Embedding-to-embedding method based on autoencoder for solving sentence analogies

Weihao MAO

Graduate School of Information, Production, and Systems, Waseda University
EBMT/NLP Lab

July 17th, 2023

Table of contents

Background

Analogy is a relationship between four objects A, B, C, and D. It is read as "A is to B as C is to D," and is written as $A : B :: C : D$.

$$\textit{please tell us about it.} : \textit{please tell me about it.} :: \textit{what do you expect us to do?} : \textit{what do you expect me to do?}$$

$$\textit{he never saw his brother again.} : \textit{he never saw his sister again.} :: \textit{he never saw his father again.} : \textit{he never saw his mother again.}$$

Background

- Analogy is a conformity of ratios between objects of the same kind
  - *ratio*
    $A : B :: C : D$
  - *conformity*
    $A : B :: C : D$

- Analogy solving
  - Find the solution to the analogical equation:
    $$A : B :: C : x \quad \Rightarrow \quad x = ?$$
  - Using predefined formula in embedding space (3CosAdd):
    $e_B - e_A = e_D - e_C \Rightarrow e_D = e_C + e_B - e_A$

Background

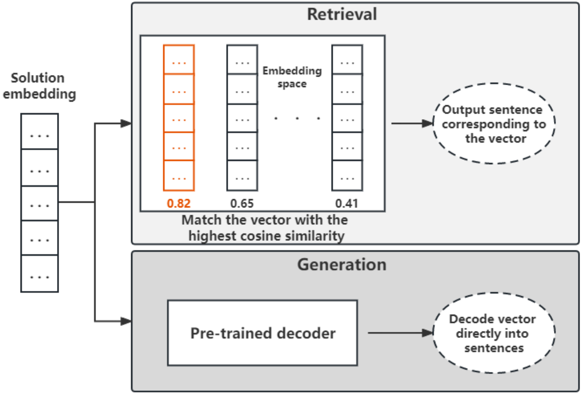


Figure: Two methods of obtaining a corresponding sentence from a given embedding.

Previous work

- Vec2Seq model proposed by (Wang and Lepage, 2020)
  - Pre-training a single-layer LSTM network as a decoder to transform the sentence vectors into corresponding sentences.
  - Designing a linear fully-connected neural network responsible for generating embeddings of the solutions of the analogy equation.

Previous work

- Vec2Seq model proposed by (Wang and Lepage, 2020)
  - Pre-trained a single-layer LSTM network as a decoder to transform sentence vectors into corresponding sentences.
  - Linear fully-connected neural network responsible for generating embeddings for the solution of an analogical equation.

Previous work

- Wang and Lepage (2020) proposed to design a small RNN-based decoder to transform sentence vectors into sentences, (the word embedding sequence is obtained from fastText [1]).

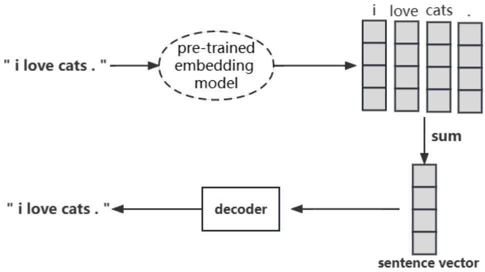


Figure: Schematic diagram of the decoding process.

[1] https://fasttext.cc/

Previous work

- Wang and Lepage (2020) experimented with three compositional methods on the known vectors as inputs to the linear regression network (LinearFCN).

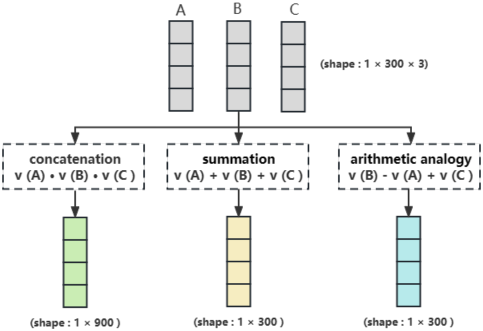


Figure: Three compositional methods on the known vectors.

Previous work

- **Limitations** of the existing model:
  - Tested on English corpora only. How about other languages?
  - Dense distribution of the sentence in the vector space. (decoders are sensitive to noise)
  - Prone to generate repetitions of words as in :
    - *i read the book day in a day.*
    - *my of my feet are taller than.*
    - *are you having having any that doing?*
  - 3CosAdd ($e_B - e_A + e_C$) assumes linear properties of the embedding space.

Previous work

- Chan et al. (2022) proposed a character-based word autoencoder to solve word morphological analogies.

- Marquer et al. (2022) proposed an analogy retrieval models (ANNr) to find the solutions of analogical equations in word vector spaces.

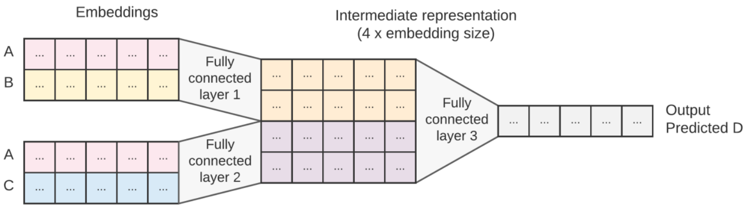


Figure: ANNr model architecture. Figure copied from (Marquer et al., 2022).

Goals

Inspired by the work of (Wang and Lepage, 2020), we design a generation-based method based on an autoencoder to address sentence analogies.

Contributions

- We have designed a more stable autoencoder architecture to reconstruct the solutions of analogical equations from the embedding space back into sentences.

Contributions

- We propose a novel model that does not rely on predefined formulas to solve analogical equations in the sentence embedding space.

Contributions

- We have achieved promising results in the generation-based approach and, to some extent, demonstrated that the effectiveness of the 3CosAdd formula decreases for longer sentences.

Vector composition method

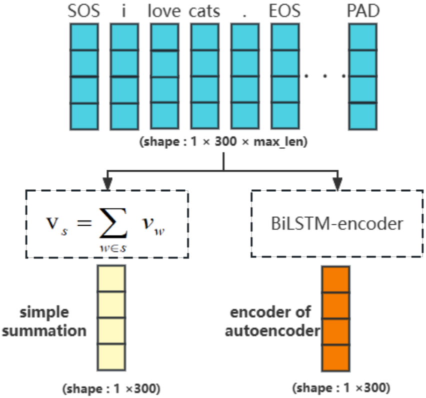


Figure: Method to convert a sequence of word embedding representations into a sentence representation.

sentence embedding method

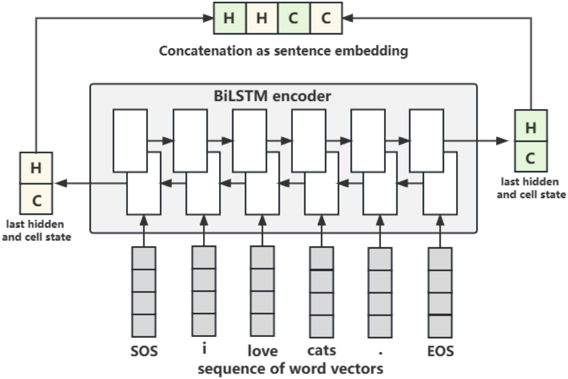


Figure: Sketch of the encoder. Figure copied from (Chan et al., 2022). The output is a sentence embedding.

Pre-training autoencoder

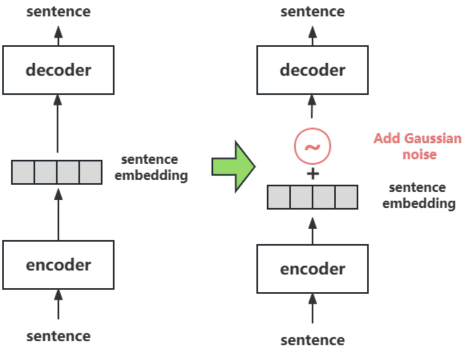


Figure: Structure of proposed autoencoder.

Offset network structure for analogy

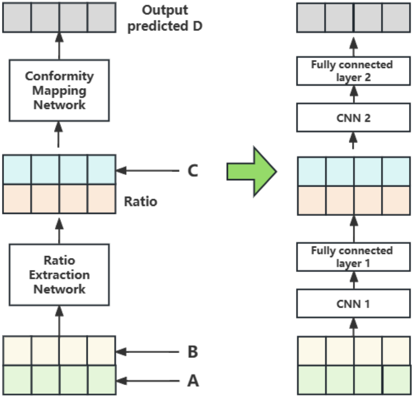
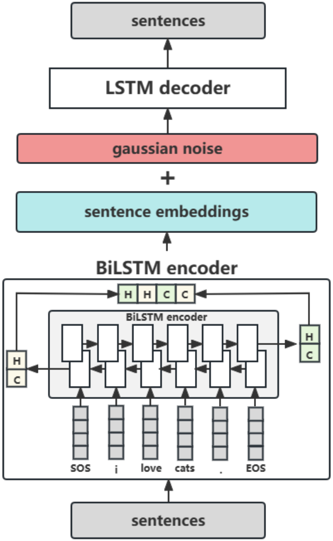
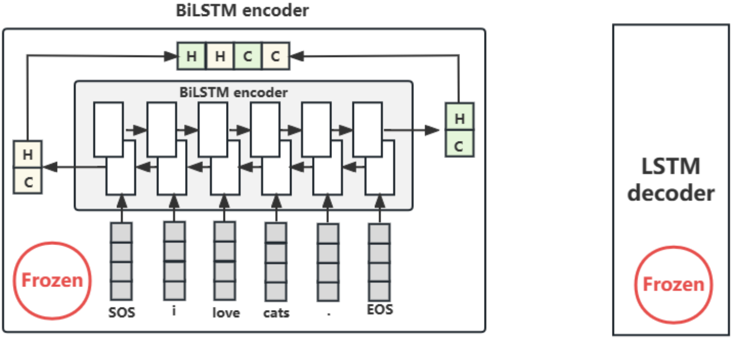


Figure: Offset network structure for analogy

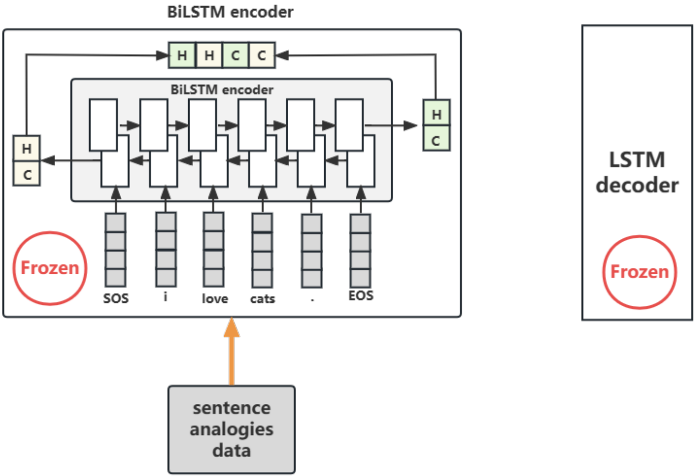
General architecture of embedding-to-embedding methods



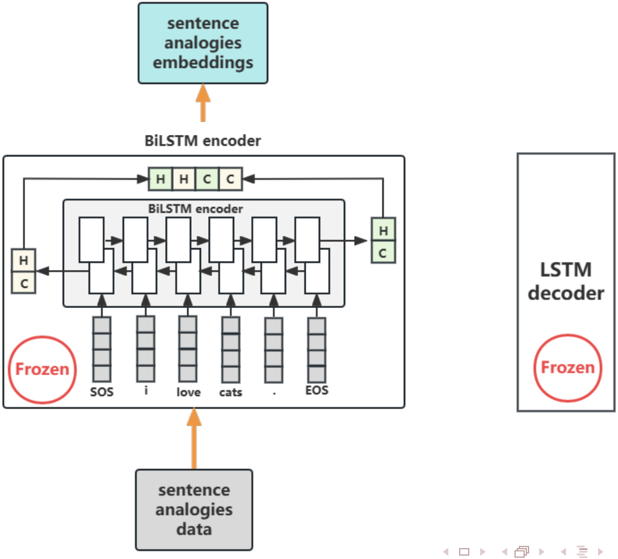
General architecture of embedding-to-embedding methods



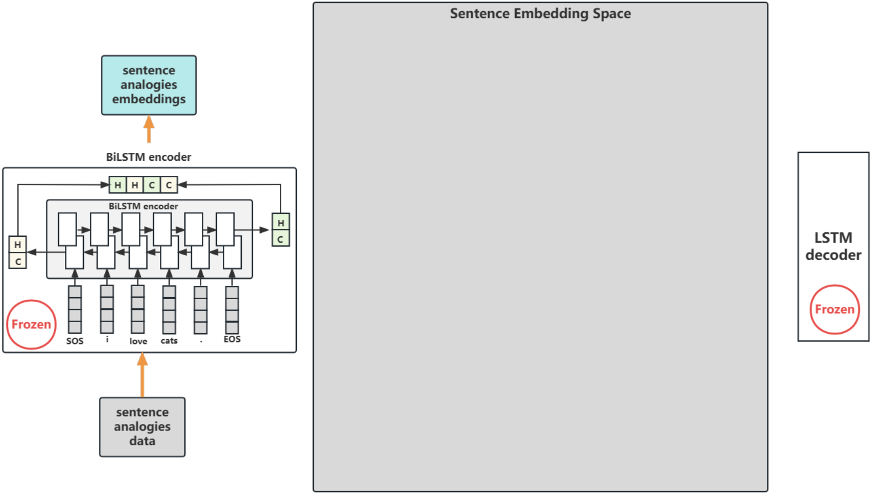
General architecture of embedding-to-embedding methods



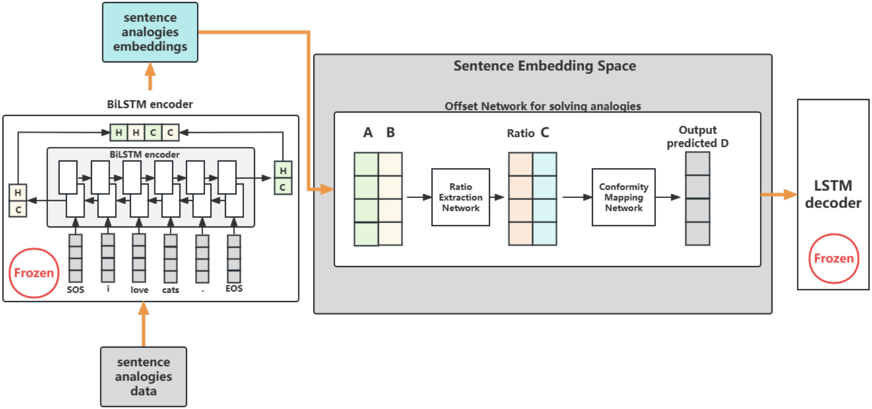
General architecture of embedding-to-embedding methods



General architecture of embedding-to-embedding methods



General architecture of embedding-to-embedding methods



Evaluation metrics

- BLEU (Papineni et al., 2002): evaluates the similarity of two sentences. Score between 0 and 100. The higher, the more similar the two sentences.
- Accuracy : ratio of exact matches to the total number of samples tested.
- Levenshtein distance : minimum number of edit operations required to convert one string into another one.

Decoding sentence embedding

- We extracted 85,000 sentences randomly from the Tatoeba [2] corpus for three languages.

data	**Number of**		
	sentences	**words/sent.**	**character/sent.**
		English	
Taining	70,000	6.6±1.7	27.6±8.4
Validation	8,750	6.5±1.7	27.3±8.2
Testing	8,750	6.5±1.7	27.3±8.2
		French	
Taining	70,000	8.7±4.9	40.0±24.9
Validation	8,750	8.7±5.0	40.0±25.3
Testing	8,750	8.7±4.9	40.0±25.0
		German	
Taining	70,000	8.7±5.0	44.4±28.0
Validation	8,750	8.7±5.0	44.6±28.3
Testing	8,750	8.6±4.9	44.3±28.0

[2] https://tatoeba.org

Decoding sentence embedding

Input Vector composition method	Model size (Mb)	BLEU	Accuracy (%)	Levenshtein distance in words	 in cahrs
		English			
simple summation	3.8	73.5 ± 0.7	62.2	1.0	4.3
encoder of autoencoder	4.4	$\mathbf{93.5 \pm 0.4}$	**91.1**	**0.1**	**0.8**
		French			
simple summation	8.8	42.2 ± 0.9	25.9	3.3	15.2
encoder of autoencoder	11.6	$\mathbf{68.5 \pm 1.1}$	**56.3**	**1.4**	**9.2**
		German			
simple summation	11.0	35.4 ± 0.8	24.0	3.7	19.1
encoder of autoencoder	13.8	$\mathbf{60.6 \pm 1.0}$	**54.0**	**2.4**	**12.5**

Table: Performance of the different models on three languages.

- In terms of accuracy, using sentence embeddings generated by the encoder of the autoencoder outperforms the simple summation approach by nearly 30% in all three languages.

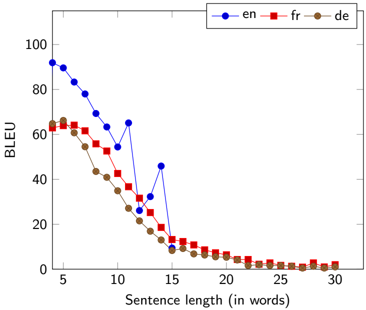
Decoding sentence embedding

Input Vector composition method	Model size (Mb)	BLEU	Accuracy (%)	Levenshtein distance in words	 in cahrs
		English			
simple summation	3.8	73.5±0.7	62.2	1.0	4.3
encoder of autoencoder	4.4	**93.5±0.4**	**91.1**	**0.1**	**0.8**
		French			
simple summation	8.8	42.2±0.9	25.9	3.3	15.2
encoder of autoencoder	11.6	**68.5±1.1**	**56.3**	**1.4**	**9.2**
		German			
simple summation	11.0	35.4±0.8	24.0	3.7	19.1
encoder of autoencoder	13.8	**60.6±1.0**	**54.0**	**2.4**	**12.5**

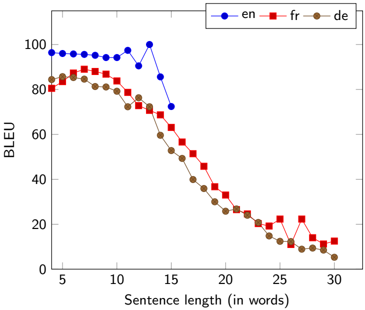
Table: Performance of the different models on three languages.

- For French and German, which have longer sentence lengths and vocabulary sizes, two to three times larger than that of English, the decoding performance decreases slightly.

Decoding sentence embeddings



(a) simple summation



(b) encoder of autoencoder

Figure: Performance of models on sentences with different lengths in three different languages

Solving sentence analogies

- Semantico-formal analogy set (Lepage, 2019), which contains 5,607 sentence analogies in English.

Data	**Number of**			
	analogies	**sentences**	**words/sent.**	**character/sent.**
Training	3,364	3,185	7.1±1.2	27.0±5.7
Validation	1,122	1,769	7.1±1.1	26.6±5.6
Testing	1,121	1,667	7.0±1.1	26.3±5.6
Total	5,607			

Table: Semantico-formal analogy set from Tatoeba

Solving sentence analogies

- Experimental settings:
  - Decoder model: single-layer LSTM

Experiment name	Composition method	Model for solving analogies
sum-FCN	simple summation	LinearFCN
enc-FCN	encoder of autoencoder	LinearFCN
enc-Offset	encoder of autoencoder	Offset nerwork
enc-ANNr	encoder of autoencoder	ANNr model

Table: Experiment names and structures

Solving sentence analogies

Experiment name	BLEU	Accuracy (%)	Levenshtein distance in words	in cahrs
sum-FCN	91.0±1.3	82.5	0.3	1.3
enc-FCN	**92.0**±**1.3**	**84.6**	**0.2**	**1.0**
enc-Offset	89.1±1.6	78.2	0.4	1.8
enc-ANNr	80.3±2.2	73.1	0.6	2.7

Table: Performance of the different models on semantico-formal analogy set.

From the perspective of

- obtaining sentence embeddings:
  encoder of autoencoder > simple summation
- solving analogies: FCN > Offset > ANNr

Solving sentence analogies

- Formal analogy set: We extracted about 10,000 sentence formal analogies from Tatoeba in three languages using the Nlg package (Fam and Lepage, 2018).

data	**Number of**			
	analogies	**sentences**	**words/sent.**	**character/sent.**
		English		
Taining	8,000	18,515	5.7±1.7	22.7±8.1
Validation	1,000	3,639	5.5±1.7	22.1±7.9
Testing	1,000	3,666	5.6±1.7	22.2±8.1
		French		
Taining	8,000	14,803	7.0±2.7	29.7±12.3
Validation	1,000	3,482	7.0±2.9	30.1±12.8
Testing	1,000	3,478	7.0±3.0	30.1±13.4
		German		
Taining	8,000	12,729	6.1±2.0	29.2±10.9
Validation	1,000	3,226	6.1±2.0	28.6±10.5
Testing	1,000	3,232	6.1±1.9	28.5±10.4

Solving sentence analogies

Experiment name	BLEU	Accuracy (%)	Levenshtein distance in words	Levenshtein distance in chars
English				
sum-FCN	**91.0±1.8**	**90.8**	**0.3**	**1.0**
enc-FCN	89.6±2.1	88.6	0.4	1.3
enc-Offset	80.6±2.2	76.1	0.7	2.4
French				
sum-FCN	64.3±2.6	46.2	1.7	7.5
enc-FCN	**71.8±2.2**	**57.9**	**1.4**	**5.4**
enc-Offset	70.6±2.2	56.1	1.5	6.2
German				
sum-FCN	73.6±2.3	62.3	0.9	3.8
enc-FCN	**84.1±2.1**	**78.8**	**0.6**	**2.6**
enc-Offset	77.0±2.3	69.0	0.8	3.6

Table: Performance of the different models on formal analogy set in three languages.

- When sentences are short, the FCN network performs better than the Offset network.

Solving sentence analogies

Experiment name	BLEU	Accuracy (%)	Levenshtein distance in words	Levenshtein distance in chars
		English		
sum-FCN	**91.0±1.8**	**90.8**	**0.3**	**1.0**
enc-FCN	89.6±2.1	88.6	0.4	1.3
enc-Offset	80.6±2.2	76.1	0.7	2.4
		French		
sum-FCN	64.3±2.6	46.2	1.7	7.5
enc-FCN	**71.8±2.2**	**57.9**	**1.4**	**5.4**
enc-Offset	70.6±2.2	56.1	1.5	6.2
		German		
sum-FCN	73.6±2.3	62.3	0.9	3.8
enc-FCN	**84.1±2.1**	**78.8**	**0.6**	**2.6**
enc-Offset	77.0±2.3	69.0	0.8	3.6

Table: Performance of the different models on formal analogy set in three languages.

- The longer the average length of sentences, the worse the performance: French $<$ German $<$ English

Performance on longer sentences

- The FCN network (in conjunction with the formula from 3CosAdd to process embeddings as inputs) and the Offset network are close in performance when the sentences are long.
- 3CosAdd relies on a fixed formula and cannot learn from the dataset. It is effective for simple short sentence analogies but may not perform well for longer sentences.

Performance on longer sentences

- The FCN network (in conjunction with the formula from 3CosAdd to process embeddings as inputs) and the Offset network are close in performance when the sentences are long.
- 3CosAdd relies on a fixed formula and cannot learn from the dataset. It is effective for simple short sentence analogies but may not perform well for longer sentences.

Performance on longer sentences

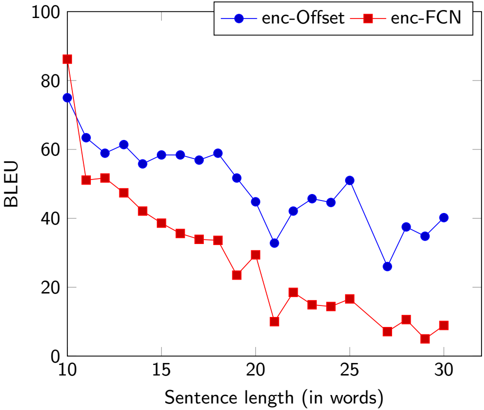


Figure: Performance of models on sentences with different lengths in French.

Conclusion

- We proposed an auto-encoder architecture that internally removes noise (see Page 16) to generate sentence embeddings and reconstruct sentences, achieving high accuracy in decoding sentence embeddings.

Conclusion

- We devised an embedding-to-embedding method and a model (see Page 17) that learns analogies from datasets in the sentence embedding space without relying on any predefined formula.

Conclusion

- Our experiments demonstrate that this approach performs better than a model relying on the 3CosAdd formula, especially in cases where the sentences are longer.

Future work

- Explore more advanced encoder-decoder architectures that are better suited for decoding longer sentences.
- Generating more meaningful sentence embeddings specifically designed for analogies.

Thank you for your attention.

Sample of analogous data set

you 're my friend.	:	*you 're an angel.*	::	*she 's my friend.*	:	*she 's an angel.*
tom is outgoing.	:	*tom had jeans on.*	::	*he is outgoing.*	:	*he had jeans on.*
french is his mother tongue.	:	*it 's his first day at school.*	::	*french is her mother tongue.*	:	*it 's her first day at school.*

Kevin Chan, Shane Peter Kaszefski-Yaschuk, Camille Saran, Esteban Marquer, and Miguel Couceiro. Solving Morphological Analogies Through Generation. In *IJCAI-ECAI Workshop on the Interactions between Analogical Reasoning and Machine Learning (IARML@IJCAI-ECAI 2022)*, volume 3174 of *Proceedings of the IJCAI-ECAI Workshop on the Interactions between Analogical Reasoning and Machine Learning (IARML@IJCAI-ECAI 2022)*, pages 29–39, Vienna, Austria, July 2022. Miguel Couceiro and Pierre-Alexandre Murena. URL `https://hal.inria.fr/hal-03674913`.

Rashel Fam and Yves Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1171`.

Yves Lepage. Semantico-formal resolution of analogies between sentences. In *the 9th Language and Technology Conference (LTC 2019)*, page 57–61, May 2019.

Esteban Marquer, Safa Alsaidi, Amandine Decker, Pierre-Alexandre Murena, and Miguel Couceiro. A deep learning approach to solving morphological analogies. In Mark T. Keane and Nirmalie Wiratunga, editors, *Case-Based Reasoning Research and Development*, pages 159–174, Cham, 2022. Springer International Publishing. ISBN 978-3-031-14923-8.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Liyan Wang and Yves Lepage. Vector-to-sequence models for sentence analogies. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 441–446, 10 2020. doi: 10.1109/ICACSIS51025.2020.9263191.