Université de Lorraine - UFR Math-Info
Master SCA – M1
2017-2018

Fiche de projet tutoré / Project form

Fusion de phrases / Sentence Fusion

Encadrement / Supervisors
- encadrant principal / main supervisor : Claire Gardent, CNRS/LORIA, Nancy (France)
- autres encadrants / other supervisors : Shashi Narayan, Informatics, University of Edinburgh (UK)

Description / Description

Sentence fusion is a text-to-text generation task which takes related sentences as input and merges these into a single output sentence.

Sentence fusion is used in the context of multi-document summarization, where the input sentences typically come from multiple documents describing the same event. It is used in the context of abstractive summarisation where simple sentences may need to be fused into a single sentence to allow for a more fluent, more readable abstract. Finally, it can be useful for question answering to generate more complete answers out of several complementary sentences. Many current QA systems use various parallel answer-finding strategies, each of which may produce an N-best list of answers (e.g., [maybury 2004]). For instance, in response to the question (1a), the system might extract the answer sentences in (1b-c):

(1a) *What causes RSI?*
(1b) *RSI can be caused by repeating the same sequence of movements many times an hour or day.*
(1c) *RSI is generally caused by a mixture of poor ergonomics, stress and poor posture.*

(2) *RSI can be caused by a mixture of poor ergonomics, stress, poor posture and by repeating the same sequence of movements many times an hour or day.*

These two incomplete answers might be fused into a more complete answer such as (3).

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

Previous work on sentence fusion [Barzilay, 2003] typically consists of a pipeline integrating

three components: alignment, fusion and generation. First, the dependency structures of the input sentences are aligned to find the common information in both sentences. On the basis of this alignment, the common information is framed into an fusion tree capturing the shared information, which is then realized in natural language by generating all traversals of the fusion tree and scoring their probability using an n-gram language model.

The aim of this tutored project is to investigate deep learning approaches to sentence fusion and to evaluate them using the Split-and-Rephrase dataset [Narayan et al., 2017].

The split-and-rephrase dataset consists of 1,100,166 pairs of the form < (MC, TC), {(M1,T1), … ,(Mn , Tn)} > where TC is a complex sentence and T1 … Tn is a sequence of texts with semantics M1, … , Mn expressing the same content MC as TC. For instance, the complex sentence (3a) is associated with the meaning representation (3b) and with three simpler sentences (4a-c) whose joint meaning is (3b).

(3a) *John Clancy is a labour politican who leads Birmingham, where architect John Madin, who designed 103 Colmore Row, was born*.

(3b) { Birmingham | leaderName | John_Clancy_(Labour_politician), John_Madin | birthPlace | Birmingham, 103_Colmore_Row | architect | John_Madin }

(4a) *Labour politician, John Clancy is the leader of Birmingham.*
{ Birmingham | leaderName | John_Clancy_(Labour_politician) }

(4b) *John Madin was born in Birmingham.*
{ John_Madin | birthPlace | Birmingham }

(4c) *He was the architect of 103 Colmore Row.*
{ 103_Colmore_Row | architect | John_Madin }

Using the split-and-rephrase dataset, the aim of the project is to explore and compare the following neural models:

BL: An attention-based sequence-to-sequence model augmented with coverage [Tu et al. 2016] where the input sentences S1, .. Sn are concatenated and encoded using a single encoder[1].

Multi: A Multi-encoder-decoder models in which each each input sentence Si is encoded using a separate encoder [Zoph, 2016][2]. This will require either modifying Zoph's model so that it can handle more than two encoders (Zoph's model assumes that there are only two sources) or taking a simpler approach by splitting S1, … ,Sn into two chunks and using the existing model.

Sem: Using the input sentence semantics in addition to the input sentences to enrich the encoder with semantic information. This can be done eg by concatenating the input

1 https://github.com/tuzhaopeng/NMT-Coverage
2 https://github.com/isi-nlp/Zoph_RNN

sentences with their semantics and using a single encoder or by using a multi-encoder-decoder model as described in [Zoph, 2016].

More generally, the aim is to compare a semantic-based approach integrating semantic knowledge about shared entity and predicates into a sequenc-to-sequence model with uni and multi-source sequence to sequence models which only takes into account the input sentences.

Livrables et échéancier / Deliverable and schedule

- November-December: Reading on Sentence Fusion and Deep Learning
    - Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In Proceedings of NAACL-HLT.
    - Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP.
    - R. Barzilay, K. McKeown, and M. Elhaded. Information fusion in the context of multi-document summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), Maryland, 1999.
    - R. Barzilay. Information Fusion for Multi-document Summarization. Ph.D. Thesis, Columbia University, 2003.
    - Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.
    - Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 76–85, Berlin, Germany, August 2016. Association for Computational Linguistics

    - Coursera course: https://www.coursera.org/specializations/deep-learning (you can take it for free if you do not need the certificate)

- 15 January - 15 February (Baseline)
    - Choose deep learning framework
    - Format data for deep learning model
    - Implement basic sequence to sequence model
    - Implement evaluation code
    - Get results for Baseline

- 15 February - 15 March (Multi-encoder)
    - Study Zoph's multi-source encoder
    - Implement multi-source encoder for sentence fusion
    - - Get results for Multi-encoder approach

- 15 March - 30 April (Integrating Semantics)
    - Decide how to integrate semantics into the model (multi-source or simple

concatenation)
    - Implement model with semantics
    - Get results

- 1 May - 31 May
    - Write Report and prepare slides for presentation

Bibliographie / References

- [Badhanau et al., 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473
- [Barzilay et al., 2003] R. Barzilay, K. McKeown, and M. Elhaded. Information fusion in the context of multi-document summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), Maryland, 1999
- [Barzilay, 2003] R. Barzilay. Information Fusion for Multi-document Summarization. Ph.D. Thesis, Columbia University, 2003
- [Barzilay et al., 2004] M. Maybury. New Directions in Question Answering. AAAI Press, 2004.
- [Narayan et al., 2017] S. Narayan, C. Gardent, S. Cohen and A. Shimorina. Split and Rephrase. Proceedings of EMNLP 2017.
- [Luong et al., 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP.
- [Tu et al., 2003] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), , pages 76–85, Berlin, Germany, August 2016.
- [Zoph et al., 2016] Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In Proceedings of NAACL-HLT.