Context Dependent Claim Detection

Ran Levy Yonatan Bilu Daniel Hershcovich Ehud Aharoni Noam Slonim
IBM Haifa Research Lab / Mount Carmel, Haifa, 31905, Israel
{ranl,yonatanb,danielh,aehud,noams}@il.ibm.com

Abstract

While discussing a concrete controversial topic, most humans will find it challenging to swiftly raise a diverse set of convincing and relevant claims that should set the basis of their arguments. Here, we formally define the challenging task of automatic claim detection in a given context and discuss its associated unique difficulties. Further, we outline a preliminary solution to this task, and assess its performance over annotated real world data, collected specifically for that purpose over hundreds of Wikipedia articles. We report promising results of a supervised learning approach, which is based on a cascade of classifiers designed to properly handle the skewed data which is inherent to the defined task. These results demonstrate the viability of the introduced task.

1 Introduction

The ability to argue in a persuasive manner is an important aspect of human interaction that naturally arises in various domains such as politics, marketing, law, and health-care. Furthermore, good decision making relies on the quality of the arguments being presented and the process by which they are resolved. Thus, it is not surprising that argumentation has long been a topic of interest in academic research, and different models have been proposed to capture the notion of an argument (Toulmin, 1958; Freeley and Steinberg, 2008). A fundamental component which is common to all these models is the concept of a *claim* (or *conclusion*). Specifically, at the heart of every argument lies a single claim, which is the assertion the argument aims to prove. Given a concrete topic, or context, most humans will find it challenging to swiftly raise a diverse set of convincing and relevant claims that should set the basis of their arguments. The purpose of this work is to formally define the challenging task of automatic claim detection *in a given context*, to outline a preliminary solution to this task, and to assess its performance over annotated real world data, collected specifically for this purpose.

In his classical argument model, Toulmin defined a claim as a conclusion whose merit must be established (Toulmin, 1958). Since we are interested not in detecting claims in general (Mochales Palau and Moens, 2009; Teufel, 1999), but rather in detecting claims that are specifically relevant to a pre-defined concrete context, we suggest a definition with a more functional flavor. In practice, we found this definition easy to convey to human labelers, and consequently feasible to capture by automatic detection methods. In particular, we define the following two concepts:

- **Topic** – a short phrase that frames the discussion.
- **Context Dependent Claim (CDC)** – a general, concise statement that directly supports or contests the given Topic.

Given these definitions, as well as a few more detailed criteria to reduce the variability in the manually labeled data, human labelers were asked to detect CDCs for a diverse set of Topics, in relevant Wikipedia articles. The collected data, that were used to train and assess the performance of the statistical models, are now freely available upon request for academic research.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1489–1500, Dublin, Ireland, August 23-29 2014.

The distinction between a CDC and other related texts can be quite subtle, as illustrated in Table 1. For example, automatically distinguishing a CDC like S1 from a statement that simply defines a relevant concept like S2, from a claim which is not relevant enough to the given Topic like S3, or from a statement like S4 that merely repeats the given Topic in different words, is clearly challenging. Further, CDCs can be of different flavors, ranging from factual assertions like S1 to statements that are more of a matter of opinion (Pang and Lee, 2008) like S5, adding to the complexity of the task. Finally, our data suggest that even if one focuses on Wikipedia articles that are highly relevant to the given Topic, only $\approx 2\%$ of their sentences include CDCs. Moreover, as illustrated in Table 2, detecting the exact CDC boundaries is far from trivial, as in a typical single Wikipedia sentence there are many optional boundaries to consider. Thus, we are faced with a large number of candidate CDCs, of which only a tiny fraction represents positive examples, that might be quite reminiscent of some of the negative examples. Nonetheless, as we demonstrate, a supervised learning approach – which is based on a cascade of classifiers, carefully designed to properly handle the exceptionally skewed data – can address these difficulties to attain promising results.

Topic: The sale of violent video games to minors should be banned		
S1	Violent video games can increase children's aggression	V
S2	Video game addiction is excessive or compulsive use of computer and video games that interferes with daily life	X
S3	Many TV programmers argue that their shows just mirror the violence that goes on in the real world	X
S4	Violent video games should not be sold to children	X
S5	Video game publishers unethically train children in the use of weapons	V

Table 1: Examples for CDCs and for statements that should not be considered as CDCs. The V and X indicate if the candidate is a CDC for the given Topic, or not, respectively.

Topic: The sale of violent video games to minors should be banned	
S1	Because violence in video games is interactive and not passive, critics such as Dave Grossman and Jack Thompson argue *that* **violence in games hardens children to unethical acts**, calling first-person shooter games "murder simulators", although no conclusive evidence has supported this belief.

Table 2: A CDC is often only a small part of a single Wikipedia sentence – e.g., the part marked in bold in this example. Detecting the exact CDC boundaries represents an additional challenge.

In summary, the key contribution of this work is three fold: we define the new task of Context Dependent Claim Detection; introduce a novel manually labeled benchmark dataset, collected specifically for this task; and outline an automatic solution for which we report first results over these data. These results are encouraging, demonstrating the viability of the introduced task.

2 Task Definition and Related Work

We assume that we are given a Topic and a relatively small set of relevant free-text articles, provided either manually or by automatic retrieval methods (Macdonald et al., 2010; Zhang et al., 2007). Our goal is to automatically pinpoint CDCs within these documents. We further require that the detected CDCs are reasonably well phrased, so that they can be instantly and naturally used in a discussion about the given Topic.This task, which we term *Context Dependent Claim Detection (CDCD)*, can be of great practical importance in decision support and persuasion enhancement, in various domains where relevant massive corpora are available for mining.

CDCD can be seen as a sub-task in the emerging wider field of argumentation mining that involves identifying argumentative structures within a document, as well as their potential relations

(Mochales Palau and Moens, 2009; Cabrio and Villata, 2012; Wyner et al., 2012). However, CDCD has several distinctive key features. Most importantly, as implied by its name, a CDC is defined with respect to a given context – the input Topic. Thus, identifying general characteristics of a claim-like statement as done in (Mochales Palau and Moens, 2009) is not sufficient, since one should further identify the relevance of the candidate claim to the Topic. In addition, we do not restrict ourselves to a particular domain nor to structured data (Mochales Palau and Moens, 2009), but rather consider free-text Wikipedia articles in a diverse range of subject matters. Moreover, in CDCD we require pinpointing the exact claim boundaries, which do not necessarily match a whole sentence or even a clause in the original text, thus adding a significant burden to the task, compared to classical tasks that are focused on sentence classifications (Guo et al., 2011).

CDCD also shares some relations with Argumentative Zoning (Teufel, 1999; Guo et al., 2011). There, the aim is to divide the text of a scientific article into "zones", each characterized by the rhetorical nature of its content. However, our work is not limited to scientific literature that often has a more objective and less persuasive style. Further, as mentioned, we go beyond sentence classification, aiming to detect the exact claim boundaries, and require detecting only claims relevant to a given Topic, rather than just any claim mentioned in a given article.

Finally, another important line of research is the Textual Entailment (TE) framework (Dagan et al., 2009). In this framework, a text fragment, T, is said to entail a textual hypothesis H if the truth of H can be most likely inferred from T. While TE can be an important underlying utility in CDCD, and perhaps vice versa, the tasks are quite different. For example, common instances of TE are rephrases or summarizations of a sentence; however these cannot serve to support or contest a given Topic, as they merely repeat it (Table 1, S4). Furthermore, TE focuses on factual assertions, which can be true or false, whilst CDC may represent a relevant opinion that perhaps does not have a strict truth value associated to it (Table 1, S5). More generally, TE is typically focused on declarative statements. However, persuasion and argumentation often have an emotional aspect and thus may involve additional sentence types. Correspondingly, in our framework it is quite natural that the Topic, or the associated CDCs, will correspond to imperative sentences, or even to exclamatory sentences.

3 Data

Our supervised learning approach relies on labeled data that were collected as described below. A detailed description of the labeling process is given in (Aharoni et al., 2014). Due to the high complexity of the labeling task, we worked with in-house labelers which were provided with detailed guidelines, and went through rigorous training.

At the core of the labeling guidelines, we outlined the definition of a CDC as *a general, concise statement that directly supports or contests the given Topic*. In practice, the labelers were asked to label a text fragment as a CDC if and only if it complies with *all* the following five criteria:

- **Strength** – Strong content that directly supports/contests the Topic.
- **Generality** – General content that deals with a relatively broad idea.
- **Phrasing** – The labeled fragment should make a grammatically correct and semantically coherent statement.
- **Keeping text spirit** – Keeps the spirit of the original text.
- **Topic unity** – Deals with one topic, or at most two related topics.

The guidelines further included concrete examples, taken from Wikipedia articles, to clarify these criteria. When in doubt, the labelers were naturally asked to make a judgment call. The labelers work was carefully monitored, and they were provided with detailed feedback as needed.

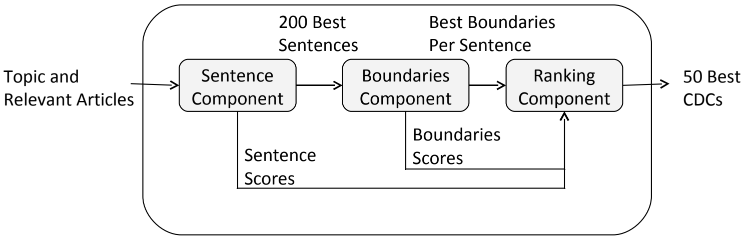
We selected at random 32 debate motions from *http://idebate.org/debatabase*, covering a wide variety of topics, from atheism to the US responsibility in the Mexican drug wars. Each motion served as a single Topic and went through a rigorous labeling process, consisted of three stages. First, given a Topic, 5 labelers searched Wikipedia independently for articles that they believe contain CDCs. Next, each of the articles identified in this search stage was read by 5 labelers, that worked independently to detect candidate CDCs. Finally, each of the candidate CDC proposed in the previous stage, was examined by 5 labelers that independently decided whether to confirm or reject the candidate. For the purposes of this work, we only considered candidate CDCs that were confirmed by a majority, i.e., by at least three labelers participating in the confirmation stage. The resulting labeled CDCs correspond to claims that can be naturally used in a discussion about the given Topic.

Through this process, for the 32 examined Topics, a total of 326 Wikipedia articles were labeled, yielding a total of 976 CDCs. Thus, even when considering articles that are presumably relevant to the given Topic, on average only 2 out of 100 sentences include a CDC. On the other hand, it should be noted that it was not clear to begin with that Wikipedia articles will contain CDCs that satisfy our relatively strict labeling guidelines. Nonetheless, on average, the labeling process yielded around 30 CDCs per Topic. Finally, the average Kappa agreement between pairs of labelers in the confirmation stage was 0.39, which is a relatively high agreement considering the complexity of the labeling task and the inherent elusiveness of the involved concepts.

4 Technical Approach

Our CDCD approach is designed as a cascade, or funnel, of three components (depicted in Figure 1), which receives as input a Topic along with relevant articles and should output the CDCs contained therein. The purpose of the funnel is to gradually focus on smaller and smaller CDC-containing text segments, while filtering out irrelevant text. Thus, the cascade divides the high level CDCD problem into smaller and more tangible problems – given an article, detect sentences that include CDCs; given a sentence, detect the exact CDC boundaries; given a set of CDC candidates, rank them so that true candidates are on top.

Figure 1: High level design of our CDCD approach. The indicated numbers are the ones used in our experiments, and in general should be determined based on the data and use case.



To appreciate the need for this cascade, let us first consider the scale of this detection problem. In our labeled data, per Topic we have an average of 10 relevant Wikipedia articles that contain at least 1 CDC. Each article contains an average of 155 sentences, each sentence spans on average 23 words, i.e., ≈ 200 sub-sentences, each of which may represent a candidate CDC. Thus, in principle, for each Topic we consider around 300,000 candidate CDCs, of which typically only 30 represent positive examples. By breaking the problem into independent sub-problems, at each stage the skew between positive examples and negative examples is less daunting, thus easier to handle by classical machine learning techniques. In addition, since much surplus text is filtered along the cascade, "downstream" components typically examine much smaller amounts of text, and thus can plausibly make use of more demanding algorithms. Finally, this conceptual separation naturally allows to develop features tailored individually to each task; for example, the grammatical correctness of a text fragment is clearly relevant for boundaries detection, while being irrelevant when classifying whole sentences.

In general, each component was developed independently within the classical supervised learning paradigm. Namely, numeric features are extracted from binary-labeled text segments, and are used to train a classifier. Next, this classifier is used to assign a score to each incoming test candidate and high-scoring candidates are passed on to the next component. In addition, rule-based filters might be used to discard some of the candidates. Note, while developing a "downstream" component we implicitly assumed that the previous "upstream" components have worked perfectly. Hence, for example, the *train-*

ing data for the boundary-detection component comprised only of sentences that truly contain CDCs. In what follows, we discuss each component in greater detail.

4.1 Sentence Component

The **Sentence Component** is responsible for detecting CDC-sentences, that is, to determine whether a candidate sentence contains a CDC or not. Some sentences contain more than one CDC but this is not very common. Hence, we consider this as a binary classification problem. The component receives an average of 1500 sentences per Topic and passes the top scoring 200 sentences to the next component. Specifically, we used Logistic Regression (LR) classifier due to its efficiency and its model interpretability, and focused our efforts on developing highly discriminative features for this classifier.

Since our focus is on detecting claims that are relevant to a given Topic, we naturally developed two main types of features – **Context features**, which examine the relation between the candidate sentence and the Topic; and **Context-free features**, which rely solely on the content of the candidate sentence, aiming to capture the probability it includes a “claim like” statement. For computing the context-features, we use the topic as it appears in debatabase (see Section 3). Specifically, the most dominant features we identified included:

MatchAtSubject: Cosine similarity between the Topic and the subjects of the candidate sentence – namely, all tokens that are marked as the subject of some sub-tree in the sentence ESG parse tree (McCord et al., 2012).

ExpandedCosineSimilarity: Cosine similarity between the Topic and semantic expansions of the candidate sentence. We use WordNet (Miller, 1995) to expand nouns into synonyms, hypernyms and hyponyms.

ESGFeatures: Binary features obtained from the ESG parser (McCord et al., 2012). The most prominent among them is an indicator of whether the sentence contains the token ”that” that is assigned the ”conjugate” feature by the ESG – see, for example, the emphasized “that” in the example in Table 2. Other features include: verb in present tense, infinitive verb, year, and named location.

SubjectivityScore: A classifier-based score that captures the degree of subjectivity in the sentence (Raykar et al., 2014).

Sentiment: Ratio of sentiment words in the sentence, based on a list of sentiment words from (Hu and Liu, 2004).

In addition to these Context features and Context-free features, we also developed a feature that represents a mix of these two types, that was proven essential to our performance, and relied on an extension of the Sequential Pattern Matching (SPM) algorithm (Srikant and Agrawal, 1996). Specifically, for this **SequentialPatternMatch** feature, each sentence token was encoded as a tuple describing several attributes for that token – e.g., the token’s text, the token’s POS tag, and various binary indicators, indicating if the token is a sentiment word, if it is mentioned in the given Topic, if it is included in an automatically learned lexicon of “claim words”, and if it is identified by a NER utility (Finkel et al., 2005). A variant of the SPM algorithm (Srikant and Agrawal, 1996) was then used to detect patterns that characterize CDC-sentences, and these patterns were added to the features examined by the LR classifier. Specifically, each of these feature values was set to 1 if a candidate sentence had a match with the relevant pattern, and to 0 otherwise. For example, in Table 2, the word ”that” is encoded as [that,IN,CDC] implying it is included in the “claim words” lexicon with POS tag IN; the word ”games” is encoded as [games,NNS,Topic] implying it is mentioned in the Topic with POS tag NNS; and the word ”unethical” is encoded as [unethical,JJ,Sentiment] implying it is a sentiment word with POS tag JJ. Correspondingly, in this sentence there is a match to the sequential pattern: [that,IN], [Topic], [Sentiment], which is one of the patterns detected automatically by our algorithm, as characterizing CDC-sentences. A more detailed description of this extended SPM approach will be given elsewhere.

It is worthwhile mentioning that one can envision this component as being broken up into two: One component that detects general claim-sentences, regardless of whether or not they relate to the Topic, based on the context-free features; Another component will detect relation to the Topic, regardless of whether or not the sentence is a claim. (Or some variation of this setting.)

The problem with this approach, as we see it, is that it greatly complicates the annotation guidelines and the associated annotation work. That is, without a topic, it is less clear how to define what a claim is, and deciding when a sentence is related to the topic is bound to be highly subjective. Furthermore, taking this approach would require adding additional detection and confirmation stages, lowering the amount of collected annotated data. For these reasons we have adopted the combined approach, even though it makes error analysis more difficult - without manual analysis it is not clear whether the errors are sentences which do not contain claims or which are unrelated to the topic, or both.

4.2 Boundaries Component

The **Boundaries Component** is responsible for detecting the exact CDC boundaries within CDC-sentences. Notice, that our definition of a CDC and the associated labeling guidelines – that gave rise to our ground-truth data – imply that in free text articles a CDC often do not correspond to an easily identified sub-tree in the sentence parse tree. For example, since we are interested in detecting focused claims the labelers are often led to mark a concise claim rather than a compound claim as in the following sentence – *"The argument of deprivation states that* **abortion is morally wrong** *because it deprives the fetus of a valuable future"*. Note that choosing the boundaries from *"abortion"* to *"future"* would have included two distinct claims. Similarly, since the labelers are guided to prefer more general versions of the CDC, as long as the original text spirit is kept, determining where the CDC should start could be quite a subtle decision. Thus, the exact CDC boundaries often rely on the semantics of the text, and not just on its grammatical structure. Correspondingly, identifying the exact CDC boundaries is far from trivial.

Based on similar considerations to those mentioned above, we divide this component into two sub-components.

Boundaries Coarse Filter: This sub-component is based on a Maximum Likelihood probabilistic model that given a sentence, selects the 10 sub-sentences whose boundaries most probably correspond to a CDC. Specifically, given a sentence, for each of its sub-sentences[1] we consider the token preceding it; the token with which it starts; the token with which it ends; and the token following it, where a token here can be a word or a punctuation mark. Given these four tokens, the algorithm estimates the probability that this sub-sentence represents a CDC. For practical purposes, the probability is estimated naively, by assuming that each token is independent of the others. In addition, the Boundaries Coarse Filter employs simple rules to filter out trivial cases such as sub-sentences that do not contain a verb and a noun, or sub-sentences for which the parse root is tagged as a sentence-fragment.

Boundaries Fine-Grained Filter: This sub-component is based on a LR classifier that selects one sub-sentence out of the 10 provided by the Boundaries Coarse Filter. Here as well we considered Context-free features and Context features, where the former type were typically weighted as more dominant by the LR classifier. Importantly, though, the Context-free features examined by this sub-component relied on the division of the entire sentence, as implied by the examined boundaries. Specifically, the candidate boundaries induce a division of the containing sentence into three parts: prefix, candidate body, and suffix, where the prefix and/or suffix might be empty. The features are then calculated for each of these three parts independently. Thus, for example, the presence of the word "that" in the prefix as opposed to its presence in the candidate body, will increase or decrease the confidence of the examined boundaries, respectively. In addition, the LR classifier considered features derived from the probabilistic model defined by the Boundaries Coarse Filter, that also aim to assess the probability that the examined boundaries yield a CDC.

Next, we elaborate on some of the dominant features examined by the Boundaries Fine-Grained Filter.

CDC-Probability features: These features indicate the conditional probability that the examined boundaries define a CDC, given the tokens around and within these boundaries. For example, the **Word-Before-Word-After** numeric feature, denoted $P(ta, tb)$, is defined as follows. Let $\{t_1, \ldots, t_n\}$ represent the list of tokens in a sentence, where a token is a word or a punctuation mark, then $P(ta, tb)$ is the probability that the sub-sentence $\{t_i, \ldots, t_j\}$ represents a CDC, given that $t_{i-1} = t_a, \ t_{j+1} = t_b$, as

[1] Here, a "sub-sentence" is any consecutive sequence of three tokens or more, that is included in the examined sentence.

estimated from our training data. Similarly, the **Word-Before-First-PoS** feature is based on the estimated conditional probability that the candidate defined by the examined boundaries is a CDC, given the token before the boundaries, t_{i-1}, and the POS-tag of the first token within the boundaries, t_i. Other features of this type include the conditional probability based on the presence of single tokens within the boundaries, and the initial score assigned to the examined boundaries by the Boundaries Coarse Filter.

Sentence-Probability features: These features aim to indicate the probability that the examined boundaries induce a grammatically correct sentence. For this purpose we examine a set of 100,000 presumably grammatically correct sentences, taken from a separate set of Wikipedia articles, and estimate the probability of each word to appear in a given position in a valid sentence. Next, given the examined boundaries, we ask for each of its first three tokens and each of its last three tokens, what is the probability of having a grammatically correct sentence, given that the observed token is in its observed position.

ModifierSeparation: The ESG parser (McCord et al., 2012) describes the modifiers of its parsed tokens, such as the object of a verb. Typically, a token and its modifier should either be jointly included in the CDC, or not included in it. This notion gave rise to several corresponding features.

Parse Sentence Match: These are binary features that indicate whether the examined boundaries correspond to a sub-tree whose root is labeled ”S” (sentence) by the Stanford parser (Socher et al., 2013) or by the ESG parser (McCord et al., 2012), while parsing the entire surrounding sentence.

”that-conj” matches CDC: A binary feature indicating whether in the ESG parsing we have a subordinator ”that” token, whose corresponding covered text matches the examined boundaries.

DigitCount: Counts the number of digits appearing in the sentence – before, within, and after the examined boundaries.

UnbalancedQuotesOrParenthesis: Binary features, indicating whether there is an odd number of quote marks, or unbalanced parenthesis, within the examined boundaries.

4.3 Ranking Component

The **Ranking Component** is responsible for the final scoring of the CDC candidates. It is also based on a LR classifier, that considers the scores of all previous components, as well as additional features described below. A simpler alternative could have been to rely solely on the initial sentence component ranking. However, since CDCs often correspond to much smaller parts of their surrounding sentence, considering the scores of all previous components is more effective. In contrast to the components described above, for which the training set is fully defined by the labeled data, the Ranking Component needs be trained also on the output of its “upstream” components, since it relies on the scores produced by these components.

In addition, the Ranking Component is using the following features:

CandidateComplexity, a score based on counting punctuation marks, conjunction adverbs (e.g., ”likewise”, ”therefore”), sentiment shifters (e.g., ”can not”, ”undermine”) and references, included in the candidate CDC.

Sentiment, **ExpandedCosineSimilarity** and **MatchAtSubject**, as in the Sentence Component above, estimated specifically for the CDC candidate.

5 Experiments

We describe the results of running the cascade of aforementioned components, in the designed order, in a Leave-One-Out (LOO) fashion, over 32 Topics. In each LOO fold, the training data consisted of the labeled data for 31 Topics, while the test data consisted of articles that included at least one CDC for the designated test Topic.

The **Sentence Component** was run with the goal of selecting 200 sentences for the test Topic, and sorting them so that CDC-containing sentences are ranked as high as possible. As shown in Table 3, the mean precision and recall of this component, averaged across all 32 folds, were 0.09 and 0.73, respectively. When looking at the top scoring 50 sentences per Topic, the mean precision and recall are 0.18 and 0.4, respectively. As evident by the last row in Table 3, these results are way beyond a trivial random selection of sentences, indicating that the Sentence Component is capturing a strong statistical signal associated with CDCs.

	Precision	Recall	Precision @ 50	Recall @ 50
mean	0.09	0.73	0.18	0.4
std	0.05	0.19	0.10	0.21
min	0.01	0.27	0.02	0.10
max	0.18	1.00	0.40	1.00
rand	0.02	0.13	0.02	0.03

Table 3: Sentence component. Last line indicates the expected values had selection been made at random.

Next, we employed the two sub-components of the **Boundaries Component**. First, the Boundaries Coarse Filter selected 10 sub-sentences for each candidate sentence. Recall that for sentences which actually contain CDCs, the aim is to have this CDC kept among the selected 10 sub-sentences. As shown in Table 4, this happens for 98% of the CDC-containing sentences (see Table 4). In other words, if the examined sentence included a CDC, the Boundaries Coarse Filter almost always included this CDC as one of the top 10 sub-sentences it suggested for that sentence. In the second step, for each candidate sentence, the Boundaries Fine-Grained Filter sorted the 10 sub-sentences proposed by the Boundaries Coarse Filter, aiming to have the CDC – if one exists – at the top. As indicated in Table 4, if indeed a CDC was present amongst the 10 sub-sentences sorted by this component, then it was ranked first in 50% of the cases. These results as well are clearly way beyond what is expected by random sorting, indicating a strong statistical signal that was properly captured by this component.

	Boundaries Coarse Filter Recall	Boundaries Fine-Grained Filter Recall
mean	0.98	0.50
std	0.39	0.16
min	0.67	0.25
max	1.00	1.00
rand	0.04	0.004

Table 4: Boundaries component - The left column relates to the fraction of sentences where the labeled CDC is among the top 10 candidates ranked by the Coarse Filter. The right column relates to to the fraction of sentences where the labeled CDC is identified correctly by the Fine-Grained Filter. The last row indicates the expected values had selection been made at random.

Finally, the **Ranking Component** combines the scores generated in the previous steps, as well as additional features, to set the final order of CDC candidates. The goal of this component – similar to that of the entire CDCD task – is to select 50 CDC candidates with high precision. Note, that on average, there are around 30 labeled CDCs per Topic. Thus, on average, the maximal precision at 50 should be around 0.6. As indicated in Table 5, our final precision at 50, averaged across all 32 folds, was 0.12, which is again way beyond random performance. Focusing at our top predicions naturally results with even higher precision – for example, the precision of our top 5 predictions was on average 0.23.

It should be noted that the analysis presented here is fairly strict. A predicted CDC is considered as True Positive if and only if it precisely matches a labeled CDC, that was confirmed as such by at least three labelers. Thus, for example, if a predicted CDC was confirmed by only two out of five annotators, it will be considered as an error in the analysis above. Furthermore, if the predicted CDC has a significant overlap with a labeled CDC, it will still be considered as an error, even if it represents a grammatically correct variant of the labeled CDC, that was simply less preferred by the labelers due to relatively minor considerations. Thus, although we still need to quantify the frequency of these “weak” errors, it is clear that for most practical scenarios, the performance of our system are above the strict numbers described here.

	Precision @ 5	Precision @ 10	Precision @ 20	Precision @ 50
mean	0.23	0.20	0.16	0.12
std	0.21	0.20	0.11	0.07
min	0.00	0.00	0.00	0.00
max	0.80	0.60	0.50	0.32
rand	0.00008	0.00008	0.00008	0.00008

Table 5: Ranking component

Category	Number of candidates
Accept	27
Accept with corrections	3
Generality failed	9
Strength failed	145
Text Spirit failed	1
Multiple candidates	5
Repeats topic	5
Incoherent	37
Not a sentence	17

Table 6: Number of ”false” claims in each rejection category

6 Error Analysis

We present an analysis of the errors (using a slightly earlier version of the system). The analysis covered the same 32 Topics described above, where for each Topic we analyzed the errors among the top 10 predictions. In total there were 249 sentences which did not exactly match the annotated data. Each of these seemingly-erroneous CDC candidates was then given to 5 annotators, who had to confirm or reject it and select a rejection reason. The goal of the analysis is to understand the types of errors the system makes as well as to obtain feedback on text spans that were not originally detected by the labelers (possible misses). Specifically, the labelers were instructed to choose one of the options in the following list:

Accept - The candidate should be accepted as is.

Accept with correction - The candidate should be accepted with minor corrections.

Generality failed - The candidate is too specific.

Multiple Candidates - The candidate contains more than one claim.

Repeats Topic - The candidate simply reiterates the topic (or its negation) rather than claim something about it.

Strength failed - The candidate does not directly and explicitly supports or contests the topic.

Text Spirit failed - The candidate does not keep the spirit of text in which it appeared.

Incoherent - The candidate is not a coherent claim.

Not a sentence - The candidate is not grammatical.

A majority vote was used to obtain the final answer. Table 6 gives the number of candidates in each category. As can be seen, about 10% of the candidates were actually accepted in this round. Most of the errors were attributed to ”Strength Failed”, which is a fairly wide category. In future analysis we plan to break it down into more specific sub-categories. Table 7 gives some examples of candidates generated by the system (which do not exactly match the annotated data) and their corresponding categories.

7 Discussion and Future Work

We introduced the CDCD task which is scientifically challenging, and moreover, potentially invaluable for various novel applications. We outlined a machine learning approach to address this task, which is designed based on a cascade of classifiers for handling the special difficulties of this task, and in

Category	Topic	Candidate claim
Accept	The sale of violent video games to minors should be banned	Some researchers believe that while **playing violent video games leads to violent actions**, there are also biological influences that impact a person's choices.
Accept with corrections	Democratic governments should require voters to present photo identification at the polling station	Proponents of a similar law proposed for Texas In March 2009 also argued that **photo identification was necessary to prevent widespread voter fraud**.
Generality failed	Parents should be allowed to genetically screen foetuses for heritable diseases	While psychological stress experienced during a cycle might not influence an IVF outcome, it is possible that **the experience of IVF can result in stress that leads to depression**.
Strength failed	Physical eduaction should be compulsory	**Physical education trends have developed recently to incorporate a greater variety of activities**.
Strength failed	Parents should be allowed to genetically screen a for heritable diseases	However, **the trade-off between risk of birth defect and risk of complications from invasive testing is relative and subjective**; some parents may decide that even a 1:1000 risk of birth defects warrants an invasive test while others wouldn't opt for an invasive test even if they had a 1:10 risk score.
Strength failed	Parents should be allowed to genetically screen foetuses for heritable diseases	This has made international news, and had led to accusations that **many doctors are willing to seriously endanger the health and even life of women in order to gain money**.
Multiple candidates	Wind power should be a primary focus of future energy supply	**The use of wind power reduces the necessity for importing electricity from abroad and strengthens the regional economy**.
Repeats topic	Affirmative action should be used	More recently, a Quinnipiac poll from June 2009 finds that 55% of Americans feel that **affirmative action should be abolished**, yet 55% support affirmative action for disabled people.
Incoherent	Bribery is sometimes acceptable	**The difference with bribery is that this is a tri-lateral relation**.
Incoherent	Parents should be allowed to genetically screen foetuses for heritable diseases	Having this information in advance of the birth means that **healthcare staff as well as parents can better prepare themselves for the delivery of a child with a health problem**.
Not a sentence	A mandatory retirement age should be instituted	**Mandatory retirement is the age** at which persons who hold certain jobs or offices are required by industry custom or by law to leave their employment, or retire.

Table 7: Example sentences for each rejection category

particular the inherently skewed ratio between positive examples and negative examples. We assessed the performance of the proposed approach over a novel benchmark dataset, carefully developed for this task. Our results verify the soundness of our definitions, and the validity of the introduced CDCD task.

In future work we intend to expand the collected labeled data and to generate new versions of this benchmark, that will be further released for academic research. In parallel, we intend to explore various ways to improve the accuracy of our predictions. One intriguing direction, highlighted by examining our data, is the possibility of defining different CDC types. For example, it might be that developing separate classifiers for factual CDCs – like S1 in Table 1, and other classifiers designed to detect more subjective CDCs – like S5 in Table 1, will yield better performance, assuming that each of these two types has a distinguished statistical signature. Similarly, it might be that developing domain-orineted statistical models will further enhance the quality of the CDC predictions.

In this work we analyzed labeled data in which for a given Topic, the relevant articles were manually identified. Combining a CDCD solution with automatic opinion retrieval techniques (Macdonald et al., 2010; Zhang et al., 2007) would be a natural next step towards developing an even more powerful CDCD system. Moreover, while compelling arguments start with high quality and relevant claims, they must include reliable evidence to support the validity of the introduced claims. Thus, combining a CDCD system with a system that automatically detects such supportive evidence , may give rise to a new generation of automatic argumentation methods. In principles, such methods may detect relevant CDCs in some articles, and support these CDCs with evidence detected within other articles, or even within entirely different corpora, ending up with automatically generated arguments, that were never explicitly proposed before in this form by humans. Developing successful solutions for the CDCD task is a fundamental step in pursuing this vision.

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 52nd Annual Meeting of the Association for Computational Linguistics, First Workshop on Argumentation Mining*. Association for Computational Linguistics, June.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea, July. Association for Computational Linguistics.

I. Dagan, B. Dolan, B. Magnini, and D. Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(04).

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.

A. Freeley and D. Steinberg. 2008. *Argumentation and Debate*. Cengage Learning.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 273–283, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Knowledge Discovery and Data Mining*, pages 168–177.

Craig Macdonald, Rodrygo L.T. Santos, Iadh Ounis, and Ian Soboroff. 2010. Blog track research at trec. *SIGIR Forum*, 44(1):58–75, August.

M. C. McCord, J. W. Murdock, and B. K. Boguraev. 2012. Deep parsing in watson. *IBM J. Res. Dev.*, 56(3):264–278, May.

George A. Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009),*, pages 98–109. ACM.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Vikas Raykar, Mitesh Khapra, Amrita Saha, Priyanka Agrawal, and Shantanu Godbole. 2014. Subjectivity detection. work in progress.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing With Compositional Vector Grammars. In *ACL*.

Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. pages 3–17.

Simone Teufel. 1999. Argumentative zoning: Information extraction from scientific text. Technical report.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor J. M. Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *COMMA*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 43–50. IOS Press.

Wei Zhang, Clement Yu, and Weiyi Meng. 2007. Opinion retrieval from blogs. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 831–840, New York, NY, USA. ACM.