# Extending the Comparative Argumentative Machine: Multilingualism and Stance Detection

Irina Nikishina,[1,*] Alexander Bondarenko,[2,*] Sebastian Zaczek,[1]
Onno Lander Haag,[2] Matthias Hagen,[2] Chris Biemann[1]

[1] Universität Hamburg [2] Friedrich-Schiller-Universität Jena

**Abstract** The comparative argumentative machine CAM can retrieve arguments that answer comparative questions—questions that ask which of several to-be-compared options should be favored in some scenario. In this paper, we describe how we equipped CAM with a better answer stance detection (i.e., a better detection of which option "wins" a comparison) and with system variants to support non-English requests. As for the improved answer stance detection, we develop RoBERTa-based approaches and experimentally show them to be more effective than previous feature-based and LLM-based stance detectors. As for the multilingualism, in a proof of concept, we compare two approaches to support Russian requests and answers: (1) translating the original English CAM data and (2) using an existing replica of CAM on native Russian data. Comparing the translation-based and the replica-based CAM variants in a user study shows that combining their answers seems to be the most promising. For individual questions, the retrieved arguments of the two variants are often different and of quite diverse relevance and quality. As a demonstrator, we deploy a first multilingual CAM version that combines translation-based and replica-based outputs for English and Russian and that can easily be extended to further languages.

**Keywords:** Answering Comparative Questions · Argumentation Machines · Answer Stance Detection · Cross-Language Argument Retrieval

## 1 Introduction

Decision making is part of everyday life, yet it can involve a complex and time-consuming process when pro / con arguments on the potential alternatives need to be gathered and weighed (e.g., 'Should I buy or rent a house?').

There are many ways to gather arguments or opinions on some comparison objects: asking other people but also using web search engines, LLM-based systems, specialized product comparison websites, research prototypes, etc. One research prototype that was developed for open comparative questions (i.e., not just focusing on products) and that should be more effective than skimming through a search engine's classical "ten blue links" is the comparative argumentative machine CAM [33].[1] The web interface of CAM takes some comparison

[*] These authors contributed equally.

[1] http://ltdemos.informatik.uni-hamburg.de/cam/

objects and aspects as input, retrieves (argumentative) sentences relevant to the comparison, detects the sentences' comparative stances (i.e., which comparison object is favored), and presents a tabular result. As the original CAM only supports English inputs and results, recently, a replica of CAM for Russian comparisons—the RuCAM system—has been developed [22].

An important component of CAM and RuCAM is stance detection to determine for each retrieved sentence which comparison object is favored and which object is the overall "winner" in the retrieved sentences (e.g., 'buying' vs. 'renting' in the house example). Accurately grouping the retrieved sentences in CAM's and in RuCAM's tabular result presentation with respect to the favored objects ensures that users are not misled and can come to "correct" conclusions. Still, with F1 scores of 0.85 and 0.82, respectively, CAM's and RuCAM's current rule- and XGBoost-based [8] or rule- and BERT-based [13] stance detection seem to leave some room for further effectiveness improvement. Our first research question thus is: **(RQ1)** Can advanced BERT models like RoBERTa improve the effectiveness of CAM's and RuCAM's answer stance detection?

To address RQ1, we fine-tune several RoBERTa-based models [20,9] on the 5,759 English sentences of the CompSent-19 training set [29] using the masking approach of Bondarenko et al. [4]. In our experiments, our new stance detectors achieve F1 scores of 0.91 for English and 0.87 for Russian and thus are more effective than CAM's and RuCAM's current detectors. Interestingly, a multilingual XLM-RoBERTa model turned out to be as effective as an English-only model so that the same stance detector can be used in CAM and in RuCAM.

Developed as a replica of the original CAM system, RuCAM uses the Russian and not the English part of the Common Crawl.[2] Still, another possibility to support some non-English language would have been to simply translate the original CAM's inputs and results. As there was no comparison of these two ideas yet and as we aim for a single multilingual CAM system, our second research question is: **(RQ2)** What are the strengths and weaknesses of machine translation-based and replica-based "localization" of CAM?

To address RQ2, we use Russian as the target language and compare machine-translated CAM responses and RuCAM responses in manual analyses and in a user study. The results of our manual analyses indicate that CAM and RuCAM retrieve rather different results of varying relevance and quality (translated CAM results tend to be more relevant while the RuCAM results tend to be of higher quality). Therefore, a combination of translation-based and replica-based results seems to be a promising direction for a multilingual CAM system.

Our code and data are publicly available.[3] As a demonstrator of a multilingual CAM system, we equip the existing CAM and RuCAM backends with a new interface for multilingual translation- and replica-based search in English and Russian[4] that can easily be extended to support further languages.

[2] https://commoncrawl.org/
[3] https://github.com/webis-de/RATIO-24
[4] https://cam-multi.ltdemos.informatik.uni-hamburg.de

## 2 Related Work

This section provides an overview of the CAM system and the existing approaches for comparative stance detection.

### 2.1 CAM Overview

Comparative information needs in web search were first addressed by developing simplistic search interfaces where two to-be-compared products were entered separately in a left and a right search box [25,35]. The search results were presented to the searcher as side-by-side two standard "ten blue links" lists for each product. Later, the comparative argumentative machine (CAM) was developed to tackle open-domain comparisons (not just products) [33]. The CAM's web interface takes as input user-specified two comparison objects (e.g., 'buy a house' and 'rent a house') and optional comparison aspect(s) (e.g., 'risk'). Then, using Elasticsearch,[5] CAM retrieves comparative arguments (e.g., 'It is less risky to rent a house than to buy') relevant to the user input from the DepCC corpus [30] containing about 14 billion English sentences coming from the Common Crawl corpus (if no relevant arguments are found, CAM will respectively notify the user). For each retrieved argument, its comparative stance is detected so that the arguments can be grouped into two columns (i.e., whether one or the other object is preferred according to individual arguments) for the final result presentation. In the CAM's output, the arguments are ranked based on the Elasticsearch relevance score. Additionally, the final comparison score is shown to the user, which determines the overall "winner" of the comparison. The score combines the stance detector's confidence and the Elasticsearch score and is summed up over all the retrieved arguments for each comparison object [33]. The design of the CAM system is also shown in Figure 1.

The comparative stance in the context of CAM is defined as ternary label: (1) First comparison object "wins" a comparison, i.e., it performs better or is more suitable for the comparison aspect compared to the second object (e.g., 'It is less risky to buy a house than to rent a house'), (2) second object "wins" a comparison (e.g., 'It is less risky to rent a house than to buy a house'), or (3) none "wins" or no comparison is present; such statements are excluded from the CAM's final result presentation [33].

A user study with CAM showed that the study participants were able to answer comparative questions faster and more accurately compared to a standard web search [33]. Later, RuCAM [22], a Russian version of the CAM system (that supports the English language only), was developed that replicates the original CAM pipeline using a Russian part of the Common Crawl corpus [28].

### 2.2 Comparative Stance Detection

One of the important (Ru)CAM components is a comparative stance detector that allows to place the retrieved arguments on the correct "winning" side. More

[5] https://www.elastic.co

generally, stance detection is the task of identifying the author's viewpoint (attitude, opinion) towards a target, which can be a debate topic, an entity, or a claim [23]. Earlier works mostly focused on the stance detection in online debates [19] and proposed rule-based [2,24,41,42] and feature-based classifiers using, for instance, SVM [15,3,36,42] or Naïve Bayes [2,15,31]. Later, neural network architectures like CNN [43,16], LSTM [45,44], and transformer-based models like BERT [13] became state-of-the-art approaches [1,34,39].

The aforementioned works mostly focused on the stance detection towards a single target. Our task is to detect the stance given two comparison objects, i.e., two stance targets (e.g., 'buy a house' vs. 'rent a house'). For detecting a comparative stance, i.e., a "winning" comparison object in English sentences, different feature-based classifiers were tested [29], e.g., logistic regression [12], SVM [11], XGBoost [8], etc. Trained on 5,759 and tested on 1,440 English sentences (CompSent-19 dataset [29]), the most effective stance detector (XGBoost with InferSent embeddings [10]) achieved a micro-avg. F1 of 0.85 (3 labels: first / second object "wins" or no comparison).

Later, on the same dataset, a stance detector was tested that employed multi-hop graph attention over a dependency graph sentence representation [21]. Each sentence was represented by its dependency graph, which, for simplicity, was then converted from the original directed graph into an undirected graph. Embeddings for each sentence word (node in the dependency graph) were calculated using BERT [13]. Then, Graph Attention Networks [40] were used to embed the relation between the comparison objects. Finally, a feed-forward layer with a softmax function was added to project the embedding vectors into classes for prediction. The proposed approaches achieved a micro-avg. F1 of 0.87, outperforming the previous XGBoost-based stance detector.

Recently, large language models like LlaMa-2 [38], GPT-3.5 Turbo [26], and GPT-4 [27] using zero-shot and few-shot prompting were tested [18]. In addition to rigorous prompt engineering, the authors designed a retry message to tackle the cases when an LLM returned malformed answers, i.e., answers violating the predefined format suitable to extract the predicted stance. Interestingly, all tested LLMs did not improve over the aforementioned stance detectors.

In this paper, we fine-tune the RoBERTa [20] and XLM-RoBERTa models [9] following the idea of masking the comparison objects with special tokens [4]. The evaluation results show that our stance detectors are more effective than previous feature-, neural-, and LLM-based approaches, achieving a micro-avg. F1 of 0.91 for the Engish and 0.87 for the Russian languages.

## 3 Improving Comparative Stance Detection

The current CAM implementation allows to choose between a rule-based comparative stance detector that uses the handcrafted list of cue words and an XGBoost-based classifier [29,33]. The latter one is more effective and achieves a micro-avg. F1 of 0.85 (3 labels: first/second object "wins" or no comparison).

**Table 1.** Stance detection effectiveness of different approaches tested on English sentences from the CompSent-19 dataset (class distribution: 'no comparison' 73%, 'first object wins' 19%, 'second object wins' 8%) [29]. Reported are F1 scores per stance class and a micro-averaged F1.

| | | Stance label | | | |
|---|---|---|---|---|---|
| **Stance detector** | **Ref.** | *First* | *Second* | *None* | **Micro-avg.** |
| Rule-based | [29] | 0.65 | 0.44 | 0.90 | 0.82 |
| GPT-3.5 Turbo (few-shot) | [18] | 0.68 | 0.48 | 0.90 | 0.84 |
| LLaMa-2 70B (few-shot) | [18] | 0.75 | 0.60 | 0.91 | 0.85 |
| XGBoost + InferSent | [29] | 0.75 | 0.43 | 0.92 | 0.85 |
| GPT-4 (few-shot) | [18] | 0.78 | 0.65 | 0.91 | 0.86 |
| ED-GAT$_{\text{BERT}}$ | [21] | 0.78 | 0.56 | 0.93 | 0.87 |
| RoBERTa-masked | our | 0.86 | **0.70** | 0.94 | **0.91** |
| XLM-RoBERTa-masked | our | **0.87** | 0.69 | **0.95** | **0.91** |

To address our first research question, we fine-tune the English RoBERTa [20] and multilingual XLM-RoBERTa [9] models on the CompSent-19 train set (5,759 English sentences) [29]. Following the idea by Bondarenko et al. [4], we mask the comparison objects using the special masking tokens: `[FIRST OBJECT]` and `[SECOND OBJECT]`, before fine-tuning.[6]

We compare the effectiveness of our stance detectors with several approaches from previous work that were tested on the CompSent-19 test set (1,440 English sentences; we again mask the comparison objects). The results in Table 1 show that our stance detectors achieve a convincing micro-avg. F1 of 0.91 and are more effective than previous existing feature-based and LLM-based approaches.

*Multilingual Stance Detection.* To test our comparative stance detector for the Russian language, we use a dataset of 1,208 manually labeled Russian sentences [22]. Due to a relatively small number of labeled examples, we use the whole dataset to test our multilingual stance detector based on XLM-RoBERTa that was fine-tuned on English sentences. We also test on the original test set for a more fair comparison with the stance detectors for Russian from previous work [22]. Our stance detector achieves a micro-avg. F1 of 0.87 on the test set and 0.83 on the full dataset (cf. Table 2), which is more effective than the rule-based approach and fine-tuned RuBERT [22].

An interesting observation is that for English sentences, fine-tuning a multilingual RoBERTa is as effective for stance detection as fine-tuning an English-only model. We thus suggest using for practical application a single fine-tuned

[6] Hyperparameters were selected using a 5-fold cross-validation on the train set. Models: roberta-large and xlm-roberta-large, batch size: 16, learning rate: 0.00003, training epochs: 5. Fine-tuning was performed using Colab's Tesla T4 GPU (RoBERTa) and NVIDIA GeForce RTX 4090 (XLM-RoBERTa).

**Table 2.** Stance detection effectiveness of different approaches tested on Russian sentences (class distribution: 'no comparison' 70%, 'first object wins' 21%, 'second object wins' 9%) [22]. Reported are F1 scores per stance class and a micro-averaged F1. For comparison with the original stance detectors for Russian [22], we additionally report the results on the original test set.

| | | Stance label | | | |
|---|---|---|---|---|---|
| **Stance detector** | **Ref.** | *First* | *Second* | *None* | **Micro-avg.** |
| *Test set (119 sentences)* | | | | | |
| Rule-based | [22] | 0.34 | 0.33 | 0.82 | 0.69 |
| RuBERT-based | [22] | 0.57 | 0.38 | 0.91 | 0.82 |
| XLM-RoBERTa-masked | our | **0.71** | **0.53** | **0.93** | **0.87** |
| *Full dataset (1,208 sentences)* | | | | | |
| Rule-based | [22] | 0.41 | 0.27 | 0.74 | 0.62 |
| XLM-RoBERTa-masked | our | **0.68** | **0.53** | **0.90** | **0.83** |

XLM-RoBERTa to detect the stance in both the English and Russian languages.

## 4 Adapting CAM to Russian

One of the limitations of CAM is its restriction on the English language. To address our second research question, we explore the use of machine translation to translate CAM output to the target Russian language. We then compare the translation-based approach with the existing replica-based system, RuCAM [22].

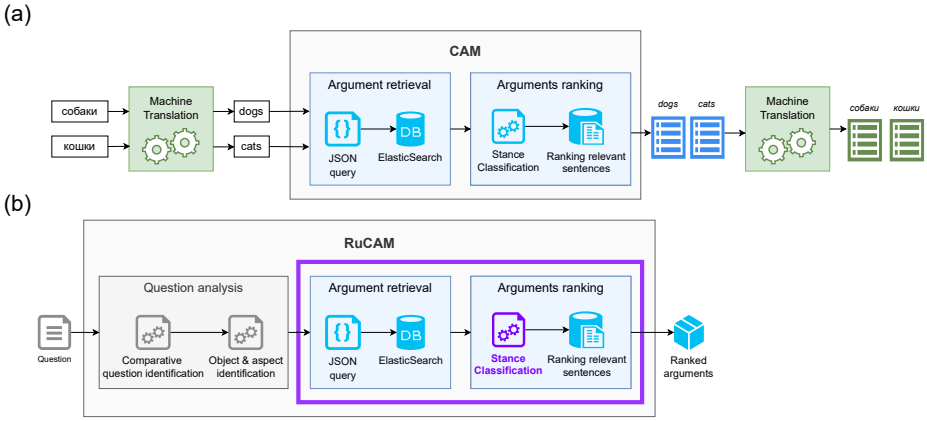
### 4.1 Translation-based System

As the translation model, we use OPUS-MT [37]. While OPUS-MT could be replaced with any system, we motivate its use since it is open access and easy to implement using the Huggingface's transformers library.

The overall CAM architecture extended with the translation modules is presented in Figure 1(a). First, we added a translation step to the CAM's input of the comparison objects and aspect(s) from Russian to English,[7] so that we could retrieve English sentences from the Elasticsearch index. Afterwards, we translate the retrieved arguments from English to Russian for the final result presentation.[8] This system does not require any additional data and indexing.

[7] https://huggingface.co/Helsinki-NLP/opus-mt-ru-en
[8] https://huggingface.co/Helsinki-NLP/opus-mt-en-ru

**Figure 1.** (a) The architecture of the CAM system extended with translation modules; new modules are in green; (b) The architecture of the RuCAM system [22]. The modules that we used for the language adaptation comparison are in purple. The updated stance classification model is also highlighted in purple.




### 4.2 RuCAM

To compare with the alternative pipeline for language adaptation, we consider RuCAM [22]: the Russian Comparative Argumentative Machine[9] that implements its own Elasticsearch index based on the Open Super-large Crawled Aggregated corpus (OSCAR) [28] containing 21 billion Russian sentences from the Common Crawl corpus. RuCAM also accepts natural language questions as input; however, we skip previous steps (comparative question identification and object and aspect detection) and query the system directly with two comparison objects and, optionally, aspect(s). We also replace the original stance classification model with our fine-tuned XLM-RoBERTa-masked model to improve the quality of the system output. Figure 1(b) presents the overall architecture of RuCAM, highlighting in purple the modules used for the comparison as well as the updated stance classification model.

### 4.3 System Comparison on Retrieval Effectiveness

To evaluate the two adaptation techniques, we set up a manual annotation following the methodology from the Touché 2020–2022 shared tasks on argument retrieval [5,6,7]. For the user study, we use the Touché 2022 dataset [6] that contains 50 comparative questions, each labeled with two comparison objects that we also translated into Russian (e.g., 'Should I buy or rent a house?') The English CAM found matches (i.e., relevant arguments) for 40 object pairs, and the RuCAM had matches for 46 pairs. Then, we provided five volunteer annotators with the annotation guidelines from the Touché tasks and asked to label the

[9] http://rucam.ltdemos.informatik.uni-hamburg.de

**Table 3.** Relevance-wise (R) and quality-wise (Q) retrieval effectiveness of the replica-based and translation-based systems. The nDCG scores are calculated for the two ranked lists separately (the first or the second object "wins" a comparison split after stance detection) and for all the retrieved results before stance detection (overall).

| System | First object | | | | Second object | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nDCG@5 | | nDCG@10 | | nDCG@5 | | nDCG@10 | | nDCG@5 | | nDCG@10 | |
| | R | Q | R | Q | R | Q | R | Q | R | Q | R | Q |
| Replica-based | 0.84 | **0.93** | 0.90 | **0.96** | 0.83 | **0.93** | 0.90 | **0.96** | 0.83 | **0.93** | 0.90 | **0.96** |
| Transl.-based | **0.91** | 0.81 | **0.97** | 0.91 | **0.85** | 0.79 | **0.94** | 0.90 | **0.88** | 0.78 | **0.95** | 0.91 |

retrieved arguments based on: (1) the relevance to the comparison object pairs as not relevant (label 0), relevant (label 1), and highly relevant (label 2), and (2) the argument quality (rhetorical well-writtenness) as low quality (label 0), sufficient quality (label 1), and high quality (label 2). In total, the labeled dataset comprises 1,238 arguments (at most 10 arguments for each object pair).[10]

To calibrate the annotators' interpretations of the guidelines, we conducted an initial Fleiss' $\kappa$ test in which each annotator had to label the same 15 arguments for 3 object pairs (5 arguments for each pair). The observed Fleiss' $\kappa$ values were 0.734 for argument relevance (substantial agreement) and 0.45 for argument quality (moderate agreement). Furthermore, after the initial $\kappa$ test, we organized a follow-up discussion with all the annotators to clarify potential misinterpretations, e.g., the cases where one of the objects is ambiguous (e.g., 'Milk tastes better than cow's milk in the supermarket.'; it is not clear whether the first object refers to 'goat milk' or not) or the argument is too long and contains both the comparison as well as the unrelated text (e.g., 'Comp JAVA Program Description: I will not say for sure that NetBeans is better than Eclipse, I believe that each development environment has its own strengths and weaknesses.'). Afterwards, each annotator independently judged the results for disjoint subsets of the topics (i.e., each unique object pair was assigned to one annotator only).

Using the resulting manual labels for the argument relevance and quality, we calculate nDCG@5 and nDCG@10 scores [17] for each comparison object separately (since CAM and RuCAM split the arguments according to the "winning" object) and overall scores for all arguments retrieved for all object pairs (see Table 3). We assume that the scores are comparable, as both systems retrieve arguments from Common Crawl. Apparently, different corpora might have different numbers of relevant arguments because of social and cultural differences.

The results show that both relevance and quality nDCG scores are relatively high for both systems. However, the argument relevance in a translation-based system is consistently higher. This might be explained either by a more effective stance detector for English or a more effective Elaticsearch retrieval for English,

[10] The labeled dataset and annotation guidelines are available in the GitHub repository: https://github.com/webis-de/RATIO-24

**Table 4.** Similarity scores between the arguments retrieved by the translation-based and replica-based systems. The respective embedding models used to calculate a cosine similarity are given in parentheses.

| Metric | Score |
|---|---|
| ROUGE-1 | 0.257 |
| ROUGE-2 | 0.046 |
| Cos. sim. (sentence-transformers/LaBSE) | 0.404 |
| Cos. sim. (ai-forever/sbert_large_nlu_ru) | 0.635 |
| Cos. sim. (DeepPavlov/rubert-base-cased-sentence) | 0.656 |

possibly due to linguistic differences (Russian is a highly inflected language while English is weakly inflected). On the other hand, the argument quality-wise nDCG scores for translated arguments are lower than those for arguments retrieved in the original Russian language. This result can be highly correlated with the quality of the machine translation system.

### 4.4 Measuring Argument Similarity

To understand whether the arguments retrieved by the two systems are lexically and semantically similar and whether there is a need to replicate the whole system instead of translating, we calculate similarity scores between arguments retrieved by two systems for 78 comparison objects (from 39 pairs).

The scores are reported in Table 4. To calculate ROUGE-1 and ROUGE-2, we tokenize and lemmatize the arguments since Russian is a highly inflected language. The resulting ROUGE scores are relatively low (ROUGE-1 is 0.257 and ROUGE-2 is 0.046), indicating that the translated texts are lexically quite dissimilar to the texts retrieved by the RuCAM system.

To calculate cosine similarity scores, we use the three following sentence embedding models: (1) multilingual LaBSE [14],[11] (2) Russian BERT large uncased,[12] and (3) Sentence RuBERT,[13] where sentence representations are mean-pooled token embeddings analogous Sentence-BERT [32]. The highest mean cosine similarity score between different embeddings is 0.656, which is borderline, however, similarities might dramatically differ for the arguments from different pairs. For example, the arguments for the pairs "Chinese vs. Western medicine" have a mean similarity of 0.763, and for the "morning vs. afternoon sun", the mean score is 0.290 (examples for these two cases are in Tables 5 and 6).

We also looked at how pairs are ranked according to the ROUGE-1 and cosine similarity metrics of their arguments. First of all, our goal was to check to what extent two metrics are related when identifying similar arguments from two systems. We measured the Spearman's correlation coefficient between ROUGE-1 and cosine similarity, which showed a weak correlation of 0.371 ($p$-value = 0.02).

[11] https://huggingface.co/sentence-transformers/LaBSE
[12] https://huggingface.co/ai-forever/sbert_large_nlu_ru
[13] https://huggingface.co/DeepPavlov/rubert-base-cased-sentence

**Table 5.** Example arguments for the object pair 'Chinese medicine vs. Western medicine' that has the highest mean cosine similarity 0.763 among all the object pairs ('rubert-base-cased-sentence' embeddings). For the demonstration purpose, we translated Russian arguments into English using OPUS-MT.

| Translation-based | Replication-based (RuCAM) |
|---|---|
| The amazing thing is that with Traditional Chinese Medicine I always get better faster than all of my colleagues who are relying on Western medicine. | "I think more and more Western doctors are realizing today that Chinese medicine is effective," says Dr. Li. |
| Chinese medicine is superior to Western medicine. | Chinese medicine has outstripped Western medicine in some respects. |
| As for the treatment of Nephrotic syndrome, by large, Chinese medicine is superior to Western medicine. | In Chinese medicine, attention is paid to hidden factors, whereas Western medicine pays more attention to visible indicators. |
| What I am saying is Chinese medicine is a better method of healthcare than Western medicine. | In Chinese medicine, for example, kidneys are given much more attention than in Western medicine. |
| I am a firm believer that Chinese medicine is better than Western in many cases. | Chinese medicine has coped with what European medicine has not coped with. |

From Table 7, one can also see that pairs with the most similar and dissimilar arguments do not overlap much, especially between the top-10 object pairs: 'Chinese medicine vs. Western medicine', 'steel knives vs. ceramic knives', and 'Google vs. Yandex search'. According to cosine similarity, more general or common knowledge concepts get higher scores, while for the ROUGE-1 metric, top-10 similar arguments are for companies, brands, and specific topics like programming or medicine. Surprisingly, 'kids vs. adults', 'rain water vs. tap water', and 'skiing vs. snowboarding' appear at the top of cosine similarity scores but at the bottom of the list for the ROUGE-1 score. 'BMW vs. Audi', 'Kenya vs. Tanzania', and 'morning sun vs. afternoon sun' are object pairs that were shown to be different by both metrics. Secondly, our goal was to see, how similar were the arguments from two systems regarding both metrics. Manual analysis of the pairs that were scored differently by ROUGE-1 and cosine similarity showed that high ROUGE-1 scores indeed represent similar arguments, while low cosine similarity scores for those cases can be explained by the unequal number of arguments for each language that increases the impact of the outliers.

Thus, we conclude that a good approach for extending CAM is to combine the translation- and replica-based systems: the results show that the lexical similarity of the arguments from both systems is quite low, while the similarity according to semantic representations is borderline. We also analyzed the arguments to understand whether the dissimilarities could be explained by cultural differences present in the source languages. We identified the following main trends:

(a) The retrieved arguments address different aspects of the culture and everyday life of the source language speakers. For example, when comparing car

**Table 6.** Example arguments for the object pair 'morning sun vs. afternoon sun' that has the lowest mean cosine similarity 0.290 among all the object pairs ('rubert-base-cased-sentence' embeddings). For the demonstration purpose, we translated Russian arguments into English using OPUS-MT.

| Translation-based | Replication-based (RuCAM) |
|---|---|
| And remember: morning sun is cooler than afternoon sun. | Gerberas can be grown in full sun, but it is better in the morning sun and in the midday shade. |
| The morning sun is cooler and gentler than the afternoon hot sun. | The location is sunny, but the bright afternoon sun is less useful, shaded. |
| Morning sun is better than afternoon sun. | The morning sun is best with reflected light the rest of the time. |
| Early morning sun is better than late afternoon sun since the flowers last longer under cooler conditions. | Hot summer sun Many rhododendrons tolerate the morning sun better, although there are some species and varieties that do not tolerate the sun at all. |
| Experienced gardeners know it, morning sun is cooler than afternoon sun. | The morning sun is always preferable to the midday sun, which can burn plants. |

brands, English arguments tend to care more about *safety*, *engine capacities*, and *technology*, while Russian arguments pay attention to *price*, *repair costs*, *wear and tear*, and car *modifications* present on the Russian car market.

(b) Cultural bias occurs in both more specific and more generic comparisons. For instance, for the 'IELTS vs. TOEFL' object pair (more specific comparison), English arguments focus on *complexity* and the test's *specific features*, whereas Russian arguments mainly discuss the *certificate's recognition in other countries*. For the 'skiing vs. snowboarding' pair (more generic), English arguments discuss the *learning rate* and *complexity*, whereas Russian arguments care more about *adrenaline*, *safety* and which sport is *better for families*.

(c) However, for some more generic comparisons like 'football vs. basketball' or 'Western medicine vs. Chinese medicine', the arguments mostly compare the same aspects like *effectiveness*, *popularity*, and often express *personal preferences*.

The aforementioned examples highlight that the provenance of retrieved arguments (the language in particular) significantly influences their diversity and introduces potential cultural nuances. In the process of adapting the CAM system to a new language, meticulous consideration should be given to various facets, including the translation quality, the cultural predisposition inherent in the source language, and the preferences of the target users—whether they seek responses tailored to a specific language and culture or a more expansive overview. However, in general, a recommended strategy is to merge the outputs of translated arguments and arguments in the target language, thereby enhancing the topical coverage.

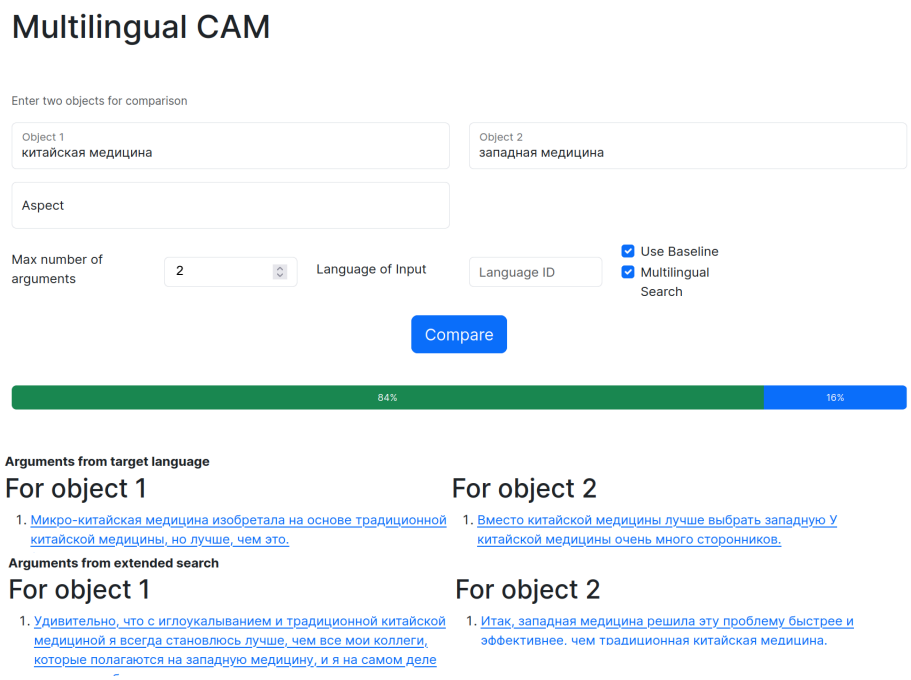
**Table 7.** Object pairs and similarity scores between the retrieved arguments by two systems: translation-based and replica-based. The object pairs are sorted in descending order of the similarity scores. Highlighted in green are the pairs that get high/low/medium scores by the two similarity metrics. Highlighted in red are the pairs showing discrepancies in two similarity metrics.

| Cosine Similarity | | ROUGE-1 | |
|---|---|---|---|
| Chinese vs. Western medicine | 0.763 | cow milk vs. goat milk | 0.393 |
| Apple vs. Google | 0.744 | rain water vs. tap water | 0.362 |
| PHP vs. Python | 0.736 | London vs. Paris | 0.335 |
| Linux vs. Windows | 0.732 | Chinese vs. Western medicine | 0.330 |
| artificial sweeteners vs. white sugar | 0.728 | skiing vs. snowboarding | 0.328 |
| steel knives vs. ceramic knives | 0.721 | kids vs. adults | 0.321 |
| Ibuprofen vs. Aspirin | 0.718 | steel knives vs. ceramic knives | 0.307 |
| hybrid vs. diesel | 0.701 | Google vs. Yahoo search | 0.302 |
| Google vs. Yahoo search | 0.700 | train vs. plane | 0.301 |
| OpenGL vs. Direct3D | 0.687 | Internet Explorer vs. Firefox | 0.292 |
| ASP vs. PHP | 0.677 | artificial sweeteners vs. white sugar | 0.287 |
| NetBeans vs. Eclipse | 0.674 | basketball vs. football | 0.287 |
| Xbox vs. PlayStation | 0.669 | cats vs. dogs | 0.286 |
| laptop vs. desktop | 0.661 | IELTS vs. TOEFL | 0.284 |
| Canon vs. Nikon | 0.650 | Apple vs. Google | 0.284 |
| electric stove vs. gas stove | 0.645 | electric stove vs. gas stove | 0.270 |
| IELTS vs. TOEFL | 0.643 | Ibuprofen vs. Aspirin | 0.265 |
| cow milk vs. goat milk | 0.626 | OpenGL vs. Direct3D | 0.251 |
| quicksort vs. merge sort | 0.621 | gas vs. charcoal | 0.249 |
| Family Guy vs. The Simpsons | 0.619 | Xbox vs. PlayStation | 0.249 |
| basketball vs. football | 0.608 | Linux vs. Windows | 0.246 |
| MAC vs. PC | 0.607 | pasta vs. pizza | 0.246 |
| Adidas vs. Nike | 0.596 | ASP vs. PHP | 0.235 |
| Ford vs. Toyota | 0.592 | hybrid vs. diesel | 0.229 |
| gas vs. charcoal | 0.592 | laptop vs. desktop | 0.224 |
| train vs. plane | 0.591 | PHP vs. Python | 0.223 |
| London vs. Paris | 0.590 | NetBeans vs. Eclipse | 0.211 |
| pasta vs. pizza | 0.586 | Python vs. R | 0.211 |
| Pepsi vs. Coca-cola | 0.585 | Boeing vs. Airbus | 0.211 |
| Internet Explorer vs. Firefox | 0.583 | MAC vs. PC | 0.208 |
| cats vs. dogs | 0.581 | Family Guy vs. The Simpsons | 0.203 |
| Boeing vs. Airbus | 0.573 | quicksort vs. merge sort | 0.188 |
| kids vs. adults | 0.567 | Ford vs. Toyota | 0.185 |
| Python vs. R | 0.561 | Adidas vs. Nike | 0.181 |
| rain water vs. tap water | 0.560 | Canon vs. Nikon | 0.175 |
| skiing vs. snowboarding | 0.559 | morning sun vs. afternoon sun | 0.168 |
| BMW vs. Audi | 0.528 | BMW vs. Audi | 0.154 |
| Kenya vs. Tanzania | 0.305 | Pepsi vs. Coca-cola | 0.151 |
| morning sun vs. afternoon sun | 0.290 | Kenya vs. Tanzania | 0.127 |

## 5 Multilingual CAM

To showcase the combined approach, we develop a demonstration of multilingual CAM that allows to search for arguments in English CAM and Russian RuCAM via their respective APIs. The interface of a combined system is shown in Figure 2. It accepts a pair of comparison objects and an optional comparison aspect in either language and retrieves arguments in both languages when the option 'Multilingual Search' is selected. Otherwise, the answers are provided in the input languages. Optionally, the user can specify the input language (e.g., when searching for "BMW" written in Latin script in the Russian texts); otherwise,

**Figure 2.** The multilingual CAM interface. Shown are the results for a comparison 'Chinese medicine vs. Western medicine' in Russian. The output combines arguments in the target language (upper part) and translated from English (lower part). The 'Chinese medicine'-object (left-hand) "wins" a comparison (see the bar in the middle).




the language is identified automatically. The output arguments are grouped into two blocks: those that come from the corpus of an input language and (in the case of the multilingual search option) translated from another language.

Our first prototype of multilingual CAM currently has several technical limitations. First, it depends on the successful response from the CAM or RuCAM API. Second, it relies on the machine translation module, which may incorrectly translate the user input, resulting in a failure to find relevant arguments. Therefore, future work should focus on overcoming the aforementioned shortcomings by locally hosting the retrieval corpora and deploying other translation models.

## 6 Conclusion

In this paper, we improved the answer stance detection of CAM and RuCAM—systems that can answer comparative questions in English and Russian—by fine-tuning RoBERTa-based models. Furthermore, we compared the replica-based RuCAM approach of "localizing" CAM to the Russian language to a simple machine translation-based CAM variant. Our analyses showed that translating

CAM's inputs and outputs also yields decent effectiveness scores with respect to result relevance and quality. However, we also found that the results retrieved by the two systems (translation-based and replica-based) are lexically and semantically quite dissimilar as, for instance, the Russian results from the replica-based RuCAM system that uses native Russian data can be more culture-specific and might take into account uncommon and unexpected aspects. Therefore, combining the results of the translation-based and of the replica-based CAM variants could yield more diverse arguments for comparisons.

As a demonstrator of a multilingual CAM system, we implemented an interface to combine the results of translation- and replica-based CAM systems. In a user study, we found that, for instance, the perceived result quality is highly dependent on the translation quality; in our study, translated results were perceived as more relevant but of a lower quality than the results retrieved in the target language—often also related to the translation quality of the actual search terms (comparison objects and aspects).

## Limitations

Our current work focused on two rather high-resource languages (English and Russian) so that our findings and conclusions may not be applicable to lower-resource languages. In future research, we thus plan to also analyze CAM-like systems in other languages.

Furthermore, our study results depend on two restricting factors: (1) the choice of the machine translation model, and (2) potential biases of the manual annotations. To alleviate the first factor, we relied on previous work and preferred publicly available translation models that can be easily deployed. As for the second factor, while annotation bias cannot be fully avoided, we ensured that our annotators understood and followed the guidelines by conducting pilot annotations with a follow-up discussion of possible misinterpretations. In the future, larger studies with a bigger group of human annotators are necessary for more robust conclusions.

Finally, CAM and RuCAM operate on large document collections, in which the amount of relevant data cannot be controlled or measured. To more closely study sociocultural questions in the context of comparison analyses, other more focussed collections might be better suited.

**Acknowledgements** This work has been partially supported by the DFG (German Research Foundation) through the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999).

## References

1. Allaway, E., McKeown, K.R.: Zero-shot stance detection: A dataset and model using generalized topic representations. In: Proceedings of the 2020 Conference on

Empirical Methods in Natural Language Processing, EMNLP 2020. pp. 8913–8931. ACL (2020). https://doi.org/10.18653/V1/2020.EMNLP-MAIN.717
2. Anand, P., Walker, M.A., Abbott, R., Tree, J.E.F., Bowmani, R., Minor, M.: Cats rule and dogs drool!: Classifying stance in online debate. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2011. pp. 1–9. ACL (2011), https://aclanthology.org/W11-1701/
3. Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., Slonim, N.: Stance classification of context-dependent claims. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. pp. 251–261. ACL (2017). https://doi.org/10.18653/V1/E17-1024
4. Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., Hagen, M.: Towards understanding and answering comparative questions. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM 2022. pp. 66–74. ACM (2022). https://doi.org/10.1145/3488560.3498534
5. Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument retrieval. In: Proceedings of the 11th International Conference of the CLEF Association, CLEF 2020. pp. 384–395. Springer (2020). https://doi.org/10.1007/978-3-030-58219-7_26
6. Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2022: Argument retrieval. In: Proceedings of the 13th International Conference of the CLEF Association, CLEF 2022. pp. 311–336. Springer (2022). https://doi.org/10.1007/978-3-031-13643-6_21
7. Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument retrieval. In: Proceedings of the 12th International Conference of the CLEF Association, CLEF 2021. pp. 450–467. Springer (2021). https://doi.org/10.1007/978-3-030-85251-1_28
8. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016. pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. pp. 8440–8451. ACL (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.747
10. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017. pp. 670–680. Association for Computational Linguistics (2017). https://doi.org/10.18653/V1/D17-1070
11. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
12. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, NAACL-HLT 2019. pp. 4171–4186. ACL (2019). https://doi.org/10.18653/V1/N19-1423
14. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022. pp. 878–891. ACL (2022). https://doi.org/10.18653/V1/2022.ACL-LONG.62
15. Hasan, K.S., Ng, V.: Stance classification of ideological debates: Data, models, features, and constraints. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013. pp. 1348–1356. AFNLP / ACL (2013), https://aclanthology.org/I13-1191/
16. Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., Inui, K.: Tohoku at SemEval-2016 Task 6: Feature-based model versus convolutional neural network for stance detection. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016. pp. 401–407. ACL (2016). https://doi.org/10.18653/V1/S16-1065
17. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002). https://doi.org/10.1145/582415.582418
18. Kang, I., Ruan, S., Ho, T., Lin, J.C., Mohsin, F., Seneviratne, O., Xia, L.: LLM-augmented preference learning from natural language. arXiv 2310.08523 (2023), https://doi.org/10.48550/arXiv.2310.08523
19. Küçük, D., Can, F.: Stance detection: A survey. ACM Comput. Surv. **53**(1), 12:1–12:37 (2021). https://doi.org/10.1145/3369026
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv 1907.11692 (2019), http://arxiv.org/abs/1907.11692
21. Ma, N., Mazumder, S., Wang, H., Liu, B.: Entity-aware dependency-based deep graph attention network for comparative preference classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. pp. 5782–5788. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.512
22. Maslova, M., Rebrikov, S., Artsishevski, A., Zaczek, S., Biemann, C., Nikishina, I.: RuCAM: Comparative Argumentative Machine for the Russian Language. In: Analysis of Images, Social Networks and Texts — 11th International Conference, AIST 2023. LNCS, vol. 14486. Springer (2024), https://link.springer.com/chapter/10.1007/978-3-031-54534-4_6
23. Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 Task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016. pp. 31–41. ACL (2016). https://doi.org/10.18653/V1/S16-1003
24. Murakami, A., Raymond, R.: Support or oppose? Classifying positions in online debates from reply activities and opinion expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010. pp. 869–875. CIPS (2010), https://aclanthology.org/C10-2100/
25. Nadamoto, A., Tanaka, K.: A comparative web browser (CWB) for browsing and comparing web pages. In: Proceedings of the Twelfth International World Wide Web Conference, WWW 2003. pp. 727–735. ACM (2003). https://doi.org/10.1145/775152.775254
26. OpenAI: Introducing ChatGPT (2022), https://openai.com/blog/chatgpt
27. OpenAI: GPT-4 technical report. ArXiv 2303.08774 (2023), https://doi.org/10.48550/arXiv.2303.08774

28. Ortiz Suárez, P.J., Sagot, B., Romary, L.: Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora, CMLC-7 2019. pp. 9–16. IDS (2019). https://doi.org/10.14618/ids-pub-9021
29. Panchenko, A., Bondarenko, A., Franzek, M., Hagen, M., Biemann, C.: Categorizing comparative sentences. In: Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019. pp. 136–145. ACL (2019). https://doi.org/10.18653/V1/W19-4516
30. Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S.P., Biemann, C.: Building a web-scale dependency-parsed corpus from common crawl. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018. ELRA (2018), http://www.lrec-conf.org/proceedings/lrec2018/summaries/215.html
31. Rajadesingan, A., Liu, H.: Identifying users with opposing opinions in twitter debates. In: Proceedings of the 7th International Conference, SBP 2014. LNCS, vol. 8393, pp. 153–160. Springer (2014). https://doi.org/10.1007/978-3-319-05579-4_19
32. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019. pp. 3980–3990. ACL (2019). https://doi.org/10.18653/V1/D19-1410
33. Schildwächter, M., Bondarenko, A., Zenker, J., Hagen, M., Biemann, C., Panchenko, A.: Answering comparative questions: Better than ten-blue-links? In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019. pp. 361–365. ACM (2019). https://doi.org/10.1145/3295750.3298916
34. Schiller, B., Daxenberger, J., Gurevych, I.: Stance detection benchmark: How robust is your stance detection? Künstliche Intell. **35**(3), 329–341 (2021). https://doi.org/10.1007/S13218-021-00714-W
35. Sun, J., Wang, X., Shen, D., Zeng, H., Chen, Z.: CWS: a comparative web search system. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006. pp. 467–476. ACM (2006). https://doi.org/10.1145/1135777.1135846
36. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006. pp. 327–335. ACL (2006), https://aclanthology.org/W06-1639/
37. Tiedemann, J., Thottingal, S.: OPUS-MT - building open translation services for the world. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020. pp. 479–480. European Association for Machine Translation (2020), https://aclanthology.org/2020.eamt-1.61/
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv 2302.13971 (2023), https://doi.org/10.48550/arXiv.2302.13971
39. Vamvas, J., Sennrich, R.: X-stance: A multilingual multi-target dataset for stance detection. In: Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020. CEUR Workshop Proceedings, vol. 2624. CEUR-WS.org (2020), https://ceur-ws.org/Vol-2624/paper9.pdf

40. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representations, ICLR 2018. OpenReview.net (2018), https://openreview.net/forum?id=rJXMpikCZ
41. Walker, M.A., Anand, P., Abbott, R., Grant, R.: Stance classification using dialogic properties of persuasion. In: Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics, NAACL-HLT 2012. pp. 592–596. ACL (2012), https://aclanthology.org/N12-1072/
42. Walker, M.A., Anand, P., Abbott, R., Tree, J.E.F., Martell, C.H., King, J.: That is your evidence?: Classifying stance in online political debate. Decis. Support Syst. **53**(4), 719–729 (2012). https://doi.org/10.1016/J.DSS.2012.05.032
43. Wei, W., Zhang, X., Liu, X., Chen, W., Wang, T.: pkudblab at SemEval-2016 Task 6 : A specific convolutional neural network system for effective stance detection. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016. pp. 384–388. ACL (2016). https://doi.org/10.18653/V1/S16-1062
44. Yu, N., Pan, D., Zhang, M., Fu, G.: Stance detection in Chinese microblogs with neural networks. In: Proceedings of the 5th Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016. LNCS, vol. 10102, pp. 893–900. Springer (2016). https://doi.org/10.1007/978-3-319-50496-4_83
45. Zarrella, G., Marsh, A.: MITRE at SemEval-2016 Task 6: Transfer learning for stance detection. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016. pp. 458–463. ACL (2016). https://doi.org/10.18653/V1/S16-1074