Identifying Argument Components through TextRank

Georgios Petasis and **Vangelis Karkaletsis**
Software and Knowledge Engineering Laboratory
Institute of Informatics & Telecommunications
National Center for Scientific Research (N.C.S.R.) "Demokritos"
Athens, Greece.
{petasis,vangelis}@iit.demokritos.gr

Abstract

In this paper we examine the application of an unsupervised extractive summarisation algorithm, TextRank, on a different task, the identification of argumentative components. Our main motivation is to examine whether there is any potential overlap between extractive summarisation and argument mining, and whether approaches used in summarisation (which typically model a document as a whole) can have a positive effect on tasks of argument mining. Evaluation has been performed on two corpora containing user posts from an on-line debating forum and persuasive essays. Evaluation results suggest that graph-based approaches and approaches targeting extractive summarisation can have a positive effect on tasks related to argument mining.

1 Introduction

Argumentation is a branch of philosophy that studies the act or process of forming reasons and of drawing conclusions in the context of a discussion, dialogue, or conversation. Being an important element of human communication, its use is very frequent in texts, as a means to convey meaning to the reader. As a result, argumentation has attracted significant research focus from many disciplines, ranging from philosophy to artificial intelligence. Central to argumentation is the notion of argument, which according to (Besnard and Hunter, 2008) is a set of assumptions (i.e. information from which conclusions can be drawn), together with a conclusion that can be obtained by one or more reasoning steps (i.e. steps of deduction). The conclusion of the argument is often called the claim, or equivalently the consequent

or the conclusion of the argument. The assumptions are called the support, or equivalently the premises of the argument, which provide the reason (or equivalently the justification) for the claim of the argument. The process of extracting conclusions/claims along with their supporting premises, both of which compose an argument, is known as argument mining (Goudas et al., 2015; Goudas et al., 2014) and constitutes an emerging research field.

Several approaches have been already presented for addressing various subtasks of argument mining, including the identification of argumentative sentences (i.e. sentences that contain argumentation components such as claims and premises), argumentation components, relations between such components, and resources for supporting argument mining, like discourse indicators and other expressions indicating the presence of argumentative components. Proposed methods mostly relate to supervised machine learning exploiting a plethora of features (Goudas et al., 2015), including the combination of several techniques, such as the work presented in (Lawrence and Reed, 2015).

One of the difficulties associated to argument mining relates to the fact that the identification of argument components usually depends on the context in which they appear in. The locality of this context can vary significantly, based not only on the domain, but possibly even to personal writing style. On one hand, discourse indicators, markers and phrases can provide a strong and localised contextual information, but their use is not very frequent (Lawrence and Reed, 2015). On the other hand, the local context of a phrase may indicate that the phrase is a fact, suggesting low or no *argumentativeness* at all, while at the same time, the same phrase may contradict to another phrase several sentences before or after the phrase in question, constituting the phrase under ques-

94

tion an argumentative component (Carstens and Toni, 2015). While it is quite easy to handle local context through suitable representations and learning techniques, complexity may increase significantly when a broader context is required, especially when relations exist among various parts of a document.

In this paper we want to examine approaches that are able to handle interactions and relations that are not local, especially the ones that can model a document as a whole. An example of a task where documents are modelled in their entirety, is document summarisation (Giannakopoulos et al., 2015). Extractive summarisation typically examines the importance of each sentence with respect to the rest of the sentences in a document, in order to select a small set of sentences that are more "representative" for a given document. A typical extractive summarisation system is expected to select sentences that contain a lot of information in a compact form, and capture the different pieces of information that are expressed in a document. The main idea behind this paper is to examine whether there is any potential *overlap between these sentences* that summarise a document, and *argumentation components* that can exist in a document. Assuming that in a document the author expresses one or more claims, which can be potentially justified through a series of premises or support statements, it will be interesting to examine whether any of these argumentation components will be assessed as significant enough to be included in an automatically generated summary. Will a summarisation algorithm capture at least the claims, and characterise them as important enough to be included in the produced summary?

In order to examine if there is any overlap between extractive summarisation and argument mining (at least the identification of sentences that contain some argumentation components, such as claims), we wanted to avoid any influence from the documents and the thematic domains under examination. Ruling out supervised approaches, we examined summarisation algorithms that are either unsupervised, or can be trained in different domains than the ones they will be applied on. Finally, we opted for an unsupervised algorithm, TextRank (Mihalcea and Tarau, 2004), a graph-based ranking model, which can be applied on extractive summarisation by exploiting "similarity"

among sentences, based on their content overlap. We conducted our study on two corpora in English. The first one is a corpus of user generated content, compiled by Hasan and Ng (2014) from online debate forums on four topics: "abortion", "gay rights", "marijuana", and "Obama". The second corpus, compiled by Stab and Gurevych (2014), contains 90 persuasive essays on various topics. Initial results are promising, suggesting that there is an overlap between extractive summarisation and argumentation component identification, and the ranking of sentences from TextRank can help in tasks related to argument mining, possibly as a feature in cooperation with an argumentation mining approach.

The rest of the paper is organised as follows: Section 2 presents an brief overview of approaches related to argument mining, while section 3 presents our approach on applying TextRank for identifying sentences that contain argumentation components. Section 4 presents the experimental setting and evaluation results, with section 5 concluding this paper and proposing some directions for further research.

2 Related Work

A plethora of approaches related to argument mining consider the identification of sentences containing argument components or not as a key step of the whole process. Usually labeled as "argumentative" sentences, these approaches model the process of identifying argumentation components as a two-class classification problem. In this category can be classified approaches like (Goudas et al., 2015; Goudas et al., 2014; Rooney et al., 2012), where supervised machine learning has been employed in order to classify sentences into argumentative and non-argumentative ones.

However, there are approaches which try to solve the argument mining problem in a completely different way. Lawrence et al. (2014) combined a machine learning algorithm to extract propositions from philosophical text, with a topic model to determine argument structure, without considering whether a piece of text is part of an argument. Hence, the machine learning algorithm was used in order to define the boundaries and afterwards classify each word as the beginning or end of a proposition. Once the identification of the beginning and the ending of the argument propositions has finished, the text is marked from each

starting point till the next ending word.

Another interesting approach was proposed by Graves et al. (2014), who explored potential sources of claims in scientific articles based on their title. They suggested that if titles contain a tensed verb, then it is most likely (actually almost certain) to announce the argument claim. In contrast, when titles do not contain tensed verbs, they have varied announcements. According to their analysis, they have identified three basic types in which articles can be classified according to genre, purpose and structure. If the title has verbs then the claim is repeated in the abstract, introduction and discussion, whereas if the title does not have verbs, then the claim does not appear in the title or introduction but appears in the abstract and discussion sections.

Another field of argument mining that has recently attracted the attention of the research community, is the field of argument mining from online discourses. As in most cases of argument mining, the lack of annotated corpora is a limiting factor. In this direction, Hasan and Ng (2014), Houngbo and Mercer (2014), Aharoni et al. (2014), Green (2014), Stab and Gurevych (2014), and Kirschner et al. (2015) focused on providing corpora spanning from online posts to scientific publications that could be widely used for the evaluation of argument mining techniques. In this context, Boltužić and Šnajder (2014) collected comments from online discussions about two specific topics and created a manually annotated corpus for argument mining. In addition, they used a supervised model to match user-created comments to a set of predefined topic-based arguments, which can be either attacked or supported in the comment. In order to achieve this, they used textual entailment features, semantic text similarity features, and one "stance alignment" feature.

One step further, Trevisan et al. (2014) described an approach for the analysis of German public discourses, exploring *semi-automated* argument identification by exploiting discourse analysis. They focused on identifying conclusive connectors, substantially adverbs (i.e. "hence", "thus", "therefore"), using a multi-level annotation. Their approach consists of three steps, which are performed iteratively (manual discourse linguistic argumentation analysis, semi-automatic text mining (PoS-tagging and linguistic multi-level annotation) and data merge) and their re-

sults show the argument-conclusion relationship is most often indicated by the conjunction because followed by "since", "therefore" and "so".

Ghosh et al. (2014) attempted to identify the argumentative segments of texts in online threads. Expert annotators have been trained to recognise argumentative features in full-length threads. The annotation task consisted of three subtasks: In the first subtask, annotators had to identify "Argumentative Discourse Units" (ADUs) along with their starting and ending points. Secondly, they had to classify the ADUs according to the "Pragmatic Argumentation Theory" (PAT) into "Callouts" and "Targets". As a final step, they indicated the link between the "Callouts" and "Targets". In addition, a hierarchical clustering technique has been proposed that assess how difficult it is to identify individual text segments as "Callouts".

Levy et al. (2014) defined the task of automatic claim detection in a given context and outlined a preliminary solution, aiming to automatically pinpoint context dependent claims (CDCs) within topic-related documents. Their supervised learning approach relies on a cascade of classifiers. Assuming that the articles examined are relatively small set of relevant free-text articles, they provided either manually or automatic retrieval methods. More specifically, the first step of their approach is to identify sentences containing CDCs in each article. As a second step a classifier is used in order to identify the exact boundaries of the CDCs in sentences identified as containing CDCs. As a final step, each CDC is ranked in order to isolate the most relevant CDCs to the corresponding topic.

Finally, Carstens and Toni (2015) focus on extracting argumentative relations, instead of identifying the actual argumentation components. Despite the fact that few details are provided and their approach seems to be concentrated in pairs of sentences, the presented approach is similar to the approach presented in this paper in the sense that both concentrate on relations as the primary starting point for performing argument mining.

3 Extractive Summarisation and Argumentative Component Identification

3.1 The TextRank Algorithm

TextRank is a graph-based ranking model, "which can be used for a variety of natural language

processing applications, where knowledge drawn from an entire document is used in making local ranking/selection decisions" (Mihalcea and Tarau, 2004). The main idea behind TextRank is to extract a graph from the text of a document, using textual fragments as vertices. What constitutes a vertex depends on the task the algorithm is applied on. For example, for the task of keyword extraction vertices can be words, while for summarisation the vertices can be whole sentences. Once the vertices have been defined, edges can be added between two vertices according to the "similarity" among text units represented by vertices. Again, "similarity" depends on the task. As a last step, an iterative graph-based ranking algorithm (a slightly modified version of the PageRank algorithm (Brin and Page, 1998)) is applied, in order to score vertices, and associate a value (score) to each vertex. These values attached to each vertex are used for the ranking/selections decisions.

In the case of (extractive) summarisation, TextRank can be used in order to extract a set of sentences from a document, which can be used to form a summary of the document (either through post-processing of the extracted set of sentences, or by using the set of sentences directly as the summary). In such a case, the following steps are applied:

- The text of a document is tokenised into words and sentences.
- The text is converted into a graph, with each sentence becoming a vertex of the graph (as the goal is to rank entire sentences).
- Connections (edges) between sentences are established, based on a "similarity" relation. The edges are *weighted* by the "similarity" score between the two connected vertices.
- The ranking algorithm is applied on the graph, in order to score each sentence.
- Sentences are sorted in reversed order of their score, and the top ranked sentences are selected for inclusion into the summary.

The notion of "similarity" in TextRank is defined as the overlap between two sentences, which can be simply determined as the number of common words between the two sentences. Formally, given two sentences S_i and S_j, of sizes N and M respectively, with each sentence being represented by a set of words W such as $S_i = W_1^i, W_2^i, ..., W_N^i$ and $S_j = W_1^j, W_2^j, ..., W_M^j$, the similarity between S_i and S_j can be defined as (Mihalcea and Tarau, 2004):

$$Similarity(S_i, S_j) = \frac{|W_k|W_k \in S_i \& W_k \in S_j|}{log(|S_i|) + log(|S_j|)}$$

In our experiments we have used a slightly modified similarity measure which employs TF-IDF (Manning et al., 2008), as implemented by the open-source TextRank implementation that can be found at (Bohde, 2012).

3.2 Argumentative Component Identification

The main focus of this paper is to evaluate whether there is any overlap between argument mining and automatic summarisation. An automatic summarisation algorithm such as TextRank is expected to rank highly sentences that are "recommended" by the rest of the sentences in a document, where "recommendation" suggests that sentences address similar concepts, and the sentences recommended by other sentences are likely to be more informative (Mihalcea and Tarau, 2004), thus more suitable for summarising a document. In a similar manner, a claim is expected to share similar concepts with other text fragments that either support or attack the claim. At the same time, these fragments are related to the claim with relations such as "support" or "attack". Thus, there seems to exist some overlap between how arguments are expressed and how TextRank selects and scores sentences. In the work presented in this paper we will try to exploit this similarity, in order to use TextRank for identifying sentences that contain argumentation components.

In its initial form for the summarisation task, TextRank segments text into sentences, and uses sentences as the text unit to model the given text into a graph. In order to apply TextRank for claim identification, we assume that an argument component can be contained within a single sentence, essentially ignoring components that are expressed in more than one sentences. A component can also be expressed as a fragment smaller than a sentence: in this case we want to identify whether a sentence contains a component or not. As a result, we define the task of component identification as the identification of sentences that contain an argumentation component.

In order to identify the sentences that contain argumentation components, we tokenise a document into tokens and we identify sentences. Then we apply TextRank, and we extract a small number (one or two sentences) from the top scored sen-

tences. If the document contains an argumentation component, we expect the sentence containing the component to be included in the small set of sentences extracted by TextRank.

4 Empirical Evaluation

In order to evaluate our hypothesis, that there is a potential overlap between automatic summarisation (as represented by extractive approaches such as TextRank) and argument mining (at least claim identification), we have applied TextRank on two corpora written in English. The first corpus has been compiled from online debate forums, containing user posts concerning four thematic domains (Hasan and Ng, 2014), while the second corpus contains 90 persuasive essays on various topics (Stab and Gurevych, 2014).

4.1 Experimental Setup

The first corpus that has been used in our study has been compiled and manually annotated as described in (Hasan and Ng, 2014). User generated content has been collected from an online debate forum[1]. Debate posts from four popular *domains* were collected: "abortion", "gay rights", "marijuana", and "Obama". These posts are either in favour or against the domain, depending on whether the author of the post supports or opposes abortion, gay rights, the legalisation of marijuana, or Obama respectively. The posts were manually examined, in order to identify the *reasons* for the stance (in favour or against) of each post. A set of 56 reasons were identified for the four domains, which were subsequently used for annotating the posts: for each post, segments that correspond to any of these reasons were manually annotated.

We have processed the aforementioned corpus, and we have removed the posts where the annotated segments span more than a single sentence, keeping only the posts where the annotated segments are contained within a single sentence. The resulting number of posts for each domain are shown in Table 1. The TextRank implementation used in the evaluation has been written in Python[2], and is publicly available through (Bohde, 2012).

Each post is associated with one (and in some cases more than one) segment that expresses the

main reason for the author to be in favour or against the domain. In order to examine whether there is an overlap between argument mining and summarisation, we have applied TextRank on each post, and we have examined whether the single, top ranked sentence by TextRank, contains the segment marked as the reason. In case the segment is contained in the top ranked sentence returned by TextRank, the post is classified as correctly identified. If the reason segment is not contained in the returned sentence, the post is characterised as an error. Evaluation results are reported through *accuracy* (proportion of true results among the total number of cases examined).

Finally, two experiments were performed, with the only difference being the number of sentences selected from TextRank to form the summary. During the first experiment (labelled as E_1), only a single sentence was selected (the top-ranked sentence as determined by TextRank), while during the second experiment (labelled as E_2) we have selected the *two* top-ranked sentences.

The main motivation for selecting the corpus compiled by Hasan and Ng (2014) was the fact that most of its documents have been manually annotated with a single claim, which was associated with a text fragment that most of the times is contained within a sentence. Having a single sentence as a target constitutes the evaluation of an approach such as the one presented in this paper easier, as the single sentence that represents the main claim of the document can be compared to the top-ranked sentence by the extractive summarisation algorithm. A corpus that has similar properties, in the sense that there is a "major" claim represented by a text fragment that is contained within a sentence, has been compiled by Stab and Gurevych (2014). This corpus contains 90 documents that are persuasive essays, and have been manually annotated with an annotation scheme that includes a "major" claim for each document, a series of arguments that support or attack the major claim, and series or premises that underpin the validity of an argument. Despite being a smaller corpus than the first corpus used for evaluation, having only 1675 sentences, that fact that it contains only 90 documents suggests that its documents are slightly larger than the posts of the first document by (Hasan and Ng, 2014). The average length of a persuasive essay is 18.61 sentences in this second evaluation corpus, which is larger than the aver-

	"abortion"	"gay rights"	"marijuana"	"Obama"	all domains
Number of posts	398	403	352	298	1451
Mean post size (in sentences)	11.15	8.22	6.52	6.50	8.25
Mean argumentative sentences number	1.80	1.65	1.97	1.39	1.71

Table 1: Number of posts per domain in the first corpus used for evaluation (Hasan and Ng, 2014).

Domain	Total Posts	E_1 Correct Posts	E_1 Accuracy	E_2 Correct Posts	E_2 Accuracy
"abortion"	398	150	0.37	226	0.57
"gay rights"	403	160	0.40	235	0.58
"marijuana"	352	168	0.47	239	0.68
"Obama"	298	159	0.53	208	0.70
all domains	1451	631	0.44	919	0.63

Table 2: Baseline results (for experiments E_1 and E_2) – first evaluation corpus (Hasan and Ng, 2014).

Experiment	Total	Correct Essays	Accuracy
E_1	90	3	0.03
E_2	90	10	0.11
E_1 (all claims)	90	30	0.33
E_2 (all claims)	90	48	0.53

Table 3: Baseline results (for experiments E_1 and E_2) – second evaluation corpus (Stab and Gurevych, 2014).

age post size of 8.25 sentences of the first corpus (Table 1). As a result, the second corpus that will be used for evaluation (Stab and Gurevych, 2014) provides the opportunity to evaluate TextRank on larger documents, where the selection of the sentence that represents the "major" claim is potentially more difficult, as the set of potential candidate sentences is larger. Finally, there is a single "major" claim for each persuasive essay, and the mean number of all claims (including the "major" claim) is 5.64 per persuasive essay.

4.2 Baseline

As a baseline approach, a simple random sentence extractor has been used. The sentences contained in each document (post for the first and essay for the second evaluation corpus respectively) were randomly shuffled by using the Fisher-Yates shuffling algorithm (Fisher and Yates, 1963). Then we extract a small number (the first or the two first sentences) from the sentences as randomly shuffled, simulating how we apply TextRank for identifying the sentences that contain argumentation components. The results obtained from this random shuffle baseline are shown in Table 2 for

the first evaluation corpus (Hasan and Ng, 2014), while the results for the second evaluation corpus (Stab and Gurevych, 2014) are presented in Table 3.

4.3 Evaluation Results

As described in the experimental setting, we have performed two experiments. During the first experiment (E_1) we have generated a "summary" of a single sentence (the top-ranked sentence by TextRank), while for the second experiment (E_2) we have selected the two top-ranked sentences as the generated "summary". In both experiments, each post is characterised as correct if the reason segment is contained in the extracted "summary"; otherwise the post is characterised as an error. The evaluation results are shown in Table 4 for experiment E_1 and Table 5 for experiment E_2.

As can be seen from Tables 4 and 5, TextRank has achieved better performance (as measured by accuracy) than our baseline in both experiments, E_1 and E_2. For experiment E_1, accuracy has increased from 0.44 (of the baseline) to 0.51, while in experiment E_2, accuracy has increased from 0.63 to 0.71, when considering all four domains. In addition, TextRank has achieved better performance for all individual domains than the baseline, which randomly selects sentences. Another factor is document size: the mean size of posts (measured as the number of contained sentences) seems to vary between the four domains, ranging from 6.5 sentences for domains "Obama" and "marijuana" to 11 sentences for domain "abortion". TextRank has exhibited better performance than the baseline even for the domains with larger

Domain	Total Posts	Correct Posts	Accuracy	Accuracy (Baseline)
"abortion"	398	171	0.43	0.37
"gay rights"	403	201	0.50	0.40
"marijuana"	352	199	0.56	0.47
"Obama"	298	175	0.59	0.53
all domains	1451	746	0.51	0.44

Table 4: Evaluation Results (for experiment E_1) – first evaluation corpus (Hasan and Ng, 2014).

Domain	Total Posts	Correct Posts	Accuracy	Accuracy (Baseline)
"abortion"	398	267	0.67	0.57
"gay rights"	403	269	0.67	0.58
"marijuana"	352	273	0.78	0.68
"Obama"	298	223	0.75	0.70
all domains	1451	1032	0.71	0.63

Table 5: Evaluation Results (for experiment E_2) – first evaluation corpus (Hasan and Ng, 2014).

Experiment	Total Essays	Correct Essays	Accuracy	Accuracy (Baseline)
E_1	90	14	0.16	0.03
E_2	90	26	0.29	0.11
E_1 (all claims)	90	47	0.52	0.33
E_2 (all claims)	90	67	0.74	0.53

Table 6: Evaluation Results – second evaluation corpus (Stab and Gurevych, 2014).

posts, such as "abortion". Of course, as the size of documents increases the task of selecting one or two sentences becomes more difficult, and this is evident by the drop in performance (for both TextRank and the baseline) for domains "abortion" and "gay rights" when compared to the rest of the domains.

Results are similar for the second evaluation corpus of persuasive essays, as is shown in Table 6. Again TextRank has achieved better performance than the baseline for both experiments, E_1 and E_2. The overall performance of both TextRank and the baseline is lower than the first corpus, mainly due to the increased size of persuasive essays compared to posts (having an average size of 18.61 and 8.25 sentences respectively). For the second corpus an additional experiment has been performed, which expands the set of claims that have to be identified, from only the "major" claim, to all the claims (including the "major" one) in an essay. This experiment (labelled as "E_1 (all claims)" and "E_2 (all claims)" in Table 6) examines whether the top-ranked sentence (experiment "E_1 (all claims)") by TextRank is a claim, or whether the first two sentences as ranked

by TextRank contain a claim (experiment "E_2 (all claims)"). As expected, the performance of both TextRank and the baseline has been increased, as this is an easier task. The mean number of all claims (including the "major" claim) is 5.64 per persuasive essay.

Regarding the overall performance of the summarisation algorithm and its use for identifying a sentence containing an argumentation component, TextRank has managed to achieve a noticeable increase in performance over the baseline, despite the fact that it is an unsupervised algorithm, requiring no training or any form of adaptation to the domain. This suggests that an algorithm that models a document as a whole can provide positive information for argument mining, even if the algorithm has been designed for a different task, as is the case for TextRank variation used, which targets extractive summarisation. In addition, the evaluation results suggest that there is some overlap between argument mining and summarisation, leading to the conclusion that there are potential benefits for approaches performing argument mining through the synergy with approaches that perform document summarisation.

5   Conclusions

In this paper we have applied an unsupervised algorithm for extractive summarisation, TextRank, on a task that relates to argument mining, the identification of sentences that contain an argumentation component. Motivated by the need to better address relations and interactions that are not local within a document, we have applied a graph-based algorithm, which models a whole document having sentences as its basic text unit. Evaluation has been performed on two English corpora. The first corpus contains user posts from an on-line debating forum, which has been manually annotated with the reasons each post author uses to declare its stance, in favour or against, towards a specific topic. The second corpus contains 90 persuasive essays, which has been manually annotated with claims and premises, along with a "major" claim for each essay. Evaluation results suggest that graph-based approaches and approaches targeting extractive summarisation can have a positive effect on tasks of argument mining.

Regarding directions for further research, there are several axes that can be explored. Our evaluation results suggest that TextRank achieved better performance than the baseline for documents between 6 and 11 sentences, and it would be interesting to evaluate further its performance on longer documents. At the same time, the performance of TextRank depends on how "similarity" between its text units is defined; alternative "similarity" measures can be considered, even supervised ones that measure distance according to information obtained from a domain, or information obtained for a specific task. Even an external knowledge base can be explored, providing distances closer to semantic similarity. Finally, a third dimension is to examine alternative extractive summarisation algorithms, in order to clarify further whether other summarisation algorithms can have a positive impact for argument mining, similar to the results achieved by TextRank.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June. Association for Computational Linguistics.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Josh Bohde. 2012. Document Summarization using TextRank. `http://joshbohde.com/blog/document-summarization`. [Online; accessed 12-May-2016].

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland, June. Association for Computational Linguistics.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, June. Association for Computational Linguistics.

R.A. Fisher and F. Yates. 1963. *Statistical tables for biological, agricultural, and medical research*. Hafner Pub. Co.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June. Association for Computational Linguistics.

George Giannakopoulos, Jeff Kubina, Ft Meade, John M Conroy, MD Bowie, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. Multiling 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 270.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 287–299. Springer.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and the social web. *International Journal on Artificial Intelligence Tools*, 24(05):1540024.

Heather Graves, Roger Graves, Robert Mercer, and Mahzereen Akter. 2014. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99, Baltimore, Maryland, June. Association for Computational Linguistics.

Nancy Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar, October. Association for Computational Linguistics.

Hospice Houngbo and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, Maryland, June. Association for Computational Linguistics.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, June. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO, June. Association for Computational Linguistics.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In G. Michael Youngblood and Philip M. McCarthy, editors, *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23-25, 2012*. AAAI Press.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In Junichi Tsujii and Jan Hajic, editors, *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Bianka Trevisan, Eva Dickmeis, Eva-Maria Jakobs, and Thomas Niehr. 2014. Indicators of argument-conclusion relationships. an approach for argumentation mining in german discourses. In *Proceedings of the First Workshop on Argumentation Mining*, pages 104–105, Baltimore, Maryland, June. Association for Computational Linguistics.