Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Five Years of Argument Mining: a Data-driven Analysis

Elena Cabrio and **Serena Villata**
Université Côte d'Azur, CNRS, Inria, I3S, France
elena.cabrio@unice.fr ; villata@i3s.unice.fr

Abstract

Argument mining is the research area aiming at extracting natural language arguments and their relations from text, with the final goal of providing machine-processable structured data for computational models of argument. This research topic has started to attract the attention of a small community of researchers around 2014, and it is nowadays counted as one of the most promising research areas in Artificial Intelligence in terms of growing of the community, funded projects, and involvement of companies. In this paper, we present the argument mining tasks, and we discuss the obtained results in the area from a data-driven perspective. An open discussion highlights the main weaknesses suffered by the existing work in the literature, and proposes open challenges to be faced in the future.

1 Introduction

If you had the dream that one day, in the broad Artificial Intelligence (AI) area, Natural Language Processing (NLP) researchers and Knowledge Representation and Reasoning (KRR) researchers were able to sit down together at the table of a joint panel, discussing on how to make progress and realize automated argument detection, then this paper is for you. This is the story of a research area called *Argument Mining* (AM), and how it has become an important topic in AI.

Few approaches to what is now called argument mining started to appear around 2010, when the first methods to mine (different connotations of) *arguments* from natural language documents were proposed: [Teufel *et al.*, 2009] introduced the definition of argumentative zoning for scientific articles, and [Mochales and Moens, 2011] proposed a way to detect arguments from legal texts. Since these seminal approaches, the need for automated methods to mine arguments and the relations among them from natural language text was brought to light, but it was only briefly touched upon. The parallel advances, from the formal point of view in the research field of computational models of argument, and from the point of view of the computational techniques for learning and understanding human language content in the NLP and the Machine Learning fields, boosted the almost contemporary organization of two events in 2014 targeting open discussions about the challenge of mining arguments from text. Both the workshop on Argument Mining[1] co-located with ACL, and the workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing [2] we organized, shared the same goal: bringing together the communities of NLP and of formal argumentation to jointly work towards the definition of the new research area of *argument mining*. Since then, two Dagstuhl Seminars have been organized on such topic[3], the Argument Mining workshop holds every year, two tutorials on AM have been given at IJCAI-2016[4] and ACL-2016[5], three ESSLLI courses[6] in 2017, and AM has became a topic in major AI and NLP conferences. All these clues prove its growing importance in AI.

Argument mining involves several research areas from the AI panorama: NLP provides the methods to process natural language text, to identify the arguments and their components (i.e., premises and claims) in texts and to predict the relations among such arguments, KRR contributes with the reasoning capabilities upon the retrieved arguments and relations so that, for instance, fallacies and inconsistencies can be automatically identified in such texts, and Human-Computer Interaction guides the design of good human-computer digital argument-based supportive tools.

The goal of this paper is to provide an overview of the existing approaches in the AM literature, mainly focusing on recent developments in NLP. With respect to the two former state-of-the-art contributions [Peldszus and Stede, 2013; Lippi and Torroni, 2016b], we adopt a different perspective, and we propose a data-driven analysis of the existing work in AM, structuring it around precise axes, i.e., application scenarios, algorithms, features, and produced resources for systems evaluation.

In the remainder, Section 2 provides the task definition, Section 3 discusses the existing work. Section 4 investigates the weaknesses of current approaches and open challenges.

[1] https://goo.gl/kF4Eep

[2] https://goo.gl/ttVUZk

[3] I.e., Debating Technologies (https://goo.gl/osqEY3) and Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (https://goo.gl/jS1Co6)

[4] https://goo.gl/kd4456

[5] http://acl2016tutorial.arg.tech/

[6] https://goo.gl/Cw1FLC

2 The Argument Mining Framework

Argument(ation) mining has been defined as "the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand" [Habernal and Gurevych, 2017]. Two stages are crucial in the argument mining framework:

Arguments' extraction : The first stage is the identification of arguments within the input natural language text. This step may be further split in two different stages such as the detection of argument components (e.g., claim, premises) and the further identification of their textual boundaries. Many approaches have recently been proposed to address such task, that adopt different methods like Support Vector Machines (SVM) [Mochales and Moens, 2011; Lippi and Torroni, 2016c; Stab and Gurevych, 2017; Niculae *et al.*, 2017; Bar-Haim *et al.*, 2017], Naïve Bayes classifiers [Duthie *et al.*, 2016], Logistic Regression [Levy *et al.*, 2014].

Relations' prediction : The second stage consists in predicting what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues. The relations between the arguments may be of heterogeneous nature, like *attacks* and *supports*. They are used to build the argument graphs, in which the relations connecting the retrieved arguments (i.e., the nodes in the graph) correspond to the edges. Different methods have been employed to address this task, from standard SVMs to Textual Entailment [Cabrio and Villata, 2013]. This stage is also in charge of predicting, in structured argumentation, the internal relations of the argument's components, such as the connection between the premises and the claim [Stab and Gurevych, 2017].

To clarify such tasks, let us consider the following example from the political debate of the Campaign "Trump - Clinton" on September 2016.[7] The first task of the argument mining framework consists in detecting the arguments from the text. In the example below, we highlight the arguments that can be identified (premises underlined and claims in bold):

A_1: *She talks about solar panels. We invested in a solar company, our country.* ***That was a disaster****. They lost plenty of money on that one. Now, look, I'm a great believer in all forms of energy, but we're putting a lot of people out of work.*

A_2: *Well,* ***I'm really calling for major jobs*** *because the wealthy are going create tremendous jobs. They're going to expand their companies. They're going to do a tremendous job.*

It appears evident that the argumentative sentences "in the wild", i.e., in natural language text as the ones reported in the examples, are pretty far from the prototypical argumentation patterns usually investigated in KRR, increasing the complexity of the task.

[7]Debate extracted from the Commission on Presidential Debates (http://debates.org).

Let us consider now another example from an online debate about *Random sobriety tests for drivers*[8], where we identify again premises and claims.

A_3: *Little evidence random alcohol tests deter drunk driving. There is a dearth of research regarding the deterrent effect of checkpoints. The only formally documented research regarding deterrence is a survey of Maryland's "Checkpoint Strikeforce" program. The survey found no deterrent effect: "* ***To date, there is no evidence to indicate that this campaign, which involves a number of sobriety checkpoints and media activities to promote these efforts, has had any impact on public perceptions, driver behaviors, or alcohol-related motor vehicle crashes and injuries.*** *This conclusion is drawn after examining statistics for alcohol-related crashes, police citations for impaired driving, and public perceptions of alcohol-impaired driving risk."*

A_4: ***Random breath testing doesn't necessarily lower drunk driving.*** *Many countries have had random testing for some time and have seen no real fall in drink driving figures.*

A_5: ***Random sobriety tests for drivers are effective at deterring drunk driving.***

Given these three arguments, the relations among them have to be predicted. Let us consider that the two relations we aim at identifying are *attack* (a negative relation between two arguments, e.g., a contradiction) and *support* (a positive relation between two arguments) only. In this case, we have that argument A_3 supports argument A_4, and argument A_4 attacks argument A_5.

It is important to underline at this point that argument mining differs from well known *opinion mining* (or *sentiment analysis*): while opinion mining focuses on understanding *what* users think about a certain topic or product, argument mining revolves around *why* users have a certain opinion about a topic or product.

Both the main argument mining tasks require high-quality annotated corpora to train and to evaluate the performances of automated approaches. The reliability of an annotated corpus is guaranteed by the calculation of the inter-annotator agreement that measures the degree of agreement in performing the annotation task among the involved annotators. For instance, when building a dataset for relation prediction, the statistical measure to be used to calculate the inter-rater agreement among the labels assigned by the annotators is the Cohen's kappa coefficient which takes into account also agreement occurring by chance. The equation for κ is $\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$ where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement. If the raters are in complete agreement then κ = 1, if there is no agreement among the raters other than what would be expected by chance, κ = 0. For NLP tasks, the agreement is considered as significant when $\kappa >$0.6.[9]

[8]http://www.debatepedia.com/en/index.php/Debate:_Random_sobriety_tests_for_drivers

[9]For more details about inter-annotator agreement, we refer the reader to [Artstein, 2017].

3 Data-driven Analysis

This section provides an overview of the main recent contributions in the argument mining research area. Due to space constraints, we made the choice of selecting and focusing on the most recent research papers published in main AI and NLP conferences, minimizing the redundancy in the same authors' citations. For older (till 2015) contributions, we refer the reader to Lippi and Torroni [2016b].[10] We propose a data-driven approach to the analysis of the existing work in the area, structuring it around precise axes, i.e., coarse-grained application scenarios (Sections 3.1, 3.2, 3.3, and 3.4), most performing algorithms (Table 1), most commonly used features (Table 2), and released datasets (Table 3).

3.1 Education

In the education field, argument mining has been applied to two genres of text, namely, student essays written in response to controversial topics, and scientific articles.

Persuasive essays. A persuasive essay explains a specific topic and attempts to persuade the audience that the speaker's point of view is the most informed, logical and valid perspective on the topic. This makes such kind of texts an excellent playground to test AM tasks. For instance, [Stab and Gurevych, 2017] propose an approach to identify argument components using sequence labeling at the token level, and apply a joint model for detecting argumentation structures (optimized using Integer Linear Programming). They also build an annotated corpus of persuasive essays.[11]

[Eger *et al.*, 2017] use the same corpus of persuasive essays to propose a neural end-to-end AM system. Neural computational AM is at least as good as the competing feature-based Integer Linear Programming formulation, with the advantage of eliminating the need for manual feature engineering and constraint designing. Among their findings, they highlight that, even if coupling argument component detection and relation prediction is not optimal, both tasks should be treated separately, but modeled jointly, and that the relation prediction task is more challenging than the first one.

With the goal of improving the automated scoring of persuasive essays, [Nguyen and Litman, 2018] implement another end-to-end argument mining system that parses argumentative structures of free-text essays and creates argumentative features from these structures.

[Peldszus and Stede, 2015] jointly predict different aspects of the argumentation structure by combining the different subtasks prediction in the edge's weights of an evidence graph; they then apply a standard Minimum Spanning Tree decoding algorithm on a small corpus of English-German microtexts.[12] They rely on Freeman's dialectical theory using the moves of proponent and challenger in a dialectical situation as a model of the structure of the argumentation in texts.

Scientific articles. Among the earliest work that can be considered as forerunner of AM, in [Teufel *et al.*, 2009] a rhetorical-level analysis of scientific articles is introduced (*argumentative zoning*). Data annotation is based on the typical argumentation to be found in scientific articles. It reflects the attribution of intellectual ownership in scientific articles, expressions of authors' stance towards the related work, and typical statements about problem-solving processes.

3.2 Web-based Content

In the following, we present relevant contributions experimented on heterogeneous data extracted from the Web.

Wikipedia articles. IBM is putting a lot of effort in the development of debating technologies[13]. Among their contributions, [Levy *et al.*, 2014] address the task of automatically detecting *context dependent claims* in Wikipedia articles, i.e., a general, concise statement that directly supports or contests the given topic, discussed in the debate motions database[14]. As a follow up, [Rinott *et al.*, 2015] address the task of automatically detecting evidences in Wikipedia articles supporting a given claim (*context dependent evidence detection*). More recently, [Bar-Haim *et al.*, 2017] introduce the task of *claim stance classification*, decomposed into the detection of: *i)* the targets of the given topic and the claim, *ii)* the polarity (sentiment) towards each of the targets, and *iii)* whether the targets are consistent or contrastive. To evaluate this task, the Wikipedia-based IBM dataset for claim classification is extended by adding Pro/Con annotations.

[Lippi and Torroni, 2016c] present MARGOT (Mining ARGuments frOm Text),[15] a tool for argument component classification (both premises and claims) and boundaries detection. The system is tested on the IBM datasets [Bar-Haim *et al.*, 2017; Rinott *et al.*, 2015].

Microblogs and web debating platforms. [Habernal and Gurevych, 2017] propose a sequence labeling approach to identify argument components (following a modified Toulmin's model) in user-generated Web discourses, i.e., on a sample of controversial topics about education.

[Niculae *et al.*, 2017] propose a structured prediction model for AM (comparing SVMs and RNNs algorithms), jointly learning to classify elementary units and to identify the argumentative relations between them. Two datasets are used for evaluation: the Cornell eRulemaking Corpus - CDCP,[16] and the persuasive essays one [Stab and Gurevych, 2017].

[Cabrio and Villata, 2013] tackle the relation prediction task on a corpus of online debates from Debatepedia (now called `idebate.com`).[17] Starting from the opinions put forward from the users and the main issue of the debate, we investigate how Textual Entailment suites can be exploited to

[10] An updated list of external resources in Argument Mining is maintained at http://argumentationmining.disi.unibo.it/resources.html.

[11] https://goo.gl/3tXibr

[12] https://github.com/peldszus/arg-microtexts

[13] IBM Debater Datasets (https://goo.gl/MxfB7N), the EPSRC Argument Mining project (https://goo.gl/444uu8).

[14] https://idebate.org/

[15] Demo available at http://margot.disi.unibo.it/.

[16] http://joonsuk.org

[17] http://www-sop.inria.fr/NoDE/

Approaches	Component Detection		Relations prediction
	Sentence classification	**Boundaries Detection**	
SVM	[Mochales and Moens, 2011], [Duthie *et al.*, 2016] [Lippi and Torroni, 2016a; 2016c] [Habernal and Gurevych, 2017] [Bar-Haim *et al.*, 2017]	[Mochales and Moens, 2011] [Lippi and Torroni, 2016c]	[Naderi and Hirst, 2015] [Niculae *et al.*, 2017] [Stab and Gurevych, 2017] [Menini *et al.*, 2018]
P	[Villalba and Saint-Dizier, 2012] [Peldszus and Stede, 2015] [Eger *et al.*, 2017]	[Eger *et al.*, 2017]	[Villalba and Saint-Dizier, 2012] [Peldszus and Stede, 2015] [Eger *et al.*, 2017]
LR	[Levy *et al.*, 2014], [Rinott *et al.*, 2015] [Nguyen and Litman, 2018]	[Dusmanu *et al.*, 2017] [Ibeke *et al.*, 2017] [Nguyen and Litman, 2018]	[Nguyen and Litman, 2018]
RNN	[Eger *et al.*, 2017]	[Eger *et al.*, 2017]	[Niculae *et al.*, 2017] [Eger *et al.*, 2017]
ME	[Mochales and Moens, 2011], [Duthie *et al.*, 2016]	[Mochales and Moens, 2011]	
CRF	[Stab and Gurevych, 2017]		
NB	[Duthie *et al.*, 2016]		
RF		[Dusmanu *et al.*, 2017]	
TES			[Cabrio and Villata, 2013]
ML		[Levy *et al.*, 2014]	

Table 1: A comparison of the approaches applied to AM tasks. They are ordered starting from the most frequently applied methods. As for other tasks in NLP, SVMs have proved to be the most performing algorithms in different settings, and for different AM sub-tasks. The acronyms stand for: Support Vector Machine (SVM), Parsing algorithms (P), Logistic Regression (LR), Recurrent Neural Networks for language models (RNN), Maximum Entropy models (ME), Conditional Random Fields (CRF), Naïve Bayes (NB), Random Forests (RF), Textual Entailment Suites (TES) and Maximum Likelihood (ML),

predict the support (i.e., entailment) and the attack (i.e., contradiction) relations among these text snippets.

In [Khatib *et al.*, 2016], a large corpus annotated with argumentative text segments is acquired through distant supervision from the same online debate portal, and used to test a binary classifier of text argumentativeness.

Online product reviews. Argument mining techniques make it possible to capture the underlying motivations consumers express in reviews, which provide more information than a basic attitude like "I do/don't like product A". [Villalba and Saint-Dizier, 2012] discuss how the automatic recognition of arguments can be implemented on the TextCoop platform. In [Ibeke *et al.*, 2017], the authors address the task of mining contrastive opinions using a unified latent variable model on the El Capitan dataset,[18] where reviews are manually annotated with topic and sentiment labels. Analyzing arguments in user reviews suffers from the vague relation between argument mining and sentiment analysis. This is because sentiments about individual aspects of the implied claim (for/against the product) sometimes express also the reasons why the product is considered to be good or bad.

Newspaper articles. As a second scenario, [Lippi and Torroni, 2016c] evaluate MARGOT on ten newspaper articles from the New York Times, that cover various topics.[19]

Social media. In [Dusmanu *et al.*, 2017], we collected a dataset of tweets, DART, where we addressed the tasks of distinguishing argumentative tweets from non-argumentative ones. The topics of the tweets range from politics like Brexit and Grexit to the release of the new Apple Watch.

Moreover, MARGOT [Lippi and Torroni, 2016c] is applied to the comments in two Reddit threads (a sub Reddit focused on the New Hampshire primaries held on February 9th, 2016, and a sub Reddit focused on climate shift).

3.3 Legal Documents

In the legal domain, argument mining approaches have been proposed to detect premises, claims and argumentation schemes in judgments to ease the work of judges and law scholars in identifying similarities and differences among different judgments, the arguments proposed therein, and the ultimate outcome of the cases. More precisely, [Mochales and Moens, 2011] propose a system for argument component detection and inter-argument relation prediction for the legal domain. They identify premises and claims using statistical classifiers, and they define a context-free grammar to predict the relations among the different argument components. They created a corpus from the European Court of Human Rights (ECHR) judgments. Following this line of work, [Teruel *et al.*, 2018] recently present a new corpus of ECHR judgments[20] annotated with premises and claims as well as with support and attack relations among the argument components. [Grabmair *et al.*, 2015] work with a set of U.S. Court of Federal Claims cases deciding whether compensation claims comply with a federal statute establishing the National Vaccine Injury Compensation Program. The Legal UIMA system they propose extracts argument-related semantic information from such legal documents: the princi-

[18] https://github.com/eibeke/El-Capitan-Dataset

[19] https://goo.gl/mmxv9i

[20] https://github.com/PLN-FaMAF/ArgumentMiningECHR

pal argumentation roles of clauses, e.g., evidence-based finding of fact, evidence-based intermediate reasoning, and case-specific process or procedural facts.

3.4 Political Debates and Speeches

The political domain allows for intuitive applications of the argument mining framework with the final aim of detecting fallacies, persuasiveness degree and coherence in the candidate's argumentation. [Lippi and Torroni, 2016a] address the problem of argument extraction, and more precisely claim detection, over a corpus based on the 2015 UK political election debates. They aim to study the impact of the vocal features of speech on the claim detection task. The Internet Argument Corpus[21] (IAC) [Walker *et al.*, 2012] collects the posts from 4forums.com, a website for political debate. The debates have been annotated for argumentative markers like degrees of agreement with a previous post, cordiality, audience direction, combativeness, assertiveness, emotionality of argumentation, and sarcasm. [Duthie *et al.*, 2016] apply AM methods to detect the presence and polarity of ethotic arguments from UK parliamentary debates.[22] The authors also investigate how their results can be visualized to support user understanding.[23] [Naderi and Hirst, 2015] show how features based on embedding representations can improve discovering various frames in argumentative political speeches. They propose a corpus of speeches from the Canadian Parliament, and they examine the statements with respect to the position of the speaker towards the discussed topic (pro, con, or no stance). In [Menini *et al.*, 2018], we address the relation prediction task on political speeches in monological form, where there is no direct interaction between the opponents. We created a corpus, based on the transcription of speeches and official declarations issued by Nixon and Kennedy during 1960 Presidential campaign, of argument pairs annotated with the support and attack relations.[24]

Only few contributions tackle the issue of generalizing across different text types. Among them, Araucaria[25] collects arguments from heterogeneous sources, e.g., newspapers, parliamentary records, judgments and discussion fora. The annotation is based on Walton's argumentation schemes. Other work in this direction has been proposed by [Hua and Wang, 2017] and [Stab *et al.*, 2018].

4 Discussion

The aim of this survey paper is to show how the joint effort of two, rather disjoint, research communities in AI resulted in the development of a new research area: *argument mining*. This synergy among researchers from both NLP and KRR communities has lead to the conception and development of systems able to mine a variety of textual documents, e.g., legal cases, persuasive essays, online debates and tweets, to detect premises and claims, and predict the relations among them. The results obtained so far in AM have attracted the interest (and investment) of companies (e.g., IBM), and have raised high expectations for the future findings in the area.

Features
1. Syntactic and Positional
2. Lexicon
3. Topic relatedness/ semantic similarity
4. Sentiment
5. Embeddings
6. Patterns (regex)
7. Discourse
8. Bag-of-words
9. Subjectivity classifier
10. NER
11. Vocal (speech)
12. Wikipedia-based
13. PMI
14. Emoticons

Table 2: A list of the features most frequently computed for AM tasks, ordered from the most frequently used ones.

This paper has focused on the standard definition of the AM framework to highlight its main tasks. However, in the last years, a number of new challenges have been proposed in the literature. In particular, [Dusmanu *et al.*, 2017] select argumentative tweets and distinguish those conveying an opinion from those containing *factual* information, to detect their source of information (e.g., the BBC). [Habernal and Gurevych, 2016; Persing and Ng, 2017; Durmus and Cardie, 2018] focused on argument persuasion to study the relation "argument A is more convincing than argument B".

In addition, AM is strongly connected with hot topics in AI, as deep learning (heavily used in AM), fact checking and misinformation detection (the prediction of the attacks between arguments is a building block for fake news detection), and explanations of machine decisions (AM can disclose how the information on which the machine relies to make its own decisions is retrieved). Other scenarios where AM can contribute are medicine (where AM can detect information needed to reason upon randomised clinical trials), politics (where AM can provide the means to automatically identify fallacies and unfair propaganda), and for cyberbullism prevention (where AM can support the detection of repeated attacks against a person[26]).

Alas, all that glitters is not gold, and some open issues in AM should be tackled to actually attain the expectations. First of all, system performances should improve. Despite the good results obtained in some application scenarios, i.e., persuasive essays [Stab and Gurevych, 2017] (where the structure of the essays themselves eases the argument component detection task), for other kinds of documents, e.g., legal cases [Teruel *et al.*, 2018] and microtexts [Peldszus and Stede, 2015], more work is still required. It is important to underline here that also human agreement (generally viewed as the upper bound on automatic performance in annotation tasks) is affected by the complexity of the AM tasks. As a result, there still exists a gap between NLP and KRR: *(i)* NLP

[21] http://nlds.soe.ucsc.edu/software

[22] http://arg.tech/Ethan3Train, http://arg.tech/Ethan3Test

[23] https://goo.gl/P9fyzi

[24] https://dh.fbk.eu/resources/political-argumentation

[25] https://goo.gl/tU7dCr

[26] http://creep-project.eu/

	Datasets	Document source	Size	Component Detection		RP
				Sent. Clas.	BD	
Educ.	[Stab and Gurevych, 2017]	persuasive essays	402 essays	✓	✓	✓
	[Peldszus and Stede, 2015]	microtexts	112 short texts	✓		✓
Web-based content	[Bar-Haim *et al.*, 2017]	debate motions DB	55 topics	✓		
	[Rinott *et al.*, 2015]	Wikipedia, debate motions DB	58 topics, 547 articles	✓		
	[Bar-Haim *et al.*, 2017]	Wikipedia, debate motions DB	33 topics, 586 articles	✓		
	IAC	4forums.com	11,800 discussions			
	[Habernal and Gurevych, 2017]	comments, forum, blog posts	524 documents	✓		
	[Khatib *et al.*, 2016]	*i-debate*	445 documents		✓	
	NoDE	online debates	260 pairs			✓
	DART	Twitter	4,713 tweets		✓	✓
	Araucaria	newspapers, legal, debates	660 arguments	✓		
Legal	[Teruel *et al.*, 2018]	ECHR judgments	7 judgments	✓	✓	✓
	[Mochales and Moens, 2011]	ECHR judgments	47 judgments	✓	✓	✓
	[Niculae *et al.*, 2017]	eRule-making discussion forum	731 comments			✓
Politics	[Menini *et al.*, 2018]	Nixon-Kennedy Presid. campaign	5 topics (1,907 pairs)			✓
	[Lippi and Torroni, 2016a]	Sky News debate for UK elections	9,666 words	✓		
	[Duthie *et al.*, 2016]	UK parliamentary record	60 sessions	✓		
	[Naderi and Hirst, 2015]	speeches Canadian Parliament	34 sent., 123 paragr.			✓

Table 3: Available datasets for AM (sub-)tasks, grouped by their application scenario (BD=boundaries detection; RP=relation prediction).

is error-prone, and *(ii)* there is a lot of uncertainty involved in argumentation, as realized in the natural language. How to close this gap is an open research challenge: hopefully, it is getting smaller by virtue of the efforts of the AM community.

Moreover, various heterogeneous datasets have been produced since the beginning of research in AM. Because of the immaturity of a rising field, and the lack of clear definitions, each dataset has been annotated relying on slightly different definitions of argument components and of the relations holding between them, thus preventing the possibility of a straightforward alignment among datasets. While on the one side, it would be worth trying to unify existing resources, on the other side, this fact shows that AM is flexible enough to adapt to different use case scenarios, e.g., premises and claims are not the same in legal cases, persuasive essays and Twitter. In [Daxenberger *et al.*, 2017], a qualitative analysis of six different datasets used in AM is presented, to underline the different conceptualization of claims. Recently, [Schulz *et al.*, 2018] show that multi-task learning is one possible way to go. More precisely, they study whether conceptually diverse AM datasets from different domains can help deal with new AM datasets when data is limited. The question about the worthiness of unifying the existing datasets is still open and under debate. [Wachsmuth *et al.*, 2017] highlight and empirically study a related issue, i.e., the question of how different the theoretical (computational models of argument) and practical views of argumentation quality actually are. Their results show that, on the one hand, most reasons for quality differences in practice seem well-represented in the theory, but on the other hand, some quality dimensions remain hard to assess in practice, resulting in a limited agreement.

Finally, another open challenge in AM deals with multilinguality. Only very few approaches tackled the issue of applying AM methods to texts in other natural languages than English, i.e., [Peldszus and Stede, 2016] address argument component detection for German and [Basile *et al.*, 2016] tackle the relation prediction task for Italian.

Acknowledgments

The authors have received funding from EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 (MIREL).

References

[Artstein, 2017] Ron Artstein. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht, 2017.

[Bar-Haim *et al.*, 2017] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *EACL*, pages 251–261, 2017.

[Basile *et al.*, 2016] Pierpaolo Basile, Valerio Basile, Elena Cabrio, and Serena Villata. Argument mining on italian news blogs. In *CLiC-it*, volume 1749 of *CEUR Workshop Proceedings*, 2016.

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

[Daxenberger *et al.*, 2017] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the essence of a claim? cross-domain claim identification. In *EMNLP*, pages 2055–2066, 2017.

[Durmus and Cardie, 2018] Esin Durmus and Claire Cardie. Exploting the role of prior beliefs for argument persuasion. In *NAACL*, page 1035–1045, 2018.

[Dusmanu *et al.*, 2017] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In *EMNLP*, pages 2317–2322, 2017.

[Duthie *et al.*, 2016] Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *COMMA*, pages 299–310, 2016.

[Eger *et al.*, 2017] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *ACL*, pages 11–22, 2017.

[Grabmair *et al.*, 2015] Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In *ICAIL*, pages 69–78, 2015.

[Habernal and Gurevych, 2016] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *ACL*, page 1589–1599, 2016.

[Habernal and Gurevych, 2017] I. Habernal and I. Gurevych. Argumentation mining in user-generated web discourse. *Comput. Linguist.*, 43(1):125–179, 2017.

[Hua and Wang, 2017] Xinyu Hua and Lu Wang. Understanding and detecting diverse supporting arguments on controversial issues. In *ACL*, pages 203–208, 2017.

[Ibeke *et al.*, 2017] Ebuka Ibeke, Chenghua Lin, Adam Z. Wyner, and Mohamad Hardyman Barawi. Extracting and understanding contrastive opinion through topic relevant sentences. In *IJCNLP*, pages 395–400, 2017.

[Khatib *et al.*, 2016] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-domain mining of argumentative text through distant supervision. In *NAACL*, pages 1395–1404, 2016.

[Levy *et al.*, 2014] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *COLING*, pages 1489–1500, 2014.

[Lippi and Torroni, 2016a] Marco Lippi and Paolo Torroni. Argument mining from speech: Detecting claims in political debates. In *AAAI*, pages 2979–2985, 2016.

[Lippi and Torroni, 2016b] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[Lippi and Torroni, 2016c] Marco Lippi and Paolo Torroni. Margot: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303, 12 2016.

[Menini *et al.*, 2018] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Never retreat, never retract: Argumentation analysis for political speeches. In *AAAI*, pages 4889–4896, 2018.

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

[Naderi and Hirst, 2015] Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *CMNA*, pages 16–25, 2015.

[Nguyen and Litman, 2018] Huy V. Nguyen and Diane J. Litman. Argument mining for improving the automated scoring of persuasive essays. In *AAAI*, pages 5892–5899, 2018.

[Niculae *et al.*, 2017] Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured svms and rnns. In *ACL*, pages 985–995, 2017.

[Peldszus and Stede, 2013] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, 7(1):1–31, 2013.

[Peldszus and Stede, 2015] Andreas Peldszus and Mandred Stede. Joint prediction in mst-style discourse parsing for argumentation mining. In *EMNLP*, page 938–948, 2015.

[Peldszus and Stede, 2016] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *ECA*, pages 801–815, 2016.

[Persing and Ng, 2017] Isaac Persing and Vincent Ng. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *IJCAI*, pages 4082–4088, 2017.

[Rinott *et al.*, 2015] Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *EMNLP*, pages 440–450, 2015.

[Schulz *et al.*, 2018] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *NAACL*, page 35–41, 2018.

[Stab and Gurevych, 2017] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Comput. Linguist.*, 43(3):619–659, 2017.

[Stab *et al.*, 2018] Christian Stab, Tristan Miller, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *CoRR*, abs/1802.05758, 2018.

[Teruel *et al.*, 2018] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *LREC*, pages 4061–4064, 2018.

[Teufel *et al.*, 2009] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *EMNLP*, pages 1493–1502, 2009.

[Villalba and Saint-Dizier, 2012] María Paz García Villalba and Patrick Saint-Dizier. A framework to extract arguments in opinion texts. *IJCINI*, 6(3):62–87, 2012.

[Wachsmuth *et al.*, 2017] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation quality assessment: Theory vs. practice. In *ACL*, pages 250–255, 2017.

[Walker *et al.*, 2012] Marylin Walker, Jean Fox Tree, Pranav Anand, R. Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.