Context-Independent Claim Detection for Argument Mining

Marco Lippi and **Paolo Torroni**
DISI - Università degli Studi di Bologna
{marco.lippi3,p.torroni}@unibo.it

Abstract

Argumentation mining aims to automatically identify structured argument data from unstructured natural language text. This challenging, multi-faceted task is recently gaining a growing attention, especially due to its many potential applications. One particularly important aspect of argumentation mining is claim identification. Most of the current approaches are engineered to address specific domains. However, argumentative sentences are often characterized by common rhetorical structures, independently of the domain. We thus propose a method that exploits structured parsing information to detect claims without resorting to contextual information, and yet achieve a performance comparable to that of state-of-the-art methods that heavily rely on the context.

1 Introduction

Argumentation, as a discipline, has ancients roots in philosophy, where the study of argumentation may, informally, be considered as concerned with how assertions are proposed, discussed, and resolved in the context of issues upon which several diverging opinions may be held [Bench-Capon and Dunne, 2007]. In recent decades, argumentation models and techniques have been exported to fields in artificial intelligence, like multi-agent systems and artificial intelligence for legal reasoning. According to Walton [2009], there are four tasks undertaken by argumentation: identification, analysis, evaluation and invention. These are often non-trivial tasks, even for experts such as philosophers and discourse analysts, although the discipline is quite well established, so much so that informal logic textbooks have been written to introduce students to the art of argumentation [Fogelin and Sinnott-Armstrong, 1991]. The task of identification, in particular, is to identify the premises and conclusion of an argument as found in a text of discourse. The toolset of an argument analyst includes, for example, techniques to recognize *argument markers* or *cue phrases* i.e., linguistic elements like *more precisely* or *for example* that suggest the presence of elements of an argument in a sentence [Knott and Dale, 1994].

Argumentation (or *argument*) *mining* is a recent challenge that involves *automatically* identifying structured argument data from unstructured natural language corpora, by exploiting the techniques and methods of natural language processing, machine learning, sentiment analysis and computational models of argument. While the general idea and its potential applications are clear enough to justify an increasing number of research meetings and projects, some scholars point out that it is still unclear to what exactly the term "argument mining" refers [Wells, 2014]. For these reasons, research proposals vary wildly in aims and scope.

In addition, most of the proposed methods are designed to address specific domains, such as evidence-based legal documents [Palau and Moens, 2011; Ashley and Walker, 2013], personal communications and online debates [Pallotta and Delmonte, 2011; Cabrio and Villata, 2013], product reviews [Villalba and Saint-Dizier, 2012], newspaper articles and court cases [Feng and Hirst, 2011].

The *Debater* project[1] developed by IBM is one of the few known attempts to tackle an even more ambitious endeavour: to assist humans to debate and reason. The Debater vision is that of an intelligent system able to take raw information and digest and reason on that information, to understand the context, and to construct arguments pro and con *any subject*. Notice that several "expert systems" have been proposed, e.g., for teaching [Dauphin and Schulz, 2014] or to aid human reasoning by explaining plans [Caminada *et al.*, 2014] but they require a domain expert to model and represent all the relevant knowledge in terms of an argumentation system. Debater's ambition is not only to be able to address any subject, but to do so *without the intervention of a domain expert*. To this end, the IBM research team has designed a set of sophisticated classifiers that use contextual information, in order to extract valuable features for evidence, claim, and argument detection tasks *in a given context* [Levy *et al.*, 2014].

Contextual information, such as knowledge of the topic, is crucial to the performance of state-of-the-art argumentation mining tools [Levy *et al.*, 2014]. For example, Debater lets the user select a topic from a list, and returns a series of "pro" and "con" arguments. This type of solution may be restrictive in some cases. For example, if we want to dig out arguments from vastly diversified corpora, such as social media (the application potential there is enormous), then we might need

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

to seek arguments in sources where "the" topic is not at all defined, posts may contribute to several topics, and the discussion may often shift from topic to topic.

The purpose of the present work is then *to devise a method for the detection of claims in unstructured corpora, without necessarily resorting to contextual information*.

To the best of our knowledge, this is the first attempt to address such a "general" (context-independent) argumentation mining task. Our underlying hypothesis, based on the argument analysis literature, is that argumentative sentences are characterized by common structures that reflect rhetorical processes, and can therefore be extremely informative of the presence of a claim.

For instance, one could argue that a sentence such as:

The prototypical delegative democracy has been summarized by Bryan Ford in his paper, Delegative Democracy

"sounds like" a factual statement, whereas

The difficulty and cost of becoming a delegate is small

"sounds like" a claim, *independently of the topic under discussion*. At the same time, we are aware that deciding what is and what is not a claim is often matter of discussion even for human experts. Corpora labeled with information about arguments or parts thereof (such as claims) are scarce and started to appear only recently. All this contributes to making argumentation mining grow ever more challenging.

This manuscript is structured as follows. In Section 2 we define the problem. In Section 3 we describe our methodology. In Section 4 we discuss the data. In Section 5 we present experiments and results. Section 6 concludes.

2 Problem Definition

The (*Context-Independent*) *Claim Detection* problem (CD) is an information extraction task, whose goal is to identify those sentences in a document that contain the conclusive part of an argument.

According to Walton [2009], an argument is a set of statements which consists in three parts: a conclusion, a set of premises, and an inference from the premises to the conclusion. While this definition is widely accepted, these concepts have also been defined in the literature with different names: conclusions could be referred to as *claims*, premises are often called *evidence* or *reasons*, and the link between the two, i.e., the inference, is sometimes called the *argument* itself.

CD is a quite subtle task, even for humans, since the concept itself of claim could hardly be framed by a compact set of well-defined rules. Indeed, in some cases claims consist of opinions supported by some evidence facts, while in other cases they are represented by statements that describe relevant concepts [Levy *et al.*, 2014], as it happens, e.g., in legal domains. For these reasons, building a large annotated corpus for claim detection is a complex and time-consuming activity.

Problems similar to CD have been addressed by other authors in the context of argumentation mining. Palau & Moens [2011] famously built an automatic system for the extraction of argument structures in the legal domain. Cabrio & Villata [2012] focused on what we could name "inference detection," since their input data is a set of premises (or pro/con positions) and conclusions (claims) and their goal is to establish support/attack relations: but they are not concerned with claim or evidence detection. Indeed, the datasets they use do not contain non-argumentative sentences (see Section 4). Levy *et al.* [2014] assert that "at the heart of every argument lies a single claim, which is the assertion the argument aims to prove" and define a *Context Dependent Claim Detection* problem, where a topic is given as well as a relatively small set of relevant free-text articles. The goal of CDCD is to automatically pinpoint "reasonably well phrased" CDCs within these documents, which can be instantly and naturally used in a discussion about the given topic.

3 Methodology

Most of the methods proposed to address the problems overviewed in Section 2 are based on machine learning classifiers which typically rely on large sets of highly engineered features, specifically designed to address the task of interest, and very often domain-dependent.

While we concur that machine learning techniques are mandatory for this kind of application, we argue that argumentation mining is currently lacking the contribution of those machine learning algorithms that can most effectively handle structured data. In particular, information coming from NLP could be naturally encoded into structured data which in turn can be extremely informative for many argumentation mining problems, such as claim detection.

Our methodology is driven by the observation that argumentative sentences are often characterized by common rhetorical structures. To illustrate[2], consider the constituency parse trees (Figure 1, left) for a sentence containing the claim:

monarchy is unfair and elitist.

Nodes in a constituency parse tree are labeled with standard non-terminal symbols for (English) context-free grammar: for example, SBAR indicates a subordinate, VP is the verb phrase, NP the noun phrase, etc. In this case the claim is contained in a subordinate (having the SBAR tag as root) which depends on the verb *assert*. This is a common structure, as claims are often introduced by verbs such as *argue*, *believe*, *maintain*, *sustain*, *assert*, etc. In other contexts, a claim can be introduced by a colon, for example when quoting a statement, as in the following example:

He added: "A community of separate cultures fosters a rights mentality, rather than a responsibilities mentality."

Another common structure of claim includes a comparison between two concepts or arguments, as in the sentence:

Sustained strong growth over longer periods is strongly associated with poverty reduction, while trade and growth are strongly linked.

In other scenarios, claims can be reported as conclusions drawn by a set of premises, theories or evidence facts which support the argument. In that case, the supporting sources are often directly mentioned when reporting the claim, as in the following case:

[2] All the illustrations in this Section are taken from the IBM corpus used in our experiments (see Section 4).

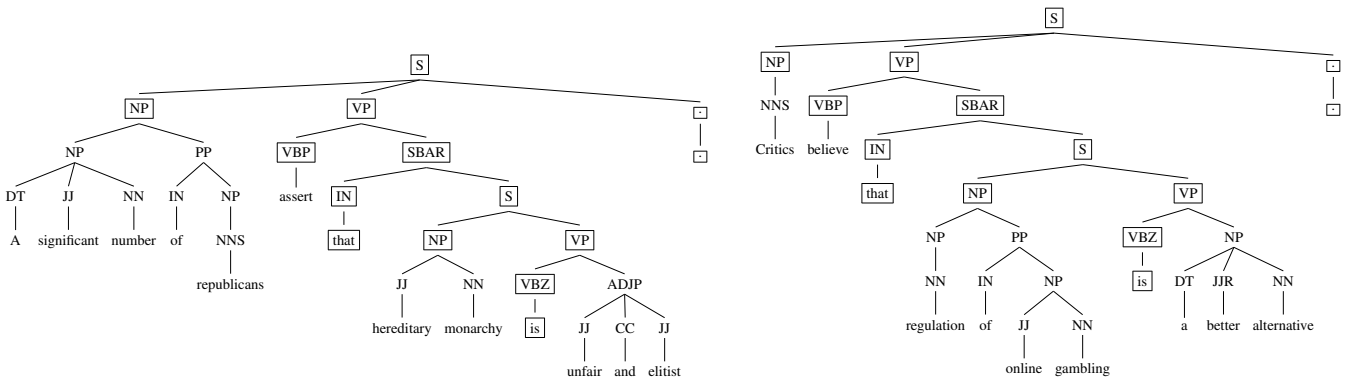

Figure 1: Constituency trees for two sentences containing claims. Boxed nodes are common elements between the two trees.

Thus, according to the theory, affirmative action hurts its intended beneficiaries, because it increases their dropout rate.

As illustrated by these examples, the structure of a sentence could be highly informative for argument detection, and in particular for the identification of a claim. Constituency parse trees are an ideal instrument for representing such information. We therefore built a claim detection system based on a Support Vector Machine (SVM) classifier which aims at capturing similarities between parse trees through Tree Kernels [Moschitti, 2006a]: Figure 1 shows two parse trees whose sentences contain two distinct claims, where boxes highlight the many common parts of their internal structure.

Kernel methods have been widely used in a variety of different NLP problems, ranging from plain text categorization up to more specific tasks like semantic role labeling, relation extraction, named entity recognition, question/answer classification and many others (see [Moschitti, 2006b; 2012] and references therein). In particular, Tree Kernels have been successfully employed in many of these applications.

When considering a classification problem, traditional classifiers such as kernel machines usually learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{X}$ is the input space, usually a vectorial space encoding attribute-value pairs, and $\mathcal{Y}$ is the output space representing the set of categories. Function f is typically learnt by minimizing a loss function over a set of N given observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. When dealing with structured data, examples $x_i \in \mathcal{X}$ are not simply represented by plain feature vectors, but they can encode complex relational structures, such as graphs or trees.

A Tree Kernel (TK) is designed so as to measure the similarity between two trees, by evaluating the number of their common substructures (or *fragments*). By considering different definitions of fragments, several TK functions are induced: for example, one could consider only complete subtrees as allowed fragments, as well as define more complex fragment structures. Intuitively, each possible tree fragment is associated to a different feature in a high-dimensional vectorial space, where the j-th feature simply counts the number of occurrences of the j-th tree fragment: the TK can therefore be computed as the dot product between two such representations of different trees. A kernel machine is then defined, which exploits the structured information encoded by the tree kernel function $K(x, z)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \cdot \phi(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) \qquad (1)$$

where ϕ is the feature mapping induced by the tree kernel K, and N is the number of support vectors. In general, the kernel between two trees T_x and T_z can be computed as:

$$K(T_x, T_z) = \sum_{n_x \in N_{T_x}} \sum_{n_z \in N_{T_z}} \Delta(n_x, n_z) \qquad (2)$$

where N_{T_x} and N_{T_z} are the set of nodes of the two trees, and $\Delta(\cdot, \cdot)$ measures the score between two nodes, according to the definition of the considered fragments.

In this work we consider the Partial Tree Kernel (PTK) [Moschitti, 2006a], which allows the most general set of fragments (called Partial Trees), being any possible portion of subtree at the considered node (see Figure 2). The higher the number of common fragments, the higher the score Δ between two nodes.

It is clear that PTK can easily and automatically generate a very rich feature set, capable of capturing structured representations without the need of a costly feature engineering process. Anyhow, it is worth remarking that the proposed TK framework allows to include in the representation of each example also a plain vector of features, which can enrich the description of the considered instance. In this case, the final kernel would be computed as the combination between a classic kernel between feature vectors (linear, polynomial, rbf, etc.) K_V and the kernel between trees K_T, e.g., with a weighted sum of the two contributions.

The kernel computed over the parse trees, possibly combined with the kernel computed over feature vectors, is used to train an SVM classifier on a set of labeled examples. The SVM classifier exploits the capability of TK functions of directly measuring structure similarity between trees and is therefore trained on sets of positive/negative example sentences represented by their parse trees (here a negative example is a sentence not containing any claim). A small portion of data is typically used in advance in order to tune the SVM parameters.

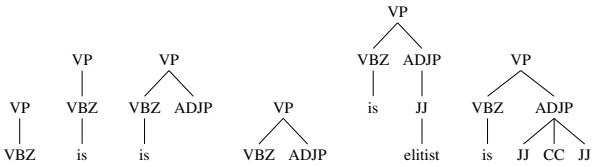

Figure 2: Examples of Partial Trees (PTs) for the right-most VP node from the constituency tree in Figure 1.

To summarize, the methodology we propose consists of the following claim detection pipeline:

1. the given text document is split into sentences using a tokenizer that detects sentence boundaries;
2. each sentence is parsed, to obtain a constituency tree;
3. sentences not containing a verb (VP tag) are discarded;
4. in order to improve generalization, words at leaves are substituted with their stemmed versions;
5. an SVM classifies each sentence as possibly containing a claim or not.

4 Data Sets

As argumentation mining is a quite recent research area, only a few labeled data sets are publicly available for the purpose of building automatic argument extractors. The construction of this kind of data sets is not trivial, as it typically involves a team of experts which must follow specific guidelines in order to build a consistent set of annotations.

A major hurdle to the exploitation of existing annotated arguments corpora for the claim detection task at hand is their lack of negative examples (non-argumentative sentences).

The University of Dundee maintains a set of corpora in the Argument Interchange Format[3] (AIF), which is very appropriate for many modeling, analysis and reasoning applications in the domain of argumentation. Yet, almost all the data sets comprised in the Dundee corpora only contain already extracted arguments, therefore lacking the non-argumentative parts of the original documents, or contain refined argument annotations as a result of a post-processing step which manipulates the original text, making it hard to automatically extract the information necessary to build a machine learning classifier. This is the case, for example, with Araucaria DB: a well-known corpus that was successfully used in several pioneering works on argumentation mining, but which could not be employed, in its current release, in our experiments.

A recent collection of annotated arguments was presented in [Cabrio and Villata, 2014], with data extracted from different sources (Depatepedia, ProCon, Wikipedia pages, and the script of a play). This benchmark, called NoDE (Natural language arguments in online DEbates), was built with the aim of analyzing the kind of link between different arguments (e.g., to distinguish attacks from supports) and, similarly to the Dundee corpora, it does not contain non-argumentative sentences. Yet, it could certainly be a very useful benchmark for the construction of an attack/support machine learning classifier.

Other annotated data sets used in previous works, such as the European Court of Human Rights (ECHR) database, specifically built for argumentation mining in legal documents [Palau and Moens, 2011], and the Vaccine/Injury Project (V/IP) [Ashley and Walker, 2013], describing juridical cases where injuries are claimed to be caused by vaccines, are regrettably not available.

A large, novel data set for argumentation mining is being developed at IBM Research in the context of the Debater project [Aharoni *et al.*, 2014]. This database consists in a collection of Wikipedia articles (taken from the 04/04/2012 dump), organized in a number of different topics, and in two sets of context-dependent claim and evidence annotations. The data set is available upon request for research purposes. The current release contains 315 articles grouped into 33 topics, in which a total of 1,332 distinct claims and 796 evidence facts have been annotated.

The IBM data set is clearly very imbalanced, as it contains over 40,000 sentences which contain neither a claim nor an evidence, and it is therefore an extremely challenging benchmark for our goal. However, it is representative of the kind of data we expect to have in our envisaged argumentation mining application scenarios.

It is important to know that the IBM data set was specifically constructed so that, for articles belonging to a given topic T, only claims related to T were annotated. Since our approach aims at extracting claims from text documents *without* knowing the topic in advance, by using the IBM data annotations our classifier could in principle receive ambiguous information regarding what is (or is not) a claim, as some of the sentences labeled as negative examples could indeed contain claims, although not topic-related. Yet, since this data set represents the largest and most heterogeneous corpus currently available for claim detection, we believe it is a very challenging benchmark for an argumentation mining system, and therefore also for our method. Moreover, in the experimental section, we will also present results, for comparison, where our classifier uses information about the topic.

A smaller publicly available data set consists in a collection of 90 persuasive essays [Stab and Gurevych, 2014a] on a variety of different topics, where a team of annotators provided information regarding claims, major claims, premises (a synonym for evidence) and attack/support relations between such components. Given the nature of both the data and the annotations, only a few sentences in this corpus are non-argumentative, resulting in a totally different benchmark with respect to the IBM dataset. Moreover, in this case, claims are not annotated as related to some specific topic.

5 Results

We first report experiments on the IBM dataset. Our goal is to detect sentences containing a claim.

A similar task was tackled by the IBM Haifa Research Group in [Levy *et al.*, 2014], where the key difference is that the system proposed by IBM intends to identify claims related (according to the annotators) to a given topic. An addi-

[3] http://corpora.aifdb.org/

Method	P@200	R@200	F_1@200	AURPC	AUROC
TK	9.8	58.7	16.8	0.161	0.808
BoW	8.2	51.7	14.2	0.117	0.771
Random Baseline	2.8	20.4	5.0	–	–
Perfect Baseline	19.6	99.3	32.7	–	–
TK + Topic	10.5	62.9	18.0	0.178	0.823
[Levy *et al.*, 2014]*	9.0	73.0	16.0	–	–

Table 1: Results obtained on the IBM Wikipedia corpus. We report precision (P), recall (R) and $F_1 = \frac{2PR}{P+R}$ when selecting the 200 most significant sentences for each topic, as well as the area under recall-precision and ROC curves. The TK classifier employs a partial Tree Kernel on Consituency Trees, which can be combined with context-dependent features (TK + Topic row). The BoW predictor employs an SVM trained using a bag-of-words of the sentence. The Perfect Baseline predicts the highest possible number of true positives. The last row reports IBM results on a slightly different version of the corpus.

tional commitment of the IBM system is to identify the exact boundaries of the claim within the sentence (CDCD, see Section 2). The system pipelines a number of classifiers, each trained by exploiting a set of highly engineered features, often obtained as a result of other additional sophisticated classifiers, such as sentiment analysis tools or modules scoring the subjectivity of a sentence. The first stage of their architecture addresses a task similar to the one we are interested in: it selects, for each topic, 200 candidate sentences that may contain a claim. Notice that [Levy *et al.*, 2014] reports on results obtained from a slightly older version of the dataset, containing 32 topics (instead of 33), and only 976 claims (instead of 1,332). Therefore, their performance measurement can only be qualitatively compared to our system.

We built an SVM classifier that exploits a kernel on the constituency parse trees obtained using the Stanford CoreNLP 3.5.0 software[4]. In this phase we do not use the information concerning the topic, and we simply rely on a single classifier using PTK. Following the same procedure adopted in [Levy *et al.*, 2014], we selected the 200 most highly ranked sentences for each topic, by employing a leave-one-topic-out procedure where, in turn, each topic is considered as test set, and the remaining ones make up the training set. We obtained an average precision (P) and recall (R) on the 33 topics equal to 9.8 and 58.7, respectively, where $P = TP/(TP+FP)$ and $R = TP/(TP+FN)$, being TP the true positives, FP the false positives, and FN the false negatives.

We remark that our performance measurements are computed over a set of labeled examples which consider *context-dependent* claims, while our system aims at detecting sentences containing claims independently of the topic. In this sense, the reported measurement certainly penalizes our approach, but still it is informative of the suitability of the method: Table 1 in fact compares the results obtained by our classifier with a random baseline (which randomly selects the 200 "best" sentences) averaged on 10 runs, and with the perfect (oracle) baseline which selects all the possible true positive examples, representing an upper bound for our method[5].

We also compare against an SVM which employs a bag-of-word (BoW) of the sentence described in terms of TF-IDF features, as typical of text categorization.

It should not be surprising that our system also detects claims which are not related to the considered topic: Table 2 shows some false positive examples, which actually appear to contain claims, even though not entirely related to the topic.

As already stated, the methodology presented in this paper allows to combine tree kernels with kernels constructed over plain feature vectors. By exploiting this principle, we can add feature vectors to the representation of each sentence, and this time we can include topic-dependent features. In particular, we trained a second SVM by adding *a single feature* which consists in the cosine similarity between the considered sentence and the given topic, both represented as bag-of-words. As shown in Table 1, the only addition of this single feature improves the performance of our approach, which is comparable with the IBM system for context-dependent claim detection described in [Levy *et al.*, 2014] (we remark the difference in the dataset versions). We also report the area under recall-precision curve (AURPC) and the area under the ROC curve (AUROC) of our classifier. When setting to zero the threshold for predicting the presence of a claim (which is the default choice in binary classification problems with SVM) the system produces a recall/precision equal to 12.1/42.7 and 13.1/48.3 in the two cases without/with topic information, respectively, heading to F_1 values of 18.1 and 20.6.

As a second benchmark, we present results on the persuasive essay data set [Stab and Gurevych, 2014a]. This is a much smaller and less representative data set for real-world applications, since almost all the sentences are annotated as containing either a premise or a claim, or both. The experiments described in [Stab and Gurevych, 2014b] report the performance of a multi-class predictor which is trained to distinguish four categories (Premise, Claim, MajorClaim, None) but works in the (arguably infrequent) scenario of knowing in advance the segmentation of the sentences into single argumentative entities (i.e., a claim or a premise).

By adopting the same kind of classifier built for the IBM data set, we performed a 10-fold cross validation on the 90 essays, and considered the union of categories Claim and MajorClaim[6] as the positive (target) class. We obtained a

[4]http://nlp.stanford.edu/software/corenlp.shtml

[5]The perfect baseline has a 99.3% recall instead of 100% because a few positive cases are discarded since the Stanford parser mispredicts the presence of a verb.

[6]MajorClaim is a specific category of essays, in fact there is one

IBM Corpus Topic	Sentence
All nations have a right to nuclear weapons	Critics argue that this would lower the threshold for use of nuclear weapons
Atheism is the only way	Some believe that a moral sense does not depend on religious belief
Endangered species should be protected	Simple logic instructs that more people will require more food
Institute a mandatory retirement age	Some theories suggest that ageing is a disease
Limit the right to bear arms	Others doubt that gun controls possess any preventative efficacy
Make physical education compulsory	Specific training prepares athletes to perform well in their sports
Multiculturalism	Indigenous peoples have the right to self-determination

Table 2: Some examples of sentences in the IBM dataset, predicted by our system to contain a claim, but actually labeled as negative examples in the corpus, owing to the context-dependent nature of the annotations.

74.6/68.4 precision/recall performance, which results in an F_1 equal to 71.4. As a qualitative comparison, for the slightly different multi-class task, Stab & Gurevych [2014b] report F_1 values equal to 63.0 and 54.0 for the MajorClaim and Claim categories, respectively: we remark that in such case a number of sophisticated features, specifically designed for the given data set, are employed (e.g., based on the essay title, on the position of the sentence within the essay, etc.) and the sentence segmentation in claim/premise candidates is given in advance, which may be an unlikely assumption in many relevant practical settings.

Finally, to obtain a *qualitative* idea of our classifier's performance, we tested the model learned on the IBM corpus on a set of 10 Wikipedia pages unrelated to the ones in the training set. To emphasize the capability of our approach to detect claims without contextual information, and to distinguish sentences which do not contain claims, we selected 5 articles on highly-controversial topics (Anti-consumerism, Effects of climate change on wine production, Delegative democracy, Geothermal heating, Software patents and free software) and 5 on non-controversial topics (Ethernet, Giardini Naxos, Iamb, Penalty kick, Spacecraft). Our system detects 34 claims in the controversial articles, and only 3 in the others. Dataset and results are available on the following website:

http://lia.disi.unibo.it/~ml/argumentationmining/ijcai2015.html

6 Conclusions

Argumentation mining is nowadays believed to have a huge potential by scholars and companies alike [Modgil *et al.*, 2013; Slonim *et al.*, 2014]. Claim detection is a key step in the process. However, this is a challenging task due to a number of factors. To the best of our knowledge, state-of-the-art solutions make strong assumptions, either on the domain (evidence-based legal documents, court cases, personal communications, product reviews, etc.), or on the format of input data, or on the knowledge available about the context. We believe that an important step forward would be to devise methods for argumentation mining that start from plain, unprocessed text, and do not assume context knowledge.

In this paper, we proposed a solution which relies on a popular technique adopted in a variety of different NLP problems, i.e., kernel methods for structured data. In particular, by focusing on the rhetoric structure of claims rather that on other context-dependent features such as sentiment, we rely on the ability of Partial Tree Kernels to generate a very rich feature set, able to capture structured representations without the need of a costly feature engineering process.

We evaluated our solution against the largest and richest known to date data set containing annotated claims. Our results are comparable with those of state-of-the-art methods, which rely on context. This is a significant achievement, if we consider that the most dominant features in state-of-the-art methods are all context dependent [Levy *et al.*, 2014].

We plan to apply the same method to the two remaining argumentation mining tasks, in particular evidence detection and argument detection. The evidence detection task appears to be similar to the claim detection, although adaptations will be of course necessary. Argument detection would require identifying relations between the sentences containing evidence and claims. Authors have addressed this task with a variety of techniques, including e.g., entailment recognition [Cabrio and Villata, 2012]. We plan instead to investigate how PTK can again be used to address this challenge, also in virtue of the well documented successful use of PTK in closely related tasks, such as sentiment analysis of social media [Agarwal *et al.*, 2011].

and only one major claim for each essay.

Acknowledgments

We sincerely thank Noam Slonim and his team at IBM, and Iryna Gurevych and Christian Stab for sharing with us their datasets, Alessandro Moschitti for providing the Partial Tree Kernel software, and Paolo Frasconi for fruitful discussions. This work was partially supported by the ePolicy EU project FP7-ICT-2011-7, grant agreement 288147. Possible inaccuracies of information are under the responsibility of the project team. The text reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

[Agarwal *et al.*, 2011] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *LSM 2011*, pages 30–38, Portland, Oregon, June 2011. ACL.

[Aharoni *et al.*, 2014] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset

for automatic detection of claims and evidence in the context of controversial topics. In *Proc. of the First Workshop on Argumentation Mining*, pages 64–68. ACL, 2014.

[Ashley and Walker, 2013] Kevin D. Ashley and Vern R. Walker. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Enrico Francesconi and Bart Verheij, editors, *ICAIL 2012, Rome, Italy*, pages 176–180. ACM, 2013.

[Bench-Capon and Dunne, 2007] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artif. Intell.*, 171(10-15):619–641, 2007.

[Cabrio and Villata, 2012] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *ACL 2012*, pages 208–212, Jeju, Korea, 2012. ACL.

[Cabrio and Villata, 2013] Elena Cabrio and Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230, 2013.

[Cabrio and Villata, 2014] Elena Cabrio and Serena Villata. NoDE: A benchmark of natural language arguments. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *COMMA 2014*, volume 266, pages 449–450. IOS Press, 2014.

[Caminada *et al.*, 2014] Martin W. Caminada, Roman Kutlak, Nir Oren, and Wamberto Weber Vasconcelos. Scrutable plan enactment via argumentation and natural language generation. In *AAMAS 2014*, pages 1625–1626, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[Dauphin and Schulz, 2014] Jeremy Dauphin and Claudia Schulz. Argteach – a learning tool for argumentation theory. In *ICTAI 2014*, pages 776–783. IEEE Press, 2014.

[Feng and Hirst, 2011] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *The 49th Annual Meeting of the ACL: Human Language Technologies, Portland, Oregon, USA*, pages 987–996. ACL, 2011.

[Fogelin and Sinnott-Armstrong, 1991] Robert J. Fogelin and Walter Sinnott-Armstrong. *Understanding Arguments: An Introduction to Informal Logic*. Harcourt Brace Jovanovich College Pubs., Fort Worth, 4th edition, 1991.

[Knott and Dale, 1994] Ailstair Knott and Robert Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18:35–62, 1994.

[Levy *et al.*, 2014] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In Jan Hajic and Junichi Tsujii, editors, *COLING 2014, Dublin, Ireland*, pages 1489–1500. ACL, 2014.

[Modgil *et al.*, 2013] Sanjay Modgil, Francesca Toni, Floris Bex, Ivan Bratko, Carlos I. Chesñevar, Wolfgang Dvorák, Marcelo A. Falappa, Xiuyi Fan, Sarah A Gaggl, Alejandro J. García, María P. González, Thomas F. Gordon, João Leite, Martin Molina, Chris Reed, Guillermo R. Simari, Stefan Szeider, Paolo Torroni, and Stefan Woltran. The added value of argumentation. In *Agreement Technologies*, Law, Governance and Technology Series 8, pages 357–404. Springer-Verlag, 2013.

[Moschitti, 2006a] Alessandro Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In Johannes Frnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *ECML 2006*, volume 4212 of *LNCS*, pages 318–329. Springer, 2006.

[Moschitti, 2006b] Alessandro Moschitti. Making tree kernels practical for natural language learning. In *EACL*, pages 113–120, 2006.

[Moschitti, 2012] Alessandro Moschitti. State-of-the-art kernels for natural language processing. In *Tutorial Abstracts of ACL 2012*, pages 2–2. ACL, 2012.

[Palau and Moens, 2011] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22, 2011.

[Pallotta and Delmonte, 2011] Vincenzo Pallotta and Rodolfo Delmonte. Automatic argumentative analysis for interaction mining. *Argument & Computation*, 2(2-3):77–106, 2011.

[Slonim *et al.*, 2014] Noam Slonim, Ehud Aharoni, Carlos Alzate, Roy Bar-Haim, Yonatan Bilu, Lena Dankin, Iris Eiron, Daniel Hershcovich, Shay Hummel, Mitesh M. Khapra, Tamar Lavee, Ran Levy, Paul Matchen, Anatoly Polnarov, Vikas C. Raykar, Ruty Rinott, Amrita Saha, Naama Zwerdling, David Konopnicki, and Dan Gutfreund. Claims on demand - an initial demonstration of a system for automatic detection and polarity identification of context dependent claims in massive corpora. In Lamia Tounsi and Rafal Rak, editors, *COLING 2014, Dublin, Ireland*, pages 6–9. ACL, 2014.

[Stab and Gurevych, 2014a] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In Jan Hajic and Junichi Tsujii, editors, *COLING 2014, Dublin, Ireland*, pages 1501–1510. ACL, 2014.

[Stab and Gurevych, 2014b] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *EMNLP 2014, Doha, Qatar*, pages 46–56. ACL, 2014.

[Villalba and Saint-Dizier, 2012] Maria Paz Garcia Villalba and Patrick Saint-Dizier. Some facets of argument mining for opinion analysis. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *COMMA 2012*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 23–34. IOS Press, 2012.

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

[Wells, 2014] Simon Wells. Argument mining: Was ist das? In Floris Bex, Floriana Grasso, and Nancy Green, editors, *CMNA14*, 2014.