Bootstrapping Generators from Noisy Data

Laura Perez-Beltrachini   Mirella Lapata

School of Informatics
University of Edinburgh
{lperez,mlap}@inf.ed.ac.uk

The Data to Text Generation Task

Input Set of Properties

Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Output Verbalisation

Robert Joseph Flaherty, (February 16, 1884 July 23, 1951) was an American film-maker. Flaherty was married to Frances H. Flaherty until his death in 1951.

The Data to Text Generation Task

Input Set of Properties

Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Output Verbalisation

Robert Joseph Flaherty, (February 16, 1884 July 23, 1951) was an American film-maker. Flaherty was married to Frances H. Flaherty until his death in 1951.

Application? Automatic generation of descriptions for Amazon

The Data to Text Generation Task

Input Set of Properties

Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Output Verbalisation

Robert Joseph Flaherty, (February 16, 1884 July 23, 1951) was an American film-maker. Flaherty was married to Frances H. Flaherty until his death in 1951.

Application? Automatic generation of descriptions for Amazon

Approach? Neural Generator trained on Loosely Related Texts

Learning Neural Generators from Loosely Related Data-Text Pairs

Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Joseph Flaherty, (February 16, 1884 – July 23, 1951) was an American film-maker who directed and produced the first commercially successful feature-length documentary film, Nanook of the North (1922). Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951. Frances worked on several of her husband's films, and received an Academy Award nomination for Best Original Story for Louisiana Story (1948).

DBpedia

Wikipedia
The Free Encyclopedia

[Lebret et al., 2016, Wiseman et al., 2017]

Learning Neural Generators from Loosely Related Data-Text Pairs

Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Joseph Flaherty, (February 16, 1884 – July 23, 1951) was an American film-maker who directed and produced the first commercially successful feature-length documentary film, Nanook of the North (1922). Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951. Frances worked on several of her husband's films, and received an Academy Award nomination for Best Original Story for Louisiana Story (1948).

Learning Neural Generators from Loosely Related Data-Text Pairs

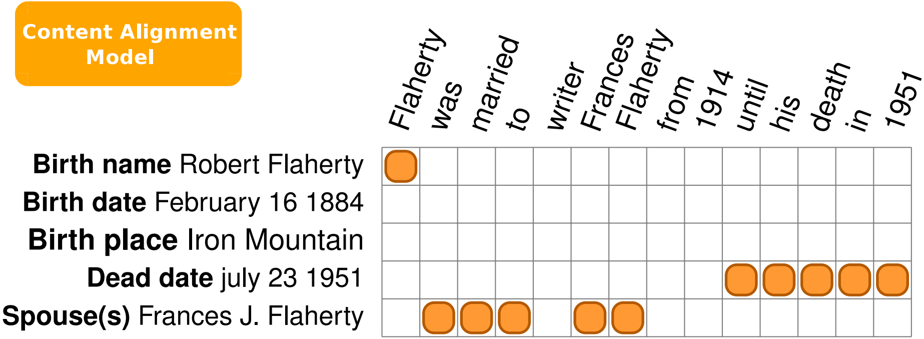
Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Joseph Flaherty, (February 16, 1884 – July 23, 1951) was an American film-maker who directed and produced the first commercially successful feature-length documentary film, Nanook of the North (1922). Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951. Frances worked on several of her husband's films, and received an Academy Award nomination for Best Original Story for Louisiana Story (1948).

Distribute attention weights & Memorised high frequency sub-sequences!

[Ghader and Monz, 2017]

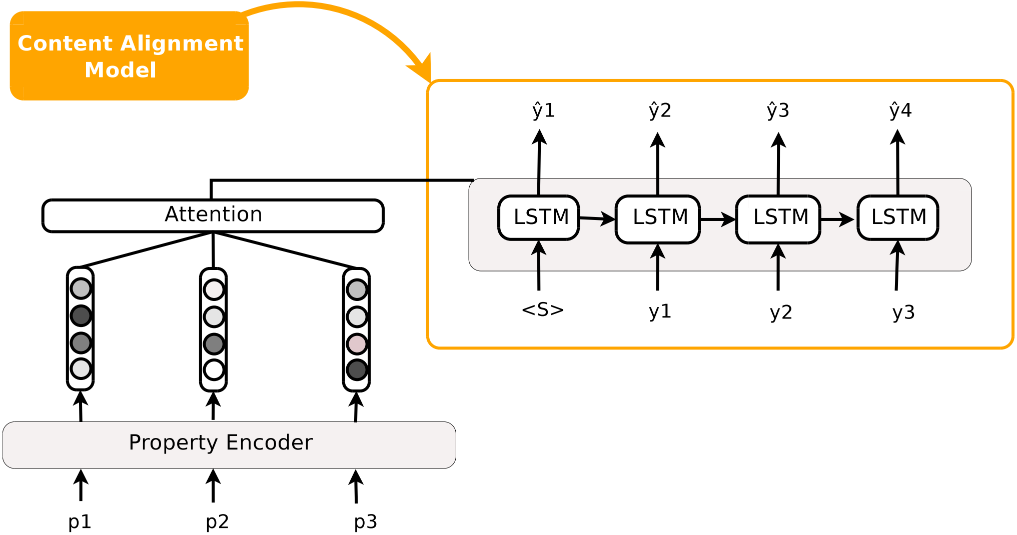
Our Approach, Pre-Train a Content Alignment Model



Content Alignment Model

	Flaherty	was	married	to	writer	Frances	Flaherty	from	1914	until	his	death	in	1951
Birth name Robert Flaherty	●													
Birth date February 16 1884														
Birth place Iron Mountain														
Dead date july 23 1951										●	●	●	●	●
Spouse(s) Frances J. Flaherty		●	●	●		●	●							

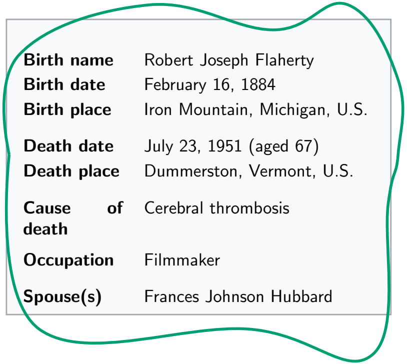
Our Approach, Use the Content Alignment Information for Training



Content Alignment Intuition

- **Multi-Instance Learning** to discover Property-Word Alignments
- Loosely related (Property Set, Text) pairs provide high level supervision

Property Set Bag



Birth name	Robert Joseph Flaherty
Birth date	February 16, 1884
Birth place	Iron Mountain, Michigan, U.S.
Death date	July 23, 1951 (aged 67)
Death place	Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

Bag Labels

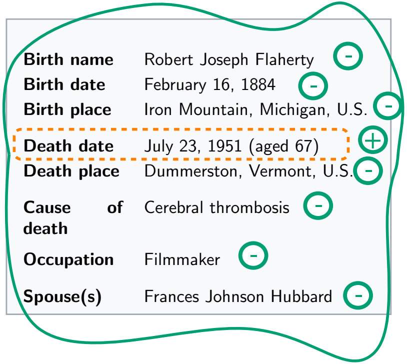
Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951.

[Keeler and Rumelhart, 1992, Karpathy and Fei-Fei, 2015]

Content Alignment Intuition

- **Multi-Instance Learning** to discover Property-Word Alignments
- Loosely related (Property Set, Text) pairs provide high level supervision

Property Set Bag



Birth name	Robert Joseph Flaherty	⊖
Birth date	February 16, 1884	⊖
Birth place	Iron Mountain, Michigan, U.S.	⊖
Death date	July 23, 1951 (aged 67)	⊕
Death place	Dummerston, Vermont, U.S.	⊖
Cause of death	Cerebral thrombosis	⊖
Occupation	Filmmaker	⊖
Spouse(s)	Frances Johnson Hubbard	⊖

Bag Labels

Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951.

[Keeler and Rumelhart, 1992, Karpathy and Fei-Fei, 2015]

Content Alignment Model

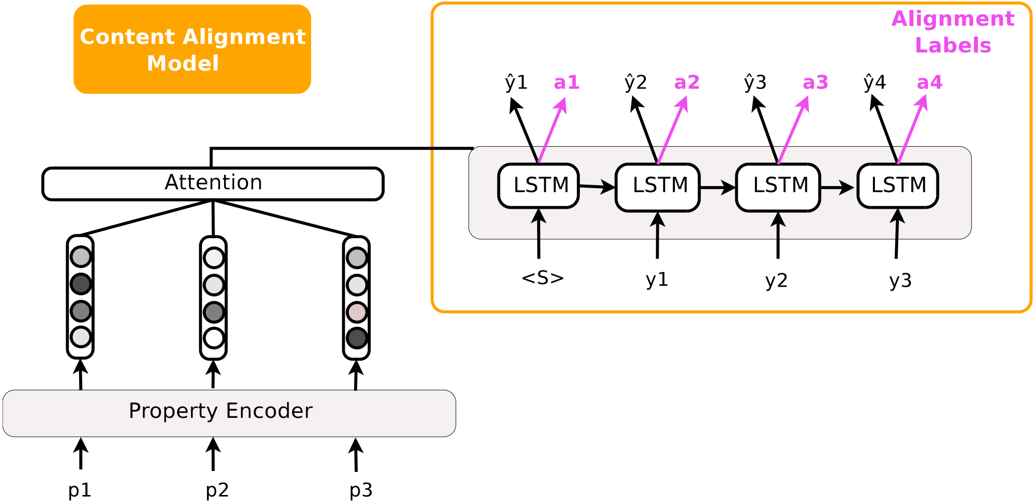
$$S_{\mathcal{P}s} = \sum_{t=1}^{|s|} max_{i \in \{1,\ldots,|\mathcal{P}|\}}\, \mathbf{p}_i \cdot \mathbf{w}_t$$

$$\begin{aligned} \mathcal{L}_{CA} = & max(0, S_{\mathcal{P}s} - S_{\mathcal{P}s'} + 1) \\ & + max(0, S_{\mathcal{P}s} - S_{\mathcal{P}'s} + 1) \end{aligned}$$

- $\mathcal{P}$ is a property set and s a sentence from the text
- each **property vector $\mathbf{p}_i$** is learned by an LSTM encoder
- **word vectors $\mathbf{w}_t$** are hidden states of an LSTM sentence encoder

[Keeler and Rumelhart, 1992, Karpathy and Fei-Fei, 2015]

Multi-Task Learning



Predicting Alignment Labels

- Words from the original text are associated with **Alignment Labels** a indicating whether the word aligns to some property in the input
- **Simultaneously predict** words and alignments, **EDMTL model**

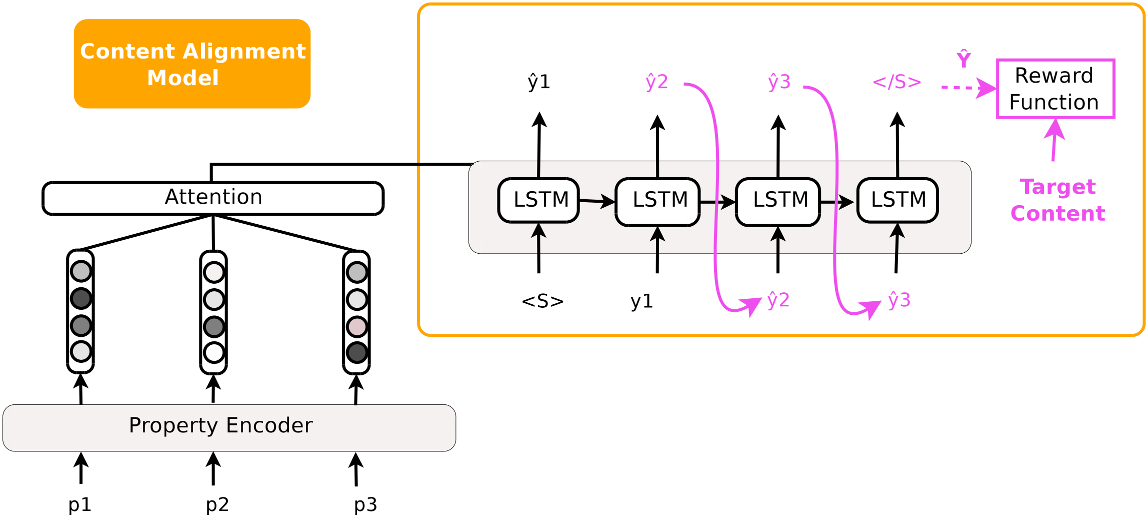
$$\mathcal{L}_{MTL} = \lambda \, \mathcal{L}_{wNLL} + (1 - \lambda) \, \mathcal{L}_{aln}$$

$$\mathcal{L}_{wNLL} = -\sum_{t=1}^{|Y|} log \, P(y_t | y_{1:t-1}, X)$$

$$\mathcal{L}_{aln} = -\sum_{t=1}^{|Y|} log \, P(a_t | y_{1:t-1}, X)$$

[Caruana, 1993]

Reinforcement Learning



Training to Optimise Content

- **Target Content** is set of words from the original text with positive alignments
- Trained to maximise a Content Precision Reward, **EDRL model**

$$r(\hat{Y}) = \gamma^{pr} \, r^{pr}(\hat{Y})$$

r^{pr} is the unigram precision of $\hat{Y}$ and **Target Content**

$$\mathcal{L}_{RL} = -\mathbb{E}_{(\hat{y}_1,\cdots,\hat{y}_{|\hat{Y}|})} \sim P_{\pi}(\cdot|X)[r(\hat{y}_1,\cdots,\hat{y}_{|\hat{Y}|})]$$

[Williams, 1992, Ranzato et al., 2016, Zhang and Lapata, 2017]

Experimental Setup

DBpedia properties and Wikipedia The Free Encyclopedia abstracts about people (**WikiBio** [Lebret et al., 2016])

- Encoder-Decoder baseline, **ED model**
- Hand-crafted templates for 50 most freqent relations, **Templ**

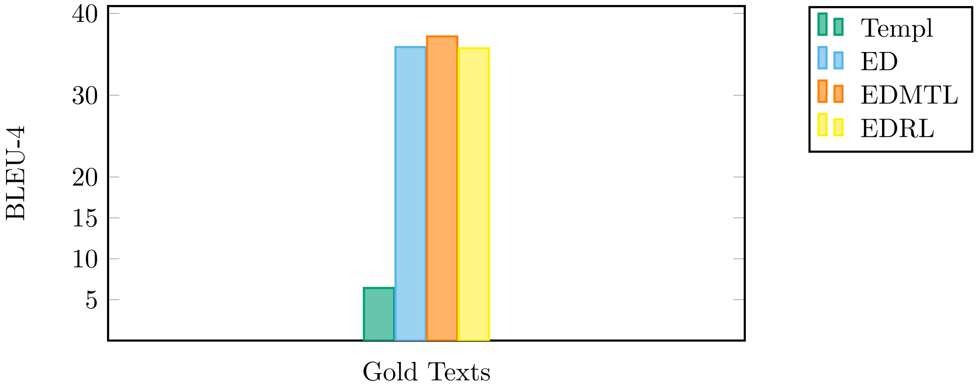
Experimental Setup

DBpedia properties and Wikipedia The Free Encyclopedia abstracts about people (**WikiBio** [Lebret et al., 2016])

- Encoder-Decoder baseline, **ED model**
- Hand-crafted templates for 50 most freqent relations, **Templ**

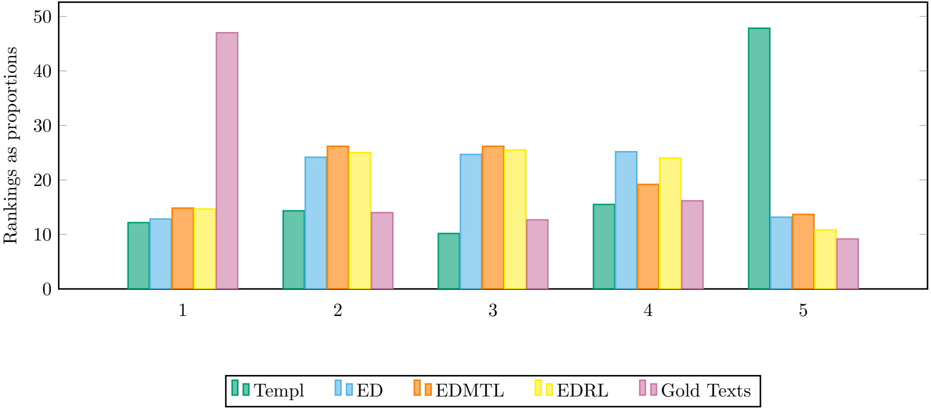
- **Automatic BLEU-4** on gold texts
- **Human ranking** of 4 models and gold texts; AMT judges
- **Ranking criteria:** (1) Is the text faithful to the set of properties?
  (2) Is the text comprehensible and fluent?

Automatic Evaluation Results



Human Evaluation Results

significance at $p < 0.05$



Example Output

Property Set:

name= dorsey burnette, **date**= may 2012, **bot**= blevintron bot, **background**= solo singer, **birth**= december 28 , 1932, **birth place**= memphis, tennessee, **death place**= {los angeles; canoga park, california}, **death**= august 19 , 1979, **associated acts**= the rock and roll trio, **hometown**= memphis, tennessee, **genre**= {rock and roll; rockabilly; country music}, **occupation**= {composer; singer}, **instruments**= {rockabilly bass; vocals; acoustic guitar}, **record labels**= {era records; coral records; smash records; imperial records; capitol records; dot records; reprise records}

Gold Text:

Dorsey Burnette (December 28 , 1932 – August 19 , 1979) was an american early rockabilly singer. He was a member of the Rock and Roll Trio.

Example Output

Property Set:

name= dorsey burnette, **date**= may 2012, **bot**= blevintron bot, **background**= solo singer, **birth**= december 28 , 1932, **birth place**= memphis, tennessee, **death place**= {los angeles; canoga park, california}, **death**= august 19 , 1979, **associated acts**= the rock and roll trio, **hometown**= memphis, tennessee, **genre**= {rock and roll; rockabilly; country music}, **occupation**= {composer; singer}, **instruments**= {rockabilly bass; vocals; acoustic guitar}, **record labels**= {era records; coral records; smash records; imperial records; capitol records; dot records; reprise records}

Gold Text:

Dorsey Burnette (December 28 , 1932 – August 19 , 1979) was an american early rockabilly singer. He was a member of the Rock and Roll Trio.

ED:

Dorsey Burnette (December 28 , 1932 – August 19 , 1979) was an american singer and songwriter. He was a member of the Rock band the band from YEAR to YEAR.

Example Output

Property Set:

name= dorsey burnette, **date**= may 2012, **bot**= blevintron bot, **background**= solo singer, **birth**= december 28 , 1932, **birth place**= memphis, tennessee, **death place**= {los angeles; canoga park, california}, **death**= august 19 , 1979, **associated acts**= the rock and roll trio, **hometown**= memphis, tennessee, **genre**= {rock and roll; rockabilly; country music}, **occupation**= {composer; singer}, **instruments**= {rockabilly bass; vocals; acoustic guitar}, **record labels**= {era records; coral records; smash records; imperial records; capitol records; dot records; reprise records}

Gold Text:

Dorsey Burnette (December 28 , 1932 – August 19 , 1979) was an american early rockabilly singer. He was a member of the Rock and Roll Trio.

EDMTL:

Dorothy Burnette (December 28 , 1932 – August 19 , 1979) was an american country music singer and songwriter. He was a member of the Rock band Roll.

Example Output

Property Set:

name= dorsey burnette, **date**= may 2012, **bot**= blevintron bot, **background**= solo singer, **birth**= december 28 , 1932, **birth place**= memphis, tennessee, **death place**= {los angeles; canoga park, california}, **death**= august 19 , 1979, **associated acts**= the rock and roll trio, **hometown**= memphis, tennessee, **genre**= {rock and roll; rockabilly; country music}, **occupation**= {composer; singer}, **instruments**= {rockabilly bass; vocals; acoustic guitar}, **record labels**= {era records; coral records; smash records; imperial records; capitol records; dot records; reprise records}

Gold Text:

Dorsey Burnette (December 28 , 1932 – August 19 , 1979) was an american early rockabilly singer. He was a member of the Rock and Roll Trio.

EDRL:

Burnette Burnette (December 28 , 1932 – August 19 , 1979) was an american singer and songwriter. He was born in memphis , Tennessee.

Conclusions

- Explicit content modelling benefits generation from loosely related Data-Text pairs
- Novel Content Selection Mechanism based on Multi-Instance Learning
- Two frameworks, MTL and RL, to instill content requirements in the training process

Future work

Modelling target document structure (e.g. phrases or sentences)

Thank you !

Questions ?

References I

Caruana, R. (1993).
Multitask learning: A knowledge-based source of inductive bias.
In *Proceedings of the 10th International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Ghader, H. and Monz, C. (2017).
What does attention in neural machine translation pay attention to?
In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Karpathy, A. and Fei-Fei, L. (2015).
Deep visual-semantic alignments for generating image descriptions.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Keeler, J. and Rumelhart, D. E. (1992).
A self-organizing integrated segmentation and recognition neural net.
In *Advances in Neural Information Processing Systems 5*, pages 496–503. Curran Associates, Inc.

Lebret, R., Grangier, D., and Auli, M. (2016).
Neural text generation from structured data with application to the biography domain.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016).
Sequence level training with recurrent neural networks.
In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico.

References II

Williams, R. J. (1992).
Simple statistical gradient-following algorithms for connectionist reinforcement learning.
Machine learning, 8(3-4):229–256.

Wiseman, S., Shieber, S., and Rush, A. (2017).
Challenges in data-to-document generation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2253, Copenhagen, Denmark.

Zhang, X. and Lapata, M. (2017).
Sentence simplification with deep reinforcement learning.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark.

Experimental Setup

Based on **WikiBio dataset**, DBpedia properties and Wikipedia The Free Encyclopedia abstracts ([Lebret et al., 2016])
train/devel/test: 165,324 / 25,399 / 23,162

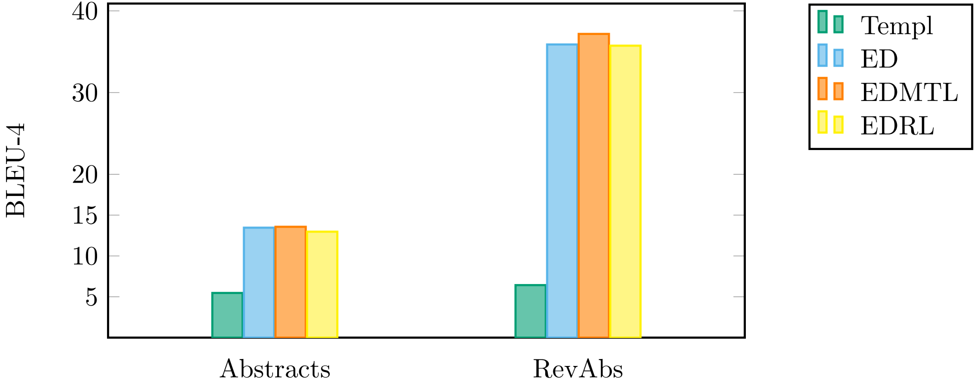
- Content Aligner optimised on development set
  - 2 annotators manually aligned 132 (Data, Sentence) pairs
  - select model with best word alignment f-score .36
  - inter-annotator agreement f-score .72
- Encoder-Decoder baseline, **ED model**
- Hand-crafted templates for 50 most freq. relations, **Templ**

Evaluation

- **Gold text collection**, **Revised Abstracts (RevAbs)**
  AMT annotators, 200 test examples, 3 revisions
- **Automatic BLEU-4** on original and revised abstracts
- **Human ranking** of 4 models and RevAbs; AMT judges, 200 revised examples, 3 judgments
- **Ranking criteria:** (1) Is the text faithful to the set of properties?
  (2) Is the text comprehensible and fluent?

Gold text collection

Automatic Evaluation Results


BLEU-4
Abstracts
RevAbs
Templ
ED
EDMTL
EDRL