Computational Models of Argument
H. Prakken et al. (Eds.)

doi:10.3233/FAIA200505

Structure or Content? Towards Assessing Argument Relevance

Marc FEGER [1], Jan STEIMANN[1] and Christian METER
Computer Networks Department
Heinrich-Heine-University Düsseldorf, Germany
firstname.lastname@hhu.de

Abstract. In this paper, we provide a detailed analysis of PageRank to determine the relevance of arguments along with content- and knowledge-based methods from the field of natural language processing. We do not only show how the cross-linking of arguments is only slightly involved in the recognition of relevance, we rather show how basic common knowledge and reader-involving methods outperform the purely structure-related PageRank. The methods we propose are based on the latest research and correlate strongly with human awareness regarding the relevance of arguments. Altogether, we show that PageRank does not fully capture the relevance of arguments and must be extended by a contextual level in order to take concepts of natural language into account at the web level, as they are unavoidably involved in argumentation.

Keywords. Argument relevance, Natural language processing, Argumentation and computational linguistics, Computational properties of argumentation, Argumentation and human-computer interaction

1. Introduction

What would you like to search for? What would you like to know? Whether simply surfing the web, doing research, shopping online, or having a discussion about if the new Star Wars film is a success, these questions are always present in our everyday lives. It is not surprising, especially in times when everyday life is more and more transferred to the digital space, that this space also adapts to the needs of the users. The web is becoming a public sphere in which opinions are sometimes shaped by objective and rational debates [1]. As a result, more and more projects like [2,3] are concerned with the systematic preparation and analysis of arguments within this digital space. The sheer unmanageable amount of data also increases the need to filter the most relevant information. For argumentation this means, similar to web documents, the embedded arguments have to be evaluated with respect to their relevance. [4] already suggested a modified version of PageRank by [5], which is supposed to abstract the relevance of arguments, similar to web pages, via their networking, possibly within the web, purely objectively and without consideration of content and logical aspects. As an example, the selection of arguments might look like the following (taken from the dataset used in [4]):

[1] Both authors contributed in equal parts to this work.

Question: *Why should peanuts be banned on board aircraft?*

Answer 1 (a_1)**:** *Peanut reactions can be life threatening. An individual doesn't have to consume the product to have a life threatening reaction. They can have contact or inhalation reactions.*

Answer 2 (a_2)**:** *Providing buffer zones to avoid contact with peanuts is a thoughtful gesture. But from a practical point of view, it does not work.*

Answer 3 (a_3)**:** *With so many food choices available, why are peanuts a necessary choice?*

Answer 4 (a_4)**:** *Restricting the ban of peanut products to certain flights is not enough.*

Although all answers take up the topic of the question, they are still of varying relevance. Compared to a_2, a_1 is more topic-related, informative and meaningful, as it addresses not only direct dangers but also implicit knowledge such as the restricted space inside aircraft. However, the comparison between a_2 and a_3 is highly subjective. Nonetheless, it can be assumed that a_2 is more relevant than a_3, since it does not contain a question and is therefore more concrete. Clearly, despite its formulation as a question, a_3 appears more informative and more concrete than a_4, since it contains a general conclusion.

In this paper we use the Webis-ArgRank-2017 data by [4] (Section 3) to investigate the impact of content- and knowledge-based methods (Section 4) on perception regarding the relevance of arguments. For this purpose the results of the pioneering work of [4] were reproduced. Besides the influence of PageRank on the relevance of the arguments, we compare them with the results obtained by using more recent and knowledge-based methods (Section 5). Our results show that PageRank and especially the evaluation of relevance exclusively by linking up arguments is not yet satisfactory. Rather, we show that simple content-based methods working with a general conceptual knowledge can achieve significantly better results (Section 6).

Consequently, our work takes up the thesis of [4] suggesting that relevance is structurally induced, and demonstrates how content-based properties co-determine inevitably the relevance of arguments, as these are unavoidably taken into account by a rational reader.

2. Related Work

In argumentation theory, relevance is considered in two parts. Local relevance describes the extent to which the premises of an argument contribute to the acceptance or rejection of the corresponding conclusion [6]. In contrast, global relevance describes the extent to which the argument contributes to the understanding of a topic [6]. [7] examined the influence of trust in an argument. Subsequently, supported by [8], it was stated that a globally relevant argument must necessarily also be locally relevant. However, a locally relevant argument need not necessarily be globally relevant. Despite the differences between the two dimensions, no sharp distinction is made in this work. Rather, first investigations of different methods are carried out to classify the methods proposed by [4]. Moreover, the relevance of an argument must be understood as a dimension of the quality of the argument itself. [9] proposed a tautological division of quality into the three main components: cogency, effectiveness and reasonableness. Whereby the two forms of relevance appear in the dimensions cogency and reasonableness. In addition, [9] notes that these dimensions can be interdependent and also branch out into sub-dimensions. Thus, global relevance is represented as a branch of reasonableness and local relevance as another branch of cogency. Reasonableness describes the extent to which an argument contributes to the understanding of a problem. Furthermore, cogency describes the extent to which the premise relevance contributes to a coherent understanding of the conclusion of an argument. Furthermore, the dependence of reasonableness on cogency is not only supported by the correlation of local and global relevance. The dependence of reasonableness and cogency is in line with the observation of [7] and [8].

In practice, a cornerstone of argument mining is understood to be the work written by [10] dealing with the collection of arguments within documents. In this context, [11] introduced the topic of argument recognition. Comments and the arguments contained therein are recognized by assigning arguments from a predefined set to the corresponding comment using similarity- and entailment-based properties. In addition, [12] identified prominent arguments in online debates by clustering using semantic- and text-based methods. Similarly, [13] used SVM and BLSTM with GloVe input layer to investigate whether the persuasiveness of an argument can be systematically captured. It was found that PageRank does not induce the persuasiveness of arguments and that a higher PageRank is associated with a lower convincingness. Likewise, [14] sketched the PageRank for capturing the global relevance of arguments. This sketch was performed by [4]. In doing so, the modified PageRank method beats a variety of intuitive comparison methods. These comparison methods covered the topics: semantics, similarity and graph structure. In addition, [15] achieved significant results regarding the finding of good counterarguments by using word- and embedding-based methods. Contiguous to this, [16] carried out an analysis of the relevance of arguments using four basic information retrieval methods: DirichletLM, DPH, BM24 and TFIDF. The results showed how the more modern methods performed significantly better and were able to capture the correlation between the relevance and the general quality of an argument.

3. Corpus

We conduct our study on the Webis-ArgRank-2017 dataset. In this dataset [4] constructed a large ground-truth argument graph as well as a ranking of a subset of arguments within this graph. This dataset serves as a first benchmark for evaluating argument relevance assessments. [4] acquired the data for constructing the graph from the Argument Web [17] storing the arguments in the Argument Interchange Format [18]. On June 2, 2016, when [4] accessed the data, the Argument Web was the largest existing argument database with structured argument corpora. It consisted of 57 corpora with 8479 argument-maps storing all information about the arguments, summing up to 49.504 argument units, describing either a premise or a conclusion, and 26.012 arguments. Duplicates and nodes which were not connected to any inference node were removed by the authors. This lead to the resulting graph with 31.080 different argument units of which 28.795 participated in 17.877 arguments. Altogether the arguments can be combined to a not necessarily coherent graph $G = (A, E)$. Each node $a_i \in A$ of the graph G describes an argument consisting of a conclusion c_i and a non-empty set of premises P_i. Thus, an argument $a_i \in A$ is represented with $a_i = \langle c_i, P_i \rangle$. An edge $(a_j, a_i) \in E \subseteq A \times A$ is given if the conclusion of a_j is used as a premise of a_i. Consequently, $P_i = \{c_1, \cdots, c_k\}, k \geq 1$.

In our work we were able to reproduce the resulting argument graph and numbers which [4] stated out in their paper. The authors made the 28.795 argument units and all following data available. We found slightly different numbers for the ground-truth-argument graph and the argument relevance benchmark dataset. Thus, [4] found 17.372 of all 31.080 argument units never to be used as a conclusion. Whereas our reproduction showed the same circumstance for 17.370 argument units. Building on this we came up with 17.096 argument units while [4] findings showed 17.093 argument to be used only once as a premise. Nevertheless, we consider the differences to be negligible since the difference is too small to have a deeper impact to a general assumption.

3.1. Benchmark Argument Ranking

In the constructed graph 3113 conclusions were part of more than one argument. Therefore, they were candidates for ranking. [4] selected 498 conclusions to be classified by two experts from computational linguistics. These experts decided for each conclusion if it contains a real claim or, e.g., if it has a personal context. Only if both experts saw a real claim the arguments has been kept. The remaining arguments were examined whether they allowed a logical inference to be drawn, if they form a valid counter-argument or if they were based on reasonable premises. The resulting benchmark dataset consists of 32 conclusions which participate in 110 arguments. These 110 arguments were then ranked by seven experts from computational linguistics and information retrieval. Each argument was ranked by how much each of its premises contributes to the acceptance or rejection of the conclusion. In terms of Kendall's τ the mean over all agreement was 0.36. The authors explain this low agreement with the general high subjectivity of argument relevance.

4. Methods

Basically, the original form of PageRank, as applied by [5], is based on the popularity of cross-linked websites. Accordingly, the relevance of a website depends on how many other relevant websites offer direct linking to this page. Furthermore, this procedure can be transferred to the argumentation graph G by replacing the web pages with argument units representing either a premise or a conclusion. Thus, the relevance of an argument, indicated by its conclusion, at a certain point in time t results from its premises and their interconnection according to G:

$$p_t(c_i) = \begin{cases} (1-\alpha)\frac{1}{|D|} + \alpha \sum_j \frac{p_{t-1}(c_j)}{|P'_j|} & : t > 0 \\ \frac{1}{|D|} & : t = 0 \end{cases} \tag{1}$$

For the initialization it should be considered that each argument is assigned the same relevance $\frac{1}{|D|}$. $|D|$ describes the number of all unique argument units in G. For each subsequent point in time $t > 0$ the relevance results from the α-weighted sum of the ground relevance $\frac{1}{|D|}$ and the linking relevance $\sum_j \frac{p_{t-1}(c_j)}{|P'_j|}$. The linking relevance reflects the importance of those arguments a_j whose conclusion is used as premise by a_i. If the conclusion c_j of a_j is used in $|P'_j|$ cases as a premise of further arguments, its relevance $p_{t-1}(c_j)$ is distributed accordingly. In addition to the *custom-made PageRank* (CPR) developed in this paper, we use the implementation of NetworkX [19] and its extension via Scipy [20] for control purposes.

4.1. Baselines Applied

Just like the baselines presented by [4], our approaches emphasize the collaboration of the premises for the respective conclusion. However, our approaches differ, not only because they user newer methods, but also because they take up aspects of content and language. For example, the *Similarity* used by [4] is intuitive, but it only covers strict word similarity. Furthermore, the *Frequency* of a premise across the arguments is used to measure relevance. Although both methods are intuitively calculable, they do not really match the judgement of a human reader of an argument. On the other hand, the methods we use are more focused on the reader and the way in which the viewer perceives an argument. Therefore, our methods take up concepts and linguistic constructs which are superior to the text as they occur in natural language. By deliberately emphasizing the dependencies within an argument, we want to tighten the baselines for the PageRank approach introduced by [4]. This is necessary because PageRank alone, through structural and without content aspects, is supposed to induce relevance in a linguistic environment, as it is the case in a debate. To make our work comparable, we adopted the intuitive baselines *Similarity*, and *Sentiment* and oriented our methods accordingly. Corresponding, the new baselines used in this paper are listed[2].

4.2. Similarity

An important aspect when comparing non-content-based methods like PageRank is the collection of content-related aspects of the data. Therefore, our methods address the so-called *semantic similarity*. Conversely, this means that the components of an argument at the word or sentence level are transferred into a corresponding vector representation. In this paper we have used Flair [21] to calculate the respective embedding in vectors.

4.3. Vector Space Models

In total, we have investigated three different ways to embed words or sentences within this work. We used GloVe [22] as a first vector representation for our model. This method produces unsupervised vector-space representations of words. Thereby a global word-word-co-occurrence statistic is learned. This kind of learning is based on linear relationships of words which correspond to a semantic similarity. Thus, vector relations as given by [23] can be described using a corresponding aggregation with, e.g., $king - man + women \approx queen$. In this paper both *GloVe with punctuation* (GWP) and *GloVe without punctuation* (GWOP) were investigated. ELMo [24] was used as the second method. Here, the embeddings are learned on the basis of a bidirectional language model. Thus, linguistic contextual properties are learned in addition to syntactic and semantic properties of the words. Apart from semantic analysis, these can also be used to answer questions and the associated textual inference. Thus, it can be determined to

[2] The code is located at: `https://github.com/hhucn/argument-relevance-paper-results`.

what extent a premise indicates a logical conclusion. Similar to GloVe, both *ELMo with punctuation* (EWP) and *ELMo without punctuation* (EWOP) were used. The third approach used was BERT [25]. This procedure also develops embeddings via bidirectional language models. However, newer techniques such as transformers etc. are added, making the embeddings given even more detailed. Similar to the previous models, we used *BERT with punctuation* (BWP) and *BERT without punctuation* (BWOP). To determine the resemblance, the Cosine-Similarity was used for each of the models mentioned, since this provided by far the most favorable results.

4.3.1. WordNet

Additionally, we have used the knowledge-based similarity function $Sim(T_1, T_2)$ introduced by [26]. This method determines the semantic similarity of two input texts T_1, T_2 by mutually picking up similar concepts. Vice-versa each individual word w of one text is compared, over the highest conceptual similarity, with the entire word concepts of the others by using the weighting $idf(w)$. Analogously, we considered in this paper a weakened variant limited to similarity from T_1 to T_2 via the average conceptual similarity across the words of T_1 to the totality of word concepts occurring in T_2. For both the *mutual knowledge-based method* (MKBM) and for the *average knowledge-based method* (AKBM), the implementation of the thesaurus WordNet [27] by NLTK [28] was used to identify the word concepts. In this thesaurus words are connected with respect to their synonyms in the form of synsets. To determine the similarity of words, the Wu-Palmer-Similarity $CoSim(c_1, c_2)$ [29] was used, which takes into account the depth of the concepts c_1, c_2 to be compared as well as the least common ancestor of both. We have made use of this similarity because, analogous to a family tree, it picks up the origin of two concepts and thus connects them to superordinate knowledge.

4.4. Sentiment

Just like [4], we have taken up the subject of sentiment, as it can certainly contribute to the persuasive power of an argument. Sentiment uses the positive tone of the premises to calculate a score for the argument. Unlike [4], a *sentiment neuronal network* (SNN) based on FastText [30], which was trained on the film ratings of IMDb, was used instead of SentiWordNet [31]. The advantage of this model architecture is its speed and simplicity. Whereby it merely consists of an input layer, which then passes the averaged feature vectors, using GloVe embeddings, to a linear classifier.

5. Results

In the subsequent section, besides the detailed analysis regarding the different implementations of PageRank and their dependency on α, the baselines given by [4] are presented in comparison with our results. It should be mentioned that the relevance of a conclusion can always be derived from the premises. For this purpose, the four different aggregations of the premise values *min*, *average*, *max*, *sum* are listed, resulting in an overall relevance of the conclusion. The relevance of *min* and *max* is determined by the smallest and the largest value of the premises. Similarly, the *sum* and *average* are used to determine the relevance of the entirety of the premises.

5.1. PageRank Comparison

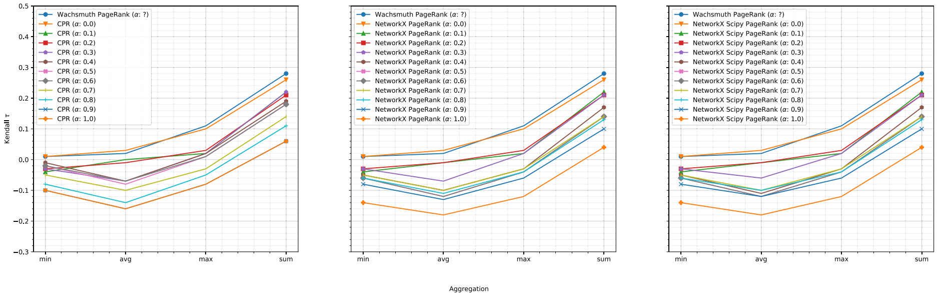

Figure 1. Development of the perception regarding the argument relevance induced by PageRank regarding all possible aggregations. CPR, NetworkX and NetworkX using Scipy were plotted against the result obtained by [4] for different α values, which regulates the influence of linking the arguments.

Since PageRank in its modified and unmodified variants always operates structurally and not content-related, a more in-depth investigation of the method as such is necessary. In the context of this, a more precise illumination of the influence of the parameter α on the supposedly induced relevance is also required. Figure 1 shows the results of the different implementations of PageRank on the argumentation graph G. The results were plotted according to the aggregations against the results obtained by [4].

As a first observation, it can be stated that the different implementations for $\alpha = 0$ achieve comparable results and, moreover, can easily keep up with the variant presented by [4]. Due to the different setups and sometimes different implementation details, slightly different results regarding the achieved τ values per aggregation are obtained. Furthermore, it can be seen that for an increasing α, which is accompanied by a higher influence of the cross-linking of arguments, the general agreement regarding relevance decreases over each aggregation, resulting in a low point for $\alpha = 1$.

5.2. Extended Baselines

As one part of this paper the baselines given by [4] were reproduced: *Random*, *Most premises*, *Frequency*, *Sentiment*, *Similarity* and *PageRank*. Their results were plotted in Figure 2 against the values obtained by the methods shown in Section 4. In contrast to the thesis stated by [4] according to which PageRank induces relevance better than frequency-based and simple content-based methods, it can be seen how methods that emphasize linguistic aspects through content-related and contextual properties achieve significantly better results in the assessment of relevance. The results achieved by using WordNet and GloVe are particularly striking.

Table 1 shows the direct numerical comparison. Altogether GWP performs with $\tau \approx 0.47$. Moreover, AKBM achieves a correlation of $\tau \approx 0.4$ and MKBM of $\tau \approx 0.34$. Likewise, the rating of the SNN for determining sentiment is $\tau \approx 0.31$. In contrast, BWP and BWOP achieve only marginally lower results than PageRank, whereas EWP performs $\tau \approx 0.28$. Therefore, results similar to PageRank are achieved. Despite the strong subjectivity of the task, the approaches mentioned above perform rather strongly compared to the average agreement among all annotators of $\tau \approx 0.36$. Similarly, GWP performs best in 16 out of 32 cases and worst in only 1 case. Not only does this result out-

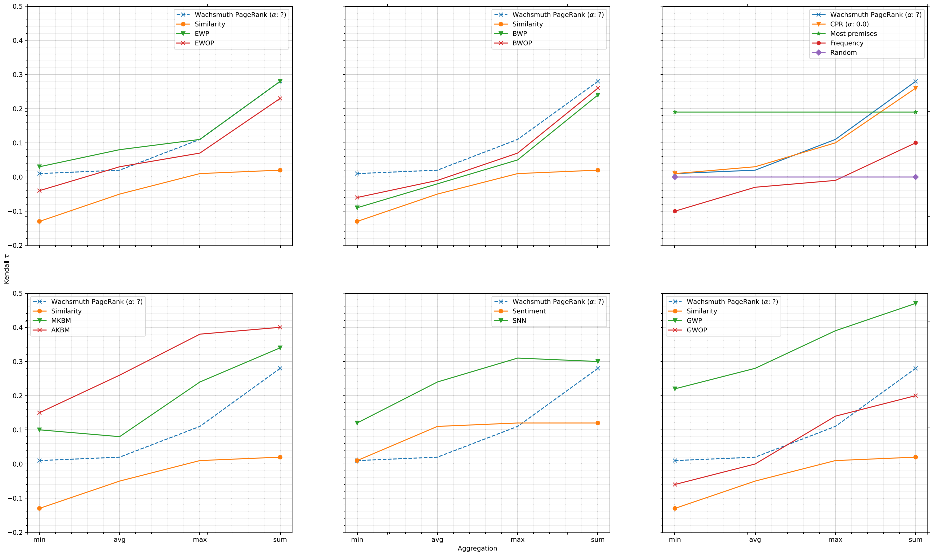

Figure 2. Direct comparison of the awareness regarding the relevance of arguments of all baseline values reported by [4] with all results obtained in this paper. *Most premises* and *Random*, which achieved slightly different values due to unequal random procedures, were adopted across all aggregations, as these can only be applied in *sum*.

perform the results of the modified PageRank reported by us, which is better in 11 cases and worse in 5 cases, but it also exceeds the results reported by [4], with 15 best and 3 worst. Furthermore, AKBM comes second with 14 best and 4 worst results followed by MKBM with 13 best and 6 worst cases. Also, SNN is slightly better than MKBM with 13 best and 5 bad cases. Likewise, EWP, EWOP, BWP and BWOP are similar to the PageRank results we previously reported.

#	Approach	(a) Minimum			(b) Average			(c) Maximum			(d) Sum			(e) Best results		
		τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*	τ	*best*	*worst*
1	PageRank	0.01	8	6	0.02	9	7	0.11	8	6	0.28	11	5	0.28	11	5
2	Frequency	-0.10	2	8	-0.03	3	9	-0.01	2	8	0.10	6	8	0.10	6	8
3	Similarity	-0.13	4	11	-0.05	5	11	0.01	6	10	0.02	6	10	0.02	6	10
4	Sentiment	0.01	6	7	0.11	9	4	0.12	6	4	0.12	9	4	0.12	9	4
5	Most premises	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.19	3	3	0.19	3	3
6	Random	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.00	5	7	0.00	5	7
7	SNN	0.12	10	6	0.24	11	5	0.31	12	5	0.30	13	5	0.31	13	5
8	GWP	**0.22**	12	5	**0.28**	13	**3**	**0.39**	**14**	**2**	**0.47**	**16**	**1**	**0.47**	**16**	**1**
9	GWOP	-0.06	5	9	0.00	6	7	0.14	8	6	0.20	8	4	0.20	8	4
10	EWP	0.03	6	9	0.08	7	8	0.11	8	8	0.28	9	5	0.28	9	5
11	EWOP	-0.04	5	9	0.03	6	8	0.07	7	8	0.23	9	6	0.23	9	6
12	BWP	-0.09	6	9	-0.02	7	8	0.05	9	8	0.24	10	5	0.24	10	5
13	BWOP	-0.06	6	9	-0.01	7	8	0.07	9	8	0.26	10	5	0.26	10	5
14	MKBM	0.10	5	7	0.08	13	6	0.24	12	8	0.34	11	9	0.34	13	6
15	AKBM	0.15	**14**	**4**	0.26	**14**	4	0.38	11	7	0.40	13	7	0.40	14	4

Table 1. Comparison of the approaches of [4] (1-6) with those used in this study (7-15). For each aggregation (a-d) the average agreement τ and the cases in which the respective approach performed best or worst over the 32 conclusions of the 110 arguments are given. (e) shows the best results of an aggregation.

6. Interpretation

Based on these results, we can embed the PageRank for the purpose of argument relevance into the current state of research. In order to achieve a better comparability of the results and procedures, we have divided these underlying methods into two categories. Thus, the vector space models working with ELMo and BERT belong to the group of direct contextual methods, which require expert knowledge of the discussion and its language usage. In this group we also include PageRank, because it includes a structural context. On the other hand, we consider those approaches working with WordNet, GloVe and Sentiment to be part of the group of indirect contextual methods. This does not mean that the mentioned methods of this group miss the underlying context completely. WordNet and GloVe, for example, take up linguistic similarity as well as higher-level concepts and work with them in the context of local evaluation of arguments. Likewise, the positive sentiment appeals to such a local context and thus emphasizes respectful interaction and the constructive effect resulting from this.

As a first finding, the methods of the directly context-related methods obtain comparable results. In some cases, PageRank even outperforms vector space models using ELMo and BERT, which could be due to the fact that the training data did not contain sufficient argumentative data, which occasionally also needs to be included in the ground-truth. However, the overall course of the results is mostly identical, indicating that similar underlying properties have been captured. We assume that the PageRank given by [4] used a α-weighting around $\alpha \approx 0$, since CPR achieved the highest agreement for $\alpha = 0$. Thus, the actual advantage of PageRank, which is to determine relevance through structure, is nearly absolutely lost, since the linking relevance is only included very poorly in the computation. Therefore, based on the results of CPR, none or at least very small portions of the structure underlying the argumentation are included in the assessment of argument relevance. The fact that the results of these methods perform similarly well accentuates the quality of the PageRank as well as the complexity of the task due to the diversity of approaches.

The methods of the indirect context-related group behave differently. Thus, all the methods mentioned perform significantly higher in terms of the awareness of relevance. GWOP, for example, without considering sentence structures, is comparable to the previous results of the context-related methods. However, GWP clearly stands out from all results by using sentence structures. The same behavior is observed for MKBM, AKBM and SNN. The better performance of those methods using WordNet and GloVe with respect to the use of sentence structures can be attributed to the local approach. Besides punctuation, both methods only consider word analogies that result from the local context. Thus, the sentence structure for the WordNet methods is taken up through idf, since the premises are considered in sentences where each sentence represents a document. For GloVe, word similarities result from their local combinations and for the WordNet methods from the synergy of the superordinate concepts. Furthermore, the positivity for Sentiment is restricted to the local context. Thus, it is not surprising that the best result for the aggregation *max* is achieved. This corresponds to the view that the most constructive premise contributes to the relevance of the argument. Overall, there seems to be an advantage of those methods which take the local context into account and thus emphasize the local relevance much better, as this reflects the reader of an argument better.

7. Conclusion

In this paper, we have shown how already existing content- and knowledge-based methods clearly exceed the PageRank as modified by [4] for determining the relevance of arguments. We were also the first to transfer the latest approaches from the field of natural language processing to the assessment of argument relevance. Additionally, we were able to embed the modified PageRank into the current state of research. We provide evidence that superordinate knowledge and concepts of natural language are more important for relevance than structural methods like PageRank, because they are more likely to be involved in convincingness. Thus, the observation of [13] suggesting counter-productive effects of PageRank on convincingness can be confirmed. Nevertheless, the PageRank should still be investigated, as it can achieve meaningful results despite its low degree of interconnectedness. Even if the properties of the arguments can be precalculated by the presented methods, their scalability on the web should still be investigated in more detail. The precalculation allows the properties to be uniquely assigned to an argument. Thus, the scaleability in the web is not limited by expensive calculations. Therefore, the presented methods could possibly keep up with the already well investigated scalability of PageRank, which also involves a precalculation phase. We therefore propose to combine content- and knowledge-based approaches with structure-emphasizing methods similar to the Hummingbird [32] algorithm used by Google, which replaced PageRank in 2013 and only partially integrates it into search queries. We are looking forward to jointly solve the existing problems and thereby paving the way for search engines to consider arguments and especially their relevance.

References

[1] D. Rasmussen, J. Habermas, C. Lenhardt, and S. Nicholsen, “Moral consciousness and communicative action.” *The Philosophical Quarterly*, vol. 43, no. 173, p. 571, 10 1993. [Online]. Available: https://doi.org/10.2307/2220013

[2] C. Meter and A. Schneider, “Various Efforts of Enhancing Real World Online Discussions,” in *ECA 2019: Proceedings of the 3rd European Conference on Argumentation*, June 2019.

[3] T. Krauthoff, C. Meter, M. Baurmann, G. Betz, and M. Mauve, “D-BAS âĂŞ A Dialog-Based Online Argumentation System,” in *Computational Models of Argument*, September 2018, pp. 325–336. [Online]. Available: http://doi.org/10.3233/978-1-61499-906-5-325

[4] H. Wachsmuth, B. Stein, and Y. Ajjour, “"PageRank" for argument relevance,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, pp. 1117–1127. [Online]. Available: http://aclweb.org/anthology/E17-1105

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[6] D. Walton, “Informal logic: A pragmatic approach, second edition,” *Informal Logic: A Pragmatic Approach, Second Edition*, pp. 1–347, 01 2008. [Online]. Available: https://doi.org/10.1017/CBO9780511808630

[7] F. Paglieri and C. Castelfranchi, “Trust, relevance, and arguments,” *Argument and Computation*, vol. 5, pp. 216–236, 2014. [Online]. Available: https://doi.org/10.1080/19462166.2014.899270

[8] C. Jacobs and S. Jackson, “Relevance and digressions in argumentative discussion: A pragmatic approach,” *Argumentation*, vol. 6, no. 2, pp. 161–176, 05 1992. [Online]. Available: https://doi.org/10.1007/BF00154323

[9] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein, “Computational argumentation quality assessment in natural language,” in *Proceedings of the 15th*

Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 04 2017, pp. 176–187. [Online]. Available: https://doi.org/10.18653/v1/E17-1017

[10] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed, "Automatic detection of arguments in legal texts," in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ser. ICAIL âĂŹ07. New York, NY, USA: Association for Computing Machinery, 2007, p. 225âĂŞ230. [Online]. Available: https://doi.org/10.1145/1276318.1276362

[11] F. Boltužić and J. Šnajder, "Back up your stance: Recognizing arguments in online discussions," in *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, 06 2014, pp. 49–58. [Online]. Available: https://doi.org/10.3115/v1/W14-2107

[12] ——, "Identifying prominent arguments in online debates using semantic textual similarity," in *Proceedings of the 2nd Workshop on Argumentation Mining*. Denver, CO: Association for Computational Linguistics, 06 2015, pp. 110–115. [Online]. Available: https://www.aclweb.org/anthology/W15-0514

[13] I. Habernal and I. Gurevych, "Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 08 2016, pp. 1589–1599. [Online]. Available: https://doi.org/10.18653/v1/P16-1150

[14] K. Al-Khatib, H. Wachsmuth, M. Hagen, J. Köhler, and B. Stein, "Cross-domain mining of argumentative text through distant supervision," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 06 2016, pp. 1395–1404. [Online]. Available: https://doi.org/10.18653/v1/N16-1165

[15] H. Wachsmuth, S. Syed, and B. Stein, "Retrieval of the best counterargument without prior topic knowledge," in *Proceedings of the 56th Annual Meeting of the Association for Computational*. Melbourne, Australia: Association for Computational Linguistics, 07 2018, pp. 241–251. [Online]. Available: https://doi.org/10.18653/v1/P18-1023

[16] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, and M. Hagen, "Argument search: Assessing argument relevance," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIRâĂŹ19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1117âĂŞ1120. [Online]. Available: https://doi.org/10.1145/3331184.3331327

[17] F. Bex, J. Lawrence, M. Snaith, and C. Reed, "Implementing the argument web," *Commun. ACM*, vol. 56, no. 10, p. 66âĂŞ73, Oct. 2013. [Online]. Available: https://doi.org/10.1145/2500891

[18] C. Chesnevar, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, S. Willmott *et al.*, "Towards an argument interchange format," *Knowl. Eng. Rev.*, vol. 21, no. 4, p. 293âĂŞ316, Dec. 2006. [Online]. Available: https://doi.org/10.1017/S0269888906001044

[19] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds., Pasadena, CA USA, 2008, pp. 11 – 15. [Online]. Available: http://conference.scipy.org/proceedings/SciPy2008/paper_2/full_text.pdf

[20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," pp. 261–272, 2020. [Online]. Available: https://doi.org/10.1038/s41592-019-0686-2

[21] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 08 2018, pp. 1638–1649. [Online]. Available: https://www.aclweb.org/anthology/C18-1139

[22] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Doha, Qatar: Association for Computational Linguistics, 10 2014, pp. 1532–1543. [Online]. Available: https://doi.org/10.3115/v1/D14-1162

[23] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPSâĂŹ16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 4356âĂŞ4364. [Online]. Available: http://arxiv.org/abs/1607.06520

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1802.05365

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[26] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, ser. AAAIâĂŹ06. AAAI Press, 2006, p. 775âĂŞ780.

[27] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 05 1998, isbn: 9780262061971.

[28] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O’REILLY, 01 2009, isbn: 978-0-596-51649-9.

[29] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” USA, p. 133âĂŞ138, 1994. [Online]. Available: https://doi.org/10.3115/981732.981751

[30] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, 04 2017, pp. 427–431. [Online]. Available: https://doi.org/10.18653/v1/E17-2068

[31] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA), 05 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[32] A. P. D. R. Yazdanifard, “How googleâĂŹs new algorithm, hummingbird, promotes content and inbound marketing,” *American Journal of Industrial and Business Management*, vol. 4, pp. 51–57, 01 2014. [Online]. Available: https://doi.org/10.4236/ajibm.2014.41009