A

Argumentation Mining: State of the Art and Emerging Trends

MARCO LIPPI and PAOLO TORRONI, DISI, University of Bologna

Argumentation mining aims at automatically extracting structured arguments from unstructured textual documents. It has recently become a hot topic also due to its potential in processing information originating from the Web, and in particular from social media, in innovative ways. Recent advances in machine learning methods promise to enable breakthrough applications to social and economic sciences, policy making, and information technology: something that only a few years ago was unthinkable. In this survey article, we introduce argumentation models and methods, review existing systems and applications, and discuss challenges and perspectives of this exciting new research area.

CCS Concepts: •**Information systems** → **Information extraction; Web mining;** •**General and reference** → **Surveys and overviews;** •**Computing methodologies** → Nonmonotonic, default reasoning and belief revision;

Additional Key Words and Phrases: Argumentation mining, artificial intelligence, computational linguistics, machine learning, knowledge representation, social media

ACM Reference Format:
Marco Lippi and Paolo Torroni, 2015. Argumentation Mining: State of the Art and Emerging Trends *ACM Trans. Internet Technol.* V, N, Article A (January YYYY), 25 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. MOTIVATION

Argumentation is a multi-disciplinary research field, which studies debate and reasoning processes, and spans across and ties together diverse areas such as logic and philosophy, language, rhetoric and law, psychology and computer science. Argumentation has come to be increasingly central as a core study within artificial intelligence [Bench-Capon and Dunne 2007], due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning. In particular, the study of argumentation in artificial intelligence gave rise to a new discipline called *computational argumentation*. Argumentation is gaining momentum in some parts of cognitive sciences too, where recent studies seem to indicate that the function of human reasoning itself is argumentative [Mercier and Sperber 2011]. Even in the (computational) social sciences, agent-based simulation models have recently been proposed, whose micro-foundation explicitly refers to argumentation theories [Mäs and Flache 2013; Gabbriellini and Torroni 2014]. An important source of data for many of the disciplines interested in such studies is the Web, and social media

---

This manuscript is an extended version of "Argument mining: a machine learning perspective" by Marco Lippi and Paolo Torroni, presented on July 26, 2015 at the *IJCAI 2015 International Workshop on Theory and Applications of Formal Argument* (TAFA-15) in Buenos Aires, Argentina.
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.
Authors' addresses: Department of Computer Science and Engineering, University of Bologna, viale Risorgimento 2, 40136 Bologna (BO). Email: marco.lippi3@unibo.it, paolo.torroni@unibo.it

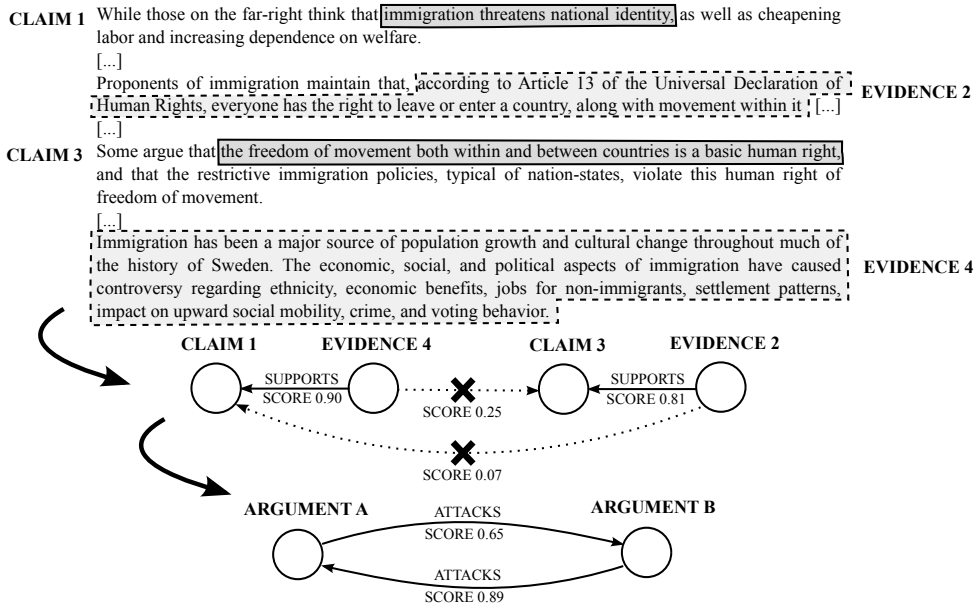
DOI: http://dx.doi.org/10.1145/0000000.0000000

in particular. Online newspapers, product reviews, blogs etc. provide an heterogeneous and ever-growing flow of information where (user-generated) arguments can be found, isolated and analyzed. The availability of such data, together with tremendous advances in computational linguistics and machine learning, created fertile ground for the rise of a new area of research called *argumentation* (or *argument*) *mining*.

The main goal of argumentation mining is to automatically extract arguments from generic textual corpora, in order to provide structured data for computational models of argument and reasoning engines.

Figure 1 shows an example of automatic extraction of arguments from text which could be performed by a fully-fledged argumentation mining system.[1] First, sentences recognized as argumentative are extracted from the input document, and argument components—in this case, claims and supporting evidence—are located within such sentences (Figure 1(a)). Subsequently, links between argument components are predicted (Figure 1(b)) in order to construct complete arguments. Finally, the connections between arguments are inferred, so as to produce a complete argument graph (Figure 1(c)).

Fig. 1. Example of argument extraction from plain text.



The growing excitement in this area is tangible. The initial studies started to appear only a few years ago within specific genres such as legal texts, online reviews and debate [Mochales Palau and Moens 2011; Saint-Dizier 2012; Cabrio and Villata 2012b]. In 2014 alone there have been at least three international events on argumentation mining,[2] while research on this topic is gaining visibility at major artificial

[1]The text and claim/evidence annotations are taken from the IBM corpus (see Section 4). In that corpus, evidence plays the role of claim-supporting premises. We will cover argument models in Section 2.

[2]The First ACL Workshop on Argumentation Mining, http://www.uncg.edu/cmp/ArgMining2014/, SICSA Workshop on Argument Mining: Perspectives from Information Extraction, Information Retrieval and Computational Linguistics http://www.arg-tech.org/index.php/sicsa-workshop-on-argument-mining-2014/, and

intelligence and computational linguistics conferences, and IBM has recently funded a multi-million cognitive computing project whose core technology is argument mining.[3]

Indeed, this is not only a scientifically engaging problem, but also one with self-evident application potential. The Web and online social networks offer 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 and Lee 2008] and social network analysis [Easley and Kleinberg 2010]. An argumentation mining 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.

This article is the first structured survey of models, methods, and applications of this exciting and rapidly developing research area. The motivation behind it is that many efforts made under the general umbrella of argumentation mining in fact aim to solve a constellation of sub-tasks, whereas a single unifying view is still missing. We thus aim to propose one such view, as well as to discuss challenges and perspectives.

This study is organized as follows. In Section 2 we introduce argumentation models found in the literature, with specific emphasis on those currently used in argumentation mining. In Section 3 we propose a general architecture for argumentation mining systems, identifying common (and uncommon) tasks and sub-tasks, and we critically discuss the main techniques used to address them, their shortcomings, and possible ways to advance the state of the art. Section 4 describes corpora and applications. These two are closely related, since argumentation mining has been applied in different domains, and for each of them a number of domain-related corpora have been constructed. One of the main hurdles for those approaching argumentation mining for the first time is indeed the lack of a comprehensive analytic study of existing corpora where new techniques can be evaluated, and that is precisely what we offer there. Section 5 proposes a subjective perspective on the major challenges that lie ahead. Section 6 concludes with a look to future applications.

2. ARGUMENTATION MODELS

The discipline of argumentation has ancient roots in dialectics and philosophy, as that branch of knowledge dedicated to the study and analysis of how statements and assertions are proposed and debated, and conflicts between diverging opinions are resolved [Bench-Capon and Dunne 2007].

Given this long-standing tradition, over the centuries argumentation has permeated many diverse areas of knowledge besides philosophy, such as language and communication, logic, rhetoric, law, and computer science. It should come to no surprise that literature is rich with argument representation models. One of the best known is due to Toulmin [1958]. Toulmin proposed that the logical microstructure of human argumentation and reasoning consists of six categories: an (incontrovertible) *datum*, which forms the basis for making a (subjective, possibly controversial) *claim*, the rule of inference (*warrant*) that links them, and other elements that serve to show how certain we are of the claim (*qualifiers*), or to set conditions for the claim to hold (*rebuttal*), or even to give a justification to the warrant (*backing*). Toulmin's model has been largely influential. However, in practice, different applications will require different argumentation structures, and the *representational fit* of Toulmin's model for use in diverse fields

the BiCi Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, http://www-sop.inria.fr/members/Serena.Villata/BiCi2014/frontiersARG-NLP.html.

[3]IBM Debating Technologies, http://researcher.watson.ibm.com/researcher/view_group.php?id=5443.

such as policy, law, design, science, philosophy and user-generated content is a matter of discussion [Newman and Marshall 1991; Habernal et al. 2014]. Similar considerations would apply to other influential models, such as IBIS [Kunz and Rittel 1970] and Freeman's [1991].

Starting from the pioneering works by Pollock [1987], Simari and Loui [1992], and Dung [1995], among others, models of argumentation have also spread in the area of AI, especially in connection with knowledge representation, non-monotonic reasoning, and multi-agent systems research, giving rise to a new field named *computational argumentation*. Here, different models have been developed along three different categories [Bentahar et al. 2010]: rhetorical, dialogical, and monological models. The two former categories highlight argumentation as a dynamic process: rhetorical models put an emphasis on the audience and on persuasive intention, whereas dialogical models describe the ways arguments are connected in dialogical structures. Monological models instead emphasize the structure of the argument itself, including the relations between the different components of a given argument.

For example, Figure 1(a) and (b) follow a monological model where the argument components are claims and evidence linked by a support relation. Figure 1(c) relates two arguments with one another, following a dialogical model based on an attack relation.

Another well-known classification in computational argumentation is the dichotomy between *abstract* argumentation and *structured* argumentation. The former is rooted in Dung's work, and it considers each argument as an atomic entity without internal structure. It is thus a dialogical model which provides a very powerful framework to model and analyze *attack* relations between arguments, as illustrated in Figure 1(c), or sets of arguments, which may or may not be *justified* according to some semantics. Structured argumentation proposes an internal structure for each argument, which could be that adopted in Figure 1(a) and (b), described in terms of some knowledge representation formalism. When the goal is to extract portions of arguments from natural language, defining the structure of an argument becomes crucial. Therefore, argumentation mining typically employs structured argumentation models. These are typically monological models; they can be however embedded also in dialogical models [Besnard et al. 2014], as shown by Figure 1.

Because there are many significant proposals for structured argumentation [Besnard et al. 2014], it is impossible to give a single formal, universally accepted definition of structured argument. An intuitive definition of argument was given by Walton 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]. In the literature, conclusions are sometimes referred to as *claims*, premises are often called *evidence* or *reasons* (or *datum*, in Toulmin's model), and the link between the two, i.e., the inference (*warrant*), is sometimes called the *argument* itself. *Argumentation* has historically referred to the process of constructing arguments and, since the advent of computational argumentation, to the process of determining the set of justified conclusions of a set of arguments. However, *argumentation mining* and *argument mining* are often used interchangeably and in a broad sense, as the field yet retains a strong element of conceptual exploration.

The task of detecting the premises and conclusion of an argument, as found in a text of discourse, is typically referred to as *identification* [Walton 2009] or *extraction*, whereas more specific sub-tasks are *claim detection* and *evidence detection* [Levy et al. 2014; Rinott et al. 2015]. Other tasks are *attribution* which refers to attributing authorship to arguments, *completion* whose goal is to infer implicit argument components such as enthymemes and tacit assumptions related to commonsense reasoning, *relation prediction* aiming at identifying inter- and intra-argument relations.

Being argumentation mining a young research domain, not only its definitions but also its approaches and targets vary widely. Some research aims at extracting the arguments from generic unstructured documents, which is a fundamental step in practical applications [Levy et al. 2014], whereas other starts from a given set of arguments and focuses on aspects such as the identification of attack/support relations between them [Cabrio and Villata 2013; Boltuzic and Snajder 2014].

There are also many related tasks, such as rhetorical characterization of sentences [Houngbo and Mercer 2014], opinionated claim analysis [Rosenthal and McKeown 2012], premise verification [Park and Cardie 2014], etc. In this survey, we mainly focus on core argumentation mining (sub-)tasks. Related tasks are briefly surveyed in Section 4 (see Table IV).

3. METHODS

Any AM system has to address a constellation of strictly inter-related tasks. Therefore, before we discuss methods, we shall first define a taxonomy to organize the tasks that go under the umbrella of AM. Next, we will survey the machine learning and natural language methods employed by existing systems based on the role they play in a typical AM system architecture. The systems developed so far implement a pipeline architecture (see Figure 2), through which they process unstructured textual documents and produce as output 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, and it will be described in a separate subsection.

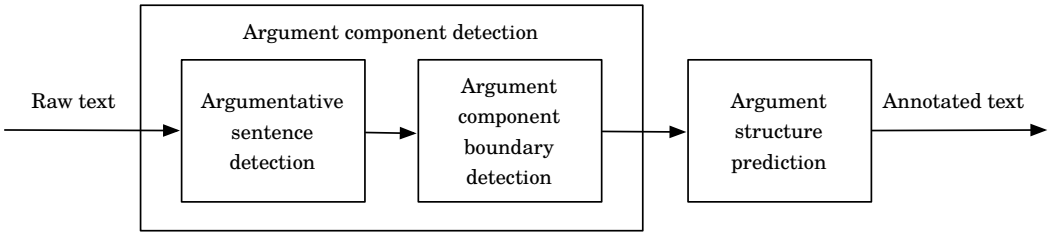


Fig. 2. Pipeline architecture of an AM system

Interestingly, the challenges faced by AM share important analogies with problems defined in neighboring areas of research. We thus conclude this section by drawing cross-disciplinary bridges between other tasks and methods in machine learning, natural language processing, discourse analysis, computational linguistics, information extraction, and knowledge representation (see Table II).

3.1. A taxonomy of problems

AM problems can be described along five orthogonal dimensions: granularity of input, genre of input, argument model, granularity of target, and goal of analysis.

The *granularity* of the processed text indicates the level of detail at which arguments (or parts thereof) are searched. Some approaches consider text portions at the level of paragraphs. This is the case, for example, with argumentative zoning [Teufel 1999], considered to be the forerunner of AM. Most of current research focuses on sentences, whereas some authors address finer-grained intra-sentence argument component boundaries.

The *genre* defines the type of input data, such as legal/law, online discussions, news, essays, etc. Until now, existing approaches have mainly focused on a single genre. Each

genre has its own peculiarities. For example, Budzynska et al. [2014] show that AM from dialogue cannot be satisfactorily addressed using dialogue-agnostic models.

Each AM system refers to a specific *argument model*, such as those surveyed in the previous section. So far, the vast majority of implemented systems use a claim/premise model.

The *target* of the mining process also varies in terms of its granularity. Some works target specific argument components such as the claim; other works target the whole argument.

Finally, the *goal* of the analysis also ranges in a spectrum of possibilities which can be used to categorize the existing approaches. The most common goals are: detection, classification, relation prediction, attribution, and completion.

In this survey we will refer to these dimensions in order to frame the AM problem and task at hand, whenever not immediately clear from the context.

3.2. Argument component detection

The typical goal of an AM system's first stage is to detect arguments (or argument components, depending on the desired target granularity) within the input text document. The retrieved entities will thus represent nodes (or parts thereof) in the final argument graph. In most of the existing systems, this problem is addressed by splitting it into two distinct sub-problems: the extraction of argumentative sentences and the detection of component boundaries. These two tasks are usually addressed in two steps, as the next two subsections will show. Yet, it is worth mentioning that not all systems necessarily follow this two-stage pipeline: in particular, [Stab and Gurevych 2014b] and [Eckle-Kohler et al. 2015] assume that the boundaries of the argument components have been previously detected by other means, thus they restrict their goal to a single classification task.

3.2.1. Argumentative sentence detection. The first step in the detection of argument components usually addresses the task of extracting those sentences in the input document that contain an argument (or part of it), and that can therefore be defined as *argumentative*. Referring to the example in Figure 1, this would correspond to identifying all the sentences that contain at least one of the argument components highlighted in gray. The problem can be easily formulated as a classification task, which could in principle be addressed by any kind of machine learning classifier. Yet, even for this simple task many different solutions may be conceived, depending on the adopted argument model and on the ultimate goal of the AM system. In general, we have three options:

(1) a binary classifier is trained to distinguish argumentative from non-argumentative sentences, leaving the task of identifying the type of argument component (e.g., a claim or a premise) to a second stage;
(2) a multi-class classifier is trained to discriminate all the argument components that exist in the adopted argument model: this assumes that a sentence can contain at most one argument component;
(3) a set of binary classifiers is trained, one for each existing argument component in the considered model, so that a sentence can be predicted to contain more than one argument component (an alternative would be to adopt a multi-label classifier, where an instance can be assigned to more than just one class).[4]

No matter what option we choose, we need to select a type of classifier, as well as the features to employ. The existing systems have used, up to now, a wide variety of classic machine learning algorithms, including Support Vector Machines (SVM) [Mochales

[4]Bishop [2006] provides a smooth introduction to classification methods.

Palau and Moens 2011; Park and Cardie 2014; Stab and Gurevych 2014b; Eckle-Kohler et al. 2015], Logistic Regression [Levy et al. 2014; Rinott et al. 2015], Naïve Bayes classifiers [Mochales Palau and Moens 2011; Biran and Rambow 2011; Park and Cardie 2014; Eckle-Kohler et al. 2015], Maximum Entropy classifiers [Mochales Palau and Moens 2011], Decision Trees and Random Forests [Stab and Gurevych 2014b; Eckle-Kohler et al. 2015]. These classifiers are trained in a supervised setting, thus on a collection of labeled examples. For each example, some representation of the text to be classified is given (e.g., in the form of a feature vector), together with the associated class (label). The training phase produces a model that can then be used to perform predictions on new (previously unseen) text.

Although several works in the literature have tried to compare some of these approaches, there is no clear evidence to tell which classifier should be preferred. In almost all the existing systems, in fact, most of the effort has been put into conceiving sophisticated and highly informative features, that have a clear and immediate impact on the performance, rather than into constructing appropriate models and algorithms for the considered problem. Thus, the approaches that have been proposed so far typically rely on simple and fast classifiers.

Even for the choice of the features, many of the existing works share several analogies, as they employ classical features for text representation. Among the most extensively used features, a still customary – although somehow naïve – choice is to adopt Bag-of-Words (BoW) representations, where a sentence S is encoded by a vector $v = \{v_1, \ldots, v_{|D|}\}$ of binary values, where D is the considered dictionary of terms, and $v_j = 1$ if word w_j in D appears in S. This model has been widely studied and extended, for example considering the well-known TF-IDF variant, which also accounts for the Term Frequency (TF) of a word in a sentence, and for the Inverse Document Frequency (IDF), which measures the rarity of a word in the vocabulary [Sebastiani 2002]. Moreover, the model can be further extended to consider also bag of bigrams and trigrams, which means to construct additional dictionaries made up by all possible pairs and triplets of terms.

In spite of its popularity, the BoW approach has two important limitations: (1) the order of the words in the sentence is ignored (being locally considered with bigrams and trigrams only) and (2) the semantic similarity between terms is not taken into account, the word *cat* and *feline* being as "distant" as *cat* and *computer*. More advanced features have therefore been developed to address such limitations, for example incorporating knowledge coming from ontologies, thesauri and lexical databases such as WordNet [Levy et al. 2014].

An additional category of features adopted by these classifiers is based upon grammatical information, coming from statistical tools such as constituency and dependency parsers and part-of-speech taggers [Manning and Schütze 2001], thus indicating the grammatical category (noun, verb, adjective, etc.) of each word in a sentence. BoW on part-of-speech tags, including bigrams and trigrams, can be built using the same approach described above for terms. Other frequently employed features include information on punctuation and verb tenses [Mochales Palau and Moens 2011; Stab and Gurevych 2014b] and discourse markers [Eckle-Kohler et al. 2015] whereas even more sophisticated ones can be obtained by external predictors, such as subjectivity scores, sentiment detectors or named entity recognition systems [Levy et al. 2014; Rinott et al. 2015]. Focusing on the genre of online debates, [Biran and Rambow 2011] employs features manually extracted from the RST Treebank [Carlson et al. 2002] in a Naïve Bayes classifier to find premises (which they call justification) supporting a given claim.

Another crucial choice in building classifiers for the detection of argumentative sentences is about whether (and how) to employ contextual information. Many ap-

proaches, in fact, make a strong use of the knowledge of the specific application domain in which they work. This is the case, for example, with the work by Palau and Moens [2011] on legal documents: a precise genre where additional features can be designed by the authors, for example, to mark the presence of specific syntactical descriptors or key-phrases, as they are very frequent in juridical language, and provide reliable cues for the detection of recurrent structural patterns. Also the system developed at IBM Research in Haifa, as a part of the Debater project, is specifically designed to work in a setting where the topic is given in advance, and argument components have to be extracted based on that information [Levy et al. 2014; Rinott et al. 2015]: these tasks go under the name of context-dependent claim detection (CDCD) and context-dependent evidence detection (CDED). Even in such cases, ad-hoc features exploiting the information about the topic are employed. Although in many cases contextual information has proven to be extremely powerful for the implementation of accurate features, it is certainly true that its use somehow limits the generalization capabilities of the AM system. As a matter of fact, domain-specific and highly engineered features are likely to overfit the data they have been constructed on, and also for this reason a crucial step forward of AM systems would be that of being tested across different corpora, genres, and application scenarios.

In an attempt to address these issues, [Lippi and Torroni 2015] have proposed an SVM-based method for context-independent claim detection (CICD), which exploits structured kernels on constituency parse trees (in particular, the Partial Tree Kernel [Moschitti 2006] was employed) to measure similarity between sentences. The constituency parse tree is very often able to capture the rhetorical structure of a sentence, which in many cases is highly indicative of the presence of a claim. In addition, tree kernels automatically construct an implicit feature space, which therefore does not need to be manually defined, and it does not require resorting to genre-specific, context-dependent information. Also [Rooney et al. 2012] exploit structured kernels, even though they only consider kernels between sequences of words and/or parts-of-speech tags.

Although the task of building general-purpose systems represents a major challenge of AM, we cannot deny that in some cases it is necessary to take into account the context in which the system will work, and this might be especially true for the domain of social media. For example, Twitter data has been widely used for opinion mining [Pan and Yang 2010; Pang and Lee 2008; Grosse et al. 2015] and microblogs certainly represent a challenge for AM as well. The inherent nature of microblogging data, which consists of very short messages, usually full of jargonistic expressions, wit, and wordplays, is likely to require specific representations and machine learning methodologies. Even within the area of social media, approaches that might be developed for microblogs will probably differ from techniques dedicated to other genres, such as forums, product reviews, blogs and news.

3.2.2. Argument component boundary detection. The goal of the second stage in the pipeline is the detection of the exact boundaries of each argument component [Stab and Gurevych 2014b], also known as argumentative discourse unit [Peldszus and Stede 2013] or argument unit [Eckle-Kohler et al. 2015]. In this *segmentation* problem one needs to decide where argument components begin and end, for each sentence that has been predicted to be argumentative, since the whole sentence may not exactly correspond to a single argument component [Habernal et al. 2014].

With reference to the example in Figure 1, CLAIM 1, EVIDENCE 2, and CLAIM 3 are portions of a single sentence each, whereas EVIDENCE 4 spans across two sentences. Notice that, in the IBM corpus, from which the example is taken, the sentence contain-

ing EVIDENCE 2 also contains another claim (not shown, for simplicity). With respect to input granularity, three non mutually-exclusive cases need to be considered:

(1) a portion of the sentence (possibly the whole sentence) coincides with an argument component;
(2) two or more argument components can be present within the same sentence;
(3) an argument component can span across multiple sentences.

The majority of the existing methods assumes only one of the above possibilities [Mochales Palau and Moens 2011; Levy et al. 2014].

Clearly, the boundary detection problem strongly depends on the adopted argument model, since different types of argument components can have specific characteristics. For example, Habernal et al. [2014] report about an annotation study of ca. 4,000 sentences using the claim/premise model, where the average claim spans 1.1 sentences and an average premise spans 2.2 sentences. The IBM corpus instead [Aharoni et al. 2014; Levy et al. 2014] (see Section 4), considers claims as short text portions, which are always entirely contained in a single sentence, while premises can span across multiple paragraphs. In that case, a maximum likelihood approach is employed to identify the most probable boundaries of context-dependent claims [Levy et al. 2014]. Rinott et al. [2015] simply consider as evidence candidates all consecutive segments up to three sentences within a paragraph.

Some works ignore the boundary detection problem. Most notably, Mochales Palau and Moens [2011] identify sub-sentences (clauses) obtained from parse trees with argument components, whereas Stab and Gurevych [2014b] and Eckle-Kohler et al. [2015] assume that sentences have already been segmented, and focus on their classification into one of four types: premise, claim, major claim, and none.

More principled approaches to segmentation could be exploited by resorting to relational and structured-output classifiers that can easily formalize the task as a sequence labeling problem [Nguyen and Guo 2007], that of assigning a class (or tag) to each word in a sentence. In this case, the classes could distinguish, for example, words within an argument component from the others. Conditional Random Fields, Hidden Markov Models and other similar methods have been successfully applied to a wide variety of problems of this kind, including for example the recognition of named entities in tweets [Ritter et al. 2011] or information extraction from social media data [Imran et al. 2013]. The main advantage of using this kind of methods relies in the possibility of performing *collective classification* on a set of examples, where instances are not treated (and thus classified) independently from one another, but the sequential order is instead taken into account, and a tag is assigned to each word in the sentence in a single, collective process. The collective classification framework has proven to be extremely powerful not only for sequences, but for any kind of relational and structured data, where relationships and dependencies among the examples have to be taken into account [Getoor 2005]. The works by [Goudas et al. 2014], [Sardianos et al. 2015], and [Park et al. 2015] represent a first attempt in this direction, by exploiting Conditional Random Fields to segment the boundaries of each argument component. In particular, Sardianos et al. [2015] also employ Recursive Neural Networks to build representations of words.

3.3. Argument structure prediction

The final and certainly most complex stage aims at predicting links between arguments, or argument components, depending on the target granularity. This 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. The problem's goal is usually referred to as *prediction*

rather than *detection*, since the target is not just a specific portion, but rather a connection (or link) between portions, of the input document. The output of this stage is a graph connecting the retrieved arguments (or parts thereof). Edges in the graph may represent different relations, such as entailment, support or conflict. With regard to the context of social media and Web documents, such a graph would be an invaluable instrument for a multitude of applications: analyzing the dynamics of social debates on the Web and the spread of influence among social networks, evaluating the acceptance of different arguments and the role of specific users within this kind of process, detecting anomalous behaviors, and even studying the most effective rhetorical strategies. The impact on social sciences, computational linguistics, formal argumentation and discourse analysis would be dramatic.

Even more than in the preceding stages of the pipeline, the formalization of the structure prediction problem is clearly influenced by the underlying argument model and by the granularity of the target. When dealing with a simple claim/premise model, in fact, the problem of predicting connections between the conclusion of an argument (the claim) and its supporting premises is formalized in a straightforward way as a link prediction task in a bipartite graph, where nodes are partitioned into two categories (the type of argument component they represent). Figure 1(b) illustrates how the AM system produces a score for each possible claim/evidence pair, representing how confident it is that there is a link between the two components. Figure 1(c) instead shows how attack relations are inferred between the detected arguments.

Clearly, more complex argument models induce a more sophisticated structure prediction task: when adopting the Toulmin model, all the components of the argument (e.g., warrant, rebuttal, backing, qualifier, etc.) have to be identified to correctly retrieve the final structure. Notice that arguments are often only partially spelled out in the text, as it usually happens, e.g., with the Toulmin model, where even the claim is sometimes left implicit [Newman and Marshall 1991]. Clearly, those argument components can not be retrieved by the first two stages of the AM pipeline. They should instead be inferred from the context. That would be yet another endeavor for the whole system, requiring a dedicated argument completion component, for the development of which no attempt has currently been made.

Current approaches to the structure prediction task 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 claim. This in turn enables using information about the claim to predict the evidence. In this case, the support links are thus obtained *by definition* when predicting the evidence. Palau and Moens [2011] have addressed the problem by parsing with a manually-built context-free grammar to predict relations between argument components. The grammar rules follow the typical rhetorical and structural patterns of sentences in juridical texts. This is a highly genre-specific approach and its direct use in other genres would be unlikely to yield accurate results. Work by Stab and Gurevych [2014b] instead proposes a classic machine learning solution, by employing a binary SVM classifier to predict links in a claim/premise model based on work by Freeman [1991]. Biran and Rambow [2011] apply the same technique mentioned earlier to premise detection also for the prediction of links between premises and claims. Finally, another important research direction adopts Textual Entailment [Cabrio and Villata 2012a], with the goal of inferring whether a support or attack relation between two given arguments holds.

Table I summarizes the methods used within the existing AM systems, highlighting which approaches have been employed at each stage. These are either techniques

Table I. A comparison of methods that have been applied within AM systems. We distinguish the three steps in the pipeline described in Section 3 (sentence classification, component boundary detection and structure prediction) and we list the algorithms that have been applied in at least one of the existing systems. The acronyms stand for: Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Maximum Entropy models (ME), Decision Trees (DT), Random Forests (RF), Recurrent Neural Networks for language models (RNN), Conditional Random Fields (CRF), Maximum Likelihood (ML), Textual Entailment Suites (TES) – which comprise suites of entailment detection algorithms [Padò et al. 2013] – and Parsing (P) using a context-free grammar.

	SC							BD		SP			
System	SVM	LR	NB	ME	DT	RF	RNN	CRF	ML	TES	P	SVM	NB
[Eckle-Kohler et al. 2015]	X		X			X							
[Lippi and Torroni 2015]	X												
[Rinott et al. 2015]		X											
[Sardianos et al. 2015]	X						X	X					
[Boltuzic and Snajder 2014]										X		X	
[Goudas et al. 2014]	X							X					
[Levy et al. 2014]		X							X				
[Stab and Gurevych 2014b]	X		X		X	X						X	
[Cabrio and Villata 2012a]										X			
[Rooney et al. 2012]	X												
[Biran and Rambow 2011]			X								X		X
[Mochales Palau and Moens 2011]	X		X	X							X		

Table II. Correspondences between AM and machine learning and natural language processing (ML-NLP) tasks.

AM	ML-NLP
Argumentative sentence detection	Sentence classification Hedge cue detection Sentiment analysis Question classification Subjectivity prediction
Argument component boundary detection	Sequence labeling Named entity recognition Text segmentation
Argument structure prediction	Link prediction Discourse relation classification Semantic textual similarity

from machine learning (SVM, LR, NB, ME, CRF, ML, DT, RF) or from computational linguistics (TES, P).

Table II instead highlights the similarities between AM sub-tasks and problems typical of machine learning and natural language processing (NLP). Argumentative sentence detection is fundamentally a sentence classification task [Kim 2014]. It thus shares analogies with NLP tasks such as hedge cue detection (determining whether sentences contain unreliable information) [Verbeke et al. 2012], subjectivity prediction [Esuli and Sebastiani 2006], question classification [Zhang and Lee 2003] and sentiment analysis [Pang and Lee 2008]. The problem of argument component boundary detection is instead a sequence labeling problem, and thus it has connections with tasks such as named entity recognition [Nadeau and Sekine 2007], and other text segmentation applications [Choi et al. 2001]. Finally, argument structure prediction is similar to link prediction tasks [Getoor and Diehl 2005], but several analogies can be drawn with relation classification in discourse analysis [Lin et al. 2009], semantic textual similarity estimation [Achananuparp et al. 2008], and different applications of textual entailment [Padò et al. 2013].

Table III. English language corpora for which there has been documented use by AM systems (top) or related applications (bottom). For each corpus, we indicate the domain and document type, the overall size, whether it contains also non-argumentative sentences (NA) and whether, at the time of writing, they are publicly available or available upon request (AV).

Reference	Domain	Document type	Size	NA	AV
[Rinott et al. 2015]	Various	Wikipedia pages	∼80,000 sent.	X	X
[Aharoni et al. 2014]	Various	Wikipedia pages	∼50,000 sent.	X	X
[Boltuzic and Snajder 2014]	Social themes	User comments	∼300 sent.		X
[Cabrio and Villata 2014]	Various	Debatepedia, etc.	∼1,000 sent.		X
[Habernal et al. 2014]	Various	Web documents	∼3,996 sent.	X	X
[Stab and Gurevych 2014a]	Various	Persuasive essays	∼1,600 sent.	X	X
[Biran and Rambow 2011]	Various	Blog threads	∼7,000 sent.	X	X
[Mochales Palau and Moens 2011]	Law	Legal Texts	∼2,000 sent.		
[Houngbo and Mercer 2014]	Biomedicine	PubMed articles	∼10,000 sent.	X	X
[Park and Cardie 2014]	Rulemaking	User comments	∼9,000 sent.		X
[Peldszus 2014]	Various	Microtexts	∼500 sent.		X
[Ashley and Walker 2013]	Law	Juridical cases	35 doc.	X	
[Rosenthal and McKeown 2012]	Various	Blogs, forums	∼4,000 sent.	X	X
[Bal and Saint-Dizier 2010]	Various	Newspapers	∼500 doc.		

4. CORPORA AND APPLICATIONS

Any attempt at AM by way of machine learning and artificial intelligence techniques clearly requires a collection of annotated documents (corpus), to be used as a training set for any kind of predictor. Constructing annotated corpora is, in general, a complex and time-consuming task, which requires to commit costly resources such as teams of experts, so that homogeneous and consistent annotations can be obtained. This is particularly true for the domain at hand, as the identification of argument components, their exact boundaries, and how they relate to each other can be quite complicated (and controversial!) even for humans (see [Mochales Palau and Ieven 2009; Habernal et al. 2014]). Moreover, different datasets have often been built with specific objectives in mind or for some particular genre, and therefore they could hardly suit to all approaches, or to all the stages in the pipeline.

Table III lists the existing corpora that have been used, up to now, in applications related to AM (we consider English corpora only). These corpora have also been analyzed by [Habernal et al. 2014] with a focus on annotation procedures, but with no emphasis on AM techniques. In this section, we provide a description of the main characteristics of these corpora, organized according to genre or application domain, together with an analysis of the systems that used them.

Corpora for structure prediction. Several annotated corpora have been constructed for the sole purpose of analyzing relations between arguments or argument components, depending on the target granularity. These corpora generally only have argumentative content, which makes them unsuitable for more general AM tasks. Some important collections of this type are maintained by the University of Dundee[5]. They aggregate many datasets with annotated argument maps, in a variety of standardized formats. These corpora include, for example, the well-known AraucariaDB[6] and several annotated transcripts of excerpts of episodes from the Moral Maze BBC radio program. Due to the goal they were built for, these corpora very often lack the non-argumentative

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

[6]The AraucariaDB has undergone several changes along subsequent versions throughout the years. In particular, a previous version was used in [Mochales Palau and Moens 2011] and [Rooney et al. 2012] to perform argument mining, and in that version also the original text was available, thus allowing also to distinguish between argumentative and non-argumentative sentences. In the current version, the original corpus is not available, and in addition some text has been modified during the tagging process, in order to better model argument maps.

parts which are necessary as negative examples for the training of some kind of discriminative machine learning classifier: basically, they assume that the sentence detection step described in Section 3.2.1 has already been made, and only argumentative sentences are available. The NoDE benchmark data base [Cabrio and Villata 2014] goes in the same direction but it has a different target granularity. It contains arguments obtained from a variety of sources, including Debatepedia[7] and ProCon[8]. It does not include non-argumentative examples. These datasets are used for tasks such as the detection of inter-argument support and attack relations, which is an instance of the structure prediction task defined in Section 3.3, with a target granularity at the argument level. The work by [Boltuzic and Snajder 2014] is in a similar vein: they consider a small corpus of documents consisting of user comments on two controversial arguments, and they develop a system for the classification of the relation between each comment and its associated argument into five different classes (strong attack, attack, strong support, support, none). Therefore, this approach can also be seen as an instance of the structure prediction task. Similarly, the German corpus presented in [Kirschner et al. 2015] is a collection of scientific publications annotated with argumentation structures (supports, attacks, details, sequence).

Legal domain. Law has been the pioneering application domain for AM, and certainly among the most successful ones, with the work by Mochales Palau and Moens [2011] on the European Court of Human Rights (ECHR) [Mochales Palau and Ieven 2009] and the AraucariaDB datasets for the extraction of claims and their supporting premises from a collection of structured legal documents. This study represents, until now, one of the few systems aiming to implement a somehow complete AM pipeline, although highly specialized on a single genre. Yet, the employed datasets are not publicly available, thus making it difficult to compare against the proposed approach. More recently, also the Vaccine/Injury Project (V/IP) [Ashley and Walker 2013] was carried out, with the goal of extracting arguments from a set of judicial decisions involving vaccine regulations. Unfortunately, also this annotated corpus is not publicly available (yet).

Biology and medicine. The development of annotated datasets from biology and medicine texts is a new trend which is attracting growing attention. It could be an extremely important step towards building ontologies and knowledge bases describing the links between either symptoms and diseases, or between genes and diseases, or even to assist personalized medicine prescriptions. In particular, [Houngbo and Mercer 2014] have proposed a system for the classification of the rhetorical category of sentences in biomedical texts, by distinguishing between four categories: introduction, method, results and discussion. Green [2014] offers a qualitative description of an ongoing process of corpus creation in this domain.

Humanities. Rhetorical, philosophical and persuasive essays constitute another interesting field for AM. A study on the integration of manual and automatic annotations on a collection of 19th century philosophical essays was proposed in [Lawrence et al. 2014]. A limited-scope but well-documented dataset was proposed by Stab and Gurevych [2014a] as a collection of 90 persuasive essays. The topics covered are very heterogeneous. Annotations with intra-sentence granularity include premises, claims and one major claim per essay. Due to the nature of the data, and to the annotation guidelines, only a few sentences in the corpus are non-argumentative. Being specifically designed for the analysis of persuasive essays, this corpus would likely not be the most appropriate choice for a training set if the goal were to generalize to other genres.

[7] http://www.debatepedia.com
[8] http://www.procon.org

In fact, for example, the "major claim" class is highly domain-specific, being often detected by employing dedicated features, such as the position of the sentence within the essay. As a matter of fact, the work described in [Stab and Gurevych 2014b] employs specific features which take advantage of some background knowledge of the application scenario. Nevertheless, it represents an interesting approach, as it implements, besides the sentence classification step, also a structure prediction task targeting attack/support relations between argument components, using an SVM classifier (see Section 3.3).

User-generated content. The social and semantic Web offers a multiplicity of different document sources with uncountable arguments. Although only a few results have yet been reported in this domain, possibly due to the heterogeneity of contents and diversity of jargon, nonetheless this trend paves the way to a variety of new interesting application domains. In this sense, AM might become the key enabling technology to make new knowledge emerge from an ocean of disorganized and unstructured content.

Indeed, the largest AM dataset to date is currently being developed at IBM Research [Aharoni et al. 2014; Rinott et al. 2015], starting from plain text in Wikipedia pages. The purpose of this corpus is to collect context-dependent claims and premises (named evidence), which are relevant to a given topic. A first version of this dataset [Aharoni et al. 2014] covered 33 topics, for a total of 315 Wikipedia articles, with evidence annotated only on 12 topics. Such dataset is large but also very unbalanced, as it contains about 2,000 argument components (claims or evidence) over about 50,000 sentences, therefore representing an extremely challenging benchmark. An approach to context-dependent claim detection on this corpus was proposed in [Levy et al. 2014], while a context-independent approach was applied in [Lippi and Torroni 2015] to the same dataset. The first system, proposed by the IBM Haifa research group, addresses both the task of sentence classification and that of boundaries detection, whereas the latter approach only considers the task of detecting sentences containing claims. Being context-independent, the work proposed by [Lippi and Torroni 2015] has been tested on more than a single corpus, reporting interesting results also on the persuasive essays dataset by Stab and Gurevych [2014a]. The work by [Rooney et al. 2012] also employs kernel methods for argumentative sentence classification, even though only considering a kernel between word and part-of-speech tag sequences, rather than constituency parse trees as in [Lippi and Torroni 2015]. A second version of the dataset was presented by Rinott et al. [2015], including 2,294 claim labels and 4,690 evidence labels collected from 547 articles, organized into 58 different topics.

There are other corpora based on user-generated content too. The work described in [Goudas et al. 2014] attempts to address several steps in the AM pipeline, including sentence classification and boundaries detection, while [Sardianos et al. 2015] focus only on the latter. The corpora used in these two works, which are in Greek, are not available. The dataset in Japanese introduced in [Reisert et al. 2014] considers premises collected in microblogs, while a collection of microtexts (originally in German, but professionally translated into English) was used in [Peldszus 2014]. The corpus presented in [Eckle-Kohler et al. 2015] consists of German news annotated with the claim/premise model.

Another well-annotated corpus was developed by [Habernal et al. 2014], to model arguments following a variant of the Toulmin model. This dataset includes 990 English comments to articles and forum posts, 524 of which are labeled as argumentative. A final smaller corpus of 345 examples is annotated with finer-grained tags. No experimental results were reported on this corpus.

In the context of the online debates genre, Biran and Rambow [Biran and Rambow 2011] have annotated a corpus of 309 blog threads collected from LiveJournal, by la-

beling claims, premises (which they call justifications) and the links between them. The corpus was employed in their experiments.

Additional datasets were recently collected from online resources, including online reviews, blogs, and newspapers. The purpose of these datasets is slightly different from, but certainly related to, AM. In particular, we mention a small collection of user reviews, in French, presented in [Saint-Dizier 2012; Villalba and Saint-Dizier 2012], and two datasets of 2,000 sentences each, developed by Rosenthal and McKeown [2012], with the purpose of extracting so-called *opinionated claims*. These consist in 285 LiveJournal blogposts and 51 Wikipedia discussion forums, respectively.

Other AM-related tasks. Table IV displays the existing systems related to AM and the task they were built for. On the bottom half of the table we group items that do not exactly fit with the classic AM pipeline, although they have a clear relation with AM. These refer to the following tasks:

— argumentative opinion analysis, with the ⟨TextCoop⟩ platform [Saint-Dizier 2012; Villalba and Saint-Dizier 2012], which basically constructs arguments from opinions and supportive elements such as illustrations and evaluative expressions, by using a set of handcrafted rules that explicitly describe rhetorical structures;
— opinionated claim mining [Rosenthal and McKeown 2012], a problem that is closer to sentiment analysis, where the aim is to detect assertions containing some belief, of whose truth a user is attempting to convince an audience;
— the classification of the rhetorical category of sentences [Houngbo and Mercer 2014], specifically whether a sentence within a scientific document is part of an introductory, experimental, methodological or conclusive section;
— dialogical argument mining [Budzynska et al. 2014] with the ⟨TextCoop⟩ platform, and a theoretical underpinning in inference anchoring theory [Budzynska and Reed 2011];
— predicting whether and to what extent premises can be verified [Park and Cardie 2014; Park et al. 2015];
— argument scheme classification [Feng and Hirst 2011], that is, whether an argument is proposed by an example, from consequences, from cause to effect, etc.;
— the classification of argument graph structures in microtexts [Peldszus 2014];
— automatically essay scoring [Ong et al. 2014] using argumentation structures extracted with an ontology.

A web page is maintained at http://argumentationmining.disi.unibo.it/ with a list of pointers to the corpora discussed in this survey.

Table V summarizes and compares all the known systems that address at least one of the tasks in the AM pipeline described in Section 3. Interestingly, they all adopt a basic *claim/premise* argument model. For all the approaches, we report the tasks they address, whether they are context-independent, and whether they have been tested on multiple corpora or genres. At present, no known AM system covers all the stages of the AM pipeline, although all those summarized here certainly present interesting ideas. The aim to build more general, context-independent systems, and the need to deploy ways for evaluating novel approaches on a variety of genres, as opposed to one single genre, are two major challenges for the whole research field.

Finally, we report on the performance achieved by the existing systems in their respective tasks, with the caveat that measurements obtained on different corpora, different tasks, or different experimental setups are clearly not comparable. Mochales Palau and Moens [2011] report a 73% and 80% accuracy on the argumentative sentence classification task in a previous version of the AraucariaDB and on the ECHR corpus, respectively. With respect to the classification of argument components, they

Table IV. A complete list of systems that address at least one task of the AM pipeline (top) or that address AM-related tasks (bottom). For each system, we report the task(s) they address.

System	Task
[Eckle-Kohler et al. 2015]	Claim/premise mining
[Lippi and Torroni 2015]	Claim mining
[Rinott et al. 2015]	Evidence mining
[Sardianos et al. 2015]	Boundary detection
[Boltuzic and Snajder 2014]	Structure prediction for arguments
[Goudas et al. 2014]	Claim/premise mining
[Levy et al. 2014]	Claim mining
[Stab and Gurevych 2014b]	Claim/premise mining and structure prediction
[Cabrio and Villata 2012a]	Textual entailment for arguments
[Rooney et al. 2012]	Claim/premise mining
[Biran and Rambow 2011]	Evidence mining and structure prediction
[Mochales Palau and Moens 2011]	Claim/premise mining and structure prediction
[Park et al. 2015]	Premise verifiability categorization
[Budzynska et al. 2014]	Dialogical argument mining
[Park and Cardie 2014]	Premise verifiability categorization
[Houngbo and Mercer 2014]	Rhetorical category sentence classification
[Ong et al. 2014]	Automated essay scoring
[Peldszus 2014]	Argument graph structure classification
[Rosenthal and McKeown 2012]	Opinionated claim mining
[Villalba and Saint-Dizier 2012]	Argumentative opinion analysis
[Feng and Hirst 2011]	Argument scheme classification

Table V. A comparison of all the existing systems which implement at least one of the stages in the classic AM pipeline. We indicate whether it performs Component Detection (CD), its sub-tasks Sentence Classification (SC) and Boundaries Detection (BD), Structure Prediction (SP), whether it is context-independent (CI), and whether it has been tested on multiple corpora (MC) and multiple domains (MD).

System	CD	SC	BD	SP	CI	MC	MD
[Eckle-Kohler et al. 2015]	X				X		
[Lippi and Torroni 2015]	X	X			X	X	X
[Rinott et al. 2015]	X	X					
[Sardianos et al. 2015]	X		X		X		
[Boltuzic and Snajder 2014]				X	X		X
[Goudas et al. 2014]	X	X	X		X		
[Levy et al. 2014]	X	X	X				
[Stab and Gurevych 2014b]	X	X		X			
[Cabrio and Villata 2012a]				X	X		X
[Rooney et al. 2012]	X	X			X		
[Biran and Rambow 2011]	X	X		X	X		
[Mochales Palau and Moens 2011]	X	X		X		X	

achieve a precision/recall of 77%/61% for conclusions, and 70%/66% for premises on ECHR. In [Rooney et al. 2012], the argumentative sentence classification task on the same version of the AraucariaDB achieves a 65% accuracy. Biran and Rambow [2011] report a 34% F_1 on the task of premise classification on their corpus of LiveJournal blog threads. Levy et al. [2014] report a 17% F_1 on claim mining on the IBM corpus. A similar figure (18%) is reported by Lippi and Torroni [2015], again about claim mining, and again on the IBM corpus, though on a slightly different version. However, the two experimental setups are different, and also the two tasks are different, because the second approach does not use topic information. [Lippi and Torroni 2015] also reports an 71% F_1 on the persuasive essays corpus [Stab and Gurevych 2014a]. On the same corpus, Stab and Gurevych [2014b] achieve an F_1 equal to 63%, 54% and 83%, respectively, for the detection of major claims, claims and premises (thus on a different multi-class problem). The same article reports a 52% F_1 for the classification of support relations. In [Goudas et al. 2014], a 77% F_1 on argumentative sentence classi-

fication is shown on a Greek corpus of data from social media, with a 42% F_1 for the task of detecting boundaries. In [Sardianos et al. 2015], an F_1 between 15% and 20% is reported for boundaries detection in different settings, on a different Greek corpus. For structure prediction, Cabrio and Villata [2012a] report a 67% accuracy using textual entailment, whereas Boltuzic and Snajder [2014] report an F_1 ranging from 70% to 81% on different problem formulations and on a completely different corpus. Biran and Rambow [2011] report a 42% F_1 for claim-premise link prediction.

5. CHALLENGES AND PERSPECTIVES

Argumentation mining certainly shares some analogies with opinion mining and sentiment analysis. Yet, as highlighted also in [Habernal et al. 2014], while the goal of opinion mining is to understand *what people think about something*, the aim of argumentation mining is to understand *why*, which implies looking for causes and reasons rather than just for opinions and sentiment. The distinguishing element of an argument stays in its inherent structure, that is in the specification of its premises, its claim, and the inferential process that connects the former to the latter. In this sense, the great ambition of argumentation mining is that of moving from opinion mining and sentiment analysis to the next level: the analysis of those reasoning processes that bring humans to rationally accept or reject an opinion, argument, or theory (although we are well aware that influencing a real audience is not simply a matter of presenting a set of rational, deductive arguments, see [Crosswhite et al. 2004]).

Within this context, there are some aspects of the AM problem which, up to now, have been only marginally considered but which could provide a crucial contribution to the development of the research field. We will present them in the remainder of this section.

5.1. Dealing with big data

Many disciplines are nowadays attracted by the so-called *big data challenge*, that is, exploiting the exceptional amount of information and knowledge now available through the Internet, for the most diverse and heterogeneous tasks. Clearly, this is a huge opportunity also for argumentation mining. A multitude of information sources from the Social and Semantic Web can provide argumentative statements with different characteristics, coming from social network posts, forums, blogs, product reviews, comments to newspapers articles. Clearly, dealing with very large data collections raises the issue of the scalability of argumentation mining systems, which in many cases have been tested on small corpora only. Moreover, the Web could help solving another key issue in argumentation mining, namely, the limited availability of annotated corpora. In fact, an interesting line of research would be exploiting crowdsourcing assessment to annotate very large corpora. In several contexts other than argumentation mining, such as image classification [Nowak and Rüger 2010] and object detection [Deng et al. 2013], the extraordinary computational power of the crowd has been used to construct labeled databases on the large scale. One of the biggest existing datasets on sentiment analysis has also been constructed using a crowdsourcing mechanism [Socher et al. 2013]. Many studies have been and still are conducted on how to effectively leverage the crowd, and consistency problems will necessarily have to be addressed. The potential behind this approach is certainly enormous, as motivated by the increasing number of existing crowdsourcing platforms, such as Amazon Mechanical Turk[9]. As for argumentation mining, one additional challenge is given by the subtlety of the task asked to users, which has a very negative impact on inter-annotation agreement and is the reason why initial attempts in this direction have

[9]https://www.mturk.com/mturk/welcome

been unsuccessful [Habernal et al. 2014]. In this sense, visual tasks such as image tagging are certainly more intuitive and easy to define, whereas detecting the boundaries of an argument component might be a tricky assignment. The challenge here is to identify meaningful annotation tasks that can result in an acceptable inter-annotation agreement between non-expert annotators.

5.2. Dealing with unsupervised data

When dealing with very large datasets, a crucial issue is that of employing fast and efficient machine learning algorithms. Due to the difficulties and costs in creating large and complete annotated corpora, a seemingly unavoidable alternative to crowdsourcing is to employ machine learning technologies capable of dealing also with unsupervised or semi-supervised data. Deep learning techniques, which have recently obtained breakthrough results in many AI domains, including natural language processing, certainly represent one of the most appealing choices in this direction. The ability to deal with large and unsupervised corpora is indeed one of the most crucial aspects of deep networks, and among the chief reasons of their impressive success. Unsupervised learning is in fact employed in many deep architectures as a pre-training phase whose aim is to extract hierarchical sets of features directly from the data, while the interaction with supervisions (the symbolic layer) only applies at a further stage. In very recent years, a huge amount of deep learning research has focused on language tasks, producing a wide variety of sophisticated systems dedicated to specific applications (e.g., see [LeCun et al. 2015] and references therein). Many of these systems are capable of capturing advanced semantic information from unstructured text, and within this context, the Web is an invaluable mine of multi-faceted and heterogeneous information. A very successful approach in this direction is given by the so-called *word embeddings*, sometimes also called simply *word vectors*, that basically consist in automatically learned feature spaces encoding high-level, rich linguistic similarity between terms [Mikolov et al. 2013]. Recurrent Neural Tensor Networks [Socher et al. 2013] and Tree-Structured Long Short-Term Memory Networks [Tai et al. 2015] and also a language-specific version of Convolutional Neural Networks [Kim 2014], are examples of deep architectures capable of dealing with input sequences of arbitrary length, which have recently achieved state-of-the-art results in language-specific applications, including sentiment analysis. Therefore, they could be naturally adapted to perform some of the steps in the AM pipeline, such as sentence classification. [Sardianos et al. 2015] presents a first approach to exploiting word vectors (for the Greek language) within an argumentation mining system.

5.3. Dealing with structured and relational data

Another important shortcoming of most of the existing argumentation mining approaches is a proper handling of structured and relational data (see also [Moens 2014]). In the last decade, machine learning has run across a so-called *relational revolution* [Getoor 2005], in order to extend methodologies and algorithms to deal with this highly informative kind of data, such as trees, graphs, or sequences. Structured-output approaches such as conditional random fields or structured support vector machines, for example, can perform collective classification, which means that predictions on new (previously unseen) examples can be produced collectively, by taking into account the inherent structure of data, such as sequentiality information or networking relationships. This is the case, for example, of paragraphs within a textual document, consecutive utterances in a dialogue, or links within a social network. As an attempt in this direction, conditional random fields were applied in [Goudas et al. 2014] and [Sardianos et al. 2015] to address the problem of argument component segmentation. The problem of predicting relations between premises and claims, or between different arguments,

can instead be easily modeled as a task of link prediction within a graph, where nodes represent arguments or argument components. Here another major contribution will likely come from statistical relational learning, a research field which aims at combining first-order logic with statistical machine learning (i.e., symbolic and sub-symbolic representations). Whereas statistical machine learning and graphical models can naturally deal with uncertainty in data, on the other hand the expressive power of first-order logic can be exploited to model background knowledge of a given domain, and to describe relations between data instances. This kind of approach has been successfully applied to a variety of tasks that show several similarities with AM. For example, link discovery in social and biological networks [Getoor and Diehl 2005] is basically a task of link prediction, thus resembling structure prediction in a graph of arguments; information extraction and entity resolution in textual corpora [Culotta et al. 2006; Poon and Domingos 2007] share analogies with argumentative sentence classification, while sequence tagging and sentence parsing [Poon and Domingos 2009] could offer interesting perspectives on the modeling of structured textual data.

6. CONCLUSION

Looking to the future, the development of powerful argumentation mining tools paves the way to lots of new, exciting applications across many disciplines, in social sciences and humanities, as well as life sciences and engineering. Decision making and policy learning [Milano et al. 2014], for example, could employ automatically extracted arguments in order to improve models and support strategic choices. Engineering workflow processes have already exploited argumentation models for the automated evaluation of alternative design solutions [Baroni et al. 2015], and argumentation mining could be an additional asset in the process.

Market analysis and customer profiling could greatly benefit from the analysis of arguments provided by users over the Web. In this scenario, an interesting perspective is also that of analyzing social behavior and social interaction, as well as the dialectics and rhetoric of the proposed arguments. Here, argumentation mining might unlock innovative ways of organizing, supporting and visualizing online debates, for example by clustering posts, and proposing new rankings based on, say, the argumentative force and mutual position of agreement of the parties involved and their contributed posts. There have already been several attempts in this direction by parts of the computational argumentation community [Modgil et al. 2013; Gabbriellini and Torroni 2015], but lacking natural argument analysis tools, they mostly had to rely on the collaborative attitude of expert users. Studies on opinion diffusion [Guille et al. 2013] could also certainly benefit from the development of AM systems: in that case, the standard AM pipeline should be integrated with tools coming from social network analysis [Scott 2012] in order to exploit network structure, communication patterns, and information about users and their connections. In a similar context, another recent trend in AI that could be exploited by AM is that of recommendation systems, where there have been several attempts to combine sentiment analysis instruments with collaborative filtering [Leung et al. 2006; Kawamae 2011]. On a slightly different perspective, a theoretical analysis of the potential application of argumentation techniques to sentiment analysis for economic and business documents is given in [Hogenboom et al. 2010].

Argumentation mining could also provide a crucial incentive to the development of reasoning engines over arguments originated from the Web, and act as an enabling technology for the Argument Web vision [Bex et al. 2013], that of a URI-addressable structure of linked argument data making it possible to follow a line of argument across disparate sites and multimedia resources.

Intelligence analysis is another interesting domain where argumentation mining systems could provide useful support, especially in the context of human-machine com-

munication and multi-agent systems [Lindahl et al. 2007]. Narrative understanding and computer storytelling represent additional challenging application scenarios for the research field, with specific reference to the construction of services oriented to educational purposes.

Finally, building machines capable of human-level reasoning has long been one of the grand challenges of AI, one which unfortunately clashes with the cost of building and maintaining structured knowledge in open domains from unstructured data. We witnessed how statistical inference methods in DeepQA systems such as IBM Watson have already reached such a level of sophistication to compete and win against champion players at Jeopardy! [Fan et al. 2012]. Argumentation mining could be the next breakthrough towards this vision. It could be the key for a new generation of AI systems able to combine statistical and logical inference frameworks together: able, that is, to extract arguments from human-made, unstructured, free text, and then to reason from them logically and so produce new arguments, and thus new knowledge.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for insightful and constructive comments, and Noam Slonim, Andreas Peldszus, Chris Reed, Iryna Gurevych, Christian Stab, Serena Villata, Kevin Ashley, Patrick Saint-Dizier, Joonsuk (Jon) Park, Hospice Houngbo, Georgios Petasis and Ruty Rinott for offering material and clarifications regarding the argumentation mining corpora surveyed here.

REFERENCES

Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. 2008. The evaluation of sentence similarity measures. In *Data warehousing and knowledge discovery*. Springer, 305–316.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 64–68. http://acl2014.org/acl2014/W14-21/pdf/W14-2109.pdf

Kevin D. Ashley and Vern R. Walker. 2013. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *ICAIL 2012, Rome, Italy*, Enrico Francesconi and Bart Verheij (Eds.). ACM, 176–180. DOI:http://dx.doi.org/10.1145/2514601.2514622

Bal Krishna Bal and Patrick Saint-Dizier. 2010. Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association.

Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6, 1 (2015), 24–49. DOI:http://dx.doi.org/10.1080/19462166.2014.1001791

Trevor J. M. Bench-Capon and Paul E. Dunne. 2007. Argumentation in artificial intelligence. *Artificial Intelligence* 171, 10-15 (2007), 619–641. DOI:http://dx.doi.org/10.1016/j.artint.2007.05.001

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (2010), 211–259. DOI:http://dx.doi.org/10.1007/s10462-010-9154-1

Philippe Besnard, Alejandro Javier García, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Ricardo Simari, and Francesca Toni. 2014. Introduction to structured argumentation. *Argument & Computation* 5, 1 (2014), 1–4. DOI:http://dx.doi.org/10.1080/19462166.2013.869764

Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the Argument Web. *Commun. ACM* 56, 10 (Oct. 2013), 66–73. DOI:http://dx.doi.org/10.1145/2500891

Or Biran and Owen Rambow. 2011. Identifying Justifications in Written Dialogs by Classifying Text as Argumentative. *Int. J. Semantic Computing* 5, 4 (2011), 363–381. DOI:http://dx.doi.org/10.1142/S1793351X11001328

Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

Filip Boltuzic and Jan Snajder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 49–58.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards Argument Mining from Dialogue. In *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014 (Frontiers in Artificial Intelligence and Applications)*, Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti (Eds.), Vol. 266. IOS Press, 185–196. DOI:http://dx.doi.org/10.3233/978-1-61499-436-7-185

K. Budzynska and C. Reed. 2011. *Whence inference?* Technical Report. University of Dundee.

Elena Cabrio and Serena Villata. 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)*. Association for Computational Linguistics, Jeju, Korea, 208–212.

Elena Cabrio and Serena Villata. 2012b. Natural Language Arguments: A Combined Approach. In *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*, Luc De Raedt, Christian Bessière, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas (Eds.), Vol. 242. IOS Press, 205–210. DOI:http://dx.doi.org/10.3233/978-1-61499-098-7-205

Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation* 4, 3 (2013), 209–230. DOI:http://dx.doi.org/10.1080/19462166.2013.862303

Elena Cabrio and Serena Villata. 2014. NoDE: A Benchmark of Natural Language Arguments. In *COMMA 2014 (Frontiers in Artificial Intelligence and Applications)*, Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti (Eds.), Vol. 266. IOS Press, 449–450.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. Technical Report LDC2002T07. Linguistic Data Consortium, Philadelphia. Web Download.

Freddy YY Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent semantic analysis for text segmentation. In *In Proceedings of EMNLP*. Citeseer.

Jim Crosswhite, John Fox, Chris Reed, Theodore Scaltsas, and Simone Stumpf. 2004. Computational Models of Rhetorical Argument. In *Argumentation Machines (Argumentation Library)*, Chris Reed and Timothy J. Norman (Eds.), Vol. 9. Springer Netherlands, 175–209. DOI:http://dx.doi.org/10.1007/978-94-017-0431-1_6

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 296–303.

Jia Deng, Jonathan Krause, and Fei-Fei Li. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE, 580–587. DOI:http://dx.doi.org/10.1109/CVPR.2013.81

Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–358. DOI:http://dx.doi.org/10.1016/0004-3702(94)00041-X

David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets. Reasoning About a Highly Connected World*. Cambridge University Press. http://www.cs.cornell.edu/home/kleinber/networks-book/

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 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)*. Association for Computational Linguistics, Lisbon, Portugal, to appear.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining.. In *EACL*, Vol. 6. 2006.

J. Fan, A. Kalyanpur, D.C. Gondek, and D.A. Ferrucci. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development* 56, 3.4 (May 2012), 5:1–5:10. DOI:http://dx.doi.org/10.1147/JRD.2012.2186519

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

James B Freeman. 1991. *Dialectics and the macrostructure of arguments: A theory of argument structure*. Vol. 10. Walter de Gruyter.

Simone Gabbriellini and Paolo Torroni. 2014. A New Framework for ABMs Based on Argumentative Reasoning. In *Advances in Social Simulation - Proceedings of the 9th Conference of the European Social Simulation Association, ESSA 2013, Warsaw, Poland, September 16-20, 2013 (Advances in Intelligent Systems and Computing)*, Bogumil Kaminski and Grzegorz Koloch (Eds.), Vol. 229. Springer, 25–36. DOI:http://dx.doi.org/10.1007/978-3-642-39829-2_3

Simone Gabbriellini and Paolo Torroni. 2015. Microdebates: structuring debates without a structuring tool. *AI Communications* (2015).

Lise Getoor. 2005. Tutorial on Statistical Relational Learning. In *ILP (Lecture Notes in Computer Science)*, Stefan Kramer and Bernhard Pfahringer (Eds.), Vol. 3625. Springer, 415. DOI:http://dx.doi.org/10.1007/11536314_26

Lise Getoor and Christopher P Diehl. 2005. Link mining: a survey. *ACM SIGKDD Explorations Newsletter* 7, 2 (2005), 3–12.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument Extraction from News, Blogs, and Social Media. In *Artificial Intelligence: Methods and Applications*, Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles (Eds.). LNCS, Vol. 8445. Springer International Publishing, 287–299. DOI:http://dx.doi.org/10.1007/978-3-319-07064-3_23

Nancy Green. 2014. Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 11–18. http://acl2014.org/acl2014/W14-21/pdf/W14-2102.pdf

Kathrin Grosse, María P. González, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. 2015. Integrating argumentation and sentiment analysis for mining opinions from Twitter. *AI Communications* 28, 3 (2015), 387–401. DOI:http://dx.doi.org/10.3233/AIC-140627

Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM SIGMOD Record* 42, 2 (2013), 17–28.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing. Forlì-Cesena, Italy, July 21-25, 2014 (CEUR Workshop Proceedings)*, Elena Cabrio, Serena Villata, and Adam Wyner (Eds.), Vol. 1341. CEUR-WS.org. http://ceur-ws.org/Vol-1341/paper4.pdf

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In *Advances in Conceptual Modeling–Applications and Challenges*. Springer, 200–209.

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

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical Extraction of Disaster-relevant Information from Social Media. In *Proceedings of the 22Nd International Conference on World Wide Web Companion (WWW '13 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1021–1024. http://dl.acm.org/citation.cfm?id=2487788.2488109

Noriaki Kawamae. 2011. Predicting future reviews: sentiment analysis models for collaborative filtering. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 605–614.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1746–1751.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. In *Proceedings of the Second Workshop on Argumentation Mining*. Association for Computational Linguistics, 1–11.

Werner Kunz and Horst WJ Rittel. 1970. *Issues as elements of information systems*. Institute of Urban and Regional Development, University of California, Vol. 131. Berkeley, California, US.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 79–87. http://acl2014.org/acl2014/W14-21/pdf/W14-2111.pdf

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 531 (2015), 436–444.

Cane WK Leung, Stephen CF Chan, and Fu-lai Chung. 2006. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In *Proceedings of the ECAI 2006 workshop on recommender systems*. Citeseer, 62–66.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *COLING 2014, Dublin, Ireland*, Jan Hajic and Junichi Tsujii (Eds.). ACL, 1489–1500. http://www.aclweb.org/anthology/C14-1141

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 343–351.

Eric Lindahl, Stephen O'Hara, and Qiuming Zhu. 2007. A multi-agent system of evidential reasoning for intelligence analyses. In *AAMAS*, Edmund H. Durfee, Makoto Yokoo, Michael N. Huhns, and Onn Shehory (Eds.). IFAAMAS, 279.

Marco Lippi and Paolo Torroni. 2015. Context-Independent Claim Detection for Argument Mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, Qiang Yang and Michael Wooldridge (Eds.). AAAI Press, 185–191. http://ijcai.org/papers15/Abstracts/IJCAI15-033.html

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. MIT Press.

Michael Mäs and Andreas Flache. 2013. Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PLoS ONE* 8, 11 (11 2013), e74516. DOI:http://dx.doi.org/10.1371/journal.pone.0074516

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34 (4 2011), 57–74. Issue 02. DOI:http://dx.doi.org/10.1017/S0140525X10000968

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

Michela Milano, Barry O'Sullivan, and Marco Gavanelli. 2014. Sustainable Policy Making: A Strategic Challenge for Artificial Intelligence. *AI Magazine* 35, 3 (2014), 22–35. DOI:http://dx.doi.org/10.1609/aimag.v35i3.2534

Raquel Mochales Palau and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments? A case study on the legal argumentation of the ECHR. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009), Barcelona, Spain, 8-12 June 2009*. ACM, 21–30. DOI:http://dx.doi.org/10.1145/1568234.1568238

Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law* 19, 1 (2011), 1–22. DOI:http://dx.doi.org/10.1007/s10506-010-9104-x

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. 2013. The added value of argumentation. In *Agreement Technologies*. Springer-Verlag, 357–404. DOI:http://dx.doi.org/10.1007/978-94-007-5583-3_21

Marie-Francine Moens. 2014. Argumentation mining: Where are we now, where do we want to be and how do we get there?. In *Post-proceedings of the forum for information retrieval evaluation (FIRE 2013)*.

Alessandro Moschitti. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *Machine Learning: ECML 2006*, Johannes Frnkranz, Tobias Scheffer, and Myra Spiliopoulou (Eds.). LNCS, Vol. 4212. Springer Berlin Heidelberg, 318–329. DOI:http://dx.doi.org/10.1007/11871842_32

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

Susan E. Newman and Catherine C. Marshall. 1991. *Pushing Toulmin Too Far: Learning From an Argument Representation Scheme*. Technical Report. Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94034.

Nam Nguyen and Yunsong Guo. 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th international conference on Machine learning*. ACM, 681–688.

Stefanie Nowak and Stefan Rüger. 2010. How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*. ACM, New York, NY, USA, 557–566. DOI:http://dx.doi.org/10.1145/1743384.1743478

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 24–28. http://acl2014.org/acl2014/W14-21/pdf/W14-2104.pdf

Sebastian Padò, Gil Noh, Asher Stern, Rui Wang, and Robert Zanol. 2013. Design and Realization of a Modular Architecture for Textual Entailment. *Journal of Natural Language Engineering* 1 (12 2013), 1–34.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *Knowledge and Data Engineering, IEEE Transactions on* 22, 10 (Oct 2010), 1345–1359. DOI:http://dx.doi.org/10.1109/TKDE.2009.191

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. DOI:http://dx.doi.org/10.1561/1500000011

Joonsuk Park and Claire Cardie. 2014. Identifying Appropriate Support for Propositions in Online User Comments. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, 29–38. http://www.aclweb.org/anthology/W/W14/W14-2105

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015. Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments. In *Proceedings of the Second Workshop on Argumentation Mining*. Association for Computational Linguistics.

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

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7, 1 (2013), 1–31. DOI:http://dx.doi.org/10.4018/jcini.2013010101

John L. Pollock. 1987. Defeasible Reasoning. *Cognitive Science* 11, 4 (1987), 481–518. DOI:http://dx.doi.org/10.1207/s15516709cog1104_4

Hoifung Poon and Pedro Domingos. 2007. Joint Inference in Information Extraction. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, 2007, Vancouver, Canada*. AAAI Press, 913–918.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised Semantic Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–10. http://dl.acm.org/citation.cfm?id=1699510.1699512

Paul Reisert, Junta Mizuno, Miwa Kanno, Naoaki Okazaki, and Kentaro Inui. 2014. A Corpus Study for Identifying Evidence on Microblogs. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*. Association for Computational Linguistics and Dublin City University, 70–74. http://aclweb.org/anthology/W14-4910

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 440–450.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1524–1534. http://dl.acm.org/citation.cfm?id=2145432.2145595

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying Kernel Methods to Argumentation Mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23-25, 2012*, G. Michael Youngblood and Philip M. McCarthy (Eds.). AAAI Press. http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/view/4366

Sara Rosenthal and Kathleen McKeown. 2012. Detecting Opinionated Claims in Online Discussions. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*. IEEE Computer Society, 30–37.

Patrick Saint-Dizier. 2012. Processing natural language arguments with the<TextCoop>platform. *Argument & Computation* 3, 1 (2012), 49–82. DOI:http://dx.doi.org/10.1080/19462166.2012.663539

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. (2015), 56–66. http://www.aclweb.org/anthology/W15-0508

John Scott. 2012. *Social network analysis*. Sage.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1 (2002), 1–47. DOI:http://dx.doi.org/10.1145/505282.505283

Guillermo R. Simari and Ronald P. Loui. 1992. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence* 53, 23 (1992), 125 – 157. DOI:http://dx.doi.org/10.1016/0004-3702(92)90069-A

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631. Citeseer, 1642.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *COLING 2014, Dublin, Ireland*, Jan Hajic and Junichi Tsujii (Eds.). ACL, 1501–1510. http://www.aclweb.org/anthology/C14-1142

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *EMNLP 2014, Doha, Qatar*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 46–56.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *CoRR* abs/1503.00075 (2015). http://arxiv.org/abs/1503.00075

Simone Teufel. 1999. Argumentative zoning. *PhD Thesis, University of Edinburgh* (1999).

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

Mathias Verbeke, Paolo Frasconi, Vincent Van Asch, Roser Morante, Walter Daelemans, and Luc De Raedt. 2012. Kernel-based logical and relational learning with kLog for hedge cue detection. In *Inductive logic programming*. Springer, 347–357.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some Facets of Argument Mining for Opinion Analysis. In *COMMA 2012 (Frontiers in Artificial Intelligence and Applications)*, Bart Verheij, Stefan Szeider, and Stefan Woltran (Eds.), Vol. 245. IOS Press, 23–34. DOI:http://dx.doi.org/10.3233/978-1-61499-111-3-23

Douglas Walton. 2009. Argumentation Theory: A Very Short Introduction. In *Argumentation in Artificial Intelligence*, Guillermo Simari and Iyad Rahwan (Eds.). Springer US, 1–22. DOI:http://dx.doi.org/10.1007/978-0-387-98197-0_1

Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 26–32.