Split and Rephrase

Shashi Narayan, Claire Gardent, Shay B. Cohen and Anastasia Shimorina

Split and Rephrase

John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.

Labour politician, John Clancy is the leader of Birmingham.

John Madin was born in Birmingham.

He was the architect of 103 Colmore Row.

Split and Rephrase

John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.

Labour politician, John Clancy is the leader of Birmingham.
John Madin was born in Birmingham.
He was the architect of 103 Colmore Row.

John Clancy is a labor politican who leads Birmingham.
The architect of 103 Colmore Row was born here.
His name was John Madin.

Our Contributions

Split-and-Rephrase: A **new** sentence rewriting task

	Split	**Delete**	**Rephrase**	**Meaning-preserve**
Split-and-Rephrase	✓	✗	✓	✓

A **new benchmark** for this task

Semantically-motivated split model is a key factor in generating fluent and meaning preserving rephrasings

Split-and-Rephrase: Comparisons with Other Tasks

Compression

Paraphrasing

Fusion

Simplification

Split-and-Rephrase: Comparisons with Other Tasks

Compression

Paraphrasing

	Split	Delete	Rephrase	Meaning-preserve
Compression	✗	✓	often	✗
Split-and-Rephrase	✓	✗	✓	✓

Fusion

Simplification

(Knight and Marcu, 2000; Filippova and Strube, 2008; Cohn and Lapata, 2008; Pitler, 2010; Filippova et al, 2015)

Split-and-Rephrase: Comparisons with Other Tasks

Compression

Paraphrasing

	Split	Delete	Rephrase	Meaning-preserve
Fusion	✗	often	✓	often
Split-and-Rephrase	✓	✗	✓	✓

Fusion

Simplification

(McKeown et al., 2010; Filippova, 2010; Thadani and McKeown, 2013)

Split-and-Rephrase: Comparisons with Other Tasks

Compression

Paraphrasing

	Split	Delete	Rephrase	Meaning-preserve
Paraphrasing	✗	✗	✓	✓
Split-and-Rephrase	✓	✗	✓	✓

Fusion

Simplification

(Dras, 1999; Barzilay and McKeown, 2001; Bannard and Callison-Burch, 2005; Wubben et al., 2010; Mallinson et al., 2017)

Split-and-Rephrase: Comparisons with Other Tasks

Compression

Paraphrasing

	Split	Delete	Rephrase	Meaning-preserve
Simplification	✓	✓	✓	✗
Split-and-Rephrase	✓	✗	✓	✓

Fusion

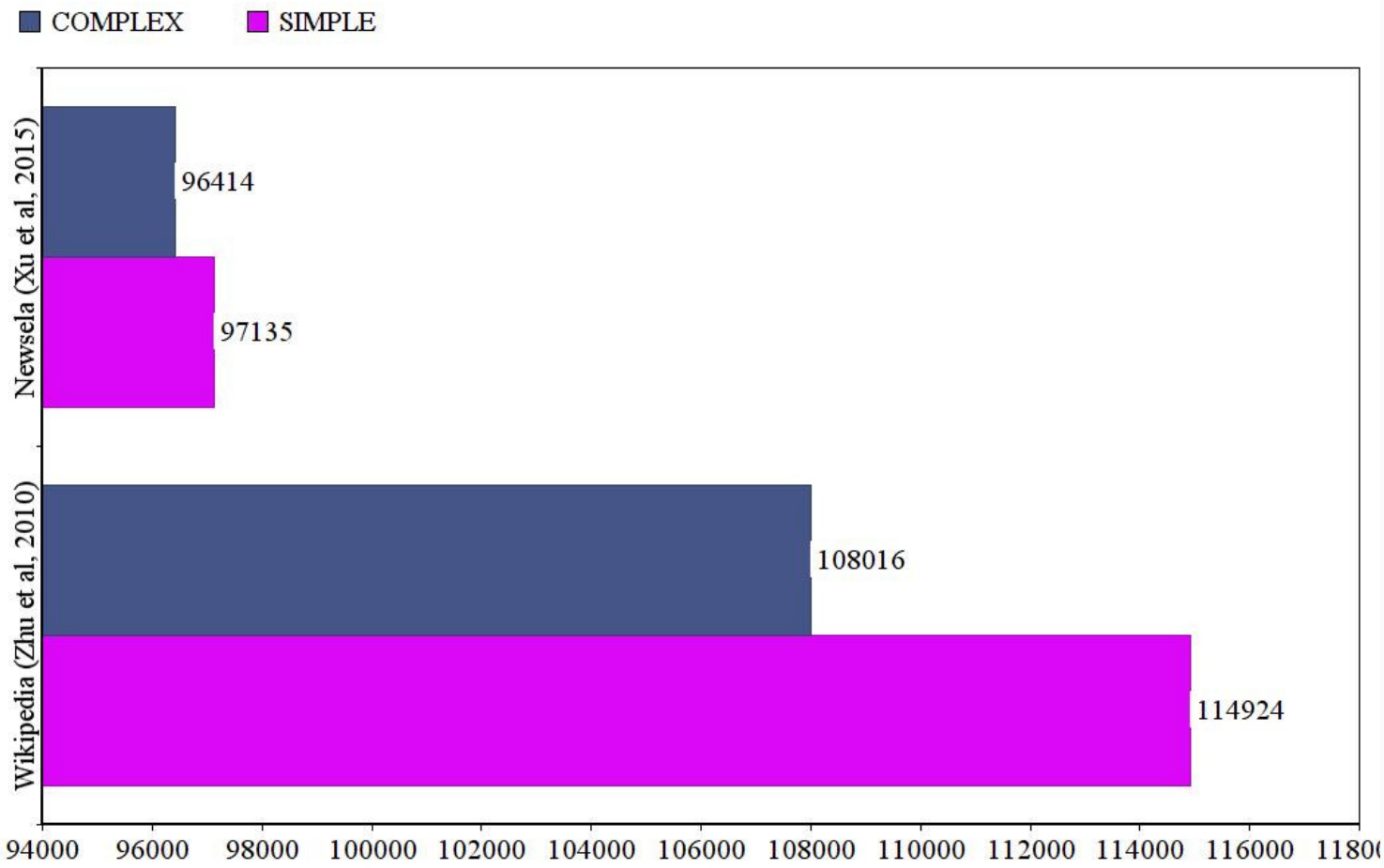
Simplification

(Zhu et al., 2010; Coster and Kauchak, 2011; Woodsend and Lapata, 2011; Wubben et al., 2012;)

(Siddharthan and Mandya, 2014; Narayan and Gardent, 2014, Xu et al., 2015; Zhang and Lapata, 2017)

Limitations of the Current Simplification Datasets

- Ill-suited for syntactic simplification related to splitting.



Split-and-Rephrase: Applications

- Shorter sentences are generally better processed by NLP systems (**NLP applications**).
- Reduced syntactic complexity will improve readability (**Societal applications**).

Split-and-Rephrase: Applications

- Shorter sentences are generally better processed by NLP systems (**NLP applications**).
- Reduced syntactic complexity will improve readability (**Societal applications**).

More beneficial than sentence simplification!

Split-and-Rephrase Benchmark

The Split-and-Rephrase Benchmark

Extracted from our large scale generation (WebNLG) corpus (Gardent et al., ACL 2017)

The WebNLG Corpus

RDF (Resource Description Framework) triple

{*Birmingham* | *leaderName* | *John_Clancy_(Labour_politician)*}

Text *Labour politician, John Clancy is the leader of Birmingham.*

Meaning representations (MRs, a set of RDF triples) paired with **one or more texts** verbalising those triples using **crowdsourcing**.

The Split-and-Rephrase Benchmark

Extracted from our large scale generation (WebNLG) corpus (Gardent et al., ACL 2017)

The WebNLG Corpus

RDF triples

$$\{\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$
$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\}$$

Text-1 *John Madin was born in Birmingham. He was the architect of 103 Colmore Row.*

Text-2 *John Madin who was born in Birmingham, was the architect of 103 Colmore Row.*

The Split-and-Rephrase Benchmark

Extracted from our large scale generation (WebNLG) corpus (Gardent et al., ACL 2017)

The WebNLG Corpus

- 13,308 MR-Text pairs, 7,049 distinct MRs, 8 DBpedia categories and 1-to-7 RDF triples in MRs.

Creating Training Corpora for Micro-Planners, Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini, ACL 2017.

The Split-and-Rephrase Benchmark

Extracted from our large scale generation (WebNLG) corpus (Gardent et al., ACL 2017)

The WebNLG Corpus

- 13,308 MR-Text pairs, 7,049 distinct MRs, 8 DBpedia categories and 1-to-7 RDF triples in MRs.

Pivot approach: Meaning representation (MR) as pivot for the extraction of paraphrases with splits.

Paraphrase Extraction with MRs as Pivot

MR

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$

$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\ \}$$

Paraphrase Extraction with MRs as Pivot

MR

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$

$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\ \}$$

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

T-2 *Labour politician, John Clancy is the leader of Birmingham.*
John Madin was born in this city.
He was the architect of 103 Colmore Row.

Paraphrase Extraction with MRs as Pivot

MR

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$

$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\ \}$$

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

T-2 *Labour politician, John Clancy is the leader of Birmingham.*
John Madin was born in this city.
He was the architect of 103 Colmore Row.

Paraphrase Extraction with MRs as Pivot

MR

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$

$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\ \}$$

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

T-2 *Labour politician, John Clancy is the leader of Birmingham.*
John Madin was born in this city.
He was the architect of 103 Colmore Row.

S-1 *Labour politician, John Clancy is the leader of Birmingham.*

Paraphrase Extraction with MRs as Pivot

MR

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{John\_Madin} \mid \textit{birthPlace} \mid \textit{Birmingham},$$

$$\textit{103\_Colmore\_Row} \mid \textit{architect} \mid \textit{John\_Madin}\ \}$$

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

T-2 *Labour politician, John Clancy is the leader of Birmingham.*
John Madin was born in this city.
He was the architect of 103 Colmore Row.

S-1 *Labour politician, John Clancy is the leader of Birmingham.*

S-2 *John Madin was born in Birmingham.*
He was the architect of 103 Colmore Row.

Paraphrase Extraction: Across and Within Entries

Across Entries {(**MR**,**T-1**), (MR-1, S-1) (MR-2, S-2)}

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

S-1 *Labour politician, John Clancy is the leader of Birmingham.*

S-2 *John Madin was born in Birmingham.*
He was the architect of 103 Colmore Row.

Paraphrase Extraction: Across and Within Entries

Across Entries {(**MR**,**T-1**), (MR-1, S-1) (MR-2, S-2)}

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

S-1 *Labour politician, John Clancy is the leader of Birmingham.*

S-2 *John Madin was born in Birmingham.*
He was the architect of 103 Colmore Row.

Within Entries {**(MR, T-1), (MR, T-2)**}

T-1 *John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.*

T-2 *Labour politician, John Clancy is the leader of Birmingham.*
John Madin was born in this city.
He was the architect of 103 Colmore Row.

The Split-and-Rephrase Benchmark

- 1,100,166 pairs of the form $\{(M_C, C), \{(M_1, S_1) \ldots (M_n, S_n)\}\}$
- 5,546 distinct complex sentences
- The vocabulary size is 3,311

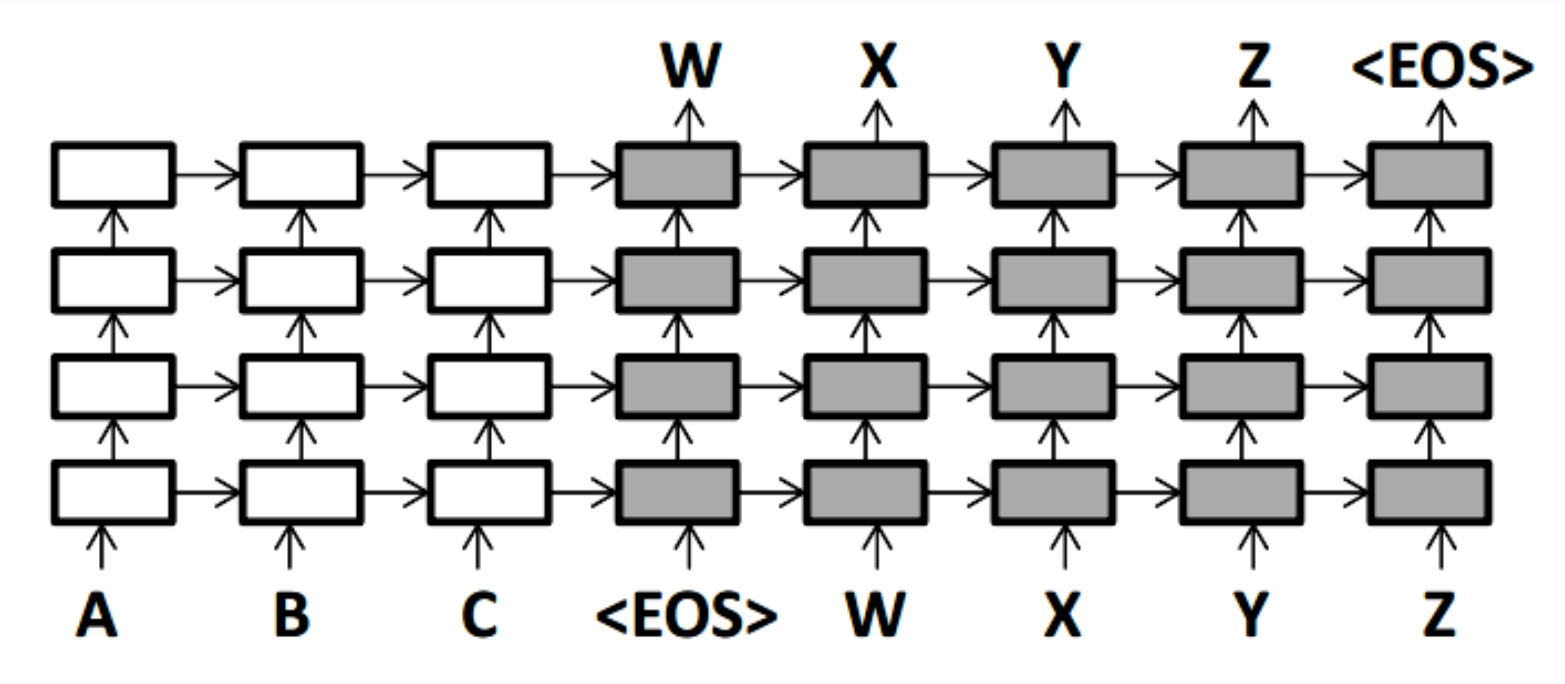
The Split-and-Rephrase Benchmark

- 1,100,166 pairs of the form $\{(M_C, C), \{(M_1, S_1) \ldots (M_n, S_n)\}\}$
- 5,546 distinct complex sentences
- The vocabulary size is 3,311
- Number of sentences in the rephrasings varies between 2 and 7 with an average of 4.99

Split-and-Rephrase Models

Encoder-decoder Framework for NMT (SEQ2SEQ)

- Optimizes $p(S|C)$

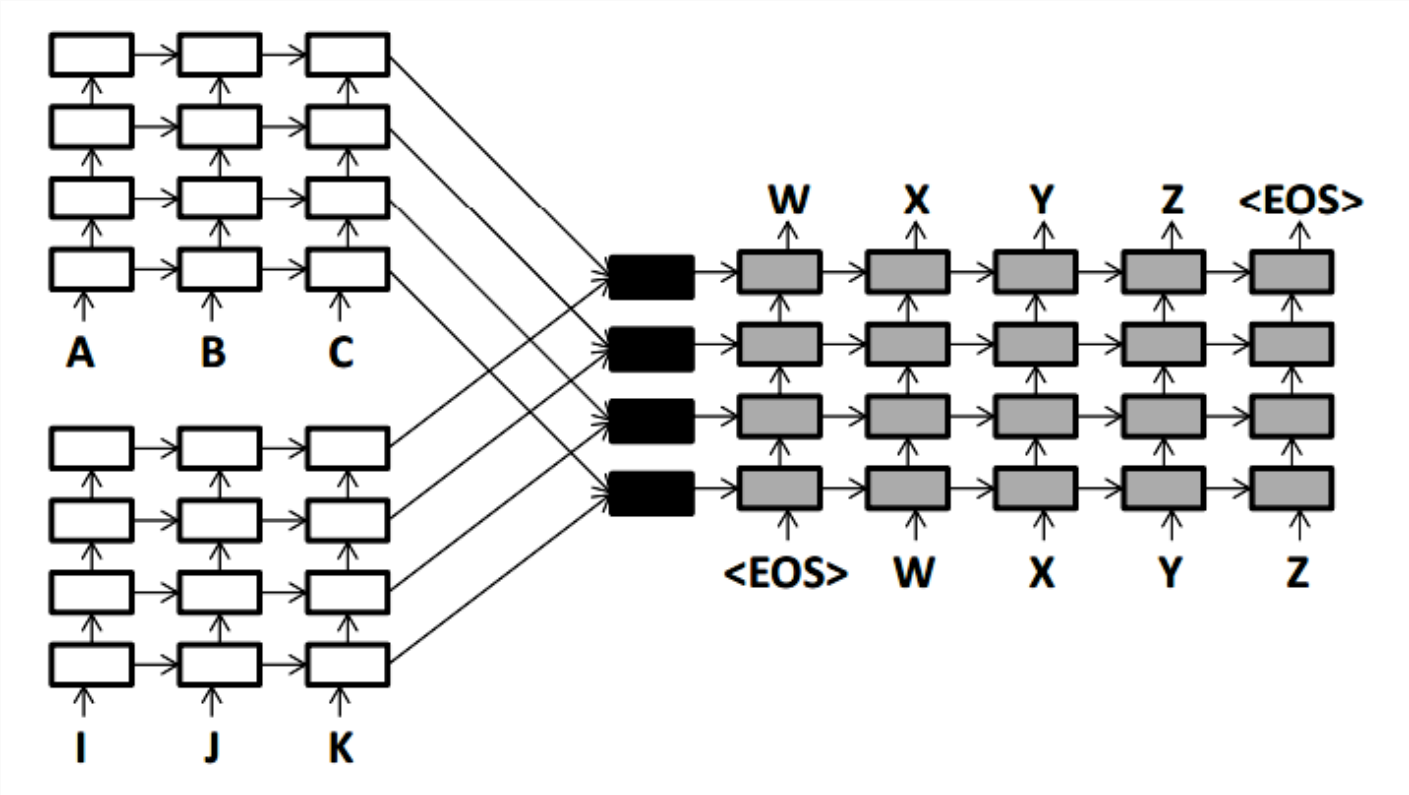


(Sutskever et al., 2011; Bahdanau et al., 2014)

Multi-source NMT (MULTISEQ2SEQ)

$$p(S|C) = \sum_{M_C} p(S|C; M_C)p(M_C|C) = p(S|C; M_C), \text{ if } M_C \text{ is known,}$$

where M_C is the meaning representation (RDF tuples) of C.



Semantically-motivated Partition and Generate

John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.

Inspired from ideas in

Hybrid Simplification using Deep Semantics and Machine Translation, Shashi Narayan and Claire Gardent, ACL 2014**.**

Semantically-motivated Partition and Generate

John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.

$$\{\ \textit{Birmingham} \mid \textit{leaderName} \mid \textit{John\_Clancy\_(Labour\_politician)},$$

$$\textit{Birmingham} \mid \textit{birthPlace} \mid \textit{John\_Madin},$$

$$\textit{John\_Madin} \mid \textit{architect} \mid \textit{103\_Colmore\_Row}\ \}$$

Semantic Representation

Semantically-motivated Partition and Generate

{ *Birmingham* | *leaderName* | *John_Clancy_(Labour_politician)*,

Birmingham | *birthPlace* | *John_Madin*,

John_Madin | *architect* | *103_Colmore_Row* }

{ *Birmingham* | *leaderName* | *John_Clancy_(Labour_politician)* }

{ *Birmingham* | *birthPlace* | *John_Madin*,

John_Madin | *architect* | *103_Colmore_Row* }

Semantically-motivated Partition and Generate

{ *Birmingham* | *leaderName* | *John_Clancy_(Labour_politician)*,

Birmingham | *birthPlace* | *John_Madin*,

John_Madin | *architect* | *103_Colmore_Row* }

{ *Birmingham* | *leaderName* | *John_Clancy_(Labour_politician)* }

Labour politician, John Clancy is the leader of Birmingham.

{ *Birmingham* | *birthPlace* | *John_Madin*,

John_Madin | *architect* | *103_Colmore_Row* }

John Madin, the architect of 103 Colmore Row, was born in Birmingham.

Semantically-motivated Partition and Generate

John Clancy is a labor politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born.

Labour politician, John Clancy is the leader of Birmingham.

John Madin, the architect of 103 Colmore Row, was born in Birmingham.

Semantically-motivated Partition and Generate

$$p(S|C; M_C) = \sum_{M_{1:n}} p(S|C; M_C; M_{1:n}) \times p(M_{1:n}|C; M_C)$$

Rephrase Partition

M_C is the meaning representation (RDF tuples) of C

$M_{1:n} = M_1, \ldots, M_n$ is the partition of M_C.

Semantically-motivated Partition and Generate

$$p(S|C; M_C) = \sum_{M_{1:n}} p(S|C; M_C; M_{1:n}) \times p(M_{1:n}|C; M_C)$$

Rephrase Partition

Learn to Partition $p(M_{1:n}|C; M_C)$

- A probabilistic model trained on the training set $\{(M_C, C), \{(M_1, S_1) \ldots (M_n, S_n)\}\}$

Semantically-motivated Partition and Generate

$$p(S|C; M_C) = \sum_{M_{1:n}} p(S|C; M_C; M_{1:n}) \times p(M_{1:n}|C; M_C)$$

Rephrase Partition

Learn to Rephrase $p(S|C; M_C; M_{1:n})$

$$p(S|C; M_C; M_1, \ldots, M_n) \approx \prod_i^n p(S_i|C; M_i), \text{ (multi-seq2seq)}$$

$$\approx \prod_i^n p(S_i|M_i), \text{ (seq2seq)}$$

Results

- Training set (4,438, 80%), Validation set (554, 10%) and Test set (554, 10%)

- We evaluate on
  - **Meaning Preservation:** Multi-reference BLEU-4 scores
  - **Splits:**
    - #S/C: Average number of sentences in the output texts
    - #Tokens/S: Average number of tokens per output sentences

Results

Model	BLEU	#S/C	#Tokens/S
INPUT	55.67	1.0	21.11

INPUT	Alan Shepard was born in New Hampshire and he served as the Chief of the Astronaut Office.

Results

Model	BLEU	#S/C	#Tokens/S
INPUT	55.67	1.0	21.11
SEQ2SEQ	48.92	2.51	10.32
MULTISEQ2SEQ	42.18	2.53	10.69

INPUT	Alan Shepard was born in New Hampshire and he served as the Chief of the Astronaut Office.
SEQ2SEQ	Alan Shepard's occupation was a test pilot. Alan Shepard was born in New Hampshire. Alan Shepard was born on Nov 18, 1923.
MULTISEQ2SEQ	Alan Shepard served as a test pilot. Alan Shepard's birth place was New Hampshire.

Results

Model	BLEU	#S/C	#Tokens/S
INPUT	55.67	1.0	21.11
SEQ2SEQ	48.92	2.51	10.32
MULTISEQ2SEQ	42.18	2.53	10.69
SPLIT-SEQ2SEQ	**78.77**	**2.84**	**9.28**
SPLIT-MULTISEQ2SEQ	77.27	**2.84**	11.63

INPUT	Alan Shepard was born in New Hampshire and he served as the Chief of the Astronaut Office.
SPLIT-SEQ2SEQ	Alan Shepard served as the Chief of the Astronaut Office. Alan Shepard's birth place was New Hampshire.
SPLIT-MULTISEQ2SEQ	Alan Shepard served as the Chief of the Astronaut Office. Alan Shepard was born in New Hampshire.

Results

Model	Task	Training Size
SEQ2SEQ	Given C, predict S	886,857
MULTISEQ2SEQ	Given C and M_C, predict S	886,866
SPLIT-MULTISEQ2SEQ	Given C and M_C, predict $M_1 \ldots M_n$	13,051
	Given C and S_i, predict S_i	53,470
SPLIT-SEQ2SEQ	Given C and T_C, predict $M_1 \ldots M_n$	13,051
	Given M_i, predict T_i	53,470

Future work

- Jointly learn to partition and rephrase

$$p(S|C; M_C) = \sum_{M_{1:n}} p(S|C; M_C; M_{1:n}) \times p(M_{1:n}|C; M_C)$$

- Coverage based encoder-decoder models

Future work

- Jointly learn to partition and rephrase

$$p(S|C; M_C) = \sum_{M_{1:n}} p(S|C; M_C; M_{1:n}) \times p(M_{1:n}|C; M_C)$$

- Coverage based encoder-decoder models
- Limitations of the Split-and-Rephrase benchmark
  - Notion of semantics simplifies with RDF triples: text is restricted to entity descriptions
  - Lexical diversity (portability to a new domain)

Where are we?

Conclusion

- We presented a new task for sentence splitting and rephrasing.
- Our experiments indicate that the semantically-motivated split model is a key factor in generating fluent and meaning preserving rephrasings.
- Our Split-and-Rephrase benchmark will be available at `https://github.com/shashiongithub/Split-and-Rephrase`.