Argumentative Ranking

Marco Lippi and **Paolo Sarti** and **Paolo Torroni**
DISI - Università degli Studi di Bologna*

Abstract

There are situations where the information we need to retrieve from a set of documents is expressed in the form of arguments. Recent advances in argumentation mining pave the way for a new type of ranking that addresses such situations and can positively reduce the set of documents one needs to access in order to obtain a satisfactory overview of a given topic. We define and implement a proof-of-concept argumentative ranking prototype, to find that the results it provides can significantly differ from, and possibly improve, those returned by an argumentation-agnostic search engine.

1 Introduction

An argument is, broadly speaking, a claim supported by evidence [22]. Argumentation is the reasoning or dialogical process of producing and evaluating such arguments. It is also the name given to the discipline that studies such processes. Arguments are present in everyday life, so much so to suggest that the need to create and use arguments to convince others is the main driver behind the evolution of human reasoning [11]. It was in fact observed that people are better at reasoning when they communicate through an argumentative context, rather than in an abstract setting. Moreover, arguments are used to convince others. Thus, persuasive communications, editorials, political debates, opinionated blogs, etc. are rich in arguments, which show themselves in diverse formats. Capturing such arguments in a way that enables us to reason from them is a cognitive process that we have been training to do well by evolution. Indeed, in many contexts where we need, for example, to form an opinion on a new topic, especially a controversial one, arguments are exactly what we are looking for.

If detecting arguments is an atavistic ability of the human mind, the automatic detection of arguments instead is a relatively new challenge for computer science. In particular, in the past few years we have witnessed great advancements in a new domain, called *argumentation mining*, which addresses the challenging task of automatically extracting structured arguments from unstructured textual corpora [9].

Therefore, while until recently the perspective of retrieving documents based on their argumentative content would have been utopic, the recent availability of argumentation mining methods and tools [1; 20; 8; 15] makes this vision suddenly more concrete.

Possible applications come to mind easily. A search engine that ranked documents based on the amount of claims about a given topic and of evidence related to such claims would be an invaluable companion for news agencies, journalists, communication departments and cabinet staff, and would be useful even to the random browser, since it would positively narrow down the set of documents that one needs to access in order to obtain a satisfactory overview of the topic.

The aim of this short speculative study is thus to introduce the concept of argumentative ranking, propose an initial portfolio of metrics that can be used to implement it, and offer a first, qualitative assessment of the potential of such a ranking by means of a proof-of-concept prototype and a controlled experiment. We show that argumentative ranking does indeed provide results that are quite different from those that are obtained by a "traditional" search engine.

This work is related to the field of *focused retrieval*, that aims to provide users with direct access to relevant information in retrieved documents [14]. Recently, the IBM Haifa Research Group also proposed a method to perform claim-oriented retrieval of Wikipedia pages [16]. Yet, such approach is only a preparatory step for claim detection, by using a set of handcrafted features that are specifically designed to select documents that are more likely to contain claims (e.g., because they contain "controversy"-related terms or are tagged with special Wikipedia annotations that indicate a controversial content). The approach we propose in this paper, instead, directly addresses the ranking problem in document retrieval, by exploiting the information coming from the claims detected by an argumentation mining system. Differently from the IBM approach, that is tailored to Wikipedia articles, our method can in principle be applied to heterogeneous documents, covering any genre and domain. Our case study is conducted on a collection of newspaper articles retrieved from the New York Times website.

*Contact: marco.lippi3@unibo.it, paolo.sarti2@studio.unibo.it, p.torroni@unibo.it.

2 Argumentation Mining

Argumentation (or argument) mining is the automatic extraction of structured arguments from unstructured textual corpora. It has been argued that building systems endowed with argument mining capabilities would pave the way to a variety of innovative applications [9]. That is confirmed by some important investments made in this area by public and private agencies.[1] This makes us believe that maturing argumentation mining technologies will advance even further in the near future.

The architecture of an argumentation mining system is defined by three crucial aspects: the *argument model* it adopts, the set of *corpora* used for training the system, and the *methodology* exploited in addressing all relevant sub-tasks.

The most popular structured argument model in literature is also the simplest possible model, whereby an argument consists of three distinct parts: a set of premises, sometimes also called *evidence*, a conclusion or *claim*, and an inference from the premises to the conclusion [22].

The works that pioneered this field were strongly connected to the available corpora. Historically, the first application domain was law [21; 12], where the idea was to identify arguments in judgments or other legal documents. Some initial datasets were collections of annotated court cases. Other important datasets are the Dundee corpora[2] and the NoDE benchmark [3], which focus on the relations between arguments. Undoubtedly, the largest available dataset to date was produced within the Debater project and is maintained by IBM Research. It consists of 547 Wikipedia articles [1; 15], organized into 58 topics, and it has been annotated with 2,294 claims and 4,690 evidence facts. Other smaller corpora are available on diverse domains such as persuasive essays [19], comments to articles and forum posts [6], and blog threads [2].

The existing argumentation mining methodologies usually implement a pipeline of subsequent stages [9], which takes in input a raw text document, and produces in output a structured document where arguments are highlighted. The first stage extracts sentences that contain an argument component (claim and/or evidence). The second stage detects the boundaries of each component. The final stage predicts the structure of argumentation, i.e., the support/attack relations between arguments or components. Because we are not interested in predicting the whole argument structure, but only in measuring the amount of arguments in a document, the first stage already provides useful output. To this end, claim/evidence detection has been addressed by a variety of tools, including structured kernel machines [17; 8], binary SVM classifiers [20; 5], logistic regression [7; 15], naïve Bayes [2; 20; 12; 5], and recursive neural networks [18].

[1] See for instance the multi-million IBM Debater project, https://www.research.ibm.com/haifa/dept/vst/mlta_data.shtml, and a large ESPRC on argumentation mining at the University of Dundee, http://www.dundee.ac.uk/news/2015/11million-ai-grant-to-mine-arguments-and-analyse-opinion.php

[2] http://www.arg.dundee.ac.uk/aif-corpora/

3 Ranking by Claims

A classifier such as those used in the first stage of the argumentation mining pipeline typically assigns a score to each sentence of a given document. In the case of claim detection, if the score is positive, the sentence is predicted to contain a claim.[3] An argumentative ranking of documents can be obtained by interpreting the sentence-level information produced by the classifier. We defined five indicators measuring the argumentative content of a document D_i:

- $\sigma_1(D_i)$: Number of sentences in D_i containing claims;
- $\sigma_2(D_i)$: Percentage of sentences in D_i containing claims;
- $\sigma_3(D_i)$: Sum of scores of sentences in D_i containing claims;
- $\sigma_4(D_i)$: Average score of sentences in D_i containing claims;
- $\sigma_5(D_i)$: Sum of scores of sentences in D_i containing claims, divided by the total number of sentences in D_i.

Each indicator $\sigma_j(D_i)$ measures a different aspect of the argumentative content of D_i. There is no absolute reason to prefer one indicator over the other. For example, it is difficult to establish a clear preference between a very short document where almost all the sentences are argumentative, and a lengthier document that contains more claims but also several non-argumentative sentences. Similarly, there are reasons for taking into account the magnitude of the scores, which could bring important additional information to the ranking, but one may also decide to ignore that, and consider simple binary information.

For want of a convincing absolute criterion, we decided to combine all these indicators in a single ranking function through a voting process. Combining different scores into a final ranking function is a typical operation in information retrieval systems (e.g., see [4; 13] and references therein). Given a corpus of M documents $\mathcal{D} = \{D_i\}_{i=1}^{M}$ related to a given query, we first computed the five scores described above $\sigma_j(D_i), j \in \{1, \ldots, 5\}$ for each document D_i, thus building five different rankings. Then, we assigned a set of points π_j to each document, based on each individual ranking, following a non-linear mapping: 25 points to the first document, 20 to the second, 16 to the third, 13 to the fourth, then 11, 10, ..., 1 point to the 5th, 6th, ..., 15th document, and 0 points to the others.[4]

The final score $S(D_i)$ of document D_i is thus obtained by summing the points obtained by the document in each of the five rankings induced by the five indicators:

$$S(D_i) = \sum_{j=1}^{5} \pi_j(D_i) \quad (1)$$

[3] In general, this threshold could be tuned so as to improve the recall or the precision of the classifier.

[4] This is the points scoring system adopted in the FIM Motorcycle Grand Prix World Championship.

4 Experiments

Quantitative evaluations of ranking systems are notoriously hard to obtain, because the key utility measure should be "user happiness", which is greatly influenced by the quality of the returned results (difficult to assess by itself), but also by independent factors, such as speed of response, interface design issues, and the size of the index [10]. We thus decided to perform a qualitative analysis of the output. To this end, we set up an experiment aimed to compare our ranking with the results retrieved by a mainstream search engine, such as Google, and identify cases where the argumentative ranking may satisfy the requests of a user.

We randomly selected 30 key phrases from the controversial topics in the IBM corpus. Of these 30 key phrases, 12 consist of a single word (e.g., abortion, austerity, gambling), and 18 of a short phrase (e.g., affirmative action, national service, wind power). We queried the Google search engine[5] with each one of the key phrases in turn, together with the expression `site:www.nytimes.com`, whose effect is to limit the scope of the search to the New York Times website. We saved the top-10 hits of each key phrase. We then implemented a simple crawler[6] in order to collect a larger set of documents from the New York Times website,[7] using the top-10 Google results as seed pages for the crawler. The crawler's policy was to follow a link if at least one of the following two conditions was met: (1) the link URL contained the searched key phrase; (2) the link was contained in a page in which the searched key phrase appeared at least once. Starting from the selected key phrases and seeds, the crawler downloaded 3,197 articles. We further discarded 11 key phrases, for which less than 20 articles could be retrieved. Table 1 provides details on the dataset.

For each article retrieved by our crawler, we run the claim detection system described in [8].

This setup enabled a qualitative comparison between the search results retrieved by a "traditional" search engine, which is mostly based on features induced by the network topology and website reputation, and the argumentation ranking approach, whose distinguishing feature is its ability to highlight argumentative content by analyzing the linguistic and semantic content of a web page.

Space restrictions allow us to comment on a few interesting cases only.[8] Let us first consider the keyword *gambling*. The top-ranked article according to our system is titled "Majority Back Referendum to Add Casinos, Poll Finds," and it does not appear among the top-10 articles retrieved by Google (see Table 2, top). This article is actually highly argumentative, as it provides many pros and cons with respect to the possibility of opening new casinos in the state of New York. In fact, among the claims retrieved by our system, we find both arguments in favor of expanding casino gambling, as in the following sentence:

Table 1: Details on the New York Times corpus developed within this work.

Key phrases	Articles	Claims/Sent.	Claims/Artic.
abortion	485	0.062	2.318
affirmative+action	85	0.106	4.553
asylum	223	0.031	1.466
austerity	172	0.054	2.366
blasphemy	22	0.033	1.091
collective+bargaining	60	0.051	2.200
contraception	53	0.097	3.396
endangered+specie	65	0.033	1.508
gambling	73	0.097	5.096
Gaza	690	0.029	1.228
Holocaust	39	0.030	1.359
Keystone+XL	126	0.051	2.048
Myanmar	296	0.037	1.368
national+service	53	0.035	2.132
nuclear+weapon	260	0.038	1.562
sex+education	43	0.052	2.698
video+game	172	0.035	2.384
wind+power	181	0.058	3.558
year+round+school	75	0.037	2.680

> *Seventy-four percent agreed that allowing the development of casinos would create thousands of jobs, and 65 percent agreed that more casinos would generate significant revenue for the state and for local governments.*

and against these new casino openings, such as this one:

> *And 55 percent agreed that developing casinos would only increase societal problems, like crime and compulsive gambling.*

This controversy is summarized by another sentence, explicitly remarking the presence of arguments in the article:

> *The poll found that voters agree with arguments both for and against expanding casino gambling.*

From Table 2 (top) we can also observe that, for this keyword, Argumentative Ranking and Google have only four top-ranked articles in common out of 10. In general, we observe that Google tends to include more news, chronicle and event-related articles, and we know that the number of backlinks plays a major role. If we consider the percentage of sentences containing claims for each article (column $\%_C$), we observe that Google does not necessarily retrieve argumentative content.

As a second example, we consider the phrase *wind+power*. Table 2 (bottom) shows the top-10 documents ranked by our system and by Google. Also in this case, our top-ranked article is not present in Google results. The article is entitled "Salvation gets cheap" and is a 2014 article containing plenty of argumentative sentences that well describe the debate around the topics of renewable energies and pollution. Some of the paragraphs detected by our systems as containing claims are:

> *Even as the report calls for drastic action to limit emissions of greenhouse gases, it asserts that the*

[5] The experiments were run on November 10–14, 2015.

[6] We used the open source library `crawler4j`.

[7] http://www.nytimes.com/

[8] All the URLs of the downloaded articles and the results of our ranking systems are available at the following website: http://argumentativeranking.disi.unibo.it.

Table 2: Titles and scores of top-10 documents ranked by our system and by Google for the keywords `gambling` (top) and `wind+power` (bottom). For each article we show the percentage of claims $\%_C$ and the overall score S. Items marked N/A were not retrieved by our crawler.

	Argumentative Ranking	$\%_C$	$S(D_i)$	Google Ranking	$\%_C$	$S(D_i)$
1.	Majority Back Referendum to Add Casinos...	0.32	94	Rein In Online Fantasy Sports Gambling	0.42	82
2.	Rein In Online Fantasy Sports Gambling	0.42	82	The Trouble With Fantasy Sports Gambling	N/A	N/A
3.	Nevada Says It Will Treat Daily Fantasy...	0.23	51	17 People in Three States Are Held in...	N/A	N/A
4.	Cash Drops and Keystrokes: The Dark...	0.13	51	The Dark World of Fantasy Sports and...	N/A	N/A
5.	Will Other Leagues Join N.B.A.? Don't Bet...	0.19	45	Cash Drops and Keystrokes: The Dark...	0.13	51
6.	N.F.L.'s Unsteady Stance on a Tricky...	0.19	39	Nevada Says It Will Treat Daily Fantasy...	0.23	51
7.	As Casino Vote Nears, Bishops Warn of...	0.38	37	Daily Fantasy Sports and the Hidden Cost...	0.14	12
8.	Seeking to Ban Online Betting, G.O.P....	0.20	36	The Perfect Predictability of Gambling...	0.07	0
9.	An Ad Blitz for Fantasy Sports Games, but...	0.14	27	Whitney Wortman and William Gambling	N/A	N/A
10.	In Sharp Pivot for N.B.A., Commissioner...	0.25	25	An Ad Blitz for Fantasy Sports Games, but...	0.14	27

	Argumentative Ranking	$\%_C$	$S(D_i)$	Google Ranking	$\%_C$	$S(D_i)$
1.	Salvation Gets Cheap	0.29	62	Wind Power Spreads Through Turbines...	N/A	N/A
2.	State of the Union Address - 2012 Transcript	0.06	50	Europe Looks Offshore for Wind Power	0.19	15
3.	Wind Power Is Poised to Spread to All States	0.46	50	Wind Power Is Poised to Spread to All States	0.46	50
4.	Tesla Ventures Into Solar Power Storage for...	0.15	46	Procter & Gamble to Run Its Factories...	0.10	0
5.	Glut of Coal-Fired Plants Casts Doubts on...	0.16	43	The Falling Cost of Wind Power	0.10	0
6.	Natural Gas: Abundance of Supply and Debate	0.22	41	Solar and Wind Energy Start to Win on...	0.17	36
7.	Texas Is Wired for Wind Power, and More...	0.16	37	Texas Is Wired for Wind Power, and More...	0.16	37
8.	Solar and Wind Energy Start to Win on...	0.17	36	Tax Credit for Wind Power	N/A	N/A
9.	A Price Tag on Carbon as a Climate Rescue...	0.11	36	A Texas Utility Offers a Nighttime Special...	0.12	0
10.	China Wins in Wind Power, by Its Own Rules	0.20	29	HP to Power Texas Data Centers With...	0.00	0

> *economic impact of such drastic action would be surprisingly small.*
>
> *On the left, you sometimes find environmentalists asserting that to save the planet we must give up on the idea of an ever-growing economy; on the right, you often find assertions that any attempt to limit pollution will have devastating impacts on growth.*
>
> *It's even possible that decarbonizing will take place without special encouragement, but we can't and shouldn't count on that.*

In this case, Argumentative Ranking and Google have only 3 top-ranked articles in common. Again, the reported statistics highlight a marked difference between the argumentative content retrieved by the two systems.

We complemented our analysis by studying the outcome of Google queries when we attached keywords such as `debate`, `argument`, and `opinion`. We obtained mixed results: while such keywords brought up the occasional article with argumentative content, we could not observe a significantly consistent improvement.

5 Conclusions

Motivated by recent advances in argumentation mining, we presented a small, speculative study aimed to define and demonstrate the usefulness of argumentative ranking. As a pilot case study we chose a set of paradigmatic, controversial topics from the IBM argumentation mining corpus, and a largely popular newspaper such as the New York Times. We compared the results obtained by our argumentative ranking system and a traditional, argumentation-agnostic search engine. We found that, in several cases, our system produces a high ranking for documents that are rich in argumentative content but are remarkably excluded from the top Google results. We believe that this new type of ranking could enable a new range of innovative applications fit to diverse domains such as journalism and politics but also law, medicine, and market analysis, as well as increase the quality of search for the random browser. Future work will include a quantitative analysis of the performance of our system, following the contributions of the recent area of focused retrieval.

References

[1] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proc. 1st Worksh. on Argumentation Mining*, pages 64–68. ACL, 2014.

[2] O. Biran and O. Rambow. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, 5(4):363–381, 2011.

[3] E. Cabrio and S. Villata. NoDE: A benchmark of natural language arguments. In *Proc. COMMA 2014*, pages 449–450. IOS Press, 2014.

[4] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. WWW '01*, pages 613–622. ACM, 2001.

[5] J. Eckle-Kohler, R. Kluge, and I. Gurevych. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proc. EMNLP*, pages 2236–2242. ACL, 2015.

[6] I. Habernal, J. Eckle-Kohler, and I. Gurevych. Argumentation mining on the web from information seeking perspective. In *Proc. Worksh. Front. Conn. Argum. Theory NLP*, CEUR-WS 1341, 2014.

[7] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In J. Hajic and J. Tsujii, editors, *COLING 2014, Dublin, Ireland*, pages 1489–1500. ACL, 2014.

[8] M. Lippi and P. Torroni. Context-independent claim detection for argument mining. In *Proc. 24th IJCAI*, pages 185–191. AAAI Press, 2015.

[9] M. Lippi and P. Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25, Mar. 2016.

[10] C. D. Manning, P. Raghavan, and H. Schütze. Evaluation in information retrieval. In *Introduction to Information Retrieval*, chapter 8. CUP, NY, 2008.

[11] H. Mercier and D. Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(02):57–74, 2011.

[12] R. Mochales Palau and M.-F. Moens. Argumentation mining. *Artif. Intell. and Law*, 19(1):1–22, 2011.

[13] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. *Inform. Process. Manag.*, 42(3):595 – 614, 2006.

[14] J. Pehcevski and J. A. Thom. Evaluating focused retrieval tasks. In *SIGIR 2007 Workshop on Focused Retrieval*, 2007.

[15] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proc.EMNLP*, pages 440–450. ACL, 2015.

[16] H. Roitman, S. Hummel, E. Rabinovich, B. Sznajder, N. Slonim, and E. Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 991–996. International World Wide Web Conferences Steering Committee, 2016.

[17] N. Rooney, H. Wang, and F. Browne. Applying kernel methods to argumentation mining. In *Proc. 25th FLAIRS*. AAAI Press, 2012.

[18] C. Sardianos, I. M. Katakis, G. Petasis, and V. Karkaletsis. Argument extraction from news. In *Proc. 2nd Worksh. on Argumentation Mining*, pages 56–66. ACL, 2015.

[19] C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In J. Hajic and J. Tsujii, editors, *COLING 2014, Dublin, Ireland*, pages 1501–1510. ACL, 2014.

[20] C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proc. EMNLP*, pages 46–56. ACL, 2014.

[21] S. Teufel. Argumentative zoning. *PhD Thesis, University of Edinburgh*, 1999.

[22] D. Walton. Argumentation theory: A very short introduction. In *Argumentation in Artificial Intelligence*, pages 1–22. Springer US, 2009.