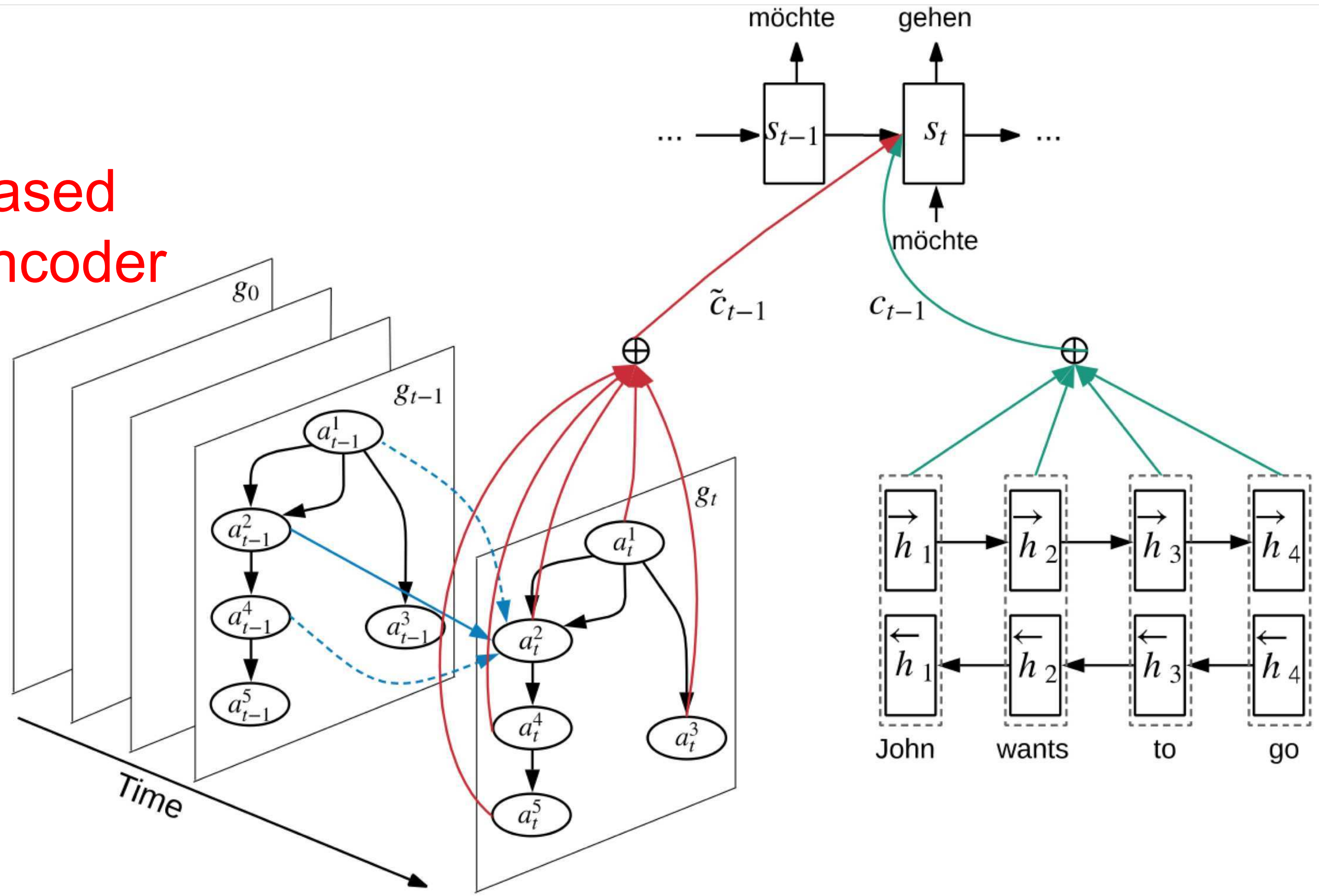
$$a_t^i, c_t^i = LSTM(m_t^i, [a_{t-1}^i, c_{t-1}^i])$$

Baseline: attentional sequence-to-sequence model



Our model: Dual2seq

GRN-based
AMR encoder



Other models

- **Dual2seq-Dep**: same with Dual2seq, but GRN is for encoding dependency trees instead of AMRs

- **Dual2seq-SRL**: same with Dual2seq, but GRN is for encoding semantic roles instead of AMRs

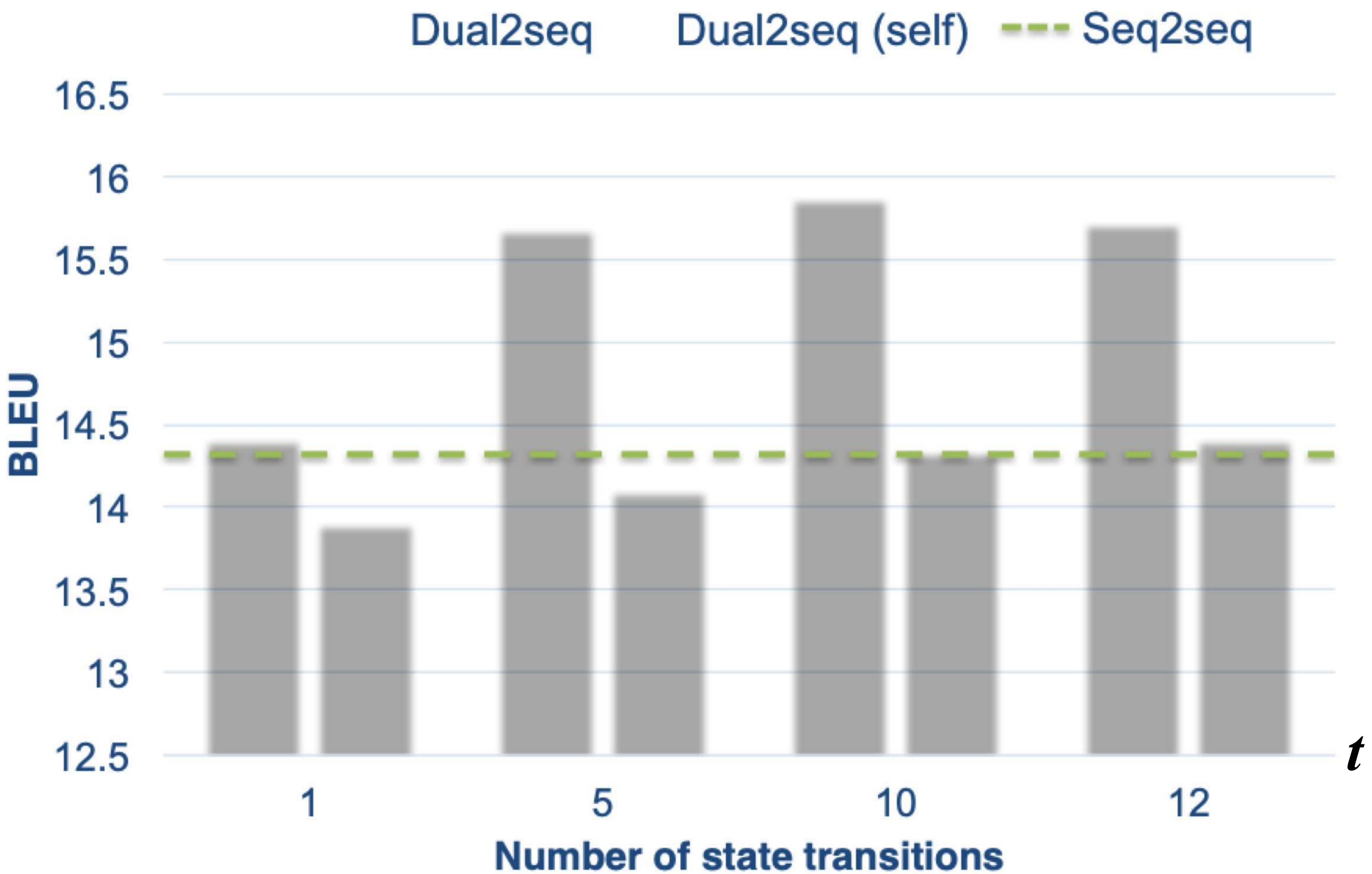
- **Dual2seq (self)**: same with Dual2seq, but GRN is for encoding source sentences, treating it as a chain graph.

- **Dual2seq-LinAMR**: use additional sequential encoder (instead of our GRN) to encode linearized AMRs.

Experiments

- Data (EN-DE):
  - Training: News commentary v11 (241K), full WMT 16 (4.5M)
  - Dev/Test: newstest2013/newstest2016
- Preprocessing:
  - Token by Moses tokenizer
  - Training sentences with length ≥ 50 are filtered
  - AMRs (JAMR), dependency trees (Stanford CoreNLP), semantic roles (IBM SIRE)
- Report cased *BLEU**, *Meteor* and *TER*↓

Development experiment



Main results

System	NC-v11			Full WMT 16		
	BLEU(%)	TER↓	Meteor(%)	BLEU(%)	TER↓	Meteor(%)
OpenNMT-tf	15.1	0.6902	30.4	24.3	0.5567	42.3
Marcheggiani et al. (Seq)	14.9	--	--	23.3	--	--
Marcheggiani et al. (Dep)	16.1	--	--	23.9	--	--
Marcheggiani et al. (SRL)	15.6	--	--	24.5	--	--
Marcheggiani et al. (both)	15.8	--	--	24.9	--	--
Seq2seq	16.0	0.6695	33.8	23.7	0.5590	42.6
Dual2seq-LinAMR	17.3	0.6530	36.1	24.0	0.5643	42.5
Duel2seq-SRL	17.2	0.6591	36.4	23.8	0.5626	42.2
Dual2seq-Dep	17.8	0.6516	36.7	25.0	0.5538	43.3
Dual2seq	**19.2**	**0.6305**	**38.4**	**25.5**	**0.5480**	**43.8**

Main results

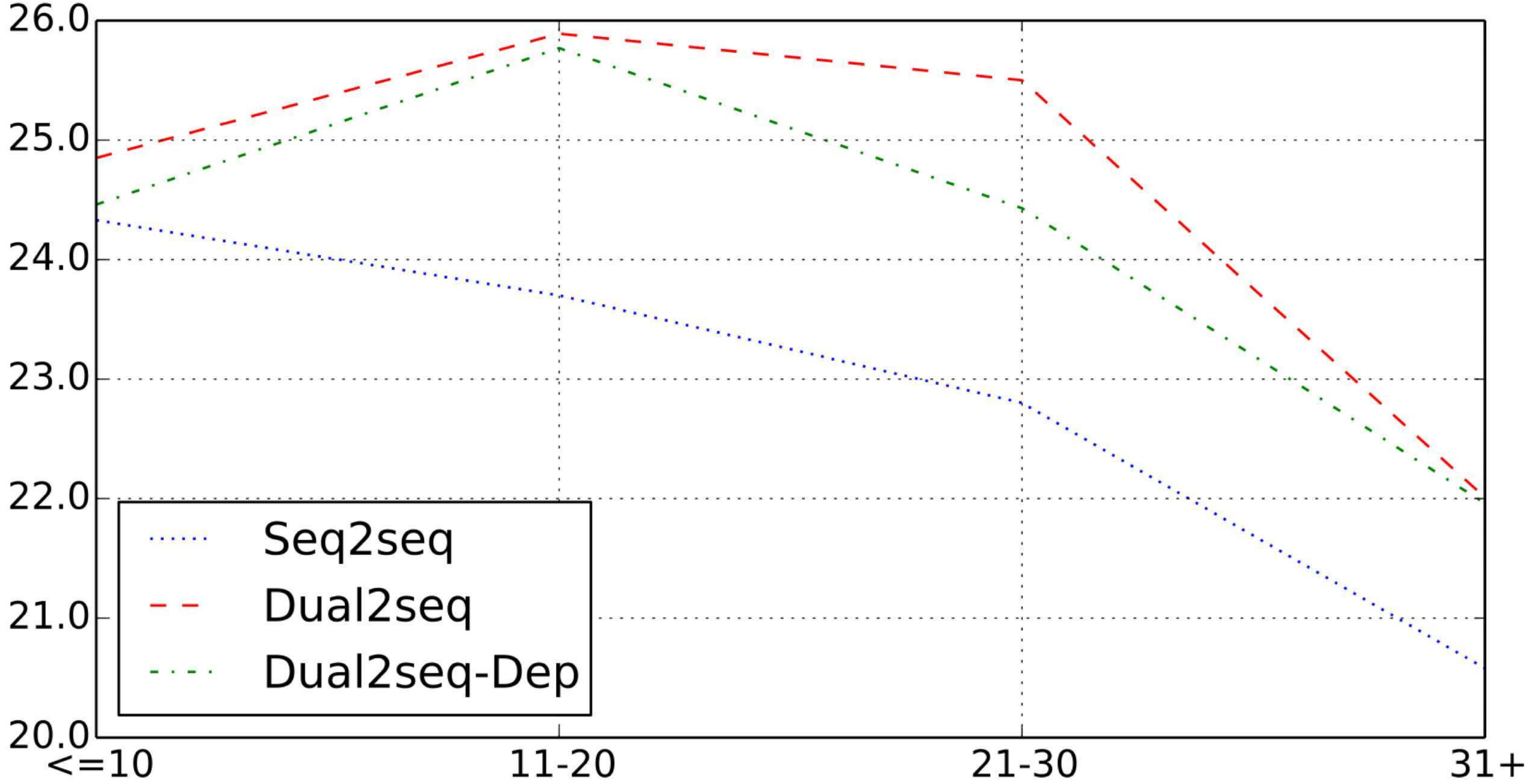
System	NC-v11			Full WMT 16		
	BLEU(%)	TER↓	Meteor(%)	**BLEU**(%)	TER↓	Meteor(%)
OpenNMT-tf	15.1	0.6902	30.4	24.3	0.5567	42.3
Marcheggiani et al. (Seq)	14.9	--	--	23.3	--	--
Marcheggiani et al. (Dep)	16.1	--	--	23.9	--	--
Marcheggiani et al. (SRL)	15.6	--	--	24.5	--	--
Marcheggiani et al. (both)	15.8	--	--	24.9	--	--
Seq2seq	16.0	0.6695	33.8	23.7	0.5590	42.6
Dual2seq-LinAMR	17.3	0.6530	36.1	24.0	0.5643	42.5
Duel2seq-SRL	17.2	0.6591	36.4	23.8	0.5626	42.2
Dual2seq-Dep	17.8	0.6516	36.7	25.0	0.5538	43.3
Dual2seq	**19.2**	**0.6305**	**38.4**	**25.5**	**0.5480**	**43.8**

+3.2 +1.8

Main results

System	NC-v11			Full WMT 16		
	BLEU(%)	TER↓	Meteor(%)	**BLEU**(%)	TER↓	Meteor(%)
OpenNMT-tf	15.1	0.6902	30.4	24.3	0.5567	42.3
Marcheggiani et al. (Seq)	14.9	--	--	23.3	--	--
Marcheggiani et al. (Dep)	16.1	--	--	23.9	--	--
Marcheggiani et al. (SRL)	15.6	--	--	24.5	--	--
Marcheggiani et al. (both)	15.8	--	--	24.9	--	--
Seq2seq	16.0	0.6695	33.8	23.7	0.5590	42.6
Dual2seq-LinAMR	17.3	0.6530	36.1	24.0	0.5643	42.5
Duel2seq-SRL	17.2	0.6591	36.4	23.8	0.5626	42.2
Dual2seq-Dep	17.8	0.6516	36.7	25.0	0.5538	43.3
Dual2seq	**19.2**	**0.6305**	**38.4**	**25.5**	**0.5480**	**43.8**

BLEU score of various sentence length



Impact of AMR Parsing Accuracy

BLEU scores of *Dual2seq* on the *little prince* data, when gold or automatic AMRs are available.

AMR Anno.	BLEU
Automatic	16.8
Gold	**17.5***

Human Evaluation

Out of 100 sentences:

Dual2seq (with AMR) better	46
Seq2seq (no AMR) better	23
Tie	31

Example Outputs

Src: Carla Hairston said she was 15 and Lamb was 20 when they met through mutual friends .
Ref: Carla Hairston sagte , sie war 15 und Lamm war 20 , als sie sich durch gemeinsame Freunde trafen .
Dual2seq: Carla Hairston sagte , sie war 15 und Lamm war 20 , als sie sich durch gegenseitige Freunde trafen .
Seq2seq: Carla Hirston sagte , sie sei 15 und Lamb 20 , als sie durch gegenseitige Freunde trafen .
Seq2seq misses reflexive pronoun in German expression "meet each other."

Src: Since then , according to local media , police vehicles are constantly coming across new refugees in Croatian Tavarnik .
Ref: Laut lokalen Medien treffen seitdem im kroatischen Tovarnik ständig Polizeifahrzeuge mit neuen Flüchtlingen ein .
Dual2seq: Seither kommen die Polizeifahrzeuge nach den örtlichen Medien ständig über neue Flüchtlinge in Kroatische Tavarnik .
Seq2seq: Seitdem sind die Polizeiautos nach den lokalen Medien ständig neue Flüchtlinge in Kroatien Tavarnik .
Seq2seq output says the police vehicles *are* refugees.

Src: Scientists have bred worms with genetically modified nervous systems that can be controlled by bursts of sound waves .
Ref: Wissenschaftler haben Würmer mit genetisch veränderten Nervensystemen gezüchtet , die von Ausbrüchen von Schallwellen gesteuert werden können .
Dual2seq: Die Wissenschaftler haben die Würmer mit genetisch veränderten Nervensystemen gezüchtet , die durch Verbrennungen von Schallwellen kontrolliert werden können .
Seq2seq: Wissenschaftler haben sich mit genetisch modifiziertem Nervensystem gezüchtet , die durch Verbrennungen von Klangwellen gesteuert werden können .
Seq2seq output says scientists breed themselves.

Conclusion

- We studied the effectiveness of AMR on neural machine translation
- We leverage a novel graph recurrent network to encode AMRs for better representations
- Experiments show the superiority of our approach over previous work

Thank you for listening!

Questions