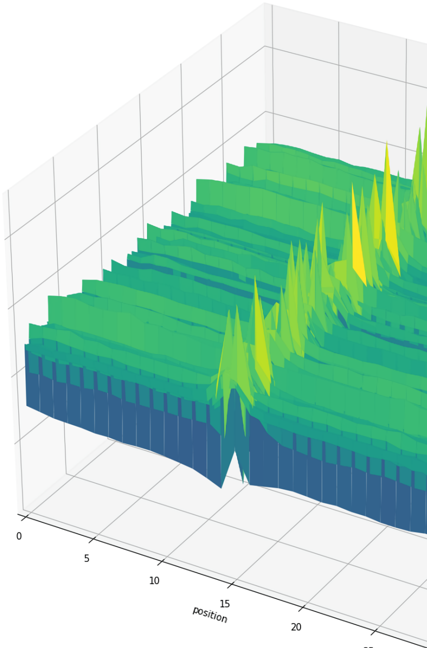
When FastText Pays Attention

Efficient Estimation of Word Representations
using Constrained Positional Weighting

Vít Novotný, Michal Štefánik, Eniafe F. Ayetiran, Petr Sojka
{witiko,stefanik.m,ayetiran}@mail.muni.cz, sojka@fi.muni.cz

Faculty of Informatics, Masaryk University

April 20, 2021



1. Introduction I
Venue

- We submitted our paper to the TACL[1] journal (ISSN: 2307-387X, https://transacl.org/).
- TACL is only 7 y. o. with no impact factor, but the 4th best h5-index at Google Scholar:

Categories > Engineering & Computer Science > **Computational Linguistics**

	Publication	h5-index	h5-median
1.	Meeting of the Association for Computational Linguistics (ACL)	135	220
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)	112	197
3.	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	90	148
4.	Transactions of the Association for Computational Linguistics	53	112

- Accepted papers (7–10 pages) are eligible for presentation at ACL conferences.
- The paper has not been accepted, but we've published it, and submitted it to arXiv.

[1] Transactions of the Association for Computational Linguistics

1. Introduction II

Background

- *Shallow log-bilinear language models (LBLs)* [1, 2, 3] provide useful word embeddings:
  1. word sense induction [4],
  2. text classification [5],
  3. question answering [6],
  4. evaluation of lexical definitions [7, 8, 9].
- *Deep attention-based models (Transformers)* [10] have redefined SOTA for 11 tasks.
- Theoretical results suggest that *anything LBLs can do, Transformers can do better*:
  - LBLs treat text as a bag of words, equivalent to linear SVMs, factorize PMI matrix [11].
  - Transformers are universal approximators [12] and (theoretically [13]) Turing-complete [14].
- Empirical results suggest that LBLs capture information missing from Transformers:
  - LBLs are *more robust on misspellings* (86% $\rightarrow$ 80%) than Transformers (93% $\rightarrow$ 70%). [15]
  - Combining Transformers + LBLs *improves dependency parsing* (61% $\rightarrow$ 77%). [16, Table 3].
- In my talk, I will:
  1. describe the evolution of the dense and sparse attention mechanisms,
  2. describe the positional LBL of Mikolov et al. [17] and relate it to dense attention,
  3. propose our constrained positional LBL and relate it to sparse attention, and
  4. evaluate the positional LBLs on LM and three novel qualitative evaluation measures.

1. Introduction III

Thinking Fast and Slow

- *The competence hierarchy* of Broadwell [18] describes the stages of human learning:

Conscious incompetence The individual lacks a skill and recognizes the deficit.
Conscious competence The individual has a skill, but execution requires concentration.
Unconscious competence The individual has a skill and they can perform it with ease.

- Kahneman [19] descibes the mind of the individual as the interplay of two systems:
  1. the fast and intuitive *System 1*, and
  2. the slow and logical *System 2*.
- NLU requires both common sense (System 1) and logical reasoning (System 2). [20]
- Peters et al. [21] show ($N = 100$) that systems 1 and 2 are mutually supportive:

Conscious incompetence System 2 is bored and directs the individual towards a new skill.
Conscious competence The individual has a skill, but execution requires system 2.
Unconscious competence The individual has a skill and they can perform it with system 1.

- Assume that System 1 = LBLs and System 2 = Transformers. Then:
  1. NLU requires both LBLs and Transformers. [16]
  2. LBLs can adopt the skills of Transformers.
  3. Improving LBLs will improve energy-efficiency and Transformers.

2. Models I

2.1. (Dense) Attention I

- Early neural machine translation (NMT) used *encoder-decoder* models [22, 23]:
  - An encoder reads and encodes a *source sentence* into a fixed-length *context vector*.
  - A decoder produces a translated *target sentence* from the context vector.
- Due to the fixed length of the context vector, NMT would deteriorate for longer sequences. [24]

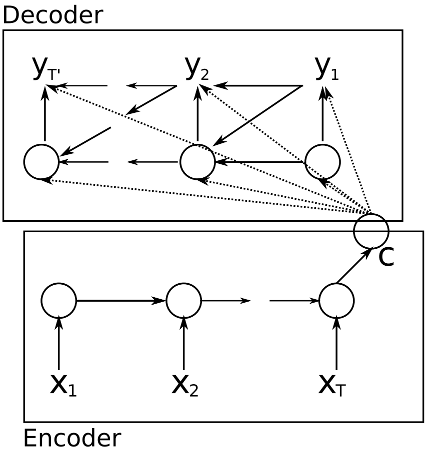
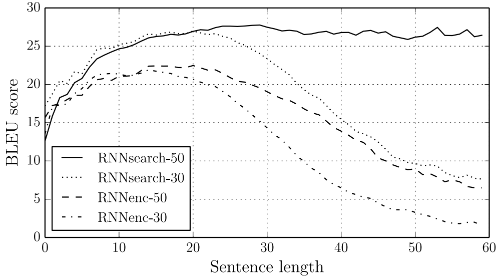


Figure: Cho et al. [23, Figure 1]

2. Models II

2.1. (Dense) Attention II

- Bahdanau et al. [25] equipped the decoder with *attention*:
  - The decoder constructs a different context vector for each target word.
  - The context vector is a weighted average of the encoder's states.



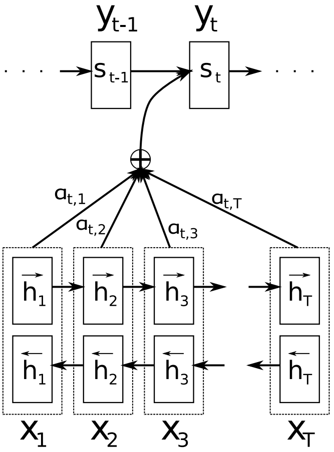
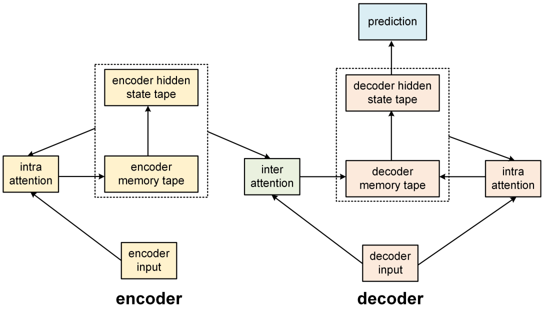


Figure: Bahdanau et al. [25, figures 1 and 2]

2. Models III

2.1. (Dense) Attention III

- Cheng et al. [26] moved attention directly into the LSTM cells.



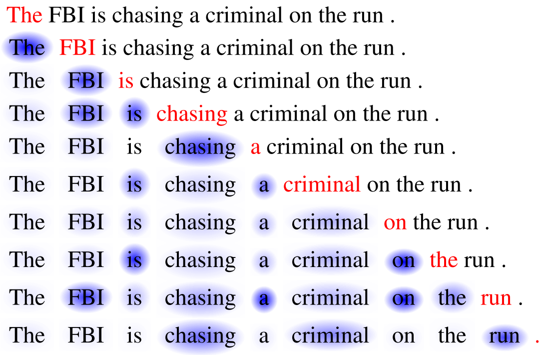


Figure: Bahdanau et al. [25, figures 1 and 3b]

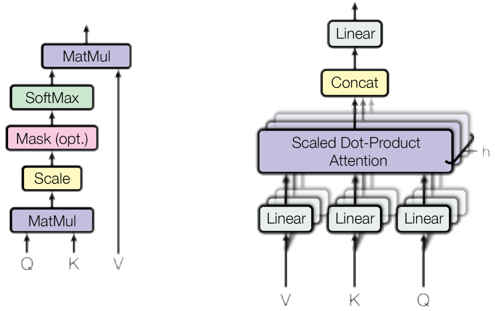
- Attention acts as a random-access memory mechanism for the RNN.

2. Models IV

2.1. (Dense) Attention IV

- Vaswani et al. [27] proposed *Transformers*:
  - Transformers replace recurrence by the vertical stacking of attention.

Scaled Dot-Product Attention Multi-Head Attention



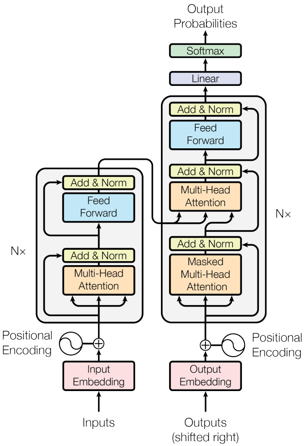
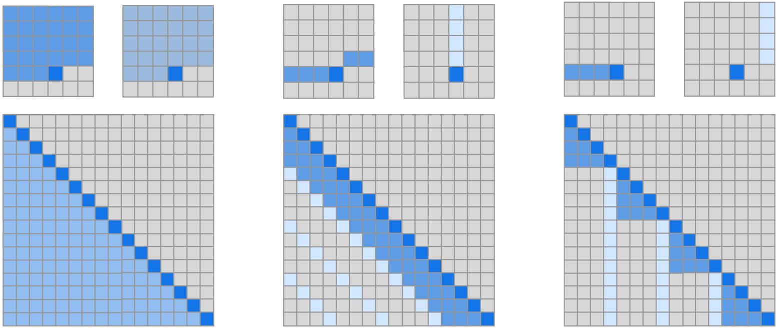


Figure: Vaswani et al. [27, figures 1 and 2]

2. Models V

2.1. (Sparse) Attention V

- Dense attention learns weights for *all pairs* of source and target words.
- Therefore, dense attention is in SPACE(n^2), where n is the size of the sequence.
- This limits the size of n and makes it *impossible to embed [28] long documents*.
- Child et al. [29] proposed to *sparsify the attention weights* to $\mathcal{O}(n\sqrt[p]{n})$ elements:

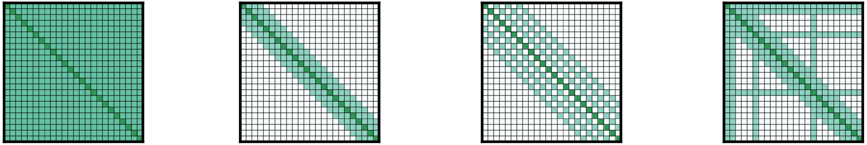


(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

2. Models VI

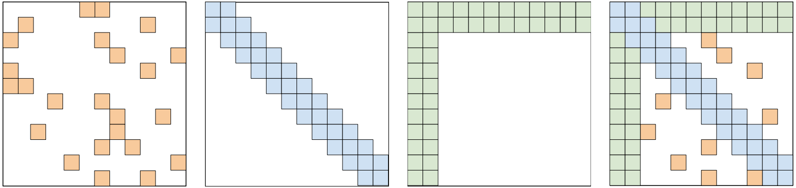
2.1. (Sparse) Attention VI

- Beltagy et al. [30] proposed other sparsification techniques, making attention SPACE(n):



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

- Zaheer et al. [31] is in SPACE(n), universal approximator, and Turing-complete:

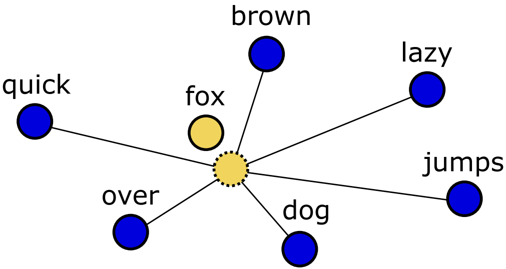


(a) Random attention (b) Window attention (c) Global Attention (d) BigBird

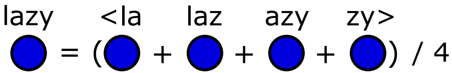
2. Models VII

2.2. Log-Bilinear Language Models I

General model [1, 32] Predicts a masked word from the mean context word vector:



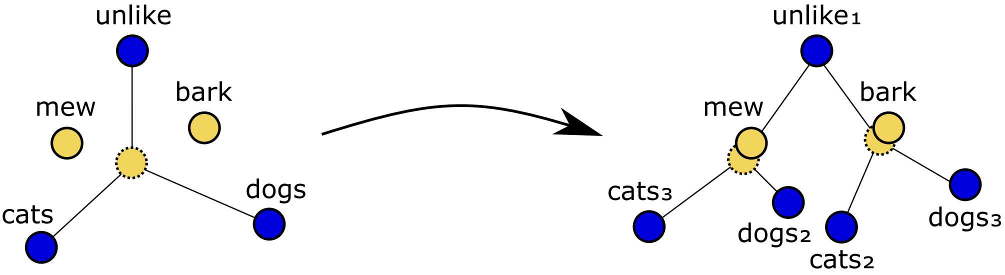
Subword model [3] Makes words share weights by modeling subword units:

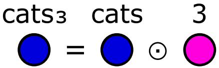


2. Models VIII

2.2. Log-Bilinear Language Models II

Positional model [17] Makes words independent on their position in the sequence:



- Position-independence is achieved through factorization: 

- More than doubles the training time compared to the subword model. [33, Table 3]
- Similarly to dense attention, the model relates different positions in the sequence.

2. Models IX

2.2. Log-Bilinear Language Models III

Constrained positional model Models contextual and fixed word meaning:  =  ⊕ **1**

- The meaning of most words is partially *context-dependent* and partially *fixed*. [34]
- The sentence "Fruit flies like ⟨*masked word*⟩." admits two interpretations:
  1. what the fly likes (adj-noun-verb-⟨*mask*⟩),
  2. how fruit flies (noun-verb-prep-⟨*mask*⟩).
  - Some masked words, such as "moisture" satisfy only the first interpretation.
  - Other masked words, such as "a vegetable" satisfy both interpretations.
- Let us now rearrange the sentence as follows: "⟨*Masked word*⟩ flies like fruit."
  - The rearranged sentence only admits the second interpretation.
  - Masked words still include "a vegetable", but no longer "moisture".
- In the constrained positional model:
  1. context-dependent features inhibit "moisture",
  2. fixed features encourage "a vegetable".
- Similar to sparse attention, the model makes it practical to use larger contexts.

3. Experiments & 4. Results I

3.3. (Hyper-)Parameter Optimization

Dataset	Number of tokens
2017 English Wikipedia	2,423,655,228
English Common Crawl	823,575,128,431

Table: Our datasets and their sizes in tokens.

Model	c	D'	Time
Subword	5		271h 55m
Positional	15	300	970h 14m
Constrained	15	60	481h 30m

Table: The optimal context window sizes c and numbers of positional features D', and training times in hours on the English Common Crawl for the subword, positional, and constrained models.

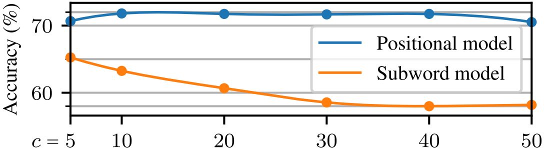


Figure: Word analogy accuracy of the subword and positional models trained on the 2017 English Wikipedia with different context window sizes c.

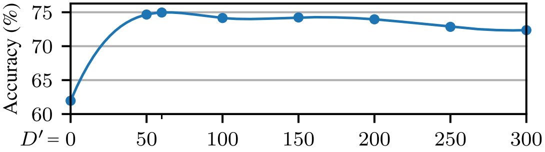


Figure: Word analogy accuracy of the constrained model trained on the 2017 English Wikipedia with different numbers of features D'.

3. Experiments & 4. Results II

3.3. Qualitative Evaluation I

Masked word prediction We show masked words w_t in the descending order of $\Pr(w_t \mid C_t)$:

$C_t^1 =$ "Unlike dogs, cats ⟨*masked word*⟩."

#	Prediction
1	cats
2	spayed
3	kennels
⋮	
1820	mew (100%)
⋮	
5581	bark (99.7%)
⋮	

$C_t^2 =$ "Unlike cats, dogs ⟨*masked word*⟩."

#	Prediction
1	kennels
2	cats
3	puppies
⋮	
4065	bark (99.9%)
⋮	
5623	mew (99.8%)

(a) Positional model

$C_t^3 =$ "Fruit flies like ⟨*masked word*⟩."

#	Prediction
1	fruit
2	flies
3	insects
⋮	
246	vegetable (99.9%)
⋮	
9036	moisture (69.6%)
⋮	

$C_t^4 =$ "⟨*Masked word*⟩ flies like fruit."

#	Prediction
1	fruit
2	insects
3	flies
⋮	
259	vegetable (99.9%)
⋮	
33465	moisture (42.8%)

(b) Constrained positional model

3. Experiments & 4. Results III

3.3. Qualitative Evaluation II

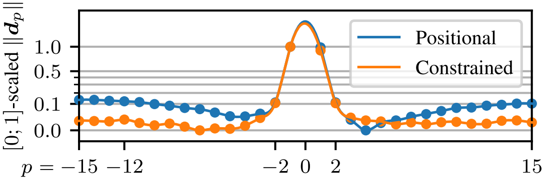


Figure: The importance of different positions p for predicting masked words in the positional and constrained positional models.

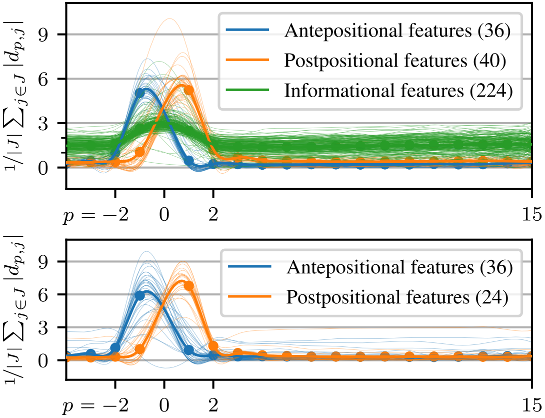


Figure: The importance of different positions p for predicting masked words in the positional (top) and constrained positional (bottom) models according to different clusters J of positional features. For each cluster J, we show its size $|J|$ in parentheses.

Importance of context words

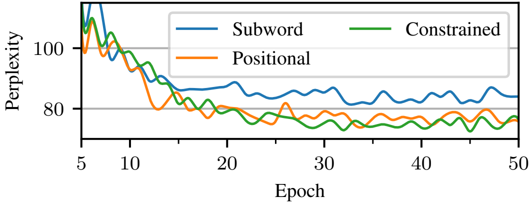
- Antepositional: “in”, “for”, “coca”
- Postpositional: “ago”, “else”, “cola”
- Informational: “finance”, “sports”, “politics”

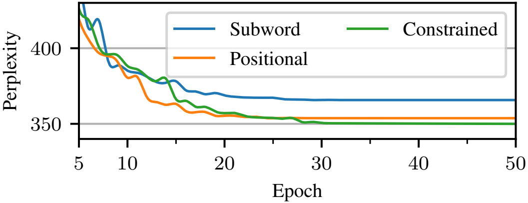
3. Experiments & 4. Results IV

3.3. Quantitative Evaluation

Subword	Positional	Constrained
360.91	347.52	*343.13*

Table: Test perplexities of RNN language models that use subword, positional, and constrained positional models as their lookup tables.





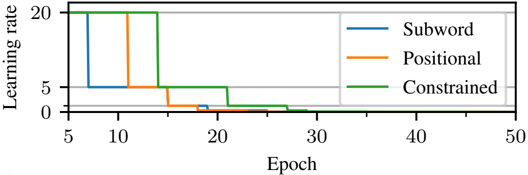


Figure: Train $\uparrow$ and validation $\nearrow$ perplexities and learning rates $\searrow$ at different epochs of RNN language models that use subword, positional, and constrained positional models as their lookup tables.

5. Conclusion & Future Work

- We have related the attention mechanism to the positional LBL of Mikolov et al. [17].
- We adapted sparse attention for our constrained positional LBL, which is:
  1. more expressive, 2. better at LM, 3. as interpretable, 4. practically fast.
- We publicly released our implementation as a couple of Python packages:
  - a low-level library, a fork of Gensim [35]: https://github.com/witiko/gensim/tree/pine,
  - a high-level user interface PInE: https://github.com/MIR-MU/pine.

- Future work should focus at:
  1. quantitative evaluation on more tasks, 2. combining LBLs with Transformers.

Thank You for Your Attention!

Bibliography I

[1] Tomáš Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781v3* (2013). URL: `https://arxiv.org/pdf/1301.3781v3.pdf`.

[2] Jeffrey Pennington et al. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[3] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. URL: `https://www.aclweb.org/anthology/Q17-1010.pdf`.

Bibliography II

[4] Eniafe Festus Ayetiran et al. “EDS-MEMBED: Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses”. In: *Knowledge-Based Systems* (2021), p. 106902. ISSN: 0950-7051. DOI: `http://dx.doi.org/10.1016/j.knosys.2021.106902`.

[5] Vít Novotný et al. “Text classification with word embedding regularization and soft similarity measure”. In: *arXiv preprint arXiv:2003.05019v1* (2020). URL: `https://arxiv.org/pdf/2003.05019v1.pdf`.

[6] Vít Novotný et al. “Three is Better than One. Ensembling Math Information Retrieval Systems”. In: *CEUR Workshop Proceedings*. Thessaloniki, Greece, 2020, p. 30. URL: `http://ceur-ws.org/Vol-2696/paper_235.pdf`.

Bibliography III

[7] Matt Kusner et al. “From Word Embeddings To Document Distances”. In: *International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 957–966. URL: `http://proceedings.mlr.press/v37/kusnerb15.html`.

[8] Vít Novotný. “Implementation Notes for the Soft Cosine Measure”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM ’18. Torino, Italy: Association for Computing Machinery, 2018, pp. 1639–1642. ISBN: 9781450360142. DOI: `10.1145/3269206.3269317`. URL: `https://doi.org/10.1145/3269206.3269317`.

Bibliography IV

[9] Tianyi Zhang et al. "BERTscore: Evaluating text generation with BERT". In: *arXiv preprint arXiv:1904.09675v3* (2020). URL: `https://arxiv.org/pdf/1904.09675v3.pdf`.

[10] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805v2* (2018). URL: `https://arxiv.org/pdf/1810.04805v2.pdf`.

[11] Omer Levy and Yoav Goldberg. "Neural word embedding as implicit matrix factorization". In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 2177–2185. URL: `http://www.cs.columbia.edu/~blei/seminar/2016_discrete_data/readings/LevyGoldberg2014.pdf`.

Bibliography V

[12] Chulhee Yun et al. "Are Transformers universal approximators of sequence-to-sequence functions?" In: *arXiv preprint arXiv:1912.10077v2* (2020). URL: https://arxiv.org/pdf/1912.10077v2.pdf.

[13] Michael Hahn. "Theoretical limitations of self-attention in neural sequence models". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 156–171.

[14] Jorge Pérez et al. "On the turing completeness of modern neural network architectures". In: *arXiv preprint arXiv:1901.03429v1* (2019). URL: https://arxiv.org/pdf/1901.03429v1.pdf.

[15] Lichao Sun et al. "Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT". In: *arXiv preprint arXiv:2003.04985v1* (2020). URL: https://arxiv.org/pdf/2003.04985v1.pdf.

Bibliography VI

[16] Kevin Clark et al. "What does BERT look at? An analysis of BERT's attention". In: *arXiv preprint arXiv:1906.04341v1* (2019). URL: https://arxiv.org/pdf/1906.04341v1.pdf.

[17] Tomáš Mikolov et al. "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/721.pdf.

[18] Martin M. Broadwell. "Teaching for learning (XVI.)". In: 20.41 (Feb. 20, 1969), 1–3a. URL: http://www.wordsfitlyspoken.org/gospel_guardian/v20/v20n41p1-3a.html.

Bibliography VII

[19] Daniel Kahneman. "Maps of Bounded Rationality: Psychology for Behavioral Economics". In: *The American Economic Review* 93.5 (2003), pp. 1449–1475. ISSN: 00028282. URL: `http://www.jstor.org/stable/3132137`.

[20] Zuzana Nevěřilová. "Improving NLP Systems with Common Sense Knowledge and Reasoning". Rigorous thesis. Brno: Masaryk university, Faculty of informatics, 2011. URL: `https://is.muni.cz/th/w0in9/` (visited on 03/30/2021).

[21] Ellen Peters et al. "Numeracy and decision making". In: *Psychological science* 17.5 (2006), pp. 407–413. DOI: `10.1111/j.1467-9280.2006.01720.x`.

[22] Ilya Sutskever et al. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014), pp. 3104–3112. URL: `https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

Bibliography VIII

[23] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078v3* (2014). URL: `https://arxiv.org/pdf/1406.1078v3.pdf`.

[24] Kyunghyun Cho et al. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259v2* (2014). URL: `https://arxiv.org/pdf/1409.1259v2.pdf`.

[25] Dzmitry Bahdanau et al. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473v7* (2016). URL: `https://arxiv.org/pdf/1409.0473v7.pdf`.

[26] Jianpeng Cheng et al. "Long short-term memory-networks for machine reading". In: *arXiv preprint arXiv:1601.06733v7* (2016). URL: `https://arxiv.org/pdf/1601.06733v7.pdf`.

Bibliography IX

[27] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. URL: `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[28] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084v1* (2019). URL: `https://arxiv.org/pdf/1908.10084v1.pdf`.

[29] Rewon Child et al. “Generating long sequences with sparse transformers”. In: *arXiv preprint arXiv:1904.10509v1* (2019). URL: `https://arxiv.org/pdf/1904.10509v1.pdf`.

Bibliography X

[30] Iz Beltagy et al. “Longformer: The long-document transformer”. In: *arXiv preprint arXiv:2004.05150v2* (2020). URL: https://arxiv.org/pdf/2004.05150v2.pdf.

[31] Manzil Zaheer et al. “Big bird: Transformers for longer sequences”. In: *arXiv preprint arXiv:2007.14062v1* (2020). URL: https://arxiv.org/pdf/2007.14062v1.pdf.

[32] Tomáš Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Bibliography XI

[33] Vít Novotný. “The Art of Reproducible Machine Learning. A Survey of Methodology in Word Vector Experiments”. In: *RASLAN 2020 Recent Advances in Slavonic Natural Language Processing* (2020), pp. 55–64. URL: `https://nlp.fi.muni.cz/raslan/raslan20.pdf#page=63`.

[34] Kent Bach. “Context dependence (such as it is)”. In: *The Continuum Companion to the Philosophy of Language* (2012), pp. 153–184. URL: `http://userwww.sfsu.edu/kbach/Bach.ContextDependence.pdf`.

[35] Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50. URL: `https://www.fi.muni.cz/usr/sojka/papers/lrec2010-rehurek-sojka.pdf`.

MUNI
FACULTY
OF INFORMATICS