Object Recognition and Generation for Comparative Question Answering Systems

Shikunova Xenia, 191 group
Scientific adviser: Klyshinsky E. S.
Scientific consultant: Nikishina I. A.

08 June 2023

Motivation

Create a **user-friendly search engine**

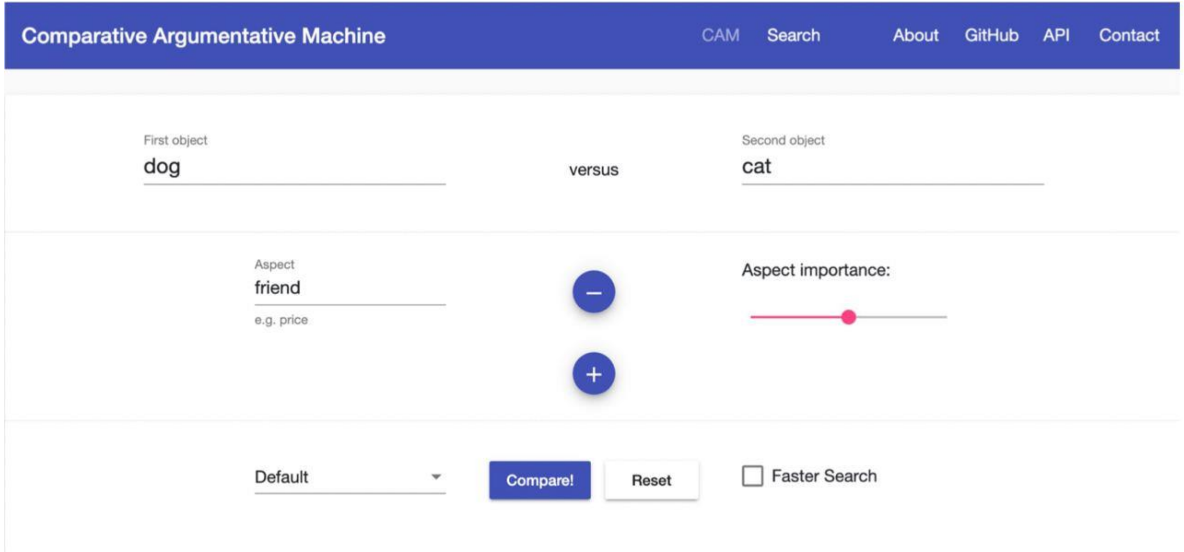
- **N**atural **L**anguage **U**nderstanding
- **N**atural **L**anguage **G**eneration

&

Create **comparative QA system**

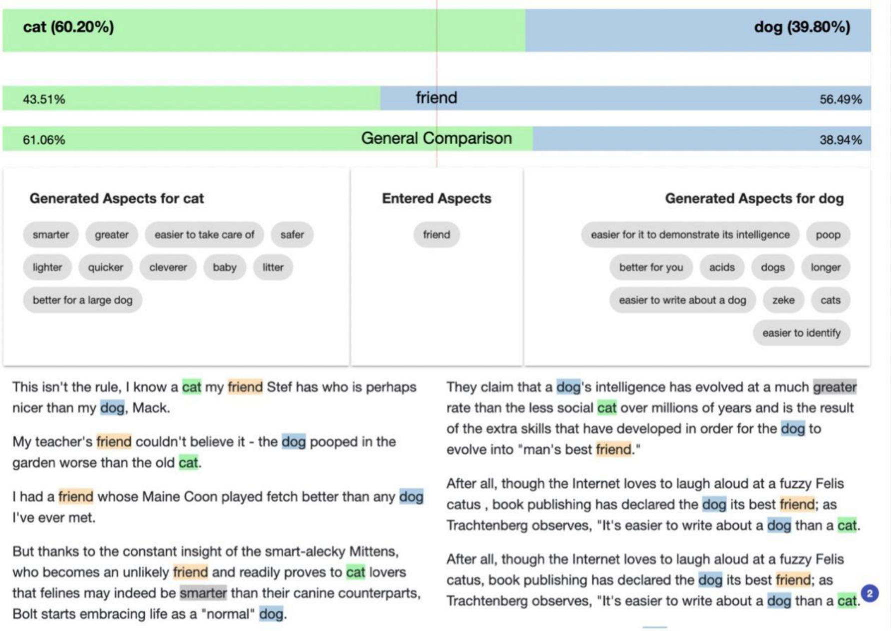
Previous works (Schildwächter et al., 2019)

Comparative **A**rgumentative **M**achine



Previous works (Schildwächter et al., 2019)

Comparative **A**rgumentative **M**achine



Previous works (Chekalina et al. 2021)

NLU and **NLG modules**

Methods: RoBERTa (sequence labeling task)

Metrics: F-score

Scores: 0.925, 0.685, 0.894 for objects, aspects, predicates

Methods: CTRL (generation), CAM-ranking, contextual extension

Metrics: ROUGE-1

Scores: 0.245, 0.229, 0.216

Tasks

Objects and **aspect** recognition

Summary generation

Test **generative** models

Try to **finetune** models

Implement **unsupervised** learning

General **methods**

Generative models

- **few-shot**
- **zero-shot**
  - Objects Recognition
- **fine-tuning**
  - Objects Recognition
  - Summarization

Pipeline

Objects and Aspect Recognition

What are the objects and the aspect of comparison in the sentence ‘[SENTENCE]’?

CAM

[INPUT]: Objects and Aspect
[OUTPUT]: List of sentences

Summarization Generation

Write a comparison of [OBJ-1] and [OBJ-2]. Summarize only relevant arguments from the list.

Objects and Aspects **Dataset**

over 3000 questions with labeled aspects and objects (Beloucif et al. 2022)

including sentences with common objects

	Aspect		**Common object**	
	YES	**NO**	**YES**	**NO**
Number of Sentences	1783	1274	727	2330
Percentages	58.3	41.7	23.8	76.2

Examples

OBJ-1 OBJ-2

What is bigger size bitmap or jpeg ?

ASPECT

OBJ-1 OBJ-2

How can you tell the difference between literal and symbolic dreams?

SHARED

Objects and Aspect **Recognition**

Models

T5
Flan-T5
LLaMA

GPT-2
GPT-Neo
GPT-J
Dolly

Metrics

Full Match
Cosine Similarity
Edit Distance

MUC
CEAF

Best score:

		FullMatch	CosSim	EdDist (found)	CosSim (found)	MUC f-score	CEAF f-score
T5	Objects	**0.649**	**0.7**	**4.64**	**0.863**	**0.725**	**0.624**
	Aspect	0.318	0.498	8.56	0,805	0.386	0.321
Flan-T5	Objects	0.437	0.679	4.4	0.827	0.724	0.592
	Aspect	**0.59**	**0.704**	**3.01**	**0.926**	**0.629**	**0.44**

Recognition **evaluation**

		FullMatch	CosSim	EdDist (found)	CosSim (found)	MUC f-score	CEAF f-score
GPT2	Objects	0.326	0.605	7.58	0.741	0.497	0.464
	Aspect	0.2	0.397	9.44	0.769	0.272	0.266
GPT-Neo	Objects	0.392	0.738	6.75	0.783	0.604	0.585
	Aspect	0.274	0.496	7.63	0.835	0.369	0.309
Dolly	Objects	0.098	0.247	10.82	0.768	0.222	0.181
	Aspect	0.425	0.477	8.01	0.936	0.445	0.243
LLaMA	Objects	0.358	0.619	7.65	0.852	0.549	0.455
	Aspect	0.412	0.575	19.52	0.867	0.475	0.327

T5 and Flan-T5 **fine-tuning**

		FullMatch	CosSim	EdDist (found)	CosSim (found)	MUC f-score	CEAF f-score
T5	Objects	0.678	**0.942**	**3.29**	0.945	0.888	0.694
	Aspect	**0.691**	**0.808**	2.33	**0.949**	0.749	0.516
Flan-T5	Objects	**0.689**	**0.942**	3.35	**0.947**	**0.895**	**0.697**
	Aspect	0.688	**0.808**	**2.06**	**0.949**	**0.753**	**0.517**

Summarization **Dataset**

over 500x3 examples of ChatGPT summarization

536 input examples each with 3 output summarizations (**0.89** precision)

Mean length INPUT	Mean length OUTPUT	Number of INPUT	Number of OUTPUT
2456	1057	536	1608

Input example

Write a comparison of "milk" and "juice". Summarize only relevant arguments from the list.

1. Yes, milk is better than juice and it's good for bones.
...
3. Watered down orange juice is worse than powdered milk in my opinion.
4. Milk is way better than sugary soda, sweet chocolate milk, or orange juice.
5. Chocolate milk is a better pick me up than juice.
6. As a parent, I think of most juice as worse than chocolate milk.
7. And I don't need to tell you that milk goes bad much quicker than orange juice.
8. Milk & water are just plain better than soda, tea or juice.\n9. Milk is a much better choice nutritionally than sodas or juice.
...
11. Listeria and other baddies grows MUCH faster in pasteurized milk than raw milk.
...
17. Barley grass juice is known to have 11x greater calcium content than milk.
19. Better Than Milk Soy Beverage Mix -- Smooth and delicious, Better Than Milk is a great cost-effective way to get a lot of soymilk for less.
...

After the summary, list the arguments you used below the text. Put citations in brackets inside the text. Do not even mention arguments that are not relevant to milk and juice.

Output example

precision **0.89**

Milk and juice are both popular beverages, but they differ in taste, nutritional value, safety, and other qualities.

Milk is often praised for its nutritional value, being a better choice than sugary sodas or juices [4][8][9]. It is also considered good for bones [1]. However, it can go bad quickly [7] and **may contain harmful bacteria if not pasteurized properly [11]**. Milk is also often compared to other beverages, like chocolate milk [5][6], which is seen as a better pick-me-up than juice.

Juice, on the other hand, **can be a refreshing and flavorful alternative to water [3]**. Some people prefer it to milk and consider watered-down orange juice worse than powdered milk [3]. Juice can also have some health benefits, like the high calcium content in barley grass juice [17]. There are also options like soy beverage mix [19] that provide a cost-effective alternative to milk.

Ultimately, the preference between milk and juice comes down to personal choices, taste, and nutritional needs. Some may prefer milk, while others choose juice to quench their thirst.

Arguments used: 1, 3, 4, 5, 6, 7, 8, 9, 11, 17, 19.

Input example

Write a comparison of "milk" and "juice". Summarize only relevant arguments from the list.

1. Yes, milk is better than juice and it's good for bones.
...
3. Watered down orange juice is worse than powdered milk in my opinion.
4. Milk is way better than sugary soda, sweet chocolate milk, or orange juice.
5. Chocolate milk is a better pick me up than juice.
6. As a parent, I think of most juice as worse than chocolate milk.
7. And I don't need to tell you that milk goes bad much quicker than orange juice.
8. Milk & water are just plain better than soda, tea or juice.\n9. Milk is a much better choice nutritionally than sodas or juice.
...
11. Listeria and other baddies grows MUCH faster in pasteurized milk than raw milk.
...
17. Barley grass juice is known to have 11x greater calcium content than milk.
19. Better Than Milk Soy Beverage Mix -- Smooth and delicious, Better Than Milk is a great cost-effective way to get a lot of soymilk for less.
...

After the summary, list the arguments you used below the text. Put citations in brackets inside the text. Do not even mention arguments that are not relevant to milk and juice.

Summarization **experiments** with **GPT-2**

Clear data

50 clear examples in train
50 default examples in test

Correct mistakes

50 similar clear and default examples in train (100)
50 default examples in test

Clear data v2.0

50 clear examples in train
10 clear examples in test

Augmentation

50 different clear and default examples in train (100)
50 default examples in test

Large train

450 default examples in train
50 default examples in test

Example of GPT-2 summarization

Football and lacrosse are both popular sports, but they have some differences and similarities.

Some argue that lacrosse is harder than soccer, but still not as tough as football [1]. Others cite lacrosse as being much safer than football [10][13], while some point to the risk of commotio cordis (a cardiac arrest caused by a sudden blow to the chest) in lacrosse [11].

In terms of comparison to other sports, some argue that lacrosse is faster than any other field game [8][14]. Moreover, lacrosse players may have an advantage if they also play other sports such as wrestling, basketball, or football [3].

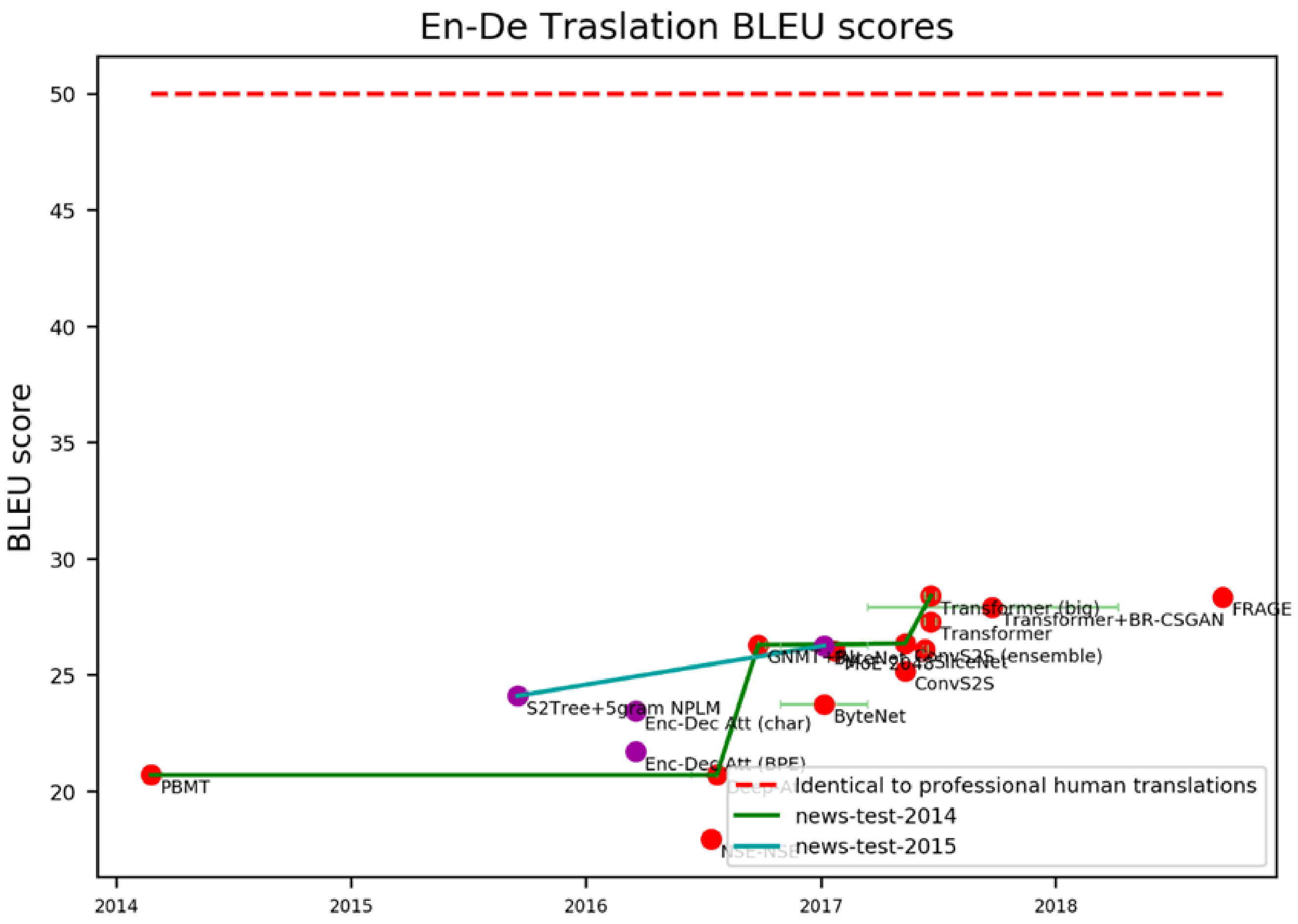
There are also arguments about the quality of equipment and apparel for both sports, with some stating that lacrosse-specific products are not sufficient for protection [6], while others praise lacrosse equipment for its superior quality [5][6].

Ultimately, the choice between football and lacrosse may depend on personal preferences and priorities. Some argue that lacrosse is a better choice for a team than football [4], while others prefer lacrosse over football [3][14].

Arguments used: 1, 2, 3, 4, 5, 6, 8, 9, 11, 13, 14.

ROUGE and **BLEU** evaluation

	ROUGE-1	ROUGE-Lsum	BLEU	length ratio
GPT-2 124M	0.495	0.429	0.255	1.029
GPT-2 345M	0.514	0.469	0.289	1.002



En-De Traslation BLEU scores
BLEU score
50
45
40
35
30
25
20
2014
2015
2016
2017
2018
PBMT
S2Tree+5gram NPLM
Enc-Dec Att (char)
Enc-Dec Att (BPE)
ByteNet
ConvS2S
ConvS2S (ensemble)
Transformer
Transformer (big)
Transformer+BR-CSGAN
FRAGE
Identical to professional human translations
news-test-2014
news-test-2015

Question

fine-tuned **Flan-T5**

CompQA-Dataset

OBJECTS
ASPECT

CAM

SENTENCES WITH PROS AND CONS

fine-tuned **GPT2**

ChatGPT-Answers

SUMMARIZATION

Interface

Referencing in summarization*

[INPUT]

1. Some people say Nokia mobile is **better** than Siemens because it's **easier to operate**. ...

7. So yes, Nokia is now **falling faster** than at its worst moment Palm, Motorola, Siemens, Windows Mobile, Ericsson, Blackberry or any other maker. ...

[OUTPUT]

Nokia and Siemens are both major technology companies ... Some argue that Nokia is **better** than Samsung in this segment **[1]** ... and that Nokia has **better engineering teams [7]** ... Arguments used: **1, 2, 3, 4, 6, 7, 9, 12, 15, 18**.

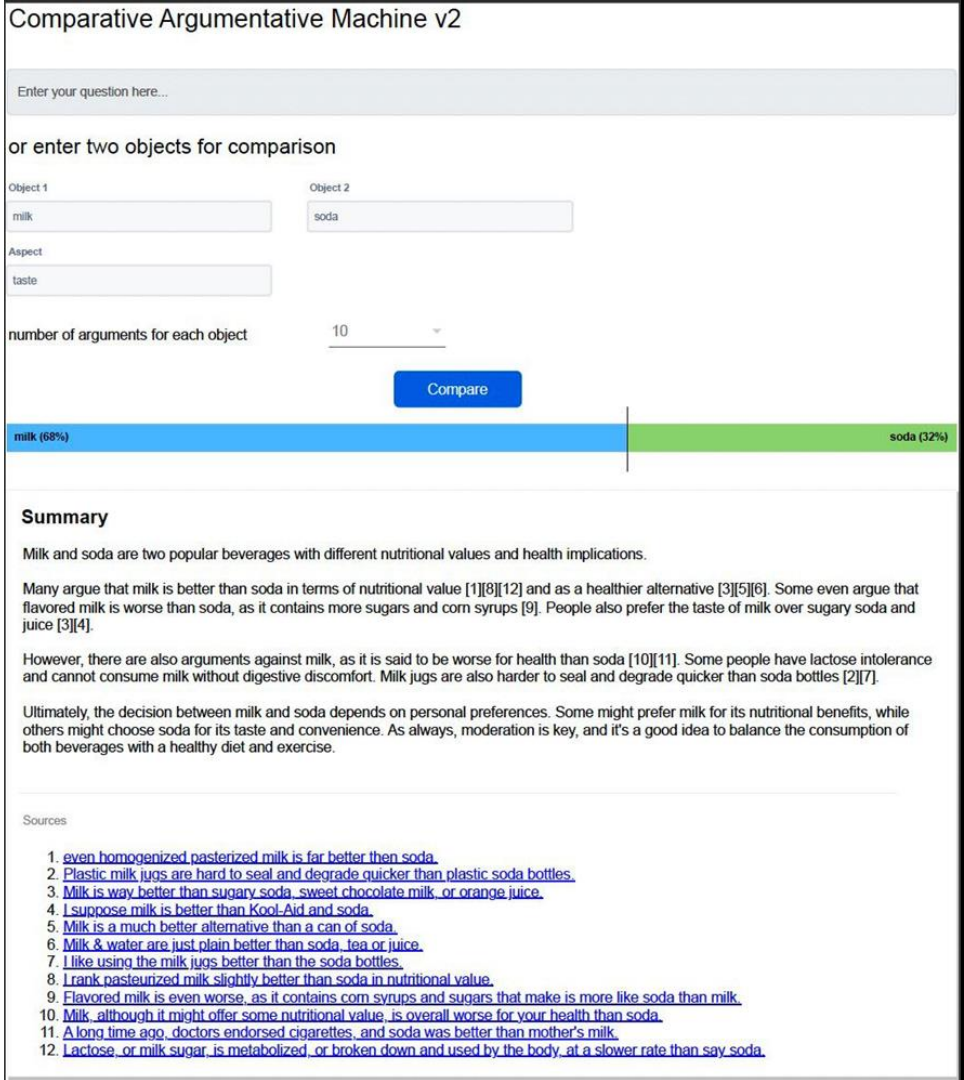
[1] no aspect mentioned

[7] mistake in polarity

mistake in list of arguments used

Web-Interface design

...sample of the future work



Also we plan to participate in **EMNLP 2023's** Demo!

https://2023.emnlp.org/calls/demos/

Conclusion

Reviewed **existing approaches**

Used various **transformer models** for objects and aspect recognition

(also with **few-shot** and **zero-shot**)

Finetuned generative models

Collected and **manually evaluated** GPT-3 summarization dataset

Did several **training experiments** for GPT-2 finetuning

References

Arora et al. 2017 – Arora J. et al. Extracting entities of interest from comparative product reviews //Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. – 2017. – С. 1975-1978.

Beloucif et al. 2022 – Beloucif M. et al. Elvis vs. M. Jackson: Who has More Albums? Classification and Identification of Elements in Comparative Questions //Proceedings of the Thirteenth Language Resources and Evaluation Conference. – 2022. – С. 3771-3779.

Black et al. 2021 – Black S. et al. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow //If you use this software, please cite it using these metadata. – 2021. – Т. 58.

Bondarenko et al. 2020a – Bondarenko A. et al. Comparative web search questions //Proceedings of the 13th International Conference on Web Search and Data Mining. – 2020. – С. 52-60.

Chekalina et al. 2021 – Chekalina V. et al. Which is better for deep learning: python or MATLAB? Answering comparative questions in natural language //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. – 2021. – С. 302-311.

Chung et al. 2022 – Chung H. W. et al. Scaling instruction-finetuned language models //arXiv preprint arXiv:2210.11416. – 2022.

Devlin et al. 2019 – Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

References

Floridi, Chiriatti 2020 – Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences //Minds and Machines. – 2020. – Т. 30. – С. 681-694.

Gao et al. 2020 – Gao L. et al. The pile: An 800gb dataset of diverse text for language modeling //arXiv preprint arXiv:2101.00027. – 2020.

Lin 2004 – Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – С. 74-81.

Liu et al. 2022 – Liu Y. et al. BRIO: Bringing order to abstractive summarization //arXiv preprint arXiv:2203.16804. – 2022.

Liu et al. 2019 – Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.

Luo 2005 – Luo X. On coreference resolution performance metrics //Proceedings of human language technology conference and conference on empirical methods in natural language processing. – 2005. – С. 25-32.

Mallick et al. 2019 – Mallick C. et al. Graph-based text summarization using modified TextRank //Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018. – Springer Singapore, 2019. – С. 137-146.

References

Panchenko et al. 2017 – Panchenko A. et al. Building a web-scale dependency-parsed corpus from CommonCrawl //arXiv preprint arXiv:1710.01779. – 2017.

Papineni et al. 2002 – Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – С. 311-318.

Peters et al. 2018 – Peters M. E. et al. Deep contextualized word representations. CoRR abs/1802.05365 (2018) //arXiv preprint arXiv:1802.05365. – 1802.

Radford et al. 2019 – Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – Т. 1. – №. 8. – С. 9.

Raffel et al. 2020 – Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //The Journal of Machine Learning Research. – 2020. – Т. 21. – №. 1. – С. 5485-5551.

Reimers, Gurevych 2019 – Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks //arXiv preprint arXiv:1908.10084. – 2019.

Reimers, Gurevych 2020 – Reimers N., Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation //arXiv preprint arXiv:2004.09813. – 2020.

References

Schildwächter et al. 2019 – Schildwächter M. et al. Answering comparative questions: Better than ten-blue-links? //Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. – 2019. – С. 361-365.

Sundheim, Sundheim 1996 – Grishman R., Sundheim B. M. Message understanding conference-6: A brief history //COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. – 1996.

Sutton, McCallum 2012 – Sutton C. et al. An introduction to conditional random fields //Foundations and Trends® in Machine Learning. – 2012. – Т. 4. – №. 4. – С. 267-373.

Touvron et al. 2023 – Touvron H. et al. Llama: Open and efficient foundation language models //arXiv preprint arXiv:2302.13971. – 2023.

Vaswani et al. 2017 – Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.

Wang, Komatsuzaki 2021 – Wang B., Komatsuzaki A. GPT-J-6B: A 6 billion parameter autoregressive language model. – 2021.