MARGOT: A Web Server for Argumentation Mining

Marco Lippi[1,*], Paolo Torroni[1]

[a]*Dipartimento di Informatica – Scienza e Ingegneria, Università degli Studi di Bologna*

Abstract

Argumentation mining is a recent challenge concerning the automatic extraction of arguments from unstructured textual corpora. Argumentation mining technologies are rapidly evolving and show a clear potential for application in diverse areas such as recommender systems, policy-making and the legal domain. There is a long-recognised need for tools that enable users to browse, visualise, search, and manipulate arguments and argument structures. There is, however, a lack of widely accessible tools. In this article we describe the technology behind MARGOT, the first online argumentation mining system designed to reach out to the wider community of potential users of these new technologies. We evaluate its performance and discuss its possible application in the analysis of content from various domains.

Keywords: argumentation mining

1. Introduction

Argumentation is a multi-disciplinary domain of knowledge that studies debate and reasoning processes. It has become increasingly central as a core study within artificial intelligence (Bench-Capon & Dunne, 2007), due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning. A recent challenge in argumentation is the automatic extraction of structured arguments from generic textual corpora. Research in this direction started a few years ago with isolated, focused studies (Mochales Palau & Moens, 2011; Saint-Dizier, 2012; Cabrio & Villata, 2012b), but it gained momentum only recently, thanks to the increasingly large-scale availability of content, paired with tremendous advances in computational linguistics and machine learning. This gave rise to a new area of research called *argumentation* (or *argument*) *mining* (henceforth, AM) (Lippi & Torroni, 2016).

The application potential of AM is evident for a wide spectrum of real-world scenarios. A natural application domain is clearly the Web, which offers a real mine of information through a variety of different sources. Currently, the techniques used to extract information from these sources are chiefly based on statistical and network analysis, as in opinion mining (Pang & Lee, 2008) and social network analysis (Easley & Kleinberg,

*Corresponding author
URL: marco.lippi3@unibo.it (Marco Lippi), p.torroni@unibo.it (Paolo Torroni)

2010). An AM system instead could enable *massive qualitative analysis* of comments posted on online social networks and specialised newspaper articles alike, providing unprecedented tools to policy-makers and researchers in social and political sciences, as well as creating new scenarios for marketing and businesses. In this sense, AM could significantly enhance many expert and intelligent systems whose application involve processing user-generated content. Bourguet et al. (2013), for example, analyze argumentation in a textual corpus on food quality in the context of a public health policy: there, AM tools could easily be inserted in the loop to enable a larger-scale analysis. Recommendation systems and collaborative filtering represent another ideal scenario, as they are increasingly often coupled with sentiment analysis instruments (García-Cumbreras et al., 2013), while the goal of AM is to bring sentiment analysis one step beyond, in order to look for causes and reasons rather than just for opinions (Habernal et al., 2014; Lippi & Torroni, 2016). Indeed, attempts in the direction of constructing argumentation-based recommender systems have been made by Teze et al. (2015). Finally, law is an important domain where expert systems may greatly benefit from AM, with the goal of automatically extracting knowledge from legal documents (Pal, 1999; Galgani et al., 2015; Lippi et al., 2015).

Interestingly, *tools* are still missing in the AM research front. The state of the art offers a myriad of isolated methods for addressing some specific AM sub-tasks, such as claim detection or support/attack relation prediction, and for well-defined genres, such as legal texts or persuasive essays: but tools, available to a wider community of users, for extracting arguments from unstructured text, have not been developed yet. This could be partly ascribed to the youth of AM as a discipline, but it certainly has to do with the fact that the concept itself of "argument" is something difficult to capture in absolute terms, thus AM methods are typically tailored to a single genre, and general-purpose AM tools are particularly difficult to produce.

The aim of this article is to present the first system designed to make AM easily accessible outside of the AM research community. Such a system, called MARGOT (Mining ARGuments frOm Text), is a concrete tool aimed at reaching out to the wider community of potential AM users. MARGOT is a Web system that exploits state of the art AM techniques and does not require any background on argumentation on the user side.

In this manuscript we describe MARGOT, provide an initial empirical assessment of its output, and discuss its potential for expanding the AM application landscape. This paper builds upon our previous work on context-independent claim detection (Lippi & Torroni, 2015), by extending it in several directions: (1) the detection of argument premises (or *evidence*, as in (Aharoni et al., 2014)); (2) the detection of boundaries of each argument component; (3) the development of an online argumentation mining tool available to scholarly as well as to layman end-users; (4) a broader experimental evaluation of the entire system, both in quantitative terms (performed on one of the widest publicly available AM corpora) and in qualitative terms (performed across different text genres).

2. Background and Related Work

An *argument* could be defined as a set of statements consisting in three parts: a set of premises, a conclusion, and an inference from the premises to the conclusion (Walton, 2009). *Argumentation* instead is the process of constructing arguments. In spite of this conceptual difference, *argumentation mining* and *argument mining* (AM) are used interchangeably to indicate the extraction of arguments from unstructured text. This is a broad and challenging task, which can be broken down into several sub-tasks such as *claim detection* and *evidence detection* (Levy et al., 2014; Rinott et al., 2015), whose focus is on the detection of argument components such as the conclusion or premises, *attribution* which refers to attributing authorship to arguments, *completion* whose goal is to infer implicit argument components, and *relation prediction* aiming at identifying inter- and intra-argument relations. There are also many other related tasks, such as rhetorical characterization of sentences (Houngbo & Mercer, 2014), opinionated claim analysis (Rosenthal & McKeown, 2012), premise verification (Park & Cardie, 2014), etc. A recent survey (Lippi & Torroni, 2016) offers a unifying view of the AM landscape, whereas previous, more specific reviews were given by Mochales Palau & Moens (2011), with a special focus on the legal domain, and by Peldszus & Stede (2013), with an emphasis on the analysis of argument diagrams.

Many different argument models exist, from Toulmin's influential model (1958) to Freeman's (1991), IBIS (Kunz & Rittel, 1970), or the aforementioned Walton's model (2009). For this reason, there is no standardization in the existing AM corpora in the literature. These are usually tailored to the domain/genre at hand, and choose the argument model accordingly. In this work, we closely follow the definition of claim and evidence given by Aharoni et al. (2014) in the presentation of the IBM corpus, whereby a (context-dependent) claim is "a general concise statement that directly supports or contests the topic", whereas a (context-dependent) evidence is a portion of text supporting a claim in the context of a given topic. However, while in the definition given by IBM the *context* (or *topic*) is given in advance, our aim is to detect claims and evidence without knowing the topic in advance. In particular, we want to avoid ending up with another detection result each time the topic is phrased differently.

The systems developed so far in AM typically implement a pipeline architecture, through which they process unstructured textual documents and produce a structured document, where the detected arguments and their relations are annotated so as to form an argument graph. Each stage in this pipeline addresses one sub-task in the whole AM problem. The typical goal of a first stage is to detect arguments or parts thereof within the input text document. The retrieved entities will thus represent nodes in the final argument graph. This problem is usually split into two distinct sub-problems: the *extraction and classification of argumentative sentences* and the *detection of component boundaries*. A final *argumentation structure prediction* stage aims at predicting links between arguments, or argument components. The output of this stage is a graph connecting the retrieved arguments or parts thereof. Edges in the graph may represent relations as diverse

as entailment, support and conflict.

As far as the methods employed, the existing systems for argumentative sentence extraction/classification have used, up to now, a wide variety of classic machine learning algorithms, including Support Vector Machines (Eckle-Kohler et al., 2015; Goudas et al., 2014; Sardianos et al., 2015; Mochales Palau & Moens, 2011), Logistic Regression (Levy et al., 2014; Rinott et al., 2015), Naïve Bayes (Eckle-Kohler et al., 2015; Stab & Gurevych, 2014b; Biran & Rambow, 2011; Mochales Palau & Moens, 2011) and Maximum Entropy classifiers (Mochales Palau & Moens, 2011), Decision Trees and Random Forests (Stab & Gurevych, 2014b). Lippi & Torroni (2016) offer a comprehensive survey. In almost all the existing systems most of the effort has been put into conceiving sophisticated and highly informative features. In most cases, such features are genre-dependent, and that is one of the reasons methods devised for one genre do not easily adapt to others. In an attempt to address these issues, we have recently proposed (Lippi & Torroni, 2015) an SVM-based method for *context-independent claim detection* (see Section 3), which exploits structured kernels on constituency parse trees to measure similarity between sentences (Moschitti, 2006a). That is also the approach implemented in MARGOT for addressing this sub-problem.

Further down in the AM pipeline, component boundary detection is a seldom addressed, but very important, segmentation problem. In current literature, many works focus on sentence-level claim/premise detection (Biran & Rambow, 2011), or assume that sentences have already been segmented into argument components (Stab & Gurevych, 2014b; Eckle-Kohler et al., 2015). The existing approaches to boundary detection employ techniques such as Conditional Random Fields (Goudas et al., 2014; Sardianos et al., 2015), Maximum Likelihood Classifiers (Levy et al., 2014), and SVM Hidden Markov Models (Habernal & Gurevych, 2015). However, detecting argument boundaries is still an open challenge.

Finally, argumentation structure prediction represents an extremely challenging task, as it requires to understand connections and relationships between the detected arguments, thus involving high-level knowledge representation and reasoning issues. Current approaches typically make several simplifying hypotheses. For example, in the corpus by Aharoni et al. (2014), an assumption is made that evidence is always associated with a given claim. This in turn enables using information about the claim to predict the evidence. Other highly genre-specific approaches have been proposed: techniques used include parsing with context-free grammars (Biran & Rambow, 2011; Mochales Palau & Moens, 2011), binary SVM classifiers predicting support/attack relations (Stab & Gurevych, 2014b), and Textual Entailment (Cabrio & Villata, 2012a). MARGOT does not yet address structure prediction.

Applications of AM have been proposed in many contexts, including social media analysis and opinion mining (Grosse et al., 2015), humanities (Stab & Gurevych, 2014a), knowledge-based systems (Aharoni et al., 2014; Rinott et al., 2015), medicine (Houngbo & Mercer, 2014), law (Ashley & Walker, 2013; Mochales Palau & Moens, 2011), information retrieval (Roitman et al., 2016) and many more. We refer to (Lippi & Torroni, 2016) for a detailed review of the state of the art in AM, including a survey of the available corpora.

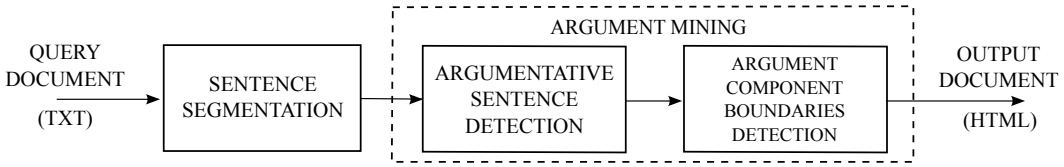


Figure 1: The pipeline of steps implemented by MARGOT.

3. Material and Methods

MARGOT performs argument mining by exploiting a combination of advanced machine learning and natural language processing techniques. The system was trained on the largest (to date) AM corpus. This dataset was developed by IBM Research in the context of the Debater, a multi-million project aiming to build a system capable of collecting and digesting information gathered from the web, reasoning upon it, and thus debating just like a human. The corpus consists of 547 Wikipedia articles (Aharoni et al., 2014; Rinott et al., 2015), organized into 58 topics, and it has been annotated with 2,294 claims and 4,690 evidence facts. Such annotations are *context-* (*topic-*) *dependent*: both claims and evidence are labeled only if they are relevant to a given topic. A *context-dependent claim* is defined (Aharoni et al., 2014) as "a general concise statement which directly supports or contests the topic," whereas *context-dependent evidence* is "a text segment which directly supports the claim in the context of a given topic." The IBM annotation guidelines also define a taxonomy of evidence facts (Aharoni et al., 2014), by distinguishing among *study* evidence (results of a quantitative analysis of data given as numbers or as conclusions), *expert* evidence (testimony by a person/group/committee/organization with some known expertise in or authority on the topic), and *anecdotal* evidence (a description of specific event instances or concrete examples). The corpus is freely distributed by IBM Research.[1]

The AM approach implemented in MARGOT follows a pipeline of subsequent stages (see Section 2) as depicted in Figure 1. MARGOT first detects *argumentative sentences*, where argumentative here means: containing at least one argument component (claim and/or evidence). As a second step, the boundaries of each component are detected.

Most current AM methods build upon machine learning classifiers that very often rely on sets of sophisticated, highly engineered and domain-dependent features, not rarely obtained as the output of some external predictor. This is the case, for example, of sentiment indicators, subjectivity scores, knowledge extracted from ontologies and thesauri (Levy et al., 2014; Rinott et al., 2015), dictionaries of specific keyphrases (Mochales Palau & Moens, 2011), features constructed from semantic role labelers and word embeddings (Habernal & Gurevych, 2015). Classic features from natural language processing, such as bag-of-words,

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

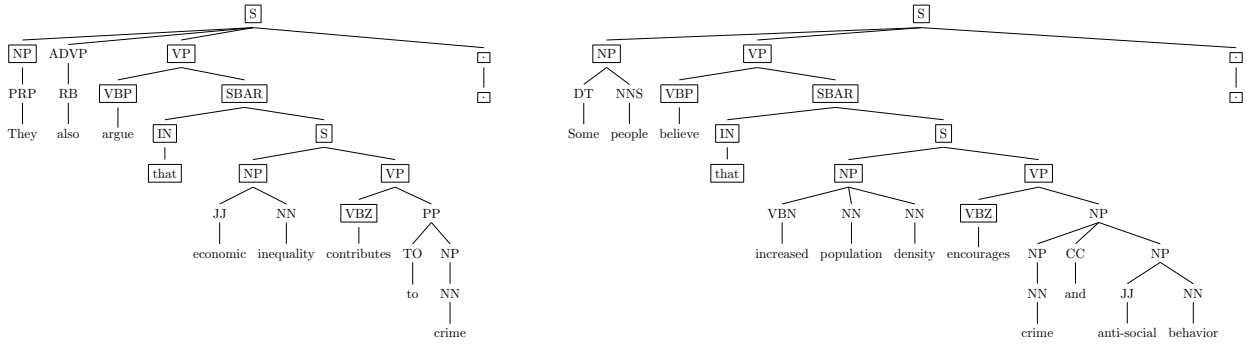


Figure 2: Constituency trees for two sentences containing claims. Boxed nodes are common elements between the two trees. Examples are taken from the IBM corpus.

part-of-speech tags, and their bigrams and trigrams, are employed as well (Mochales Palau & Moens, 2011; Stab & Gurevych, 2014b).

While many of these approaches have been successfully applied to specific domains and genres, yet the generalization of such highly specific features and models to multiple scenarios still remains an open challenge. A recent approach that goes in the direction of cross-genre and cross-topic experimentation is given by Habernal & Gurevych (2015). Such an AM system adopts Toulmin's model for an argument and is based on several sets of carefully designed features, jointly used with unsupervised features obtained with word embeddings. The authors test their system on various genres and topics (although their data set consists entirely of user-generated data), obtaining interesting results. It is hard, however, to evaluate a system across these two corpora, because the definition of claim as well as the argument model used to construct the data set employed in (Habernal & Gurevych, 2015) is different from the one adopted in the construction of the IBM corpus. In (Lippi & Torroni, 2015) we proposed a method exploiting Tree Kernels (Moschitti, 2012) to detect context-independent claims, that was applied across two distinct genres, namely Wikipedia pages and persuasive essays, showing the potential of the method to perform in different domains. MARGOT builds upon and extends the work in (Lippi & Torroni, 2015) to also address evidence detection, as well as the detection of argument component boundaries. Our methodology is driven by the observation that argumentative sentences are often characterized by common syntactic structures. To illustrate, consider the following sentences taken from the IBM corpus,[2] each containing a labeled claim (highlighted in bold):

> They also argue that **economic inequality contributes to crime**.
>
> Some people believe that **increased population density encourages crime and anti-social behavior**.

[2]All the examples are taken from the IBM corpus.

Figure 2 reports the constituency parse trees of such sentences, where boxes highlight the many common parts of their internal structure. 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:

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

Similar considerations can be made also for evidence (i.e. the premises of an argument), although clearly following different patterns. IBM defines three different types of evidence (Aharoni et al., 2014). A first type, named *study*, reports basic facts grounded on reality, mathematical truths, historical events, or results of quantitative or statistical data analyses. An example of study evidence is given by the following fragment:

> *Tropical deforestation is responsible for approximately 20% of world greenhouse gas emissions [REF]*

where [REF] indicates a reference to some bibliographic item, as it appears in Wikipedia. The second type of evidence is called *expert*, as it reports testimonies by persons, groups, committees or organizations having some known expertise in, or authority on, the topic. The following fragment represents an example of expert evidence.

> *Dr. Gary Kleck, a criminologist at Florida State University, estimated that approximately 2.5 million people used their gun in self-defense or to prevent crime each year, often by merely displaying a weapon*

Finally, a third type of evidence, named *anecdotal*, is used to indicate sentences describing specific events or instances (the "anecdotes") or concrete examples. The following fragment is an example of anecdotal evidence.

> *The Orderly Departure Program from 1979 until 1994 helped to resettle refugees in the United States as well as other Western countries*

Even in the case of evidence, features capturing information about the structure of a sentence are highly significant to perform detection. Examples are noun phrases introducing quantitative information, the presence of places, time intervals, or technical terms.

Clearly, there are sentences that can be seen as containing both claims and evidence. For example, the following fragment:

> *Most studies, however, reach the conclusion that violence in video games is not causally linked with aggressive tendencies*

is annotated in the IBM corpus both as an evidence of type study, and as a sentence containing a claim, in the portion starting after "that".

All these examples illustrate that the structure of a sentence is highly informative for argument component detection. This consideration suggests that constituency parse trees are well-suited for this task, since they encode precisely such information. Tree Kernels represent an ideal instrument to capture similarities between parse trees, and thus between argumentative sentences.

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. In particular, Tree Kernels have been successfully employed in many of these applications (see (Moschitti, 2006b, 2012) and references therein).

A Tree Kernel (TK) is designed so as to measure the similarity between two trees, by evaluating the number of their common substructures, typically named *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 with a different feature in a high-dimensional vector 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 $K(x, z)$:

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

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:

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

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. The higher the number of common fragments, the higher the score Δ between two nodes.

According to the definition of fragments, different TKs have been defined: the SubTree Kernel (STK) (Vishwanathan & Smola, 2002), the SubSet Tree Kernel (SSTK) (Collins & Duffy, 2002) and the Partial Tree Kernel (PTK) (Moschitti, 2006a). For STK a fragment is any node of the tree along with all its descendants (thus, a subtree). SSTK is more general than STK, since fragment leaves can be also non-terminal symbols, but always respecting the constraint of not breaking grammar rules. Finally, PTK is the most general, with fragments being any possible portions of subtrees at the considered node. It is clear that these kernels 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.

Whereas in (Lippi & Torroni, 2015) we employed the PTK, MARGOT uses the SSTK. The reason is that we experimentally observed very similar performance between the two kernels, with SSTK being much faster than PTK, during both training and classification. Therefore, for an online server, SSTK is to be preferred.

A second stage is then employed to detect the boundaries of the argument components. The task of boundary detection can be formulated as a sequence labeling problem, a well-known task in machine learning. Given a sentence as a sequence of N words $w_1, w_2, \ldots, w_N$, the goal is to predict a sequence of labels $y_1, \ldots, y_N$, each associated to the corresponding word. Each label y_i can take one out of two possible values, that is whether a certain word belongs to an argument component ($y_i = A$) or not ($y_i = N$). Due to its nature, a sequence labeling task is a *structured output* problem, where the goal of the prediction is a structured object, rather than a scalar value as it happens with standard classification and regression tasks. In sequence labeling, predictions must be made by taking into account the order of the items that are to be classified, so that intra-sequence relations can be exploited and all the elements of a given input sequence are jointly tagged by the classifier. Therefore, *collective classification* is usually performed, by returning the most probable configuration of targets. In this case, the typical machine learning assumption that the

examples to be classified are independent and identically distributed does not hold.

Many machine learning methods exist for sequence labeling. Here we consider SVM-HMM, that combines Structured Support Vector Machines with Hidden Markov Models (Tsochantaridis et al., 2005). SVM-HMM currently represents the state-of-the-art for many sequence labeling tasks (Nguyen & Guo, 2007). Given an observed input sequence of K elements $x = x_1, \ldots, x_K$, where $x_j \in \mathcal{X}$ is a feature vector representing the j-th element in the sequence, SVM-HMM produces a labeling sequence $\hat{y} = \hat{y}_1, \ldots, \hat{y}_K$, where $\hat{y}_j \in \mathcal{Y}$ is the label predicted for the j-th element. The labeling sequence is obtained as a result of the following maximization problem:

$$\hat{y} = \arg\max_{y} \beta^T \Phi(x, y) \tag{3}$$

where β is the vector of model parameters to be learned, and Φ a joint feature map between input and output spaces $\mathcal{X}$ and $\mathcal{Y}$. With respect to classic SVMs, the structured-output version can easily exploit dependencies between output classes via function Φ. The problem in Eq. 3 is typically addressed by a dynamic programming approach that implements a modified instance of the Viterbi algorithm for standard HMMs.

The use of SVM-HMM for argument mining has been proposed also by Habernal & Gurevych (2015), yet in that case no previous phase for argumentative sentence detection was employed. Moreover, they adopt a different argument model where the claim/evidence definitions are different from those of the IBM model. In fact, their data set contains much fewer non-argumentative sentences, which might explain why the authors did not need a preliminary stage to distinguish that category.

4. System Architecture and User Interface

The system is available as a web server, through which users can submit queries in the form of plain text. It has a simple interface (see Figure 3). Once a user asks MARGOT to find arguments on a specific text portion, the system processes the query document according to the pipeline depicted in Figure 1.

First, the text submitted to the web server is processed by the Stanford Parser (Manning et al., 2014),[3] which splits the document into sentences, and produces, for each sentence, the constituency parse tree. Then, each sentence is processed by two TK-based classifiers that detect sentences containing claims and evidence, respectively. Both classifiers take in input the parse tree and the bag-of-words feature vector of the sentence, and produce in output a score, indicating the confidence, according to the classifier, that the sentence contains a claim, or an evidence, respectively. Then, for each sentence that has been predicted to contain an argument component, a boundary detection module is applied, so as to identify the endpoints of all claims/evidence, by exploiting unigrams, bigrams and trigrams on words, part-of-speech tags, lemmas and named-entities. Stanford CoreNLP was used to generate such features. Results are then returned to the

[3] `http://stanfordnlp.github.io/CoreNLP/`

Figure 3: MARGOT's Web interface for entering the source text.

user, by displaying an HTML web page where the detected argument components are highlighted within the original document (see Figure 4). Results highlight claims in bold and evidence in italics. If a text segment is predicted to be both claim and evidence, it is highlighted in bold italics. Consistently, the examples in Section 5 of this article will follow this convention too. The remaining text, which is predicted to be neither claim nor evidence, is hidden away and can be visualized by clicking on the [...] symbol.

The user can also visualize the result of a previous query, selected at random among those stored in MARGOT's database. This is the purpose of the "Random Argument" button in the server Web interface.

5. Results and Discussion

The IBM corpus of Wikipedia pages, which was used to train MARGOT, focuses on highly argumentative topics. However, in order to detect arguments, MARGOT does not require any a-priori information about the topic,[4] and besides, our ambition is to devise a system that can be applied to a variety of input files. We thus designed a number of tests to analyze the behavior of MARGOT in a range of heterogeneous sources and genres. The generalization across different genres is certainly not trivial, and the performance of the system will certainly be influenced by the input genre. For example, MARGOT has not been trained to recognize the arguments found in philosophy classes, such as plain short syllogisms or other purely logic constructs, thus we should not expect MARGOT to distinguish between "Socrates is a man" and "Socrates

[4] We adhere to the terminology introduced by Levy et al. (2014) and also used by Rinott et al. (2015). In particular, we use the term *context-independence*, even though *topic-independence* might be more appropriate.

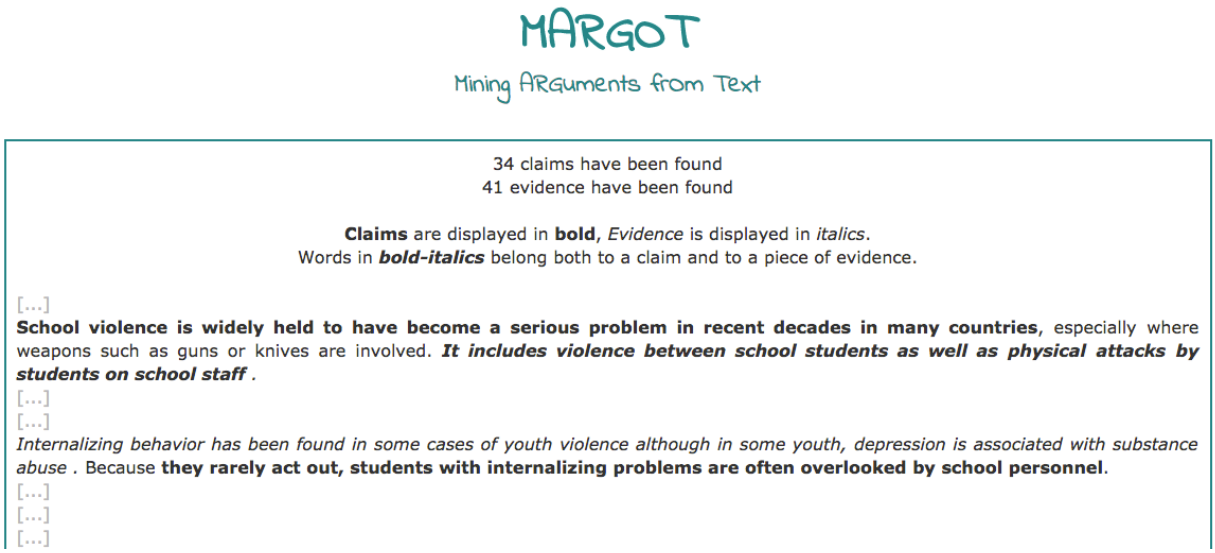


Figure 4: Sample results screen produced by MARGOT.

is mortal". It is more interesting instead to analyze the outcome of the system when processing documents that are usually found online across different genres and topics, especially controversial ones such as news editorials and political debates.

We will start our analysis of MARGOT by presenting quantitative results on the IBM corpus, in order to assess the performance of each stage in MARGOT's pipeline by employing a manually annotated data set. Subsequently, we will detail the qualitative tests run across multiple genres. For such results, we report the URLs of MARGOT's output in appendix.

5.1. IBM corpus

First, we report a quantitative evaluation of MARGOT on the IBM corpus. Following the experimental setup proposed by Levy et al. (2014) and Rinott et al. (2015), we performed a leave-one-topic-out procedure, where each of the 39 topics in the evaluation set is in turn considered as the test set, while the remaining topics make up the training set. All the performance measurements are thus obtained as a macro-average over these 39 topics, giving the idea of the capability of the system to generalize across different topics. As indicated in the IBM guidelines (Rinott et al., 2015), for all the considered tasks the remaining 19 topics are used as development set (for model selection and parameter tuning).

5.1.1. Argumentative sentence detection

For the first stage of the pipeline, we built two distinct classifiers, one to detect sentences containing claims, and the other to detect sentences containing evidence. Since we are considering binary classification problems, we can define True Positives (TP) as the number of sentences correctly identified as containing

a claim (respectively, an evidence), False Positives (FP) as the number of sentences wrongly predicted as containing a claim (respectively, an evidence), False Negatives (FN) as the number of sentences wrongly predicted as not containing a claim (respectively, an evidence). Thus, we can compute precision ($P = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), and $F_1 = \frac{2PR}{P+R}$, i.e., the harmonic mean between precision and recall. These are standard measurements in information retrieval and classification tasks. As a qualitative comparison with the work in Levy et al. (2014), we also compute P and R when selecting the 200 sentences with highest scores for each topic (P@200 and R@200). Finally, we also report the area under the ROC curves (AUROC).

In Tables 1 and 2 we compare our TK-based approach with an SVM predictor that employs bag-of-words (BoW) features of the sentence with standard TF-IDF weights. Constituency trees were obtained with the Stanford CoreNLP suite.[5] We also consider a combinations of TK with bag-of-words (TK+BoW row). We also report the results of a random baseline (RB) predictor, to give the idea of the difficulty of the task, that is largely imbalanced (less than 5% of the examples are positives). In the case of the two retrieval performance measurements (P@200 and R@200), a perfect baseline (PB) is used to show the best possible precision that can be achieved when predicting all the positive examples within the first 200 sentences (such precision is low, since many topics contain much less than 200 claims). The classifier implemented in MARGOT is the one achieving the best performance, by combining TKs with bag-of-words. Note that in (Lippi & Torroni, 2015) we directly compared against the results by Levy et al. (2014) on an older version of this corpus, showing comparable performance.

It is worth remarking that the IBM corpus only labels context-dependent argument components, while our system is not given any topic in advance. Thus, it should not surprise that our false positives include many sentences containing claims that are not strictly related to the topic considered by IBM annotators in the labeling process. Clearly, the reported precision is affected by this behavior. Table 3 shows some examples of such cases, for the task of claim detection.

5.1.2. Argument component boundaries detection

The second stage of our argumentation mining pipeline handles the task of claim/evidence boundary detection. To this aim, we employ an SVM-HMM trained to label a sentence using the claim (C) and the non-claim (N) tokens. As a feature vector representing each word w_j in the sentence, we employed bag-of-words using the original word, the part-of-speech tag, the lemma, and the named entity. To extract such information, we employed the Stanford CoreNLP suite. Such features are also extracted for all the words within a diameter D centered around w_j. Finally, we also added the following bag-of-trigrams for words, part-of-speech tags, lemmas and named entities: $[w_{j-2}w_{j-1}w_j]$, $[w_{j-1}w_jw_{j+1}]$, $[w_jw_{j+1}w_{j+2}]$. We run experiments with different values for parameter D, to assess the contribution of features coming from the neighborhood of each word, and finally chose $D = 3$.

[5] `http://corenlp.stanford.edu`.

Table 1: Results obtained on the IBM Wikipedia corpus on the task of claim detection.

Method	P	R	F_1	P@200	R@200	AUROC
BoW	10.7	50.0	16.9	9.5	56.9	0.805
SSTK	9.8	61.6	16.6	10.6	60.5	0.809
SSTK + BoW	10.5	60.1	17.5	10.6	61.1	0.816
RB	3.1	3.1	3.1	3.2	22.6	–
PB	–	–	–	21.8	100.0	–

Table 2: Results obtained on the IBM Wikipedia corpus on the task of evidence detection.

Method	P	R	F_1	P@200	R@200	AUROC
BoW	8.6	45.1	13.9	8.9	38.5	0.671
SSTK	9.8	53.2	16.1	10.9	46.8	0.718
SSTK + BoW	10.2	53.3	16.7	10.2	53.2	0.724
RB	4.9	4.9	4.9	4.9	22.5	–
PB	–	–	–	18.9	100.0	–

Table 3: Some examples of sentences in the IBM corpus predicted by MARGOT to contain a claim, but actually labeled as negative examples, owing to the context-dependent nature of the annotations.

Topic	Sentence
This house would ban gambling	A benefit of live in-play gambling is that there are much more markets
This house would embrace multiculturalism	Multicultural education is appropriate for everyone
This house would re-engage with Myanmar	Inflation is a serious problem for the economy
This house believes that opinion polls harm the democratic process	It is found that people have the same opinion of their social networks
This house would abolish the monarchy	Historically the Right has advocated preserving the wealth and power of aristocrats and nobles
This house believes that atheism is the only way	Some believe that a moral sense does not depend on religious belief
This house would make physical education compulsory	Psychological well-being is also at risk in the overweight individual due to social discrimination

We evaluate performance by using both traditional performance measurements (Precision, Recall, F_1), that simply count the number of words that are correctly classified as belonging (or not) to a claim, and measurements more suitable for sequence labeling. Traditional measurements, in fact, although somehow significant, do not capture the sequential nature of the labeling problem. Therefore, we employ also the true-positive hit rate H_T and the false-positive hit rate H_F, which have been used in similar structured output tasks (Passerini et al., 2012; Kiziltan et al., 2016). H_T is the percentage of claims that have been correctly detected (or *hit*) for at least one word, whereas H_F is the percentage of predicted claims totally disjoint from the ground truth, i.e., text portions that are wrongly interpreted as claims by the classifier (no word in common with any labeled claim). As for sentence classification, we report the macro-average

Table 4: Results on the claim/evidence boundaries detection task. Performance measurements are macro-averaged on leave-one-topic-out evaluation.

Task	P	R	F_1	H_T	H_F
Claim	63.6	71.8	66.6	85.9	24.8
Evidence	93.8	88.1	90.7	98.6	7.6

for all the metrics. Table 4 shows the results obtained for this sub-task. We achieve an H_T equal to 84.2%, which means that the vast majority of claim boundaries are correctly set to include at least some words of the claim. The percentage of claims for which all tokens are correctly detected (i.e., the correct claim is included in the predicted claim) is 54.2%. It is worth remarking that, also in this phase, some of the false positives of our system (H_F) are due to the fact that we do not exploit context-dependent information, and thus we sometimes retrieve claims that are not labeled as such in the IBM corpus, thus producing a false hit. As an example, consider the following sentence:

> Poor integration of migrant communities can give way to feelings of alienation and resentment, while well-integrated migrants demonstrate that diversity brings progress and social cohesion.

In the original corpus, the only labeled claim is "diversity brings progress and social cohesion", which is correctly labeled by our system. Yet, in addition, SVM-HMM labels as a claim also the first part of the sentence: "Poor integration of migrant communities can give way to feelings of alienation and resentment" which indeed could be considered as a claim.

5.2. Additional Wikipedia pages

As a first qualitative case study, we considered Wikipedia pages regarding either controversial or non-controversial topics. For each group, we randomly selected 10 pages, thus obtaining a test corpus of 20 documents. In order to define what is "controversial," we referred to the list provided by Wikipedia itself.[6]

In this case, MARGOT is tested on the same genre on which the system is actually trained, while further experiments validate the system across different genres. For each page, Table 5 reports the number of sentences N_S in the page, and the number and percentage of sentences predicted to contain claims and evidence, N_C, N_E, $\%_C$, $\%_E$, respectively. Not surprisingly, the results show that the number and percentage of retrieved argument components is much larger in documents regarding controversial topics (34.2% vs. 10.0% for $\%_C$). Yet, in several cases even a topic that is not particularly controversial presents some interesting argument components. For example, the following sentence, taken from "The School of Athens" (Raphael's painting) page, is predicted to contain two distinct claims:[7]

[6] In `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues` a Wikipedia page is defined as controversial if it is constantly being re-edited in a circular manner. The list of controversial issues is thus dynamic by construction. We sampled it by selecting a subset of such issues at a given point in time.

[7] Following the convention introduced earlier and adopted by the MARGOT web site, claims are displayed in bold and evidence in italics.

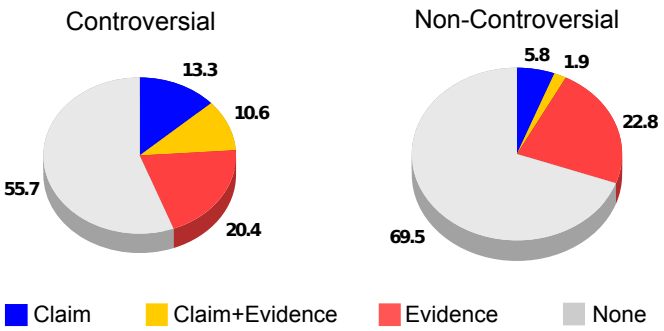


Figure 5: Percentage of sentences containing claims and/or evidence in the 20 Wikipedia articles regarding controversial (left) or non-controversial (right) topics considered in our case study.

> Commentators have suggested that **nearly every great ancient Greek philosopher can be found in the painting**, but **determining which are depicted is difficult**

For the same page, this is a piece of evidence detected by MARGOT, that could be seen as a typical example of "expert" evidence according to IBM:

> *Finally, according to Vasari, the scene includes Raphael himself, the Duke of Mantua, Zoroaster and some Evangelists [REF].*

As another example, the following claim was detected by MARGOT for the "Ethernet" article:

> **Competing proposals and broad interest in the initiative led to strong disagreement over which technology to standardize**.

Figure 5 shows a chart reporting the percentage of argumentative sentences containing claims and/or evidence, with respect to non-argumentative sentences, in the 20 Wikipedia articles associated to controversial and non-controversial topics used in this study. With respect to the numbers reported in Table 5, Figure 5 reports the micro-average instead of the macro-average (Sebastiani, 2002), that is it sums the number of argumentative sentences over all the documents, rather than computing the average of each document and finally an overall average.

Further analysis of the detected claims can be carried out also to identify claims that more frequently support or contest a topic, e.g., with the goal of automatically constructing argument maps. For example, in the "Anti-consumerism" Wikipedia pages, the impact of advertising on consumers' choices is the subject of several claims, detected by MARGOT, including the following ones:

> **Ads are then a detriment to society because they tell consumers that accumulating more and more possessions will bring them closer to self-actualization**

> **Anti-consumerists believe advertising plays a huge role in human life**

Table 5: Results obtained on 20 Wikipedia pages, regarding either controversial (top) or non-controversial (bottom) topics. N_S is the number of sentences, whereas N_C, N_E and $\%_C$, $\%_E$ are the number/percentage of retrieved claims and evidence.

Wikipedia page	N_S	N_C	N_E	$\%_C$	$\%_E$
Anti-consumerism	63	58	22	92.1	34.9
Bioethics	46	13	4	28.3	8.7
Cyberstalking	158	43	46	27.2	29.1
Deep sea mining	58	9	9	15.5	15.5
Delegative democracy	51	16	10	31.4	19.6
Effects of climate change on wine production	44	19	10	43.2	22.7
Geothermal heating	85	14	22	16.5	25.9
School violence	83	44	41	53.0	49.4
Software patents and free software	47	9	20	19.1	42.6
Vaccination and religion	62	10	32	16.1	51.6
Average	69.7	23.5	21.6	34.2	30.0
A Room with a View	119	17	41	14.3	34.5
Contrabass saxophone	26	1	5	3.8	19.2
Ethernet	136	20	39	14.7	27.9
Giardini Naxos	18	0	8	0.0	44.4
Iamb (poetry)	24	3	1	12.5	4.2
Penalty kick	83	11	29	13.3	34.9
Spacecraft	95	3	17	3.2	17.9
The School of Athens	27	5	6	18.5	22.2
Tomato sauce	61	3	3	4.9	4.9
Triple jump	35	5	6	14.3	17.1
Average	62.4	4.8	15.4	10.0	22.7

Anti-consumerists condemn advertising because **it constructs a simulated world that offers fantastical escapism to consumers, rather than reflecting actual reality**

5.3. Newspaper articles

As a second scenario, we evaluated the performance of MARGOT on ten newspaper articles, randomly extracted from the homepage of the New York Times on February 13th, 2016. The articles cover a wide collection of topics, ranging from politics to science, from health to sports. Table 6 reports our results on this corpus. This is a slightly different genre with respect to Wikipedia, yet even in this case MARGOT is capable of detecting a significant number of claims and evidence. Clearly, newspaper articles covering topics such as economy, politics, or science are much richer in arguments than mere chronicles. With respect to Table 6, the articles reporting on an ice fest in Central Park and on the suspension of a corrupted tennis umpire contain only few argument components, while the ones on climate change, interest rates, disparity in life between rich and poor are found to be highly argumentative.

Here is an example of an argumentative sentence detected for an article on Eurozone economy growth. The sentence is predicted to be evidence as a whole, while the bold part is also predicted to be a claim:

Some unsettling recent trends raise the risk that ***the Eurozone could even slip back to***

Table 6: Results obtained on 10 New York Times articles. N_S is the number of sentences, whereas N_C, N_E and $\%_C$, $\%_E$ are the number/percentage of retrieved claims and evidence.

New York Times article	N_S	N_C	N_E	$\%_C$	$\%_E$
Hillary Clinton sharpens focus after democratic debate tussles	31	3	19	9.7	61.3
Disparity in life spans of the rich and the poor is growing	53	15	32	28.3	60.4
Negative interest rates are spreading across the world: here's what...	74	30	28	40.5	37.8
How cold is it? Too cold for an ice fest	21	1	10	4.8	47.6
Eurozone economy growth	63	9	20	14.3	31.7
Tennis umpire suspended for corruption worked at U.S. Open	17	1	10	5.9	58.8
LIGO gravitational waves researchers to divide $3 Million	114	13	66	11.4	57.9
Vegetable soup built for maximum flavor...	38	4	10	10.5	26.3
Science teacher's grasp of climate change is found lacking	37	19	25	51.4	67.6
Twitter, to save itself, must scale back world-swallowing ambitions	61	29	20	47.5	32.8
Average	50.9	12.4	24.0	22.4	48.2

recession.

Political articles are also often very rich in evidence facts. An example is given by the following sentence for an article about US presidential primaries:

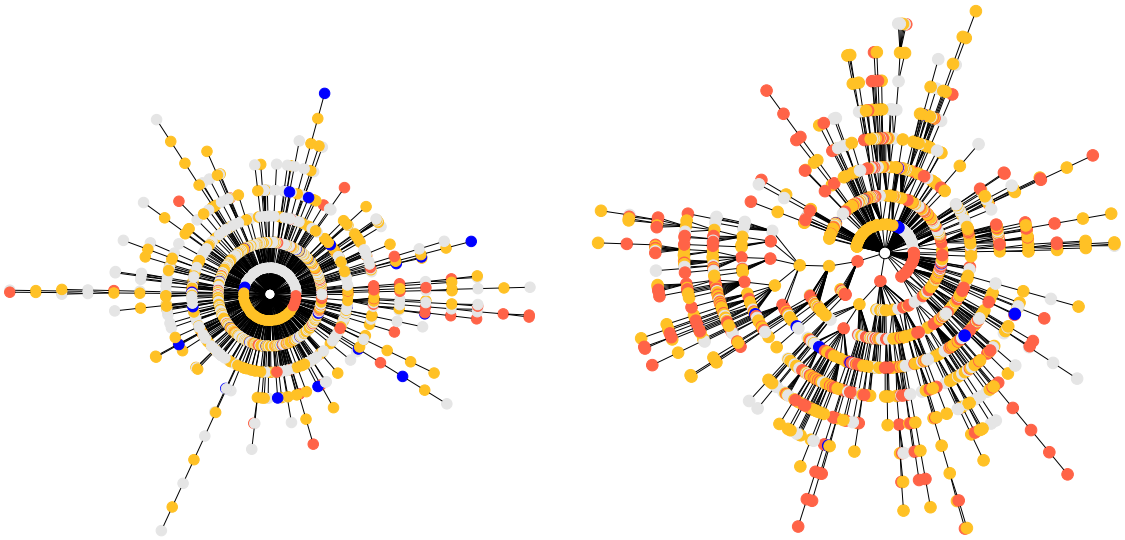
> *On Friday morning, Mrs. Clinton's team unveiled a new television ad that ties her to Mr. Obama on the issue of tightening gun laws – a perceived weakness of Mr. Sanders.*

MARGOT is a first-of-a-kind tool for automatically extracting significant, argumentation-rich information from newspaper articles. This could certainly be a valuable tool for news summarization, detection of trends and patterns in news, recommendation and user customization, cross-domain studies with information coming from social media.

5.4. Reddit comments

As a final case study, we considered the comments in two Reddit threads. In particular, we chose: (1) a subreddit focused on the New Hampshire primaries held on February 9th, 2016[8] and (2) a subreddit focused on climate shift.[9] Using the Python Reddit API Wrapper (PRAW)[10] we extracted 1,484 comments from the first subreddit and 1,520 from the second one. Note that this is a quite different text genre with respect to Wikipedia and also to newspaper articles, and thus it represents a challenging case study, aimed to gauge MARGOT's capability of generalizing across different usage scenarios. The data, in fact, consists of user-generated content written for an online discussion platform, and therefore it is full of colloquial and jargonistic expressions, and a punctuation style and syntax that considerably differ from that of MARGOT's training set. Nevertheless, from the two collections MARGOT extracted 179/803 claims, and 716/1,343

[8] `https://www.reddit.com/r/politics/comments/44zv5n/rpolitics_new_hampshire_primaries_live_thread`
[9] `https://www.reddit.com/r/science/comments/4bixv4/scientists_warn_of_perilous_climate_shift_within/`
[10] `http://praw.readthedocs.org/en/stable/`



(a) New Hampshire primaries (b) Climate shift

Figure 6: The tree of comments induced by the subreddit thread structure. Blue nodes are predicted to contain a claim, red nodes are predicted to contain evidence, yellow nodes are predicted to contain both, whereas gray nodes are predicted to contain no argument component. The inner circle corresponds to the top-level comments in the hierarchy, and the children of each node represent the replies to such comment.

pieces of evidence, respectively. These figures highlight the fact that the "climate shift" thread contains a significantly larger percentage of argument components ($\%_C$=0.47, $\%_E$=0.88) with respect to the "New Hampshire primaries" thread ($\%_C$=0.12, $\%_E$=0.48). This can be explained by a combination of different factors. For one thing, political discussions tend to be more emotionally loaded than discussions about, for instance, environmental policies or scientific phenomena, thus the former attract a great deal more of trolling and yelling, which our system was not trained to recognize as being of an argumentative nature. Another factor is the different degree of polarization in the discussions. In particular, it has been observed in Twitter conversations that if a topic is political, it is common to see two separate, polarized crowds take shape. They form two distinct discussion groups that mostly do not interact with each other (Smith et al., 2014). A similar polarization could manifest itself in Reddit threads about politics, thus we should expect to find comparatively less argumentative content there.

It is very interesting to notice that the detected arguments cover a wide spectrum of different topics. In some cases claims consist of a short, personal opinion, as in the following sentences, taken from the New Hampshire primaries subreddit:[11]

> Objectively however I say that **Kasich needs to gain ground soon to stay viable**, so...
>
> **Bernie is very polarizing**

[11] Notice that the last sentence is labeled as evidence as well as a claim.

people think Bernie is delusional

In other cases, the claim consists of a more elaborated reasoning. In the setting of New Hampshire primaries, such elaborate claims are made about the political context:[12]

> *Clinton with his centrism taught the democrats a lesson (true or not) that they haven't forgotten: that* ***the American people are actually conservative and a true progressive agenda is impossible****.*

> **parliamentary systems like Sweden makes sure that reactionary and populist parties never gain any significant power**

> **no GOP President can get elected without Ohio**

The following example is instead taken from the climate shift subreddit:[13]

> *I argued earlier that* ***the ideal thing to do would to be to ensure our species survival would be to organize a group that is dedicated to building a clean energy infrastructure for the 3rd world*** *as fast as possible*

As for evidence facts, sometimes MARGOT detects sentences that describe the trend of Reddit comments, as in the following case:

> *You're seeing 100% pro-Sanders posts because 85% of users are pro-Sanders.*

Other pieces of evidence detected by MARGOT mention instead statistics or known facts that can be used to support claims:

> *Hillary had all of the unpledged delegates (super delegates) on her side in 2008, and Obama gained massive popularity and those unpledged delegates jumped ship.*

Very often an evidence contains a reported opinion or fact, in which case the opinion is itself a claim. The following illustration is a labeled fragment of the climate shift subreddit:

> *As Martin Luther King, Jr. said, "****There is no human circumstance more tragic than the persisting existence of a harmful condition for which a remedy is readily available****".*

In the climate shift subreddit we can also observe many claims, among those detected, that are in favor of methane rather than carbon dioxide production:

[12] The first sentence is labeled as evidence as well as (partially) a claim.
[13] The whole sentence here is labeled as evidence, and its central part also as a claim.

The main concern is that **reservoirs may create anoxic conditions in their sediments that favor the production of methane rather than carbon dioxide**.

Additionally, it may be important that **reservoirs convert this plant carbon into methane rather than carbon dioxide, since it is a more potent greenhouse gas**.

In the short term, **methane is a much more potent greenhouse gas than carbon dioxide**.

This could enable further statistical, e.g. clustering-based, analysis to identify methane vs carbon dioxide as one of the most important elements in this debate.

A further, new type of analysis that could be unlocked by our system's predictions combined with information about the subreddit thread structure is a network analysis of argument components. To illustrate, we built a radial tree for each considered subreddit, by placing the top-level comments in the hierarchy (the ones without a parent) in the inner circle, and by subsequently adding a child for each reply in the thread. The two trees are depicted in Figure 6, where each node is colored according to the predicted argument components for that comment. The radial trees built with the MARGOT technology offer an innovative visualization of subreddit threads, which could facilitate new types of content and network analysis. In particular, the different visual impact of the two radial trees reflects the different argumentative structure of the two discussions. The New Hampshire primaries thread has 0.74 children for each non-argumentative node on average, 0.80 for each argumentative node, 1.72 average replies per first-level node (average tree depth). The climate shift thread has 0.67 children for each non-argumentative node on average, 0.91 for each argumentative node, and 2.87 average replies per first-level node (average tree depth). From these figures, it is clear that the climate shift discussion is more oriented to debate and to discussion among users, than the thread on politics.

6. Conclusions

MARGOT is a first-of-a-kind tool for automatically extracting significant, argumentation-rich information from a variety of genres. Possible applications of the state-of-the-art technologies underlying MARGOT range from text summarization, detection of trends and patterns in news, recommendation and user customization, and cross-domain studies with information coming from social media. There is also an added value offered by the online availability of this AM system, which was specifically designed to reach out to users that have no background in argumentation. Our aim there is to build on other previous successful experiences in popularizing computational argumentation. For example, ASPARTIX (Egly et al., 2008) was proposed as an efficient DLV-based system for computing abstract argumentation semantics, but it really made a strong contribution to the take-up of computational argumentation technologies by a wider community thanks to its Web interface,[14] and it set the de-facto standard for writing abstract argumentation

[14] `http://rull.dbai.tuwien.ac.at:8080/ASPARTIX/`

frameworks that can be fed to many existing solvers. Our ambition is to bring AM to the attention of a wider community and foster greater interest in this emerging technology. In the future we plan to extend the services offered by MARGOT, by defining an API, following the successful experience of cognate domains, for instance sentiment analysis, where many popular tools are accessible by APIs such as IBM's AlchemyAPI.[15]

In this work, we have described the technology behind MARGOT and we have discussed its performance and application potential. Among the domains traditionally addressed by expert and intelligent systems we have shown links with recommender systems, public health policy-making, and the legal domain. We wish to conclude by discussing limitations. By using tree kernels and intentionally avoiding genre-dependent features, we offer a tool that in principle could be adapted to any genre where the structure of the language is indicative of the presence of arguments. We have shown via a quantitative and qualitative analysis (see also (Lippi & Torroni, 2015)) that our approach is successful in at least some domains. However, in order to produce a truly "general-purpose" AM system, we should be able to rely on training sets that cover multiple genres, since every genre has a different style in which arguments are expressed. Unfortunately, gold standards for building and training AM systems are still rare and limited to a few domains (Lippi & Torroni, 2016). Moreover, it often so happens that argumentation is in fact context-dependent, thus the sentence "Socrates is a man" may have an argumentative force in some context, but it may not in another context. So there are intrinsic limitations in the notion itself of general-purpose AM, and thus we should not be surprised to find certain outputs of an AM system controversial or even counter-intuitive. Indeed, even those few available gold standards in this domain have a relatively low inter-annotator agreement (Habernal et al., 2014). In spite of that, AM is a challenging but worthwhile enterprise, and an AM system "for the layman," such as MARGOT, could help satisfying a long-recognized need for tools that enable users to browse, visualise, search, and manipulate arguments and argument structures (Rahwan et al., 2007), thus unlocking a great diversity of under-exploited resources that can be drawn upon in building the foundations for significant argumentation-based applications.

References

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., & Slonim, N. (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* (pp. 64–68). Association for Computational Linguistics.

Ashley, K. D., & Walker, V. R. (2013). Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In E. Francesconi, & B. Verheij (Eds.), *ICAIL 2012, Rome, Italy* (pp. 176–180). ACM. URL: http://dl.acm.org/citation.cfm?id=2514622. doi:10.1145/2514601.2514622.

Bench-Capon, T. J. M., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, *171*, 619–641. URL: http://dx.doi.org/10.1016/j.artint.2007.05.001. doi:10.1016/j.artint.2007.05.001.

Biran, O., & Rambow, O. (2011). Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, *5*, 363–381. URL: http://dx.doi.org/10.1142/S1793351X11001328. doi:10.1142/S1793351X11001328.

[15] http://www.alchemyapi.com/api/sentiment-analysis

Bourguet, J.-R., Thomopoulos, R., Mugnier, M.-L., & Abécassis, J. (2013). An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy. *Expert Systems with Applications*, *40*, 4539 – 4546. URL: http://www.sciencedirect.com/science/article/pii/S0957417413000912. doi:http://dx.doi.org/10.1016/j.eswa.2013.01.059.

Cabrio, E., & Villata, S. (2012a). Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (ACL 2012)* (pp. 208–212). Jeju, Korea: Association for Computational Linguistics.

Cabrio, E., & Villata, S. (2012b). Natural language arguments: A combined approach. In L. D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, & P. J. F. Lucas (Eds.), *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012* (pp. 205–210). IOS Press volume 242. URL: http://dx.doi.org/10.3233/978-1-61499-098-7-205. doi:10.3233/978-1-61499-098-7-205.

Collins, M., & Duffy, N. (2002). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.* (pp. 263–270). ACL.

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets. Reasoning About a Highly Connected World*. Cambridge University Press. URL: http://www.cs.cornell.edu/home/kleinber/networks-book/.

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (p. to appear). Lisbon, Portugal: Association for Computational Linguistics.

Egly, U., Gaggl, S. A., & Woltran, S. (2008). ASPARTIX: implementing argumentation frameworks using answer-set programming. In M. G. de la Banda, & E. Pontelli (Eds.), *Logic Programming, 24th International Conference, ICLP 2008, Udine, Italy, December 9-13 2008, Proceedings* (pp. 734–738). Springer volume 5366 of *Lecture Notes in Computer Science*. URL: http://dx.doi.org/10.1007/978-3-540-89982-2_67. doi:10.1007/978-3-540-89982-2_67.

Freeman, J. B. (1991). *Dialectics and the macrostructure of arguments: A theory of argument structure* volume 10. Walter de Gruyter.

Galgani, F., Compton, P., & Hoffmann, A. (2015). Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications*, *42*, 6391 – 6407. URL: http://www.sciencedirect.com/science/article/pii/S0957417415002523. doi:http://dx.doi.org/10.1016/j.eswa.2015.04.022.

García-Cumbreras, M. A., Montejo-Ráez, A., & Díaz-Galiano, M. C. (2013). Pessimists and optimists: Improving collaborative filtering through sentiment analysis. *Expert Systems with Applications*, *40*, 6758 – 6765. URL: http://www.sciencedirect.com/science/article/pii/S0957417413004442. doi:http://dx.doi.org/10.1016/j.eswa.2013.06.049.

Goudas, T., Louizos, C., Petasis, G., & Karkaletsis, V. (2014). Argument extraction from news, blogs, and social media. In A. Likas, K. Blekas, & D. Kalles (Eds.), *Artificial Intelligence: Methods and Applications* (pp. 287–299). Springer International Publishing volume 8445 of *LNCS*.

Grosse, K., González, M. P., Chesñevar, C. I., & Maguitman, A. G. (2015). Integrating argumentation and sentiment analysis for mining opinions from Twitter. *AI Communications*, *28*, 387–401. doi:10.3233/AIC-140627.

Habernal, I., Eckle-Kohler, J., & Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In E. Cabrio, S. Villata, & A. Wyner (Eds.), *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing. Forlì-Cesena, Italy, July 21-25, 2014*. CEUR-WS.org volume 1341 of *CEUR Workshop Proceedings*.

Habernal, I., & Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (pp. 2127–2137). The Association for Computational Linguistics.

Houngbo, H., & Mercer, R. (2014). An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 19–23). Association for Computational Linguistics. URL: http://acl2014.org/acl2014/W14-21/pdf/W14-2103.pdf.

Kiziltan, Z., Lippi, M., & Torroni, P. (2016). Constraint detection in natural language problem descriptions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, USA, July 9-15, 2016*.

Kunz, W., & Rittel, H. W. (1970). *Issues as elements of information systems* volume 131 of *Institute of Urban and Regional Development, University of California*. Berkeley, California, US.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., & Slonim, N. (2014). Context dependent claim detection. In J. Hajic, & J. Tsujii (Eds.), *COLING 2014, Dublin, Ireland* (pp. 1489–1500). ACL.

Lippi, M., Lagioia, F., Contissa, G., Sartor, G., & Torroni, P. (2015). Claim detection in judgments of the EU Court of Justice. In *VI International Workshop on Artificial Intelligence and the Complexity of Legal Systems (AICOL)*. Braga, Portugal. URL: http://www.aicol.eu.

Lippi, M., & Torroni, P. (2015). Context-independent claim detection for argument mining. In Q. Yang, & M. Wooldridge (Eds.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015* (pp. 185–191). AAAI Press. URL: http://ijcai.org/papers15/Abstracts/IJCAI15-033.html.

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, *16*, 10:1–10:25. URL: http://doi.acm.org/10.1145/2850417. doi:10.1145/2850417.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60). URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

Mochales Palau, R., & Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, *19*, 1–22. URL: http://dx.doi.org/10.1007/s10506-010-9104-x. doi:10.1007/s10506-010-9104-x.

Moschitti, A. (2006a). Efficient convolution kernels for dependency and constituent syntactic trees. In J. Frnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Machine Learning: ECML 2006* (pp. 318–329). Springer Berlin Heidelberg volume 4212 of *LNCS*. doi:10.1007/11871842_32.

Moschitti, A. (2006b). Making tree kernels practical for natural language learning. In *EACL* (pp. 113–120).

Moschitti, A. (2012). State-of-the-art kernels for natural language processing. In *Tutorial Abstracts of ACL 2012* ACL '12 (pp. 2–2). Stroudsburg, PA, USA: Association for Computational Linguistics.

Nguyen, N., & Guo, Y. (2007). Comparisons of sequence labeling algorithms and extensions. In *Proc. of the24th Int'l Conf. on Machine Learning, ICML 2007* (pp. 681–688). ACM volume 227 of *ACM Int'l Conf. Proceeding Series*.

Pal, K. (1999). An approach to legal reasoning based on a hybrid decision-support system. *Expert Systems with Applications*, *17*, 1 – 12. URL: http://www.sciencedirect.com/science/article/pii/S0957417499000159. doi:http://dx.doi.org/10.1016/S0957-4174(99)00015-9.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, *2*, 1–135. URL: http://dx.doi.org/10.1561/1500000011. doi:10.1561/1500000011.

Park, J., & Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 29–38). Baltimore, Maryland: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W/W14/W14-2105.

Passerini, A., Lippi, M., & Frasconi, P. (2012). Predicting metal-binding sites from protein sequence. *IEEE/ACM Trans. Comput. Biology Bioinform.*, *9*, 203–213.

Peldszus, A., & Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, *7*, 1–31.

Rahwan, I., Zablith, F., & Reed, C. (2007). Laying the foundations for a world wide argument web. *Artif. Intell.*, *171*, 897–921. URL: http://dx.doi.org/10.1016/j.artint.2007.04.015. doi:10.1016/j.artint.2007.04.015.

Rinott, R., Khapra, M., Alzate, C., Dankin, L., Aharoni, E., & Slonim, N. (2015). Show me your evidence – an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal, 17-21 September 2015* (pp. 440–450). Association for Computational Linguistics.

Roitman, H., Hummel, S., Rabinovich, E., Sznajder, B., Slonim, N., & Aharoni, E. (2016). On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 991–996). International World Wide Web Conferences Steering Committee.

Rosenthal, S., & McKeown, K. (2012). Detecting opinionated claims in online discussions. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012* (pp. 30–37). IEEE Computer Society.

Saint-Dizier, P. (2012). Processing natural language arguments with the<textcoop>platform. *Argument & Computation*, *3*, 49–82. URL: http://dx.doi.org/10.1080/19462166.2012.663539. doi:10.1080/19462166.2012.663539.

Sardianos, C., Katakis, I. M., Petasis, G., & Karkaletsis, V. (2015). Argument extraction from news. In *Proceedings of the Second Workshop on Argumentation Mining* (pp. 56–66). ACL.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*, 1–47.

Smith, M. A., Rainie, L., Himelboim, I., & Shneiderman, B. (2014). *Mapping Twitter topic networks: from polarised crowds to community clusters*. Technical Report Pew Research Center, Washington, DC. URL: http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/.

Stab, C., & Gurevych, I. (2014a). Annotating argument components and relations in persuasive essays. In J. Hajic, & J. Tsujii (Eds.), *COLING 2014, Dublin, Ireland* (pp. 1501–1510). ACL. URL: http://www.aclweb.org/anthology/C14-1142.

Stab, C., & Gurevych, I. (2014b). Identifying argumentative discourse structures in persuasive essays. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *EMNLP 2014, Doha, Qatar* (pp. 46–56). ACL.

Teze, J. C., Gottifredi, S., García, A. J., & Simari, G. R. (2015). Improving argumentation-based recommender systems through context-adaptable selection criteria. *Expert Systems with Applications*, *42*, 8243 – 8258. URL: http://www.sciencedirect.com/science/article/pii/S0957417415004509. doi:http://dx.doi.org/10.1016/j.eswa.2015.06.048.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, *6*, 1453–1484.

Vishwanathan, S. V. N., & Smola, A. J. (2002). Fast kernels for string and tree matching. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]* (pp. 569–576). MIT Press.

Walton, D. (2009). Argumentation theory: A very short introduction. In G. Simari, & I. Rahwan (Eds.), *Argumentation in Artificial Intelligence* (pp. 1–22). Springer US. doi:10.1007/978-0-387-98197-0_1.

Appendix

We report here the URLs of the pages produced by MARGOT in the experiments described in the paper.

Wikipedia pages

- http://margot.disi.unibo.it/answers/Anti-consumerism.html
- http://margot.disi.unibo.it/answers/A_Room_with_a_View.html

- http://margot.disi.unibo.it/answers/Contrabass_saxophone.html
- http://margot.disi.unibo.it/answers/Bioethics.html
- http://margot.disi.unibo.it/answers/Cyberstalking.html
- http://margot.disi.unibo.it/answers/Deep_sea_mining.html
- http://margot.disi.unibo.it/answers/Delegative_democracy.html
- http://margot.disi.unibo.it/answers/Effects_of_climate_change_on_wine_production.html
- http://margot.disi.unibo.it/answers/Ethernet.html
- http://margot.disi.unibo.it/answers/Geothermal_heating.html
- http://margot.disi.unibo.it/answers/Giardini_Naxos.html
- http://margot.disi.unibo.it/answers/Iamb_poetry.html
- http://margot.disi.unibo.it/answers/Penalty_kick.html
- http://margot.disi.unibo.it/answers/School_violence.html
- http://margot.disi.unibo.it/answers/Software_patents_and_free_software.html
- http://margot.disi.unibo.it/answers/Spacecraft.html
- http://margot.disi.unibo.it/answers/The_School_of_Athens.html
- http://margot.disi.unibo.it/answers/Tomato_sauce.html
- http://margot.disi.unibo.it/answers/Triple_jump.html
- http://margot.disi.unibo.it/answers/Vaccination_and_religion.html

New York Times articles

- http://margot.disi.unibo.it/answers/NYT_1.html
- http://margot.disi.unibo.it/answers/NYT_2.html
- http://margot.disi.unibo.it/answers/NYT_3.html
- http://margot.disi.unibo.it/answers/NYT_4.html
- http://margot.disi.unibo.it/answers/NYT_5.html
- http://margot.disi.unibo.it/answers/NYT_6.html
- http://margot.disi.unibo.it/answers/NYT_7.html
- http://margot.disi.unibo.it/answers/NYT_8.html
- http://margot.disi.unibo.it/answers/NYT_9.html
- http://margot.disi.unibo.it/answers/NYT_10.html

Reddit

- http://margot.disi.unibo.it/answers/44zv5n.html
- http://margot.disi.unibo.it/answers/4bixv4.html