Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection

Ruty Rinott[1], Lena Dankin[1], Carlos Alzate[2], Mitesh M. Khapra[3],
Ehud Aharoni[1], Noam Slonim[1]

[1]IBM Research - Haifa, Mount Carmel, Haifa, 31905, Israel,
[2]IBM Research - Ireland, Damastown Industrial Estate, Dublin 15, Ireland,
[3]IBM Research - Bangalore, India,
{rutyr,lenad,aehud,noams}@il.ibm.com
carlos.alzate@ie.ibm.com mikhapra@in.ibm.com

Abstract

Engaging in a debate with oneself or others to take decisions is an integral part of our day-to-day life. A debate on a topic (say, *use of performance enhancing drugs*) typically proceeds by one party making an assertion/claim (say, *PEDs are bad for health*) and then providing an evidence to support the claim (say, *a 2006 study shows that PEDs have psychiatric side effects*). In this work, we propose the task of automatically detecting such evidences from unstructured text that support a given claim. This task has many practical applications in decision support and persuasion enhancement in a wide range of domains. We first introduce an extensive benchmark data set tailored for this task, which allows training statistical models and assessing their performance. Then, we suggest a system architecture based on supervised learning to address the evidence detection task. Finally, promising experimental results are reported.

1 Introduction

In recent years there has been a growing interest in the area of argumentation mining (Green et al., 2014; Cardie et al., 2015; Wells, 2014). Part of this awakening is the The Debater™ project[1] whose goal is to develop technologies that will assist humans to debate and reason, e.g., by automatically suggesting arguments relevant to an examined topic. The minimal definition of such an argument (Walton, 2009) is a set of statements, made up of three parts – a claim (aka conclusion, proposition), a set of evidence (aka premises), and an inference from the evidence to the claim. Needless to say, evidence plays a critical role in a persuasive argument.

In most debate related skills, such as natural language understanding and generation, humans currently have an inherent advantage over a machine. However, in the ability to provide high quality and diverse evidence, machines have a very promising potential, being able to swiftly process large quantities of information. Nonetheless, since most of the relevant information is represented by unstructured text, successfully exploiting these resources requires the ability to identify evidence in free text. This is exactly the focus of our work. Specifically, we formally define the task of evidence detection, introduce an architecture for attacking this problem, and demonstrate its performance over dedicated manually labeled data.

Before defining the task formally, we introduce three concepts which will be used throughout this paper. These concepts were earlier defined in (Aharoni et al., 2014) and we use the same definitions here. **Topic**: a short phrase that frames the discussion. **Claim**: a general, concise statement that directly supports or contests the topic. **Context Dependent Evidence (CDE)**: a text segment that directly supports a claim in the context of the topic. The first three rows of Table 1 show examples of a topic, a claim and CDE.

For the purpose of this work, we assume that we are given a concrete topic, a relevant claim, and potentially relevant documents, provided either manually or by automatic methods (Cartright et al., 2011; Levy et al., 2014). Our task, which we term Context Dependent Evidence Detection (CDED), is to automatically pinpoint CDE within these documents. We further require that a detected CDE is reasonably well phrased, and easily understandable in the given context, so that it can be instantly and naturally used to support the claim in a discussion. Table 1 gives examples of valid CDE (V) and non-valid CDE (X) according to the definition mentioned above.

It is well recognized that one can support a claim using different types of evidence (Rieke and Sillars, 2001; Seech, 2008). Furthermore, for different use cases, different evidence types could be more suitable. Correspondingly, we develop a classification approach that is able to identify and distinguish between three common evidence types (Rieke and Sillars, 1984; Seech, 2008):

- **Study** Results of a quantitative analysis of data, given as numbers, or as conclusions. (Table 1 S1);

[1]http://researcher.ibm.com/researcher/view_group.php?id=5443

[2]Note ibuprofen is considered a PED

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 440–450,
Lisbon, Portugal, 17-21 September 2015.

Topic: Use of performance enhancing drugs (PEDs) in professional sports	
Claim A: PEDs can be harmful to athletes health	
S1: A 2006 study examined 320 athletes for psychiatric side effects induced by anabolic steroid use. The study found a higher incidence of mood disorders in these athletes compared to a control group.	V
S2: The International Agency for Research on Cancer classifies androgenic steroids as "Probably carcinogenic to humans."	V
S3: Rica Reinisch, a triple Olympic champion and world record-setter at the Moscow Games in 1980, has suffered numerous miscarriages and recurring ovarian cysts following drug abuse.	V
S4: The UN estimates that there are more than 50 million regular users of heroin, cocaine and synthetic drugs.	X
S5: FDA does not approve ibuprofen[2] for babies younger than six months due to risk of liver damage.	X
S6: Doping can ultimately damage your health.	X
Claim B: Use of PED is inline with the spirit of sport	
S7: Professor Savulescu, a philosopher and bioethicist, believes that biological manipulation embodies the sports spirit: the capacity to improve ourselves on the basis of reason and judgment.	V

Table 1: Examples for defined concepts. The V/X indicates if the candidate is a CDE to the claim above it, according to our definition.

- **Expert** Testimony by a person / group / committee / organization with some known expertise / authority on the topic. (Table 1 S2, S7);
- **Anecdotal** A description of an episode(s), centered on individual(s) or clearly located in place and/or in time. (Table 1 S3);

Examining the valid and non-valid CDEs in Table 1 it should be clear that the distinction between them is often quite subtle. For example, it is possible that a piece of text has the characteristics of a certain evidence type, but does not support the claim (see S4 in Table 1). It is also possible that a piece of text supports the claim, but is irrelevant in the context of the topic (see S5 in Table 1). It could also be the case that a piece of text entails the claim, but adds no new information to support it (see S6 in Table 1).

We present here a pipeline architecture, relying on supervised learning, to handle the different aspects of CDED which shows promising results over a variety of topics. We demonstrate that the proposed solution and features can generalize well, namely that models learned over different topics can perform reasonably well on an entirely new topic. On average, for a significant fraction of claims the proposed system succeeds to propose relevant CDE amongst its top 4 predictions, and properly determines the evidence type. Furthermore we show that we are able to automatically pinpoint claims for which the performance of the system are of even greater quality, enabling the user to obtain higher precision for these claims.

We believe that the ability to automatically provide evidence for given claims will have many practical uses, helping layman and professionals in different domains, to reach decisions and prepare for discussions, from a lawyer presenting a case in court, to a politician considering a new policy.

2 Related work

CDED is related to several other information retrieval and NLP tasks. Probably the closest of which is the relatively unexplored task of Evidence Retrieval (ER) (Cartright et al., 2011; Bellot et al., 2013). However, while ER focus is on identifying whole documents, in CDED the goal is to pinpoint a typically much shorter text segment which can be used directly to support a claim. Furthermore, ER is typically performed for factual assertions, while in CDED one may want to consider a wider range of claim types (Rieke and Sillars, 2001), cf. claim B in Table 1.

Another important line of related work is the Textual Entailment (TE) framework (Dagan et al., 2009; Glickman et al., 2005). 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 component in a CDED approach, and perhaps vice versa, the tasks are quite different. Namely, the goal of TE is detecting semantic inference while the goal of CDED is to provide evidence which can enhance the persuasion of a claim. For example, common instances of TE are rephrases or summarizations of a sentence, however they cannot serve to support a claim within a discussion, as they merely repeat it (Table 1, S6). On the other hand, an anecdotal story may have strong emotional impact that will effectively support a claim during a discussion, although the truth of the claim cannot be inferred from such evidence. Furthermore, similar to ER, TE focuses only on factual assertions, while we focus on a wider range of claims (Rieke and Sillars, 2001), cf. claim B in Table 1.

Question answering (QA) (Dang et al., 2007) also has some similar aspects to the proposed task, although aiming at a very different goal, which is to provide an explicit – typically unique and concise – answer, to a question.

The proposed CDED task should be seen as another contribution in the emerging field of argumentation mining, with several important distinct characteristics. Previous works suggested extracting full ar-

	Topics	Claims	Articles with CDE	CDE	avg. % of claims with CDE	avg. # CDE per claim
Study	30	1587	136	1018	31 (22)	2.2 (0.9)
Expert	37	1702	214	1896	46 (22)	1.9 (0.8)
Anecdotal	22	1137	70	382	17 (11)	2.0 (1.6)
Total	39	1734	274	3057	60 (17)	2.9 (3.7)

Table 2: 'Topics' indicate the number of topics included for each CDE type. This determines the number of claims considered for each type. The next columns indicate the number of articles in which at least one CDE was found; the total number of CDE detected for each type; the average percent of claims for which at least one CDE was found; and for these claims, the average number of CDE found. Note that the total number of CDE is not a simple sum of the CDE per type, as CDE can be assigned with more than one type. Standard deviations of distribution across topics are given in parenthesis where relevant.

guments (Mochales Palau and Moens, 2009), analyzing argument structure (Peldszus, 2014), and identifying relations between arguments (Cabrio and Villata, 2012; Ghosh et al., 2014). Other works focused on specific domains such as evidence-based legal documents (Mochales Palau and Moens, 2011; Ashley and Walker, 2013), online debates (Cabrio and Villata, 2012; Boltužić and Šnajder, 2014), and product reviews (Villalba and Saint-Dizier, 2012; Yessenalina et al., 2010). In addition, some works based on machine-learning techniques, used the same topic in training and testing (Rosenfeld and Kraus, 2015; Boltužić and Šnajder, 2014), relying on features from the topic itself in identifying arguments. In contrast, here, we focus on detecting an essential constituent of an argument – the evidence – rather then detecting whole arguments, or detecting other argument parts like claims (Levy et al., 2014; Lippi and Torroni, 2015). In addition, we do not limit ourselves to a particular domain, nor assume that the topic of the discussion is known in advance. Finally, we aim to pinpoint evidence in a clearly defined context, given by the pre–specified claim. Thus, the developed system should not only find pieces of text that have general evidence characteristics but further identify which of these candidates can be used to support a specific claim. Hence, as we demonstrate in our results, an essential part of a CDED system should be dedicated to model and assess the semantic relation of a candidate evidence to the given claim and topic.

3 Data

Since CDED is a new and rather complicated task, it is beneficial to examine and understand the nature of the data before moving on to developing a working solution. We therefore start by explaining the manual data annotation process, and several important observations over the resulting data.

To train and assess the classifiers in our system we rely on data collected by the procedure described in (Aharoni et al., 2014). Briefly, given a topic and a corresponding relevant claim, extracted from a Wikipedia article by human annotators, the annotators were asked to mark corresponding evidence – text segments supporting the claim. To limit the amount of time annotators spend on these tasks, labeling was restricted to the article in which the claim was found. The task was split into two stages. First, in the detection stage, five annotators read the article, and mark all CDE candidates they locate. Next, in the confirmation stage all the candidates suggested by the annotators are presented to another set of five annotators, which confirm or reject each candidate, and determine the type(s) of accepted candidates. Candidates which were confirmed by the majority of the annotators are considered CDE, and are assigned the type(s) suggested by at least three annotators.

A total of 547 Wikipedia articles associated with 58 different topics were annotated through this procedure. The topics were selected at random from *Debatabase*[3] covering a wide variety of domains, from atheism to the role of wind power in future energy supply. Out of these topics, 39 were selected at random for training and testing the classifiers included in the system. We refer to these data as the *train and test data*. The remaining 19 topics were used for tuning various feature parameters, and developing auxiliary classifiers, as described in Section 5. We refer to these data as the *held-out* data.

In the 39 topics comprising the train and test data, a total of $3,057$ distinct CDE were found in 274 articles (See Table 2). The data is highly unbalanced towards non CDE sentences. For example, for type Study, only 31% of the claims had at least one CDE. Of these 31% claims, on average, a claim was associated with 2.17 CDEs. Further, on average these 2.17 CDEs together span 1.5 sentences, whereas an average article in our data consists of 150 sentences. In other words, even for claims with at least one CDE of type Study, on average only 2% of the sentences in the claim's article are part of such Study CDE.

In general, CDE in the examined data varied in length from less than a sentence to more than a paragraph. However, 90% of these CDE were composed of segments of up to three sentences within the same paragraph. Furthermore, in 95% of the cases, CDEs were

[3] http://idebate.org/debatabase

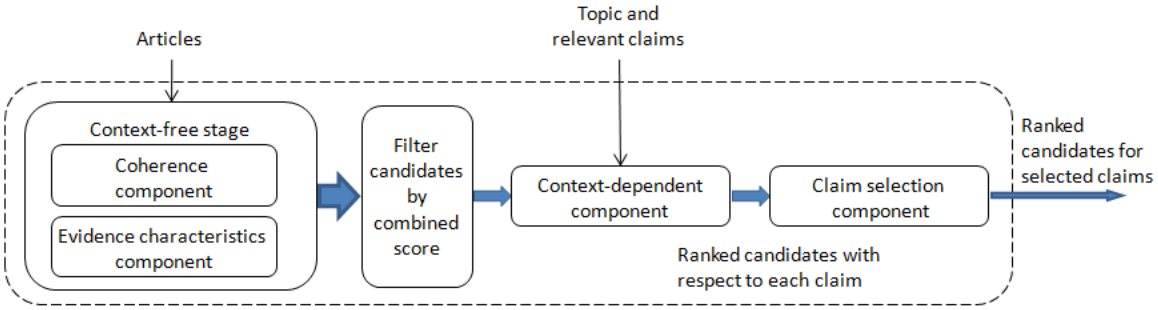

Figure 1: Schematic description of the CDED system proposed in this work.

comprised of full sentences. Examining CDEs that start or end mid-sentence, reveals that in most cases the CDE is more concise in these boundaries, but is still a valid CDE when extending the boundaries to include the full sentence. We therefore decided not to address this issue here, and we extend all CDE boundaries to full sentences.

Apparently CDE of type Study and type Expert are far more common in Wikipedia compared to Anecdotal CDE. We expect this distribution to change in other less scientifically inclined corpora.

Finally, the variance between different topics was substantial, as depicted in Table 2 (refer to the standard deviations mentioned in parenthesis). For example, the percentage of claims with Expert CDE varies from 10% in the topic *banning gambling* to 95% in the topic *US responsibility for the Mexican drug wars*. This observed variability obviously adds to the difficulty and complexity of the task.

In the experiments reported in this paper, out of the 39 topics in the train and test data, we exclude from the evaluation of each type, topics that had less than three CDE of that type. This leaves a total of 30, 37, and 22 topics for types Expert, Study, and Anecdotal, respectively.

The current work is the first to report results over these CDE data, which are more than 4 times larger compared to the data released in (Aharoni et al., 2014). These data are now freely available for research purposes [4].

4 System Architecture

The input to our system is a topic, a set of related articles and a set of relevant claims detected within these articles. Given this input, our system provides the user with a ranked list of candidate CDEs, originating from the text in the claim's article, for an automatically selected subset of the input claims.

In general, we observe that a text segment should satisfy three criteria to be considered CDE of a specific type. It must be coherent; it must have characteristics of the relevant Evidence type; and finally, of course, it should support the claim.

In addition to these observations, we note that a priori, we do not expect all claims to be supported by all CDE types (Park and Cardie, 2014). For example, opinion claims like claim B in Table 1 are expected to be less supported by Study evidence compared to factual claims, like claim A in Table 1. Moreover, as evident from Table 2, many claims do not have any associated CDE in the same article. Thus, the system performance may naturally improve if it will propose candidate CDE of a particular type, only to an automatically identified subset of the input claims.

Based on these observations, we are led to suggest an architecture which approaches CDED via a pipeline of modular components. Each of these components relies upon the results of its precedents, and is specifically designed to address a single aspect of those mentioned above. The resulting architecture is depicted in Figure 1. Briefly, in the proposed architecture, the first two components are *context-free*, i.e., focused on the general characteristics of a candidate, still not taking into account the context of the claim, nor the topic. The third component is *context-dependent*, considering the relation of the candidate to the claim and topic. Finally, the fourth component aims to identify a subset of the claims for which CDE will be proposed.

We consider all text segments composed of one, two or three consecutive sentences, included within the same paragraph as candidates (see Section 5 for more details). Given a set of such *candidate CDEs* – or simply, *candidates* – the first component, termed the *coherence component*, estimates the coherence of each candidate. For example, consider CDE S1 in Table 1. A candidate which includes only the second sentence is incoherent, as it includes critical unresolved anaphora, that cannot be understood without the previous sentence. In parallel, the second component, termed the *evidence characteristics component*, estimates the extent to which the candidate's statistical signature matches that of the examined evidence type. For example, if no quantitative analysis of data is reported, the candidate typically cannot be considered Study evidence, regardless of the claim and topic. Next, we only

[4] https://www.research.ibm.com/haifa/dept/vst/mlta_data.shtml

retain candidates for which the average score of the first two components was relatively high, aiming to further focus our attention on the most promising candidates.

The retained candidates are then considered by the *context-dependent* component which aims to determine if the examined candidate indeed supports the provided claim in the context of the topic. Thus, this component ranks all retained candidates with respect to each claim. Finally, the *claim selection* component aims to rank all input claims, according to the probability that CDEs are indeed found amongst top-ranking candidate for the claim.

Dividing the overall task into sub-tasks has several benefits. First, it allows training each component over its most suitable data, in which the signal of the relevant features is easier to capture. For example, many of the features for the *context-dependent* component aim to determine the semantic relatedness between the claim and a candidate. If one would have tried to tackle the entire CDED task simultaneously, the training data for this component would have been masked by many candidates that are highly related to the claim, although are not CDE – e.g., definitions of some aspects of the claim. These candidates would have blurred the signal that should be captured by the semantic relatedness features, as they represent candidates with negative labels that are nonetheless semantically related to the claim. By separating the tasks, we allow the *context-dependent* component to avoid this inherent difficulty, and train over much cleaner data.

Second, our pipeline allows efficient handling of the CDED task in terms of run time. Semantic relatedness features are often relatively complex and demanding in terms of run time. The significant filtering done after the *context-free* stage, reduces the number of candidates for which we have to calculate these features.

Finally, we note that some of the modular components we develop as part of the pipeline might be of interest by themselves. For example, context-free evidence detection might be useful in cases in which the claim and topic are not defined (Lippi and Torroni, 2015).

Naturally, we expect that different evidence types will have different characteristics. For example, numbers are expected to be more common in CDE of type Study compared to CDE of type Expert. Anecdotal CDE is perhaps expected to be less semantically related to the corresponding claim, as it may have a more associative relation to the claim, compared to CDE of types Study or type Expert. Correspondingly, all components are developed, trained, and assessed, independently for each CDE type.

In summary, the full flow of our system upon receiving a new topic with associated articles and claims, is as follows:

1. All articles are split into sentences, and all consecutive segments up to three sentences within a paragraph are generated as candidates.
2. Each candidate is assigned a score by the two components in the *context-free* stage and their scores are averaged. [5]
3. A dynamic programming algorithm selects a complete coverage of the article by non-overlapping candidates with the maximal average *context-free* score. The rest of the candidates are discarded.
4. The remaining candidates across all articles are sorted and only the top 15% of these candidates are retained. [6]
5. For each claim, the *context-dependent* component ranks all retained candidates within the claim's article with respect to the claim.
6. The claim selection component considers all claims and the candidates ranked with respect to each claim and assigns a score per claim. If the claim–score is below a pre–computed threshold, no candidate CDE will be presented for that claim.

All components are based on a Logistic Regression (LR) classifier, and the class probability is used as the candidate score.

5 Technical approach

In this section, we provide more technical details for each of the components in our architecture.

5.1 Coherence component

This component aims to score a candidate according to its coherence. For example, a candidate with an unresolved anaphora, or one that breaks a quotation in the middle, is expected to receive a relatively low score. As mentioned, this component considers all text segments composed of 1–3 consecutive sentences included within the same paragraph. This decision is based on the observation that such segments cover 90% of CDE in the labeled data. Reaching a full coverage requires examining segments up to 25 sentences, which would vastly increase run time, for a relatively small gain. Thus, for example, for a single paragraph with five sentences, our system will examine a total of $5 + 4 + 3 = 12$ candidates. For an article including 30 such paragraphs, a total of 360 candidates will be considered. During training, segments that conform to a labeled CDE were considered positive examples, while segments that overlap a labeled CDE, but either include additional sentence(s), or exclude part of the CDE sentences were considered negative examples.

Dominant features for this classifier included: presence of incomplete quotes; presence of contrast related conjunctive adverbs – e.g., `however`,

[5] With additional training data, we might be able to learn a more sophisticated function to combine both scores.

[6] This percentage was determined according to performance on the held-out data set. We have also experimented with methods where the threshold is score–based rather than percentage–based, which gave similar results.

nevertheless; segment length; and presence of unresolved co-references.

5.2 Evidence characteristics component

This component aims to estimate to what extent a candidate represents evidence of a certain type. The train and test data for this component consisted of all text segments composed of 1-3 consecutive sentences, included within the same paragraph. Positive examples are all labeled CDE of the corresponding evidence type. Negative examples are all candidates that do not overlap labeled CDE of the relevant type, including CDE of other types.

The dominant features for the classifier used in this component relied on the following mechanisms:

- **Lexicons** – including external lexicons (the Harvard IV-4 dictionary) and manually and automatically compiled in-house lexicons. Specifically, for each evidence type, we manually compiled a lexicon of words characterizing this type by looking at examples from the held-out data. This resulted with high–precision / low–recall lexicons. For example, for type Expert we used a lexicon of words describing persons and organizations that may have some relevant expertise, such as: economist, philosopher, court. In addition, we used the held-out data to automatically learn wider lexicons of words that are significantly associated with each type. All the in-house lexicons are described in detail in the supplementary material.

- **Named Entity Recognition (NER)**. We used the Stanford NER (Finkel et al., 2005) to extract named entities such as person and organization, and an in-house NER (Lally et al., 2012) to extract more fine grained categories such as "educational organization" and "leader".

- **Patterns**. We used regular expressions to represent features like: does that candidate contain a quote; does it contain a citation; does it contain numeric quantitative results. In addition, we generated complex regular expressions which combine the above lexicons with NER results to capture patterns indicative of different types. For example, the pattern [Person/organization, 0 to 10 wildcard words, an opinion verb - such as believe, conclude, etc.] was highly indicative of Expert evidence (cf. Table 1 S7).

- **Subjectivity classifier**. We manually labeled $1,750$ sentences, selected at random from articles in the held-out data, as either subjective or objective. Next, each sentence was represented by a concatenation of two feature vectors – (i) a bag-of-words representation, limited to a handcrafted subjectivity lexicon containing 100 words; (ii) a bag-of-patterns representation based on patterns observed as frequent in the subjective sentences, detected by a modification of the SPM algorithm (Srikant and Agrawal, 1996). An LR classifier was then trained over the labeled sentences.

5.3 Context-dependent component

The goal of this component is to estimate whether a candidate can be used to support a claim while discussing the given topic. The training data for this component are [topic / claim / CDE] triplets. Triplets in which the CDE and claim were linked in the labeled data – namely, the CDE was identified as evidence for the claim – were considered as positive examples. Negative examples were generated by combining claims and CDEs detected in the same topic and article, but that were not linked in our labeled data.

The features for the classifier used in this component can be conceptually divided into four types: (i) Semantic relatedness between the candidate and the claim (ii) Semantic relatedness between text related to the candidate and the claim (iii) Relative location of the candidate with respect to the claim and (iv) sentiment–agreement between the candidate and the claim.

In general, we rely on two methods to assess the semantic relatedness between two texts. The first is based on the cosine similarity between TF-IDF vectors representing each text. Before constructing the TF-IDF vectors each text is augmented with acronym expansions, and lexical relations (including antonym, derivationally related and pertainym) from WordNet (Miller, 1995). The second, relies on the average cosine similarity between the Word2Vec (Mikolov et al., 2013) representation of all pairs of words in the two texts, where in each pair one word is taken from the first text and the other word from the second.

For each of these two methods, we consider the semantic relatedness between the claim and: Specified slots in the candidate as detected by an in-house slot grammar parser (McCord, 1990; McCord et al., 2012); The entire candidate text; The header/sub-header of the section/subsection containing the candidate; Titles of citations referred to from the candidate.

5.4 Claim selection component

The goal of this component is to rank all claims according to the probability that the claim's article includes CDE of the relevant type, associated with the claim. The training data consisted of all claims, where positive examples included claims for which at least one CDE of the relevant type existed in the labeled data and negative examples included all remaining claims.

A thresholding mechanism on the component score is used to determine the claims for which candidates will be presented. This threshold was selected by optimizing the F1 score over the set of held-out topics.

The features used by this component exploited three types of information:

- **Claim properties**: We used the held-out data to

generate two types of lexicons. The first lexicon is generated separately per evidence type. It includes claim words that were found to be significantly associated with positive examples, namely with claims for which CDE were found. For example, for type Study, this lexicon included words such as `lead`, `result`, `development` and `significant`. The second lexicon aimed to characterize words that are significantly associated with factual claims vs. non–factual claims, with the expectation that certain evidence types might be more/less common for each of these two claim categories. For this, 550 randomly selected claims were annotated as factual/non–factual. Words identified as characterizing factual claims included `increase`, `important`, and `relate`, while words like `natural`, `freedom`, and `right` were found dominant for non-factual claims.

- **Claim's relevance to topic and article**: We expect that when an article's main topic is highly related to the claim, it will more likely include CDE for that claim. Similarly we expect that for claims at the heart of the topic, CDE is more likely to be provided. These properties are assessed by measuring the semantic relatedness between (i) the claim and the content of the claim's article and (ii) the claim and topic.

- **Properties of claim's article** : Specifically, we mainly consider the scores provided by the context-dependent component to all candidates examined in the claim's article. If the observed scores are relatively high/low, we expect the article to be more/less likely to include evidence of the considered type. Various statistics of these scores, such as the maximum score and the standard deviation are used as features aiming to capture this intuition.

6 Experimental Results

6.1 Evaluation

We evaluated our approach using the Leave-One-Out schema: for every topic, we trained the classifiers using the claims and associated CDE in all other topics and then applied the resulting models to the left out topic.

In general, we consider a candidate as true-positive if it includes all sentences included in the CDE and no additional sentences. However, for our analysis it is also interesting to separate between (i) errors in selecting the segment boundaries and (ii) errors of down the line components that are affected by these errors. Thus, we also include the *overlap* measure where we consider a candidate as true-positive if at least one sentence within it overlaps a sentence in a labeled CDE.

Our final assessment measure is the mean reciprocal rank (MRR), that is the inverse of the rank of the first CDE detected for a particular claim, averaged over all claims selected by the *claim selection component.* This is motivated by the observation that in most practical use cases, it is usually more important to be able to support many claims, than to provide all the CDE available for a single claim. We define the MRR of a claim with no CDE (errors of the claim selection component) to be 0.

Finally, we report the macro-averaged results over the different topics, that is all topics have the same weight regardless the amount of labeled claims and labeled CDE detected for them. The rational behind this is that we wish to ensure that our system does reasonably well across all topics examined. We note that micro-averaging gave overall similar results.

6.2 Comparison to Baselines

To assess the necessity and contribution of the different components we compare our full pipeline to partial pipelines, where some of the component are disabled or replaced by simple baselines. These baselines are described below.

First, we consider the **No Context-Free Stage (NCFS)** baseline which aims to assess the contribution of the *context-free* stage by skipping this stage, and passing all candidates directly to the *context-dependent* component.

Next, we consider the **Basic Claim Selection (BCS)** baseline which replaces the claim selection component. It ranks claims according to the top score of the candidate CDE for the claim. A threshold was selected on top of the training data, such that the average percentage of claims passing the threshold is equal to the average percentage of claims with CDEs in the labeled data.

Since, to the best of our knowledge, this is the first work to address CDED, there is no prior-art to compare our results to. However, to ensure that this task is indeed empirically different from related tasks, and demands a specialized pipeline to handle, we compare with two baselines that are often used in related tasks.

The **BM25** basline handles CDED as an IR task, where the claim represents the query, and all CDE candidates represent the documents in a standard IR setting. After pre-processing, which includes tokenization, stop word removal, and stemming (Porter, 1997) we use BM25 (Robertson et al., 1996) to rank all relevant candidates according to their similarity to the query, namely to the input claim.

The **W2V** baseline handles CDED as a purely semantic relatedness task using state of the art semantic relatedness measure of Word2Vec (Mikolov et al., 2013). Thus, we use the average cosine similarity between the Word2Vec representations of all words in a given candidate to all words in the claim, to rank all relevant candidates with respect to each claim.

Type	MRR				MRR overlap			
	Pipeline	NCFS	W2V	BM25	Pipeline	NCFS	W2V	BM25
Study	0.37	0.19	0.09	0.14	0.51	0.39	0.24	0.23
Expert	0.41	0.29	0.28	0.15	0.58	0.52	0.50	0.24
Anecdotal	0.18	0.04	0.04	0.04	0.31	0.11	0.11	0.11

Table 3: Macro-averaged MRR for each CDE type. Only claims with CDE in the labeled data were considered in these results.

6.3 Results

We start by assessing the proposed pipeline prior to the claim selection component. Table 3 reports the MRR following the context-dependent component when filtering out claims for which no CDE were found in the labeled data.

Impact of context free stage: Comparing the pipeline performance to the baseline using only the context dependent component (NCFS baseline), the results indicate the necessity of the context-free stage in our pipeline. That is, assessing the coherence of candidates, as well as their evidence characteristics, seems to be essential to properly address CDED. In particular, the fact that the gain is observed both in the MRR measure and in the MRR-overlap measure suggests that both the context-free components are valuable.

Impact of context dependent stage: Comparing the NCFS basline to W2V and BM25 baselines shows that for type Study, the context-dependent component alone still has an advantage over a single semantic relatedness feature. Observing feature weights learned by the LR classifier, we estimate that much of this advantage is due to also taking into consideration semantic relatedness of the claim to texts related to the candidate, namely the header of the section containing the candidate and titles of citations referred to from the candidate.

For types Expert and Anecdotal the performance of the context-dependent component are similar to those of the W2V baseline. For type Expert, this suggests that most of the signal in the context-dependent component comes from semantic relatedness between the claim and candidate CDE. Results for type Anecdotal are significantly lower. This was somewhat expected, given the smaller size of Anecdotal data available to train our classifiers (Table 2). The declined performance of the W2V and BM25 baselines for this type, further suggests that the semantic relatedness of CDE and claims for this type are less direct.

Impact of detecting segment boundaries: Comparing the *overlap* MRR measure to the exact MRR highlights that identifying the correct segment boundaries is still a challenge, and once we improve this aspect, we can expect a significant improvement in the results.

Impact of claim selection component: We next turn to assess the contribution of the claim selection component. Table 4 compares the final MRR results – at the end of the pipeline – for claims selected by the claim selection component, vs. claims selected by the BCS baseline. Additionally, to demonstrate the value of claim selection in general, we add results when considering all claims. For types Expert and Study the claim selection component shows a clear advantage over the baselines. Furthermore, the improved performance is achieved when passing a higher percentage of claims than the BCS baseline (34% vs 31% for Study and 52% vs 46% for Expert, Figure 2). Admittedly, for Anecdotal CDE the performance of claim selection are poor. For this component the small sample size for Anecdotal CDE was even more acute – there were only 151 claims with CDE of type Anecdotal – thus few positive examples to train this component.

Type	Pipeline	BCS	All claims
Study	0.25	0.16	0.12
Expert	0.34	0.23	0.20
Anecdotal	0.04	0.05	0.03

Table 4: Macro-averaged MRR over: 1) claims selected by the claim selection component, 2) claims selected by basic claim selection, and 3) all claims.

Recall that the claim selection component's threshold was tuned over the held-out data to optimize the F1 measure with respect to claims with/without CDE. However, for some applications one may favor higher precision at the expense of providing candidate CDEs for less claims. Figure 2 shows that indeed, for type Study, considering more strict thresholds of the claim selection component monotonically improves the system's overall precision, as reflected by the improved MRR. Similar results were obtained for type Expert.

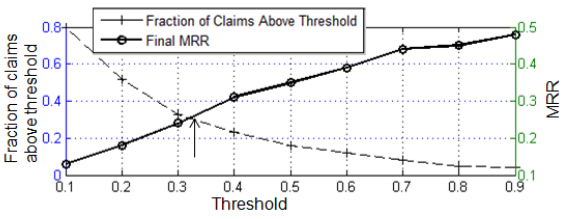

Figure 2: MRR and average fraction of passed claims as function of the claim selection threshold for type Study. Arrow indicates threshold used to obtain the results in Table 4.

6.4 Examples of System Performance

To provide some intuition for the results of our system, Table 5 shows the 4 top ranking candidate CDE

According to econometric studies, negative side effects of aid can include an unbalanced appreciation of the recipient's currency, increasing corruption, and adverse political effects such as postponements of necessary economic and democratic reforms.	X
Many econometric studies in recent years have supported the view that development aid has no effect on the speed with which countries develop.	V
An inquiry into aid effectiveness by the UK All Party Parliamentary Group (APPG) for Debt, Aid and Trade featured evidence from Rosalind Eyben, a Fellow at the Institute of Development Studies.	X
A very large part of the spend money on development aid is simply wasted uselessly. According to Gerbert van der Aa, for the Netherlands, only 33% of the development aid is successful, another 33% fails and of the remaining 33% the effect is unclear. This means that for example for the Netherlands, 1.33 to 2.66 billion is lost as it spends 4 billion in total of development aid.	V

Table 5: Top ranking candidates for the claim *aid is ineffective* in the context of the topic *trade vs. aid*

of type Study for the claim *aid is ineffective* in the context of the topic *trade vs. aid*. Among these, 2 were indeed labeled as CDE. The other two exemplify common errors of our system. Candidate 1 can be used to support a highly related claim such as *aid has negative side effects*, but does not directly support the claim under consideration. Candidate 3 mentions a relevant study, but does not present its results, hence cannot be used to support the claim.

7 Conclusions and Future Work

We have provided the definitions for the CDED task, and described a system architecture that addresses the issues at the heart of the task. We assessed the performance of the proposed approach over a novel benchmark dataset, demonstrating the validity of our architecture, and the necessity of all its components.

There are still many open issues to address and directions in which to expand the task and labeled data which we hope to address in future work.

In this paper we define CDE only in the context of *supporting* a claim. However, in many scenarios providing *counter* evidence can also be very useful. As evidence supporting and contesting a claim share many semantic and syntactic features, we believe that detecting both cases simultaneously might be easier to accomplish, although to enhance the practical use of such a solution, one may need to develop an additional component, determining the polarity of the detected CDE.

Another natural direction to pursue is expanding the documents which are considered for CDED beyond the article containing the claim. These can include additional Wikipedia articles and other resources such as newspaper archives, scientific literature, blogs, etc. This poses additional challenges in gathering labeled data, as it will require a mechanism to decide which documents to label per claim and will probably increase the number of documents to be labeled. Expanding to additional corpora will probably require development of additional features, to capture signals unique to each corpus. For example, in newspaper archives, the identity of the author might prove an important feature.

Finally, in this work we used manually identified claims and articles. Combining a CDED solution with recent works in the field of argumentation mining (Cartright et al., 2011; Levy et al., 2014; Lippi and Torroni, 2015), may give rise to a new generation of methods, that will be able to automatically construct relevant arguments on demand, for a variety of topics.

Acknowledgments

The authors would like to thank Oren Tsur, Vikas C. Raykar, Matan Orbach and Ido Dagan for many helpful discussions.

References

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

Kevin D. Ashley and Vern R. Walker. 2013. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 176–180, New York, NY, USA. ACM.

Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Veronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang. 2013. Overview of inex 2013. In *CLEF Lab Reports*, Valencia, Spain, September.

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

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *ACL (2)*, pages 208–212.

Claire Cardie, Nancy Green, Iryna Gurevych, Graeme Hirst, Diane Litman, Smaranda Muresan, Georgios Petasis, Manfred Stede, Marilyn Walker, and Janyce Wiebe, editors. 2015. *Proceedings of the Second*

Workshop on Argumentation Mining. Association for Computational Linguistics, Denver, Colorado, June.

Marc-Allen Cartright, Henry A. Feild, and James Allan. 2011. Evidence finding using a collection of books. In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, BooksOnline '11, pages 11–18, New York, NY, USA. ACM.

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

Hoa Trang Dang, Diane Kelly, and Jimmy J. Lin. 2007. Overview of the trec 2007 question answering track. In *TREC*.

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.

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

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. In Manuela M. Veloso and Subbarao Kambhampati, editors, *AAAI*, pages 1050–1055. AAAI Press / The MIT Press.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, June.

Adam Lally, John M. Prager, Michael C. McCord, Branimir Boguraev, Siddharth Patwardhan, James Fan, Paul Fodor, and Jennifer Chu-Carroll. 2012. Question analysis: How watson reads a clue. *IBM Journal of Research and Development*, 56(3):2.

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.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argumentation mining. In *Proceedings of the Twenty Fourth International Joint Conference on Artificial Intelligence*. AAAI Press.

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

Michael C. McCord. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: Proc. of the International Scientific Symposium, Hamburg, FRG*, pages 118–145. Springer, Berlin, Heidelberg.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

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.

Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

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

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, Maryland, June. Association for Computational Linguistics.

Martin F. Porter. 1997. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Richard D. Rieke and Malcolm O. Sillars. 1984. *Argumentation and the decision making process*. Harper Collins, New York, NY, USA.

Richard D. Rieke and Malcolm O. Sillars. 2001. *Argumentation and Critical Decision Making*. Longman.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1996. Okapi at trec-3. pages 109–126.

Ariel Rosenfeld and Sarit Kraus. 2015. Providing arguments in discussions based on the prediction of human argumentative behavior. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Zachary Seech. 2008. *Writing Philosophy Papers*. Cengage Learning.

Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK. Springer-Verlag.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, pages 23–34.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 1–22. Springer US.

Simon Wells. 2014. Argument mining: Was ist das? In *Proceedings of the 14th International Workshop on Computational Models of Natural Argument*, CMNA14.

Ainur Yessenalina, Yejin Choi, and Claire Cardie. 2010. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 336–341, Stroudsburg, PA, USA. Association for Computational Linguistics.