Unsupervised corpus-wide claim detection

Ran Levy
Shai Gretz*
Benjamin Sznajder
Shay Hummel
Ranit Aharonov
Noam Slonim
IBM Research - Haifa, Israel
{ranl, avishaig, benjams, shayh, ranita, noams}@il.ibm.com

Abstract

Automatic claim detection is a fundamental argument mining task that aims to automatically mine claims regarding a topic of consideration. Previous works on mining argumentative content have assumed that a set of relevant documents is given in advance. Here, we present a first corpus–wide claim detection framework, that can be directly applied to massive corpora. Using simple and intuitive empirical observations, we derive a *claim sentence query* by which we are able to directly retrieve sentences in which the prior probability to include topic-relevant claims is greatly enhanced. Next, we employ simple heuristics to rank the sentences, leading to an unsupervised corpus–wide claim detection system, with precision that outperforms previously reported results on the task of claim detection given relevant documents and labeled data.

1 Introduction

Decision making typically relies on the quality of the arguments being presented and the process by which they are resolved. A common component in all argument models (e.g., (Toulmin, 1958)) is the *claim*, namely the assertion the argument aims to prove. Given a topic of interest, suggesting a diverse set of persuasive claims is a demanding cognitive goal. The corresponding task of *automatic* claim detection was first introduced in (Levy et al., 2014), and is considered a fundamental task in the emerging field of argument mining (Lippi and Torroni, 2016). To illustrate some of the subtleties involved, Table 1 lists examples of sentences related to the topic of whether we should end affirmative action.

S1	*Opponents claim that* ***affirmative action has undesirable side-effects and that it fails to achieve its goals****.*
S2	*The European Court of Justice held that* ***this form of positive discrimination is unlawful****.*
S3	*Clearly,* ***qualifications should be the only determining factor when competing for a job****.*
S4	*In 1961, John F. Kennedy became the first to utilize the term affirmative action in its contemporary sense.*

Table 1: Example sentences for the topic 'End affirmative action': 3 sentences containing claims (in bold), and a non–argumentative sentence which is still relevant to the topic.

Previous works on claim detection have assumed the availability of a relatively small set of articles enriched with relevant claims (Levy et al., 2014). Similarly, other argument–mining works have focused on the analysis of a small set of argumentative essays (Stab and Gurevych, 2014). This paradigm has two limitations. First, it relies on a manual, or automatic (Roitman et al., 2016), process to retrieve the relevant set of articles, which is non-trivial and prone to errors. In addition, when considering large corpora, relevant claims may spread across a much wider and diverse set of articles compared to those considered by earlier works. Here, we present a first corpus–wide claim detection framework, that can be directly applied to massive corpora, with no need to specify a small set of documents in advance.

We exploit the empirical observation that relevant claims are typically (i) semantically related to the topic; and (ii) reside within sentences with identifiable structural properties. Thus, we aim to pinpoint single sentences within the corpus that satisfy both criteria.

Semantic relatedness can be manifested via a rich set of linguistic mechanisms. E.g., in Table 1,

*First two authors contributed equally.

$S1$ mentions the main concept (MC) of the topic (i.e., affirmative action) explicitly; $S2$ mentions the MC using a different surface form – 'positive discrimination'; while $S3$ contains a valid claim without explicitly mentioning the MC. Here, we suggest to use a mention detection tool (Ferragina and Scaiella, 2010), which maps surface forms to Wikipedia titles (a.k.a Wikification), to focus the mining process on sentences in which the MC is detected. Thus, we keep the potential to detect sentences in which different surface forms are used to express the MC. Moreover, using a Wikification tool can help prevent drift in the meaning of the topic. For example, consider the topic *Marriage is outdated* for which the MC is *Marriage*. Had we searched the corpus for all sentences with the word *Marriage*, we would have found many sentences that mention the term *Same sex marriage* which tends to appear more often in argumentative content within the corpus. The risk in this case, is to have the claim detection system drift towards this related but quite different topic. By using a Wikification tool, and assuming it works reasonably well, we avoid this problem. Searching for sentences with the concept *Marriage* will not return sentences in which the Wikification tool found the concept *Same sex marriage*.

However, as mentioned, semantic relatedness is not enough; e.g., $S4$ mentions the MC explicitly, but does not include a claim. To further distinguish such sentences from those containing claims, we observe that the token 'that' is often used as a precursor to a claim; as in $S1$, $S2$ and in the sentence "we observe **that** *the token 'that' is often used as a precursor to a claim*." The usage of 'that' as a feature was first suggested in (Levy et al., 2014). Thus, we use the presence of 'that' as an initial weak label, and further identify unigrams enriched in the suffixes of sentences containing 'that' followed by the MC, compared to sentences containing the MC *without* a preceding 'that'. This yields a *Claim Lexicon* (CL), from which we derive a *Claim Sentence Query* (CSQ) composed of the following ordered triplet: *that* $\rightarrow$ MC $\rightarrow$ CL, i.e., the token 'that', the MC as identified by a Wikification tool, and a unigram from the CL, in that order.

We demonstrate empirically over Wikipedia, that for sentences satisfying this query, the prior probability to include a relevant claim is enhanced compared to the background distribution. Furthermore, by applying simple unsupervised heuristics to sort the retrieved sentences, we obtain precision results outperforming (Levy et al., 2014), while using no labeled data, and tackling the presumably more challenging goal of corpus–wide claim detection. Our results demonstrate the practical value of the proposed approach, in particular for topics that are well covered in the examined corpus.

2 Related Work

Context dependent claim detection (i.e. the detection of claims that support/contest a given topic) was first suggested by (Levy et al., 2014). Next, (Lippi and Torroni, 2015) proposed the context independent claim detection task, in which one attempts to detect claims without having the topic as input. Thus, if the texts contain claims for multiple topics, all should be detected. Both works used the data in (Aharoni et al., 2014) for training and testing their models.

(Levy et al., 2014) have first described 'that' as an indicator for sentences containing claims. Other works have identified additional indicators of claims, such as discourse markers, and have used them within a rule-based, rather than a supervised, framework (Eckle-Kohler et al., 2015; Ong et al., 2014; Somasundaran and Wiebe, 2009; Schneider and Wyner, 2012).

The usage we make in this work of the word 'that' as an initial weak label is closely related to the idea of distant supervision (Mintz et al., 2009). In the context of argument mining, (Al-Khatib et al., 2016) also used noisy labels to train a classifier, albeit for a different task. They exploited the manually curated idebate.org resource to define – admittedly noisy – labeled data, that were used to train an argument mining classification scheme. In contrast, our approach requires no data curation and relies on a simple linguistic observation of the typical role of 'that' in argumentative text. Our use of the token 'that' as a weak label to identify a relevant lexicon, is also reminiscent of the classical work by (Hearst, 1992) who suggested to use lexico-syntactic patterns to identify various lexical relations. However, to the best of our knowledge, the present work is the first to use such a paradigm in the context of argument mining.

3 System Description

3.1 Sentence Level Index

Corpus-wide claim detection requires a run-time efficient approach. Thus, although the context surrounding a sentence may hint whether it contains a claim, we focus solely on single sentences and the information they contain. Correspondingly, we built an inverted index[1] of sentences for the Wikipedia May 2015 dump, covering $\sim 4.9M$ articles. After text cleaning and sentence splitting using OpenNlp[2] we obtained a sentence–level index that contains $\sim 83M$ sentences. We then used TagMe (Ferragina and Scaiella, 2010) to Wikify each sentence, limiting the context used by TagMe for disambiguation, to the examined sentence.

3.2 Topics

We started with a manually curated list of 431 debate topics that are often used in debate-related sites like idebate.org. We limit our attention to debate topics that focus on a single concept, denoted here as the MC, which is further identified by a corresponding Wikipedia page, e.g., Affirmative Action, Doping in Sport, Boxing, etc. In addition, we focus on topics that are well covered in Wikipedia, which we formally define as topics for which the query $q1 = MC$ has at least $1,000$ matches. This criterion is satisfied in $212/431$ topics, of which we randomly selected 100 as a development set (termed dev-set henceforth) and 50 topics as a test set, used solely for evaluation. The complete list of topics is given in the Supplementary Material (SM).

3.3 Claim Sentence Query (CSQ)

For the 100 dev-set topics we obtained a total of $\sim 1.86M$ sentences that match the query $q1$, hence are assumed to be semantically related to their respective topic. We refer to this set of sentences as the *q1-set*. Using 'that' as a weak label, we divide the q1-set into two classes – the sentences that contain the token 'that' before the MC, and the sentences that do not – denoted c_1 and c_2, respectively. The class c_1 consists of $\sim 183K$ sentences, hence we define the estimated prior probability of a sentence from q1-set to be included in c_1 as $P(c_1) = 0.0986$.

Based on these classes, we are interested in constructing a lexicon of claim-related words that will enable designing a query with a relatively high prior for detecting claim–containing sentences. We start with standard pre-processing including tokenization, stop-word removal, lowercasing, pos-tagging using OpenNlp, and removal of tokens mentioned in < 10 sentences in q1-set. Preliminary analysis – described in detail in the SM – suggested that we should focus on the *suffixes* of the sentences in c_1, where the suffix is defined as the part of the sentence that follows the MC. Note, that in our setting the claim is expected to occur after the token 'that' with the MC usually being the subject, hence the suffix as defined above seems like a natural candidate to search for words characteristic of claims. Formally, we define n_1 as the number of sentences in c_1 that contain w in the sentence suffix; n_2 as the number of sentences in c_2 that contain w; and $P_{suff}(c_1|w) = n1/(n1 + n2)$. Finally, we define the Claim Lexicon (CL) as the set of words which satisfy $P_{suff}(c_1|w) > P(c_1)$, namely the set of words that are characteristic of the suffixes of sentences in the class c_1. To put it differently, the set of words that, when they appear in the sentence suffix, make the sentence more likely to be in c_1 than expected by the prior.

A desirable feature of the CL is that it contains words which are indicative of claims in the general sense, i.e., in the context of many different topics. Since the resulting lexicon included some topic-specific words, mostly nouns, we applied straightforward cleansing of removing all nouns, as well as numbers, single-character tokens, and country–specific terms from the CL, ending up with a lexicon consisting of 586 words, listed in the SM.

We then use the CL to construct the *claim sentence query* (CSQ): *that* $\rightarrow$ MC $\rightarrow CL$, where CL denotes any word from the CL. We assessed the prior probability to contain a claim for sentences matching different queries by randomly selecting at most 3 sentences that match the query per dev-set topic, and annotating the resulting sentences by 5 human annotators. We find that, as expected, the prior associated with the query *that* $\rightarrow$ MC is higher than the background prior of sentences matching $q1 = MC$, 4.8% vs. 2.4%, respectively. Using the CSQ further enhances the prior to 9.8%, a factor of 4 compared to the background. Table 2 summarizes the prior and number of matches per query.

[1] See Supplementary Material (SM) for details.

[2] https://opennlp.apache.org/

Query	Prior	#Matches
MC	2.4	4872
$that \rightarrow MC$	4.8	493
$that \rightarrow MC \rightarrow CL$	9.8	74

Table 2: Summary of query evaluation. The "Prior" column shows the percentage of claim sentences estimated by the annotation experiment. The "#Matches" column shows the median number of query matches across the dev-set topics.

3.4 From CSQ to Claim Detection

Based on the sentences that match the CSQ, we are now ready to define a system that performs corpus–wide claim detection by adding sentence re-ranking, boundary detection, and simple filters.

Naturally, we are interested to present higher confidence predictions first. Remaining within the unsupervised framework, we rank the sentences by the average of two simple scores: (i) *w2v*: The CSQ only aims to ensure that the MC is present in the examined sentence. Hence, it seems reasonable to assume that considering the semantic similarity of the *entire* candidate claim to the topic will improve the ranking. Thus, we compare the word2vec representation (Mikolov et al., 2013) of each word in the sentence part following the first 'that' to each word in the MC to find the best cosine-similarity match, and average the obtained scores; (ii) *slop*: The number of tokens between 'that' and the first match to the CL. This assumes that the closer the elements appear in the sentence, the higher the probability that it contains a claim.

To perform claim detection, the claim itself should be extracted from the surrounding sentence. From the way the CSQ is constructed, it follows that the claim is expected to start right after the 'that'. The end of the claim is harder to predict. An approach to boundary detection was described in (Levy et al., 2014), but here we employ a simple heuristic, which does not require labeled data, namely ending the claim at the sentence end. Finally, sentences containing location/person named–entities after the 'that' are filtered out.

4 Results

To evaluate the performance of the proposed system we applied crowd labeling[3] on the predicted claims for all 150 topics in the dev- and test-set. For each topic we labeled the top 50 predictions, or all predictions if there were less. A prediction was considered correct if the majority of the annotators marked it as a claim[4]. The average pairwise Kappa agreement on the dev-set was 0.38, which is similar to the Kappa of 0.39 reported in this context by (Aharoni et al., 2014).

Table 3 depicts the obtained results. Using our approach – that requires no labeling and is applied over the entire Wikipedia corpus – we obtain results that outperform those reached using a supervised approach over a manually pre-selected set of articles (Levy et al., 2014) (see 'Levy' Row), though we note that we consider a different set of topics because of the restrictions we impose on the topic structure (section 3.2). In addition, the test set results are better compared to the dev-set results, suggesting that the system is able to generalize to entirely new topics.

When considering only topics for which $> K$ sentences match the CSQ, the precision increases considerably. For example, for topics that have at least 50 sentences matching the CSQ, $P@50$ is 24% and 34% in the dev- and test-set, respectively. Thus, for topics well covered in the corpus, the precision of the system is even more promising.

The precision results in table 3 are not directly comparable to "classical" argumentation mining tasks, e.g. (Stab and Gurevych, 2014), since our task involves detecting claims over a full corpus in which the ratio of positive cases is much lower (2.4% of sentences containing the MC).

	$P@5$	$P@10$	$P@20$	$P@50$
Dev	31	27	21	15
Test	32	32	28	22
Levy	23	20	16	12
	$P@5'$	$P@10'$	$P@20'$	$P@50'$
Dev	33 (94)	30 (86)	27 (70)	24 (47)
Test	33 (96)	33 (96)	31 (86)	34 (56)

Table 3: System performance in percentages. Levy - Precision as quoted in (Levy et al., 2014), $P@K$ - Precision of the top K candidates per topic, averaged over all topics (following (Levy et al., 2014)), $P@K'$ - same as $P@K$, considering only topics for which there are at least K candidate claims; number in parenthesis denotes the percentage of such topics.

5 Limitations

In this work, we only considered topics that focus on a single concept which has a corresponding

[3] via the CrowdFlower platform: www.crowdflower.com/, see details in supplementary material

[4] We require a minimum of 10 annotators per candidate. After 10 annotations, further annotations are collected until either 90% agreement is reached or 15 annotations.

Wikipedia page. Expanding the proposed framework to more complex queries, covering more than a single concept, merits further investigation. Yet, even without such an expansion, we note that controversial topics are often characterized by a corresponding Wikipedia page.

Our approach targets claims in which the MC is identified by a Wikification tool. While this allows mining claims in which the MC is expressed via different surface forms, Wikification errors also propagate to our performance. Thus, improvements in available Wikification tools are expected to improve the results of the approach. In addition, claims that do not explicitly refer to the MC are out of the radar of the proposed system, limiting its recall. Expanding the CSQ with concepts related to the MC, may mitigate this issue.

Finally, we focused on sentences matching the pattern *that* $\rightarrow$ MC. Exploring the same methodology for additional patterns characterizing claim–containing sentences is left for future work.

6 Discussion

We present an unsupervised simple framework for corpus-wide claim detection, which relies on features that are quick to compute. Exploiting the token 'that' as a weak signal, or as distant supervision (Mintz et al., 2009) for claim–containing sentences, we obtain results that outperform a supervised claim detection system applied to a limited set of documents (Levy et al., 2014). Extending this approach to other computational argumentation tasks like evidence detection (Rinott et al., 2015) is a natural direction for future work.

Notably, the system precision is clearly superior to the precision of the initial 'that' label, indicating the existence of characteristics of claim–containing sentences which may further enhance the signal embodied in this label. Thus, we hypothesize that supervised learning based on labeling the predictions of the unsupervised system can further improve the system results, e.g., by obtaining better ranking schemes and/or stronger methods to determine claim boundaries.

Finally, we demonstrated our approach over the Wikipedia corpus. We speculate that the proposed approach holds even greater potential for mining larger and more argumentative corpora such as newspapers aggregates; in particular, when considering controversial topics that are widely discussed in the media, for which it is natural to expect that relevant claims are mentioned across a very large set of typically short articles.

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.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. 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*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. 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, Ireland. Dublin City University and Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 185–191. AAAI Press.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word

representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. 2016. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 991–996, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Jodi Schneider and Adam Z Wyner. 2012. Identifying consumers' arguments in text. In *SWAIE*, pages 31–42.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.

Stephen Toulmin. 1958. The uses of argumentcambridge university press. *Cambridge, UK*.