Towards an argumentative content search engine using weak supervision

Ran Levy,* **Ben Bogin,*** **Shai Gretz,*** **Ranit Aharonov, Noam Slonim**
IBM Research
{ranl,boginb,avishaig,ranita,noams}@il.ibm.com

Abstract

Searching for sentences containing claims in a large text corpus is a key component in developing an argumentative content search engine. Previous works focused on detecting claims in a small set of documents or within documents enriched with argumentative content. However, pinpointing relevant claims in massive unstructured corpora, received little attention. A step in this direction was taken in (Levy et al., 2017), where the authors suggested using a weak signal to develop a relatively strict query for claim–sentence detection. Here, we leverage this work to define weak signals for training DNNs to obtain significantly greater performance. This approach allows to relax the query and increase the potential coverage. Our results clearly indicate that the system is able to successfully generalize from the weak signal, outperforming previously reported results in terms of both precision and coverage. Finally, we adapt our system to solve a recent argument mining task of identifying argumentative sentences in Web texts retrieved from heterogeneous sources, and obtain F_1 scores comparable to the supervised baseline.

1 Introduction

The arguments raised during a decision making process, will often determine its outcome. A common component in all argument models (e.g., (Toulmin, 2003)) is the *claim*, i.e. the assertion the argument aims to prove. The problem of automatically detecting claims supporting or contesting a given controversial topic [1] (Levy et al., 2014) is considered a fundamental task in the emerging field of computational argumentation (Lippi and Torroni, 2016; Palau and Moens, 2009). We refer to their definition of a Topic and a Claim; **Topic** - a short phrase that frames the discussion and **Context Dependent Claim** - a general, concise statement that directly supports or contests the given Topic (we henceforth use the term claim instead of Context Dependent Claim).

Previous works have focused on detecting claims within a small set of documents related to the topic (Levy et al., 2014), or within documents enriched with argumentative content (Stab and Gurevych, 2014). However, pinpointing relevant claims within massive unstructured corpora, received relatively little attention. While this problem is obviously more challenging, its potential value is also much higher. For a widely discussed topic, one should expect many relevant claims to be mentioned across a widespread set of articles in the given corpus. The remaining issue is to develop a technology to swiftly detect these claims and present the results to potential users, similarly to search engines that retrieve information in response to a query.

A step in this direction was taken in (Levy et al., 2017). They suggested a relatively strict sentence–level query (strict in the sense that it considerably limits the set of potential answers hence reduces the coverage). Their query combines three query parts that must appear in order, with possible gaps between them. The first part requires the sentence to contain the token *'that'* as it is often a precursor for a claim (e.g. <someone> argued that <claim>). The second query part requires some restriction on the scope of topics the system can handle, and assumes that each topic deals with exactly one concept (denoted

*First three authors contributed equally.

[1] We will henceforth refer to claims supporting or contesting a given controversial topic as *relevant claims*.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2066–2081
Santa Fe, New Mexico, USA, August 20-26, 2018.

MC – for main concept) and that this concept has a Wikipedia title (e.g. Affirmative Action). The second part of the query thus restricts the returned sentences to those in which the MC follows the word *'that'* (possibly with a gap). The third and final query part requires a token from a pre–specified *claim lexicon* (CL) to appear after the MC (possibly with a gap). The CL lexicon aims to characterize *claim sentences* (CS) and the process of its creation did not involve any labeled data. Table 1 shows the fifty most indicative tokens from the lexicon. Relying on this formulation led to promising precision results in the challenging task of corpus wide claim detection, albeit with low recall . Specifically, while each of the sentences in Table 2 contains a valid and relevant claim, only $S1$ satisfies their query; In contrast, $S2$ satisfies only the first part of the query ('that' preceding the MC); $S3$ satisfies the second part of the query (CL token following the MC); and $S4$ only mentions the MC. By construction, these latter three sentences, are out of the radar of Levy et al. (2017).

Claim Lexicon - partial
should, tuned, could, unconstitutional, violate, might, violated, violates, wrong, rather, valid, invalid, irrelevant, inherently, necessarily, cannot, prevail, justify, flawed, merely, corpus, ought, inevitably, cause, justifiable, unacceptable, untrue, abhorrent, unless, harmful, punished, liable, incompatible, beneficial, justifying, undecided, skimmed, indefensible, impossible, undermine, necessary, flourish, meaningless, outweigh, substantiated, refute, jeopardized, incapable, irrational, heterosexual

Table 1: Fifty most indicative words in the Claim Lexicon (starting from the most indicative)

Example Id	Sentence
S1	*He believed* **that nuclear power** *would become* **obsolete***, to be replaced by clean energy sources.*
S2	*The author concludes* **that wind energy** *has the greatest potential for near-term expansion.*
S3	*As Buckley writes, "If* **atheism** *was* **unacceptable***, superstition and fanaticism were even more so".*
S4	*Any form of* **corporal punishment** *is barbaric and has no place in a civilized polity.*

Table 2: CS examples for the topics 'We should further exploit nuclear power', 'We should further exploit wind power', 'Atheism is the only way' and 'We should prohibit corporal punishment'. The query items 'that', MC and CL are highlighted in boldface.

The main contribution of the current work is to propose a more flexible approach for corpus wide claim detection, that significantly outperforms previous work, in terms of both precision and of coverage. We also release two data sets, one of $\approx 1.5M$ sentences matching the topics in this study, and one of $2,500$ sentences predicted by our method, annotated for whether they contain a relevant claim or not [2].

We use Deep Neural Networks (DNN) trained with weak supervision that stems from different parts of the aforementioned query. Considering the list of 100 MC used by Levy et al. (2017), we first construct two weakly supervised labeled data sets, each composed of two classes. In the first, the weakly–positive class includes all sentences that mention the MC preceded by 'that'; while the weakly–negative class contains a similar number of sentences that mention the MC *without* a preceding 'that'. Our underlying assumption is that the former set will be more enriched with CS (this we first noted in (Levy et al., 2014)). However, since these two classes are trivially distinguished via the (non) presence of 'that', we train the DNN on the *suffixes* of the sentences in these data, where the suffix of a sentence is defined as the sentence part immediately following the MC.

Similarly, we construct another data set, in which the weakly–positive class includes all sentences that mention the MC followed by a token from CL; while the weakly–negative class contains a similar number of sentences that mention the MC *without* a following token from CL. Here as well, to avoid the trivial signal, we train the DNN on the *prefixes* of the sentences in these data, where the prefix of each sentence is defined as the part preceding the MC.

[2]The data sets can be downloaded from `http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`

The priors for the two positive classes as well as the strict query were estimated in (Levy et al., 2017) by performing a small labeling experiment. We present their results in table 3 in order to demonstrate that the assumptions indeed hold.

Query Name	Query	Estimated Prior
q_{MC}	MC	2.4%
q_{that}	*that* $\rightarrow$ MC	4.8%
q_{CL}	MC $\rightarrow$ CL	Not Estimated
q_{strict}	that $\rightarrow$ MC $\rightarrow$ CL	9.8%

Table 3: Estimated priors for different queries from (Levy et al., 2017).

Finally, we restrict both datasets to sentences in which the number of suffix (prefix) words is greater than 3. We assume this restriction mostly removes negative examples, and in any case will not convey a lot of information to the DNN in the learning process.

We test the performance of these DNNs over a distinct test set of 50 topics, also from (Levy et al., 2017). However, in contrast to this previous work, we consider a much more relaxed query that only requires the MC to be mentioned in the sentence. Our results clearly indicate that both DNNs were able to generalize and obtain promising precision results, that are further improved when their scores are averaged. That is, combining the predictions of a DNN trained over prefixes of sentences enriched with claims, with those by a DNN trained over suffixes of such sentences, results in a pincer–movement like approach, that successfully pinpoints a wide range of CS in a massive unstructured corpus, while using only weak supervision for training.

2 Related Work

Recently, Wachsmuth et al. (2017) suggested an argument search framework and a corresponding search engine prototype. However, the proposed system relies on arguments crawled from dedicated resources that suggest pre–written arguments for various topics, and hence, is only relevant for topics covered in these resources, and cannot be used directly over unstructured textual data. Stab et al. (2018) tackled the argument mining task in heterogeneous texts retrieved by Google search when queried with a controversial topic. They show that it is feasible to annotate the retrieved documents via crowd-sourcing and to use these labels in order to build a supervised learning system that finds arguments in the given documents. Similar to our work, sentences are treated in isolation (ignoring the document context). The only work we are aware of that tackles corpus wide claim detection, is the work by (Levy et al., 2017). Here, we demonstrate how this work can be leveraged to define weak signals for training DNNs to obtain significantly greater performance.

Several works used DNN to tackle a variety of computational argumentation tasks, such as argument mining (Eger et al., 2017), predicting argument convincingness (Habernal and Gurevych, 2016), detecting context dependent claims and evidence (Laha and Raykar, 2016) and attack and support relations between arguments (Cocarascu and Toni, 2017). However, these works used the fully–supervised learning paradigm, which is inherently demanding, especially in the context of argument mining where obtaining labeled data is notoriously difficult (Aharoni et al., 2014). In addition, Al-Khatib et al. (2016) used a distant supervision approach trained over debate portals' data, to develop a classifier for argumentative texts stored in these portals. To the best of our knowledge, the present work is the first to demonstrate the value of DNN trained solely with weak supervision (Hearst, 1992) in this challenging field.

For a good exposition on the field of argument mining refer to (Lippi and Torroni, 2016). Some notable works include (Palau and Moens, 2009) who first suggested the argument mining task, (Levy et al., 2014; Rinott et al., 2015) who focused on mining claims/evidence in the context of a user given controversial topic and several works related to specific text genres such as student essays (Stab and Gurevych, 2014), legal documents (Wyner et al., 2010; Moens et al., 2007; Grabmair et al., 2015), user comments on proposed regulations (Park and Cardie, 2014) and newspaper articles (Feng and Hirst, 2011).

3 Method

3.1 Setup and pre-processing

We follow the setup and pre-processing described in (Levy et al., 2017) – see appendix for details. We consider the same train [3] and test sets, consisting of 100 and 50 topics respectively. Next, we prepared a sentence–level index from the Wikipedia May 2017 dump, and used a simple Wikification tool (to be described in a separate publication)[4] to focus our attention on sentences that mention the MC. Filtering out sentences that mention a location/person named entity using Stanford NER (Finkel et al., 2005), after the MC, results in an average of $\approx 10K$ sentences per MC.

3.2 Claim sentence queries and weak labels

The basic query we start with, denoted q_{MC}, only requires that the MC will appear in the sentence. For the 150 topics of this study, we retrieve a total of $\approx 1.5M$ sentences matching q_{MC} (Table 4), which we release as a data set to enhance future research. Next, we consider the query q_{that}, which retrieves all sentences in which the token 'that' precedes the MC (cf. $S1$ and $S2$ in table 2). There are $\approx 1,100$ such sentences per topic (Table 4). Aiming to increase the prior of CS in the weak–positive set, for training the network, we focus on the subset of these sentences in which the token 'that' immediately precedes the MC. As a weak–negative set we consider a similar number of sentences, with similar length distribution, selected at random from the q_{MC} sentences with the additional requirement of *not* having 'that' before the MC. As explained in section 1, the corresponding DNN, termed DNN_{suff}, is trained only on the sentence suffixes.

Query Name	Query	# Sentences Per Topic
q_{MC}	MC	9,947
q_{that}	*that* → MC	1,073
q_{CL}	MC → CL	793
q_{strict}	that → MC → CL	164

Table 4: Queries used to construct weak labels. # Sentences Per Topic is averaged over the 150 topics used in this study. q_{strict} is added for reference and was not used in training the networks.

Similarly, we consider the query q_{CL}, which retrieves all sentences in which the MC is followed by a token from CL, e.g., sentences $S1$ and $S3$ in table 2. Again, these sentences as well are expected to be relatively enriched with claims. In Table 4 we see that on average we have ≈ 790 such sentences per topic. As a weak–negative set we consider a similar number of sentences, with similar length distribution, selected at random from the q_{MC} sentences with the additional requirement of *not* having a CL token after the MC. Again, the corresponding DNN, denoted by DNN_{pref} is trained only on the sentence prefixes.

Table 5 lists examples of sentences in the weak–positive and weak–negative sets used to train the networks. The part "seen" by the relevant network appears in bold, where by an anecdotal examination it is indeed possible to identify a signal in the positive sets. Table 6 summarizes the characteristics of the two datasets used to train the networks.

3.3 DNN System

For both DNN_{suff} and DNN_{pref}, we use a Bi-LSTM architecture with self-attention (Yang et al., 2016). The networks were trained on sentences retrieved for 70 of the 100 train–set topics, where sentences retrieved from the other 30 train–set topics (heldout set) were used to optimize hyper-parameters. We used Adam optimizer (Kingma and Ba, 2014) over the cross-entropy loss. The best model was

[3] The set of 100 topics was termed *dev set* in their work because there was no training involved.

[4] A Wikification tool allows retrieving sentences that mention the topic explicitly, as well as sentences which use a different surface form, as in S2 in Table 2 (wind energy surface-form linked to the wind power concept).

Network	Positive/Negative	Sentence
DNN_{suff}	Positive	*There is no good evidence that organic food* **tastes better than its non-organic counterparts**.
DNN_{suff}	Negative	*Today it is known for its remoteness, its somewhat "alternative" atmosphere, organic food* **production, and its pioneering use of wind power**.
DNN_{pref}	Positive	**Fermi did not believe that** *atomic bombs would deter nations from starting wars, nor did he think that the time was ripe for world government.*
DNN_{pref}	Negative	**In particular, fission products do not themselves undergo fission, and therefore cannot be used for** *nuclear weapons.*

Table 5: Examples from the positive and negative sets of DNN_{suff} and DNN_{pref} for the topics "Organic Food" and "Nuclear weapon". The respective prefix/suffix appears in bold.

Network	Positive sentences	Negative sentences	Part of sentence used by the network	Size of data
DNN_{suff}	*that* $\rightarrow$ MC	MC without preceding 'that'	following the MC	11,624
DNN_{pref}	MC $\rightarrow$ CL	MC without a following CL token	preceding the MC	132,856

Table 6: Characteristics of the two datasets used to train the networks. Note that the data for the suffix network is much smaller because of the restriction to sentences in which the token 'that' immediately precedes the MC.

trained with a dropout of 0.15, using a single dropout mask across all time-steps as proposed by (Gal and Ghahramani, 2016), one LSTM layer with a cell size of 128, and an attention layer of size 100. Words are represented using the 300 dimensional GloVe embeddings (Pennington et al., 2014). Inference is performed for any q_{MC} sentence by averaging the DNN_{suff} score of its suffix with the DNN_{pref} score of its prefix.

We used the heldout set to determine early stopping and to optimize the following hyper–parameters (each parameter was optimized independently): Number of layers $(1/2)$, LSTM cell size $(64/128/256/512)$, attention FF size $(50/100/200)$ and dropout rate $(0/0.05/0.1/0.15/0.2/0.25/0.3/0.35)$.

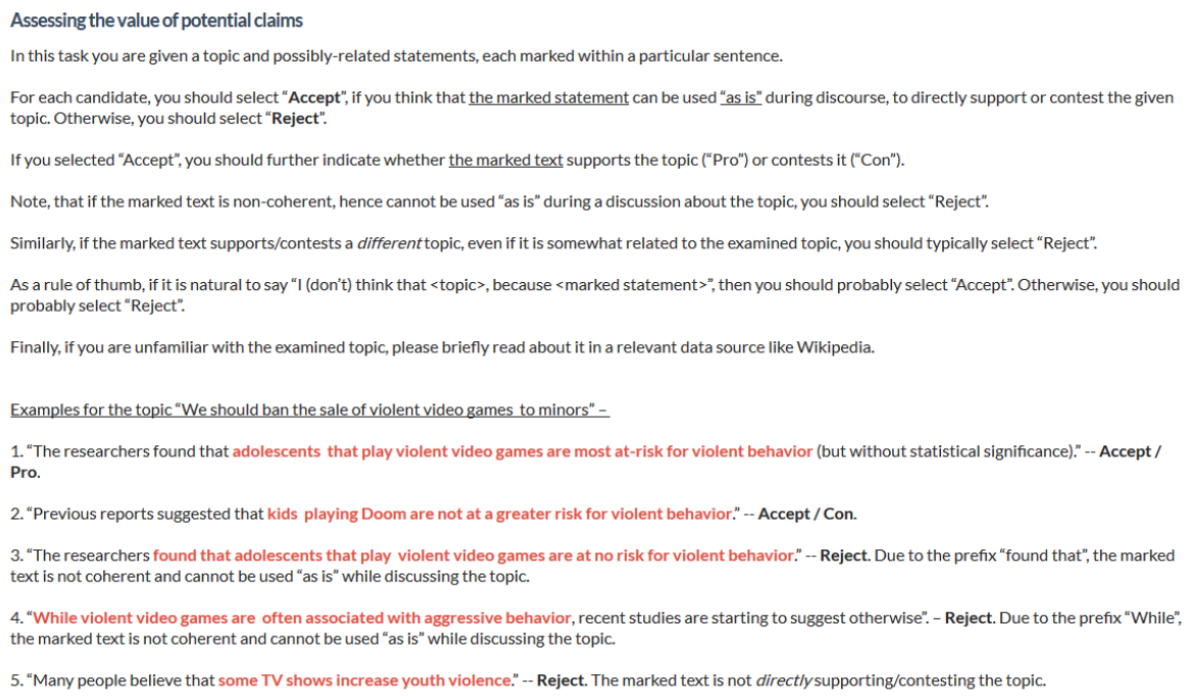
4 Data for evaluation

We labeled via crowd the top 50 predicted sentences for each of the 50 test-set topics, taking the majority vote of at least 10 workers. The guidelines are presented in figure 1. The inference is applied to all sentences containing the MC (matching q_{MC}), and hence there are always 50 predictions, that are all released along with their manual evaluation. We also label in the same manner the predictions of the system described in (Levy et al., 2017) [5]. There, since all predictions must match q_{strict}, for some topics there are less than 50 predictions. In those cases, we label all predictions.

This paper focuses on retrieving claim sentences, however, we have found that it is easier for the crowd workers to label a sentence if the phrase suggested to be the claim is highlighted. For this reason, we used an internal boundary detection component and applied it to all system versions (including the re-implementation baseline of (Levy et al., 2017)). The rest of the labeling process was done similarly to (Levy et al., 2017). Each sentence was labeled by 10-15 crowd workers per row via the Figure–Eight platform [6]. We used the MACE de–noising tool (Hovy et al., 2013) to filter labels before computing Cohen's Kappa coefficient. We averaged the Kappa coefficient across all worker pairs with at least 50 joint labeled instances. Using a threshold of 0.9 (i.e. keeping 90% of the labels) the Kappa was 0.58.

[5] We re-implemented their system since since we used a more recent Wikipedia dump and a different Wikification tool. The results we obtained are very close to the reported results.

[6] https://www.figure-eight.com/ (previously known as CrowdFlower)



Assessing the value of potential claims

In this task you are given a topic and possibly-related statements, each marked within a particular sentence.

For each candidate, you should select "**Accept**", if you think that the marked statement can be used "as is" during discourse, to directly support or contest the given topic. Otherwise, you should select "**Reject**".

If you selected "Accept", you should further indicate whether the marked text supports the topic ("Pro") or contests it ("Con").

Note, that if the marked text is non-coherent, hence cannot be used "as is" during a discussion about the topic, you should select "Reject".

Similarly, if the marked text supports/contests a *different* topic, even if it is somewhat related to the examined topic, you should typically select "Reject".

As a rule of thumb, if it is natural to say "I (don't) think that <topic>, because <marked statement>", then you should probably select "Accept". Otherwise, you should probably select "Reject".

Finally, if you are unfamiliar with the examined topic, please briefly read about it in a relevant data source like Wikipedia.

Examples for the topic "We should ban the sale of violent video games to minors" –

1. "The researchers found that adolescents that play violent video games are most at-risk for violent behavior (but without statistical significance)." -- **Accept / Pro.**

2. "Previous reports suggested that kids playing Doom are not at a greater risk for violent behavior." -- **Accept / Con.**

3. "The researchers found that adolescents that play violent video games are at no risk for violent behavior." -- **Reject.** Due to the prefix "found that", the marked text is not coherent and cannot be used "as is" while discussing the topic.

4. "While violent video games are often associated with aggressive behavior, recent studies are starting to suggest otherwise". – **Reject.** Due to the prefix "While", the marked text is not coherent and cannot be used "as is" while discussing the topic.

5. "Many people believe that some TV shows increase youth violence." -- **Reject.** The marked text is not *directly* supporting/contesting the topic.

Figure 1: Labeling guidelines exactly as they appeared in the Figure–Eight platform. Note that for the sake of this work we ignore the stance labeling (pro/con answers).

5 Results

To evaluate the performance of the network, we employ two sets of experiments. In the first we use the test-set topics in a manner similar to (Levy et al., 2017). In the second, we test our network on the UKP Sentential Argument Mining Corpus released in (Stab et al., 2018). Note that the UKP data is more inline with our goal than other argument mining tasks as it separates between sentences that support/contest a given topic from sentences that don't. A major difference between the UKP data and our test set is the source from which the sentences were taken – while we used Wikipedia, the UKP data comes from various sources, and hence it would test how well our approach generalizes to other text genres. Another important difference is in the definition of positive examples - we consider sentences containing relevant claims as positive, whereas they require that a sentence contain some supporting evidence or reasoning. The results on our test set are presented in subsection 5.1 and the results on the UKP data are presented in subsection 5.2.

5.1 Results on the Test Set

Figure 2 depicts the average number of CS (i.e., true positives) retrieved per the top $K = 10, 20, 50$ predictions. Both DNN_{pref} and DNN_{suff} seem to generalize well from the weak signal and provide comparable results to (Levy et al., 2017). More importantly, using the average score (DNN) yields the best performance, consistently outperforming (Levy et al., 2017) (with p-value < 0.005 for $K = 20, 50$ based on a two-tailed Wilcoxon test). The gap is most prominent for $K = 50$, where the DNN yields $\approx 30\%$ more CS compared to the the non–learning system that used a strict query with limited recall.

A major question is whether the learned system is able to generalize from the weak labels and identify CS that do not match the weak queries we started with. By construction, all sentences retrieved by the (Levy et al., 2017) system, match q_{strict}. From Table 7, we see that although the DNN system trained on sentences matching q_{that} or q_{CL} or both, 28% of the 2500 sentences predicted by the system, *do not match either*. Sentence $S4$ in Table 2 is an example of such a predicted sentence. The precision on those sentences, that are only known to contain the MC, is still considerably high – 0.22, and in fact comparable

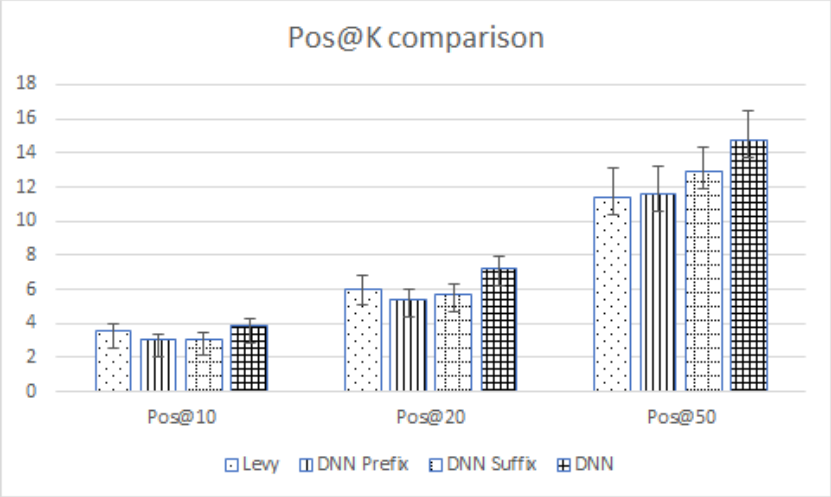


Figure 2: Results for the 50 test-set topics. $Pos@K$: The number of claim sentences (CS) out of the top K predicted sentences per topic, averaged over all test topics. Bars denote the standard error. Levy – Results reproduced on current index using the system described in (Levy et al., 2017); DNN – using the average score of DNN_{pref} and DNN_{suff}.

to the precision achieved in the restricted, low–recall system of (Levy et al., 2017). These results suggest that the DNN captures some general characteristics of CS, that are not limited to sentences that satisfy the two weak–signal queries we started with, q_{that} and q_{CL}. In addition, the precision on sentences containing one or both of the weak signals is even higher. Specifically, the precision on the subset of sentences matching q_{strict} is 0.42, a factor of two compared to the precision of (Levy et al., 2017) on this set of sentences. Thus, overall we were able to increase the potential recall from the restricted set of sentences matching q_{strict} to the full set of q_{MC} sentences, while also increasing the precision of the predictions from 0.23 to 0.3 (see the column q_{MC}).

Measure	System	q_{strict}	q^*_{that}	q^*_{CL}	q^*_{MC}	q_{MC}
Percent	Levy	100	0	0	0	100
	DNN	30	26	16	28	100
Precision	Levy	0.23	NA	NA	NA	0.23
	DNN	0.42	0.32	0.22	0.22	0.30

Table 7: Distribution and precision of predictions. q^*_{that}: matches q_{that} but not q_{CL}; q^*_{CL}: matches q_{CL} but not q_{that}; q^*_{MC}: matches q_{MC}, but not any of the other queries. Percent: out of the top 50 (or all if less available) predicted sentences matching the query. Precision: Precision of the corresponding candidates, calculated per topic, and averaged over test topics.

5.2 Results on the UKP Sentential Argument Mining Corpus

The UKP Sentential Argument Mining Corpus (Stab et al., 2018) contains a total of 25,492 labeled sentences (11,139 argumentative, 14,353 non-argumentative), divided to train (70%), validation (10%), and test sets (20%). The sentences are associated with one of 8 controversial topics – *abortion*, *cloning*, *death penalty*, *gun control*, *marijuana legalization*, *minimum wage*, *nuclear energy* and *school uniforms* – and were derived from the top 50 results of a Google query for the topic name, thus representing various genres and text types.

Adapting to the UKP dataset

In order to evaluate our method on the UKP dataset we had to adapt it to sentences that do not necessarily contain the MC. In our formulation, the MC was used as a natural point to divide the sentence into its prefix and suffix, which were then used by the appropriate networks. To overcome this difference, we applied DNN_{suff} (DNN_{pref}) to all possible suffixes (prefixes) and used the maximal score.

For a sentence S comprised of n words, $w_1, w_2, ..., w_n$, we define,

$$DNN'_{suff}(S) = max_{\{i:1..n\}} DNN_{suff}(w_i, ..., w_n)$$
$$DNN'_{pref}(S) = max_{\{i:0..n-1\}} DNN_{pref}(w_1, ..., w_{n-i})$$

The adapted scores may still be at a disadvantage because without the MC we don't have a way to select sentences that are more related to the topic. For this reason we add a similarity score $Score_{w2v}$ which is computed by taking the maximal word2vec (Mikolov et al., 2013) similarity of a word in the topic against all words in the sentence and then averaging across the words of the topic. More formally, for a topic T comprised of k words, $t_1, ..., t_k$,

$$Score_{w2v}(S) = avg_{\{i:1..k\}} max_{\{j:1..n\}} word2vec(t_i, w_j)$$

Finally, we define,

$$DNN'_{suff,w2v} = avg(DNN'_{suff}, Score_{w2v})$$
$$DNN'_{pref,w2v} = avg(DNN'_{pref}, Score_{w2v})$$

We employ the same setup that was used by (Stab et al., 2018) for the cross topic evaluation, in which the train set is comprised of the train part of all topics except for the tested topic. We use the training set only to tune the threshold from which we predict the positive class. To do so, we run the different DNN methods on the train set, compute the F_1 over all sentences, and choose the score that maximizes the F_1. This score is then used in the test set as the threshold that determines whether the network predicts a positive or not. Overall, this tuning was done 8 times, one for each train set induced by the left-out test topic.

Evaluation

The results are shown in table 8. Interestingly, our system achieves comparable results to the state of the art in the Accuracy and F_1 measures but without using human labels for training and without training on multiple text genres. These results also demonstrate the ability of the proposed method to generalize to topics that are not characterized by a single MC or that such a concept was not provided by the user. Note, the results reflect that our system and the baseline operate at different points on the precision/recall curve, choosing a different compromise between precision and recall. This is not surprising, given the choice of tuning the F_1 measure on the train set, however, it makes the comparison less obvious.

Method	Accuracy	F_1	Precision	Recall
UKP	0.69	0.66	0.75	0.52
DNN'_{suff}	0.57	0.65	0.51	0.90
$DNN'_{suff,w2v}$	0.67	0.69	0.59	0.83

Table 8: Results of the cross topic evaluation on the UKP dataset (averaged across the 8 topics). UKP method stands for the best supervised results reported in (Stab et al., 2018). From the networks combined with $w2v$ the $DNN'_{suff,w2v}$ performed best and is the one presented here.

It should be noted that the lower precision of our method may be explained by the different assumption on what an argumentative sentence is. Whereas Stab et al. (2018) reject sentences that contain claims but provide no evidence or reasoning, our network was designed to identify claims regardless of the

existence of a surrounding argument. Indeed, as mentioned in section 6.2, by sampling 50 false-positives we found that in 25% of the cases they contained relevant claims but with no evidence or reasoning.

6 Error Analysis

6.1 Test Set - 50 Topics

We analyzed the top 50 labeled predictions over three test topics for which the performance was above/near/below average (table 9).

Topic Text	Main Concept	Pos@50
We should further exploit wind power	Wind power	29
Private education brings more good than harm	Private school	13
We should protect whistleblowers	Whistleblower	9

Table 9: Test topics chosen for error analysis.

Each sentence rejected by the labelers was assigned one of the following types: **Factual** – a sentence with no argumentative content, that merely states a fact; **Different Topic** – a sentence that contains a claim for a different topic; **Other** – an assortment of problems such as bad sentence split, missing context, etc; and finally **Accept** – a sentence that should have been accepted by the labelers. The two main types of errors were Factual and Different topic, each accounting for 35% of the analyzed errors. The Accept type accounted for 18% of the rejected sentences, though this high number was mostly due to the Whistleblowers topic. We suspect that many such sentences were rejected because of bad claim boundary choices by the system [7]. Table 10 shows examples from the topic *"Private education brings more good than harm"*.

Error Type	Sentence
Different Topic	*Changes in private school enrollment is not a likely contributor to any changes in schools segregation patterns during that time.*
	In 2014 Hunt proposed that private schools should be required to form "partnerships" with local state schools if they wanted to keep their charitable status.
Factual	*Before enrolling the children, however, Mr. Brar ensured that the total cost of private school tuition would not exceed* $\$10,000$.
	The IRS announced in 1970 that private schools with racially discriminatory admissions policies would no longer receive tax exemptions
Accept	*Coaches were concerned that the private schools were winning a disproportionate amount of conference titles and had several unfair advantages.*
Other*	*It is clear that affording private education is a mere fantasy for these families.*

Table 10: Examples of sentences from the topic 'Private education brings more good than harm'. The sentences are split according to their assigned error type. * The example for the Other type was rejected because of a missing context – it is hard to judge this example without resolving the reference to "these families"

6.2 Test Set - UKP Dataset

We analyzed 50 random sentences from the UKP test set labeled as non-argumentative, on which the score of the $DNN'_{suff,w2v}$ network was higher than 0.9 (the average threshold obtained by tuning F_1 was 0.65). We add the following error type to the list above: **No Reasoning** – a sentence containing a claim with no supporting evidence or reasoning. The most frequent type of error was Factual, accounting

[7] We used a claim boundary component (Levy et al., 2014) on top of all systems in order to simplify the labeling task. This came at a cost of some CS being rejected due to errors in the boundary component.

for about 33% of the errors. The No Reasoning type accounted for about 25% of the errors, similar to the Different Topic type. Table 11 shows examples of No Reasoning sentences. These sentences contain text boundaries that are relevant claims, e.g., the boundary *the life in the womb is not human* in the first sentence, and thus are typical to sentences that our network was trained to find.

Topic	Sentence
abortion	*A question for those who believe in abortion, and that* **the life in the womb is not human**.
death penalty	**We need stricter laws and swift death penalty**.
minimum wage	*Myth:* **Raising the minimum wage will only benefit teens**.
marijuana legalization	*A small share of opponents (7%) say that while* **the recreational use of marijuana should be illegal**, *they do not object to legalizing medical marijuana.*

Table 11: Examples of sentences marked as No Reasoning from the UKP test set. The phrases marked in boldface are the suggested claim boundaries according to our analysis.

7 Discussion and Future Work

This work aims at making the first steps towards a search engine for argumentative content, by focusing on the problem of corpus wide claim detection. A variety of argument theories have been proposed throughout the years, which all agree on the importance of one argument component – the claim. Thus, properly addressing the problem of corpus wide claim detection seems like a key component in developing a full fledged argument search engine. Such an engine could add massive amounts of data to argument networks such as the world wide argument web (Rahwan et al., 2007), and further enhance decision processes in various ways. Using a similar methodology for evidence detection would be a natural way to push the boundary of existing work, e.g., (Rinott et al., 2015) from considering a pre-selected list of articles to searching full corpora. To the best of our knowledge this is the first work using weak supervision to train DNNs for argument mining, demonstrating the potential of this coupling in the field. Two directions for future work could increase the precision and coverage of our system. For increasing precision, we intend to employ a supervised approach, using labels on top of predictions from the weak–supervision approach, as it may help reach a reasonable prior of positive examples before starting the labeling effort. For the coverage, we intend to explore the same approach on top of sentences which do not necessarily contain the MC. This direction is challenging since it requires integrating a method for identifying whether a sentence is related to the topic, and would need to score sentences in which the prior for a claim is even lower.

During the error analysis on the 'Wind power' topic, we encountered the following high-scoring sentence – *"When Scratchy suggests that wind power is cheap and safe, Itchy chops Scratchy's head off with the blades of a wind turbine."*. On the one hand, Scratchy raises a legitimate claim, and on the other hand, Scratchy is a fictional character from the TV show The Simpsons. The example demonstrates a phenomenon that may be exasperated when moving from argument mining on pre-selected high-quality documents to mining large (possibly heterogeneous) text corpora – the phenomenon of claims made by unreliable sources. In extreme cases the claims made by such parties may be ridiculous or offensive and a practical search engine would need to detect and remove such claims.

Appendix A Index and Preprocessing

We processed the Wikipedia dump from May 1st, 2017. We applied text cleaning and sentence splitting using OpenNlp Sentence Detector [8] and an internal Wikification tool to wikify each sentence [9]. Starting from 5.4M articles, the sentence level index contains approximately 102M sentences. The inverted index along with the support for queries that mix surface form tokens with Wiki concepts was implemented as in (Levy et al., 2017).

We annotated the sentences retrieved by all queries using Stanford NER (Finkel et al., 2005), and removed sentences with a person/location entity after the MC (e.g., the sentence *"Yan warned Li that the Nationalist cause was doomed unless Li went to Guangdong"* for the topic *"Nationalism does more harm than good"* would be removed). This filter is motivated by our goal of retrieving general claim sentences for the topic, assuming that claims about specific entities are less interesting for potential users.

Appendix B Topics and Folds

The list of 150 topics is taken from (Levy et al., 2017) and split to dev/test in the same manner. Since here we use a learning system, we further split the dev set into a train set of 70 topics and a heldout set of 30 topics which was used to decide when to stop the learning. Tables 12 and 13 show the train topics and tables 14 and 15 show the topics of the heldout and test sets respectively.

Appendix C Released Data

We release two datasets, one containing $\approx 1.5M$ sentences matching the topics in this study based on the q_{MC} query, and one containing $2,500$ sentences predicted by our network and annotated for whether they contain a relevant claim or not (top 50 predictions across the 50 topics in the test set)[10]. The q_{MC} dataset can be found in the attached `q_mc_train.csv`, `q_mc_heldout.csv` and `q_mc_test.csv` files, according to the topics split used in the learning/evaluation process. A detailed description of this dataset appears in the `readme_mc_queries.txt` file. The system prediction dataset is in the `test_set.csv` file with a corresponding description in the `readme_test_set.txt` file.

[8] http://opennlp.apache.org/

[9] To be described in a separate publication.

[10] The datasets can be downloaded from `http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`

#	Id	Topic Text	Main Concept
1	1	We should ban the sale of violent video games to minors	Video game controversies
2	2	We should legalize doping in sport	Doping in sport
3	3	We should ban boxing	Boxing
4	4	We should abolish intellectual property rights	Intellectual property
5	5	We should protect endangered species	Endangered species
6	6	Operation Cast Lead was justified	Gaza War (2008-09)
7	7	Tower blocks are advantageous	Tower block
8	8	Private universities bring more good than harm	Private university
9	9	We should disband ASEAN	Association of Southeast Asian Nations
10	10	The free market brings more good than harm	Free market
11	11	We should ban child actors	Child actor
12	12	Religion does more harm than good	Religion
13	13	We should ban cosmetic surgery	Plastic surgery
14	14	Same sex marriage brings more good than harm	Same-sex marriage
15	15	Reality television does more harm than good	Reality television
16	16	Internet censorship brings more good than harm	Internet censorship
17	17	Socialism brings more harm than good	Socialism
18	18	We should ban beauty contests	Beauty pageant
19	19	We should adopt vegetarianism	Vegetarianism
20	20	We should adopt libertarianism	Libertarianism
21	21	The internet brings more harm than good	Internet
22	22	Science is a major threat	Science
23	23	Suicide should be a criminal offence	Suicide
24	24	Nationalism does more harm than good	Nationalism
25	25	The atomic bombings of Hiroshima and Nagasaki were justified	Atomic bombings of Hiroshima and Nagasaki
26	26	Casinos bring more harm than good	Casino
27	27	We should lower the age of consent	Age of consent
28	28	We should abolish standardized tests	Standardized test
29	29	We should ban extreme sports	Extreme sport
30	30	The alternative vote is advantageous	Instant-runoff voting
31	31	Illegal immigration brings more harm than good	Illegal immigration
32	32	We should subsidize renewable energy	Renewable energy
33	33	We should end daylight saving times	Daylight saving time
34	34	We should further exploit geothermal energy	Geothermal energy
35	35	Assisted suicide should be legalized	Assisted suicide
36	36	Security hackers do more harm than good	Hacker (computer security)
37	37	We should disband the United Nations	United Nations
38	38	We should ban hate sites	Hate speech
39	39	We should privatize future energy production	Energy development
40	40	Child labor should be legalized	Child labour
41	41	The paralympic games bring more good than harm	Paralympic Games
42	42	Chain stores bring more harm than good	Chain store
43	43	We should subsidize Habitat for Humanity International	Habitat for Humanity
44	44	We should subsidize public art	Public art
45	45	IKEA brings more harm than good	IKEA
46	46	We should ban online advertising	Online advertising
47	47	Mixed-use development is beneficial	Mixed-use development
48	48	We should ban Greyhound racing	Greyhound racing
49	49	The Israeli disengagement from Gaza brought more harm than good	Israeli disengagement from Gaza
50	50	We should not subsidize single parents	Single parent

Table 12: Train topics 1-50

#	Id	Topic Text	Main Concept
51	51	We should ban private military companies	Private military company
52	52	Coaching brings more harm than good	Coaching
53	53	We should abandon disposable diapers	Diaper
54	54	PayPal brings more good than harm	PayPal
55	55	The Internet archive brings more harm than good	Internet Archive
56	56	The 2003 invasion of Iraq was justified	2003 invasion of Iraq
57	57	Virtual reality brings more harm than good	Virtual reality
58	58	Internet cookies bring more harm than good	HTTP cookie
59	59	Magnet schools bring more harm than good	Magnet school
60	60	The right to strike brings more harm than good	Strike action
61	61	We should subsidize student loans	Student loan
62	62	We should abandon Youtube	YouTube
63	63	Ecotourism brings more harm than good	Ecotourism
64	64	Academic freedom is not absolute	Academic freedom
65	65	Homeschooling should be banned	Homeschooling
66	66	We should abolish the US Electoral College	Electoral College (United States)
67	67	Generic drugs should be banned	Generic drug
68	68	We should fight global warming	Global warming
69	69	We should fight for Quebecan Independence	Quebec sovereignty movement
70	70	We should subsidize newspapers	Newspaper

Table 13: Train topics 51-70

#	Id	Topic Text	Main Concept
1	71	The freedom of speech is not absolute	Freedom of speech
2	72	We should criminalize blasphemy	Blasphemy
3	73	Holocaust denial should be a criminal offence	Holocaust denial
4	74	Television does more harm than good	Television
5	75	We should subsidize higher education	Higher education
6	76	We should ban organic food	Organic food
7	77	Urbanization does more harm than good	Urbanization
8	78	We should adopt direct democracy	Direct democracy
9	79	We should ban lotteries	Lottery
10	80	We should close the Guantanamo Bay detention camp	Guantanamo Bay detention camp
11	81	We should abandon the insanity plea	Insanity defense
12	82	We should protect coral reefs	Coral reef
13	83	We should disband NASA	NASA
14	84	We should abolish nuclear weapons	Nuclear weapon
15	85	We should cancel the speed limit	Speed limit
16	86	Randomized controlled trials bring more harm than good	Randomized controlled trial
17	87	Anarchism brings more good than harm	Anarchism
18	88	We should subsidize public service broadcasters	Public broadcasting
19	89	We should ban labor organizations	Trade union
20	90	Pride parades bring more harm than good	Pride parade
21	91	Paternity leave brings more harm than good	Parental leave
22	92	Tabloid journalism brings more harm than good	Tabloid journalism
23	93	We should disband UNESCO	UNESCO
24	94	We should disband the National Rifle Association	National Rifle Association
25	95	Second Life brought more harm than good	Second Life
26	96	Economic sanctions bring more harm than good	Economic sanctions
27	97	Vietnam War was justified	Vietnam War
28	98	Animal slaughter is not justified	Animal slaughter
29	99	We should raise the corporate tax	Corporate tax
30	100	Division of labor is a major threat	Division of labour

Table 14: Heldout topics

#	Id	Topic Text	Main Concept
1	101	Affirmative action brings more good than harm	Affirmative action
2	102	We should ban gambling	Gambling
3	103	We should abolish the monarchy	Monarchy
4	104	Atheism is the only way	Atheism
5	105	We should further exploit wind power	Wind power
6	106	We should legalize polygamy	Polygamy
7	107	We should further exploit hydroelectric dams	Hydroelectricity
8	108	We should privatize water supply	Water supply
9	109	We should legalize prostitution	Prostitution
10	110	Zoos bring more harm than good	Zoo
11	111	Private education brings more good than harm	Private school
12	112	Recall elections are beneficial	Recall election
13	113	We should further exploit nuclear power	Nuclear power
14	114	We should abolish temporary employment	Temporary work
15	115	Surrogacy should be banned	Surrogacy
16	116	Progressive tax is beneficial	Progressive tax
17	117	We should ban alcoholic beverages	Alcoholic drink
18	118	We should ban abortions	Abortion
19	119	Astrology brings more harm than good	Astrology
20	120	Embryonic stem cell research brings more good than harm	Embryonic stem cell
21	121	We should abolish the Olympic Games	Olympic Games
22	122	We should end athletic scholarships	Athletic scholarship
23	123	Social media does more harm than good	Social media
24	124	We should disband the United Nations Security Council	United Nations Security Council
25	125	We should legalize insider trading	Insider trading
26	126	We should prohibit hydraulic fracturing	Hydraulic fracturing
27	127	We should prohibit corporal punishment	Corporal punishment
28	128	We should disband NATO	NATO
29	129	We should abolish the two-party system	Two-party system
30	130	Capital punishment brings more harm than good	Capital punishment
31	131	We should abolish term limits	Term limit
32	132	We should protect whistleblowers	Whistleblower
33	133	Twitter brings more harm than good	Twitter
34	134	ISO brings more harm than good	International Organization for Standardization
35	135	Conscientious objectors are justified	Conscientious objector
36	136	The American Bar Association brings more harm than good	American Bar Association
37	137	Digital rights management brings more harm than good	Digital rights management
38	138	We should ban the Church of Scientology	Church of Scientology
39	139	eBay brings more good than harm	EBay
40	140	We should abolish the caste system in India	Caste system in India
41	141	We should abolish infant baptism	Infant baptism
42	142	EHRs bring more harm than good	Electronic health record
43	143	Wildlife management brings more good than harm	Wildlife management
44	144	We should tax plastic bags	Plastic bag
45	145	The energy industry should be nationalized	Energy industry
46	146	We should fight protectionism	Protectionism
47	147	We should limit genetic testing	Genetic testing
48	148	We should end manned spaceflights	Human spaceflight
49	149	Extra-curricular activity should be mandatory	Extracurricular activity
50	150	We should abolish homework	Homework

Table 15: Test topics

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, June. Association for Computational Linguistics.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark, September. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada, July. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 69–78. ACM.

Ivan Habernal and Iryna Gurevych. 2016. 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 (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, August. Association for Computational Linguistics.

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.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Anirban Laha and Vikas Raykar. 2016. An empirical evaluation of various deep learning architectures for bi-sequence classification tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2762–2773, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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, August. Dublin City University and Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark, September. Association for Computational Linguistics.

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*.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Artificial intelligence and law*, pages 225–230. ACM.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Iyad Rahwan, Fouad Zablith, and Chris Reed. 2007. Laying the foundations for a world wide argument web. *Artificial intelligence*, 171(10-15):897–921.

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, September. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark, September. Association for Computational Linguistics.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Semantic processing of legal texts. chapter Approaches to Text Mining Arguments from Legal Cases, pages 60–79. Springer.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.