Datenbank Spektrum (2020) 20:131–141
https://doi.org/10.1007/s13222-020-00348-6

SCHWERPUNKTBEITRAG

Explaining Arguments with Background Knowledge

Towards Knowledge-based Argumentation Analysis

Maria Becker[1] · Ioana Hulpuş[2] · Juri Opitz[1] · Debjit Paul[1] · Jonathan Kobbe[2] · Heiner Stuckenschmidt[2] · Anette Frank[1]

Received: 8 May 2020 / Accepted: 6 June 2020 / Published online: 6 July 2020

Abstract
Most information we consume as a society is obtained over the Web. News – often from questionable sources – are spread online, as are election campaigns; calls for (collective) action spread with unforeseen speed and intensity. All such actions have argumentation at their core, and the conveyed content is often strategically selected or rhetorically framed. The responsibility of critical analysis of arguments is thus tacitly transferred to the content consumer who is often not prepared for the task, nor aware of the responsibility.

The ExpLAIN project aims at making the structure and reasoning of arguments explicit – not only for humans, but for *Robust Argumentation Machines* that are endowed with language understanding capacity. Our vision is a system that is able to deeply analyze argumentative text: that identifies arguments and counter-arguments, and reveals their internal structure, conveyed content and reasoning. A particular challenge for such a system is to uncover implicit knowledge which many arguments rely on. This requires human background knowledge and reasoning capacity, in order to explicate the complete reasoning of an argument.

This article presents ongoing research of the ExpLAIN project that aims to make the vision of such a system a tangible aim. We introduce the problems and challenges we need to address, and present the progress we achieved until now by applying advanced natural language and knowledge processing methods. Our approach puts particular focus on leveraging available sources of structured and unstructured background knowledge, the automatic extension of such knowledge, the uncovering of implicit content, and reasoning techniques suitable for informal, everyday argumentation.

Keywords Argument explicitation · Argumentative relation classification · Commonsense knowledge · KB completion · Enthymeme reconstruction

M. Becker, I. Hulpuş, J. Opitz and D. Paul contributed equally.

Maria Becker
mbecker@cl.uni-heidelberg.de

Ioana Hulpuş
ioana@informatik.uni-mannheim.de

Juri Opitz
opitz@cl.uni-heidelberg.de

Debjit Paul
paul@cl.uni-heidelberg.de

Jonathan Kobbe
jonathan@informatik.uni-mannheim.de

Heiner Stuckenschmidt
heiner@informatik.uni-mannheim.de

✉ Anette Frank
frank@cl.uni-heidelberg.de

[1] Institute for Computational Linguistics, Heidelberg University, Heidelberg, Germany

[2] Data and Web Science Group, University of Mannheim, Mannheim, Germany

1 ExPLAIN: Explaining Arguments with Background Knowledge

Argumentation is ubiquitous in political discourse as well as civic engagement. Debates are organized or develop in social media, where content consumers are often overflown with arguments of varying quality, and often what is explicitly said reflects only a part of the knowledge and reasoning that underlines the argument.

Hence, the computational analysis of argumentation is becoming an active field of research in Artificial Intelligence. Common lines of work include the identification of argument units [28, 47, 51, 53] and argumentative relations such as *support* or *attack*, *undercut* or *rebuttal* [10, 31, 39, 51], the measurement of argument quality [20, 56] and first steps towards synthesis of arguments [55]. While many natural language processing (NLP) tasks can be solved with surprising accuracy using surface features, arguments are often rhetorically framed, and thus many approaches to computational argumentation build upon linguistic discourse features when solving specific argumentation subtasks (see e.g., [21]). However, we would like to achieve a deeper understanding of the reasoning behind a line of argumentation – a challenging tasks where surface features provide little help. E.g., a lot of knowledge in argumentation is implicit [57] and calls for leveraging background knowledge sources for achieving a coherent understanding of argumentation lines. So far, only little attention has been devoted to this issue [9].

The aim of the ExpLAIN project is to endow Argumentation Machines with the capacity of *explaining arguments*. We approach this aim by performing deep natural language understanding: by integrating textual analysis with background knowledge that arguments often leave implicit, but that is crucial for making the overall argument structure coherent and understandable for a computational system, and explainable for the user of the system. In [22], we frame this vision of content-driven argument analysis by formulating the task of *Argument Explicitation*, building on insights of prior research in (computational) argumentation.

The structure of this article is as follows. Sect. 2 introduces the *Argument Explicitation* task, defines our research vision and relates it to existing research. Sect. 3 motivates the role of content and background knowledge in argument analysis and shows how the ExpLAIN project addresses knowledge-based argument analysis. We design neural models for *Argumentative Relation Classification* that integrate background knowledge and show that in this way the models achieve significant performance gains. Moreover, the way in which background knowledge is integrated into the model allows us to *interpret*, or *explain* system decisions. In Sect. 4, we investigate *arguments with implicit premises*, called *Enthymemes*. We study the characteristics of specific types of knowledge and semantic types that are left implicit in arguments and develop a classifier for commonsense knowledge relation prediction that provides a basis for reconstructing implicit premises in arguments. Sect. 5 highlights further research we conduct in view of future developments and Sect. 6 concludes.

2 Overview of Argument Explicitation

In this section, we present a framework that we propose for defining and structuring the *Argument Explicitation Task,* which we call KAME (Knowledge-aware Acceptability and Model-based Explicitation) [22]. This framework is illustrated in Fig. 1. The aim of argument explicitation is to make the argumentative structure of a text and its reasoning explicit, putting particular focus on knowledge, its interaction with discourse, and implicit information conveyed in an argument. The argument explicitation task distinguishes two levels of granularity on which explicitation takes place: (i) *macro-explicitation* – which focuses on the identification of individual *atomic arguments* (an atomic argument consist of one or more premises and exactly one conclusion) and relations between them, and (ii) *micro-explicitation* – which focuses on recognition, classification and reconstruction of argumentative units within one atomic argument. At the core of our framework resides background knowledge that is instrumental for both subtasks. However, as we discuss in Sect. 4, much of the background knowledge that arguments rely on is commonsense knowledge that is not comprehensively captured in available knowledge bases. We therefore propose a dynamic knowledge-base extension step that infers missing commonsense relations between concepts.

I. Macro-explicitation of arguments deals with the identification of individual arguments in argumentative text, and of the relations between them. On the general level, one can distinguish *support* and *attack* relations. This is similar to Dung's framework [15], but extends it with *support* relations between arguments. Our aim is to use the relations between argumentative units to isolate individual *atomic arguments* (as defined above) and the relations between them. For example, the conclusion of an atomic argument can act as a premise for another argument. Similarly, the conclusion of an argument might contradict the premise of a counter-argument, a relation between arguments that is called *rebuttal* [42]. Or, the conclusion of an argument might attack the warrant that the conclusion of a counter-argument follows from its premise, a relation called *undercut* [42]. These relations between arguments indicate the *acceptability* of arguments, which is the ability of an argument set to defend against counter-arguments [15].

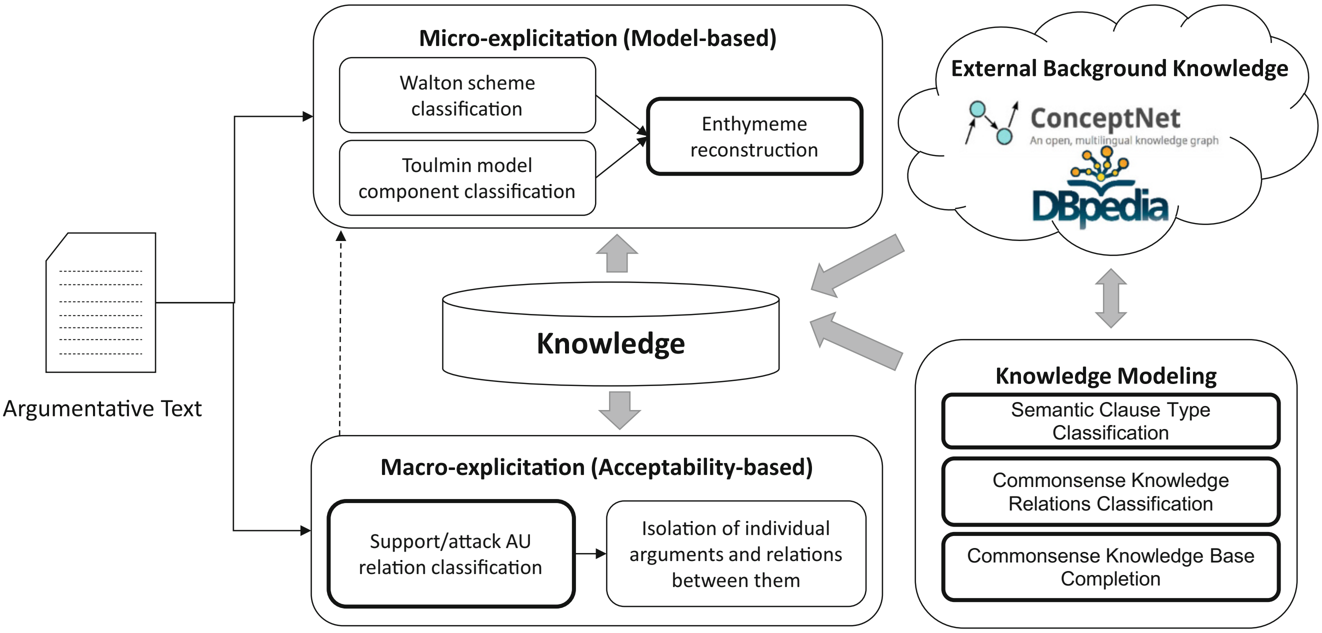
Micro-explicitation (Model-based)
Walton scheme classification
Toulmin model component classification
Enthymeme reconstruction
External Background Knowledge
ConceptNet
An open, multilingual knowledge graph
DBpedia
Knowledge
Argumentative Text
Knowledge Modeling
Semantic Clause Type Classification
Commonsense Knowledge Relations Classification
Commonsense Knowledge Base Completion
Macro-explicitation (Acceptability-based)
Support/attack AU relation classification
Isolation of individual arguments and relations between them

Fig. 1 The KAME (*Knowledge-aware Acceptability and Model-based Explicitation*) Framework. The *dotted arrow* indicates an optional dependency. *Thick lines* indicate our contributions on core aspects of KAME: (i) in *Macro-explicitation* we develop models for *support/attack* classification with background knowledge; (ii) within *Micro-explicitation* we develop resources for *Enthymeme Reconstruction*: an annotated corpus and analysis of the nature of implied knowledge; (iii) *Knowledge modeling components* for *Enthymeme Reconstruction* and *Explainable Argument Relation Classification*: on-the-fly commonsense knowledge relation prediction and situation entity classifier

For example, consider **Argument 1** [40]:

Argument 1 (alternative treatments): *Patients do often report relief of their complaints after alternative treatments. But as long as their benefits have not been scientifically proven, the health insurance companies should not cover alternative treatments.*

In this argument, the argumentative relation classification would return the relations shown in Fig. 2. The task of isolating individual arguments, that builds on top of it, would identify [unit **(2)**] **therefore** [unit **(0)**] as an atomic argument, and [unit **(1)**] as an anticipated counter-argument. The anticipated counter-argument in this case would consist of only a premise, with an implied conclusion, hence it is an enthymeme. In the micro-explication stage, using reasoning, it can be reconstructed as [unit **(1)**] **therefore NOT** [unit **(0)**]. But as we show in Sect. 4, a large part of enthymemes cannot be reconstructed by using just reasoning, as they also require background knowledge.

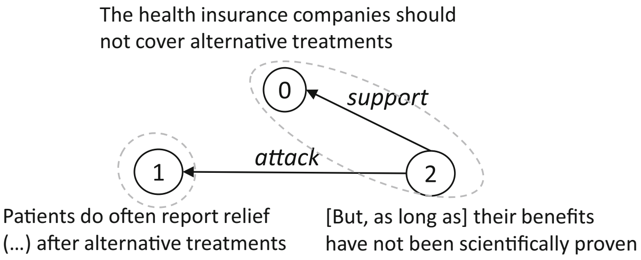
The health insurance companies should not cover alternative treatments
0
support
1
attack
2
Patients do often report relief (...) after alternative treatments
[But, as long as] their benefits have not been scientifically proven

Fig. 2 Macro-explicitation of **Argument 1:** *alternative treatments*. *Support/attack* edges indicate the relations between argumentative units. The *dotted lines* isolate the atomic arguments

Sect. 3 investigates how background knowledge contributes to the core task of argumentative relation classification. Specifically, we hypothesize that background knowledge relations that hold between entities or concepts in different argumentation units can assist information that is explicitly given in the text in accurately predicting the aforementioned relations.

II. Micro-explicitation of arguments deals with the understanding of the reasoning behind individual atomic arguments. Philosophers such as Walton [57] have identified and structured *Argumentation Schemes* as patterns of reasoning that people use in everyday, informal argumentation. Others, such as Toulmin [54] have distilled a more general model of defeasible argumentation.

These models have been researched in the AI community [16, 25, 26, 29], but most systems rely only on discourse features, are based on small training sets, and while they are able to determine the scheme of an argument, they are very limited in explaining how that scheme is instantiated. It is also unclear how additional resources can be gathered to improve results. However, this task is very important for guiding the reconstruction of arguments with implicit premises (enthymemes).

The task of *enthymeme reconstruction* has received very little attention in the AI and NLP communities [45], likely because of its complexity. As detailed in Sect. 4, in our work, we take several steps towards solving this task, by (i) conducting an annotation study to gather insights about the type of information that is usually implied but not explicitly stated in arguments, (ii) by exploiting static background knowledge sources that characterize the identified types of implicit knowledge relations between entities and concepts and (iii) by learning to predict commonsense knowledge relations that are not covered in existing knowledge bases.

3 Background Knowledge for Coarse-grained Argumentative Relation Classification

3.1 Issues with state of the art systems lacking deeper understanding of argumentative relations

Consider again **Argument 1**. When units such as **(1)** and **(2)** occur in direct textual proximity, e.g., in an essay, they are often connected with a discourse marker, as in **(1),** ***but*** **(2)** or **(1),** ***however*** **(2)**. By using such overt indicators – without deeper knowledge about reimbursement regulations in health insurance – we can easily construct such a graph, by simply projecting the indicated contrastive discourse relation onto an argumentative *attack* relation between **(2)** and **(1)**.

But in cases where such discourse markers are not explicitly given, or the units do not occur in direct proximity (e.g., they are crawled from different documents in the WWW), we can only successfully construct the graph by assessing the textual content of the units and using our knowledge about how their content relates to each other. For computers, this is not easy: as our analysis in [35] shows, removal of shallow contextual markers leads to a large degeneration in the performance of systems that predict such relations. This suggests that most state-of-the-art systems for argumentative relation classification are able to learn to identify discourse markers that are specific for particular argumentative relations, but are not able to assess the actual content of the argument. This makes systems vulnerable, since they will predict argumentative relations among random sentences as long as they are connected by a fitting rhetorical discourse marker, and – by contrast – will fail to detect meaningful relations between argumentative units if they are not part of the same discourse.

3.2 In need of knowledge! – Shallow or deep?

We distinguish two forms of knowledge for predicting argumentative relations: shallow and deep knowledge. We expect a system that successfully incorporates deep knowledge to perform well in two ways: (i) it should predict relations on unseen testing examples with high success, and (ii) it should be able to provide explanations for its decisions to justify its choices. Previously, **shallow forms of knowledge**, such as uni- or bi-gram features, have been used to assess the content of arguments. But often this is not enough to successfully predict the relation, and such a system is bound to fail when units do not appear in direct textual proximity (where discourse context cannot be exploited) [35].

Researchers have also investigated **sentiment as a proxy for deeper forms of knowledge**. For example, [52] use an auxiliary feature that reflects compositionally computed sentiment [48]. However, their feature ablation study shows that this feature does not significantly improve performance, possibly due to errors of the automatic sentiment assignment system. A more advanced approach that makes use of sentiment was proposed by [46]. They propose generalized features for *support–attack* argumentative relations that are derived from *good-for/bad-for* sentiment templates and are used as an explanatory mechanism for argumentative relations. However, while such knowledge may be useful in principle, in practice it suffers from a lack of annotated training data and large-scale lexical sentiment frame resources. Such resources may be costly to develop since, e.g., such frames need to cover or generalize over different domains, while new themes and debates are emerging as we speak. In sum, while it seems attractive in theory to exploit sentiment-based features and rules to approximate knowledge-based reasoning, in practice, this approach depends on large-scale resources and advanced models in order to accurately detect compositional sentiment and sentiment roles.

Moreover, the approach is limited to an approximation of deeper knowledge on *overt* statements and is thus not applicable to implicitness in arguments. For these reasons, we focus on deep knowledge from existing large-scale resources. Below, we elaborate on ways to gain such knowledge.

3.3 In need of knowledge! – Latent or symbolic?

One potential source of deeper knowledge is **latent knowledge** embodied in language models such as BERT [13]. Research has shown that the knowledge acquired by such models in pre-training can be leveraged by fine-tuning them to advanced semantic inference tasks [58]. Inspired by these insights, in [32] we formulate the argumentative *support vs. attack* relation prediction task as a ranking problem. We fine-tune BERT to predict the correct relation without hints from the surrounding context. Given two raw argumentative units, we connect them with two alternative discourse markers, using, e.g., *therefore* to place them into a *support* relation, as opposed to *however*, which typically indicates

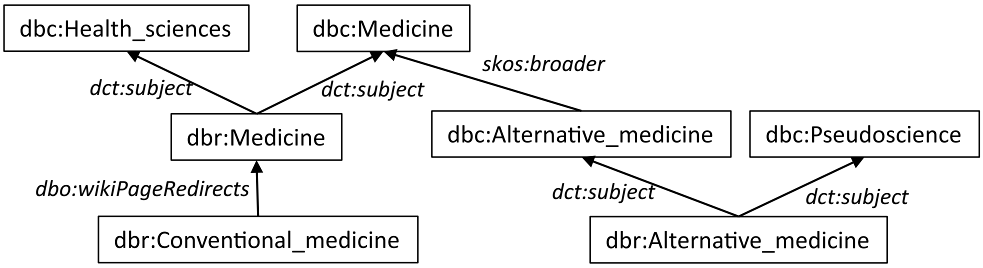

Fig. 3 DBpedia relation paths between *Conventional medicine* and *Alternative medicine*

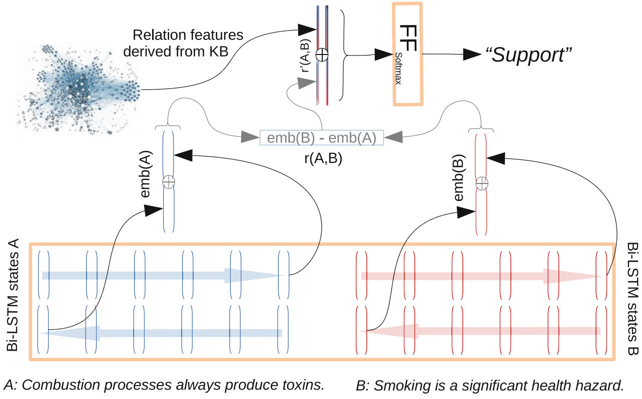
Relation features derived from KB
r'(A,B)
FF
Softmax
"Support"
emb(B) - emb(A)
r(A,B)
emb(A)
emb(B)
Bi-LSTM states A
Bi-LSTM states B
A: Combustion processes always produce toxins.
B: Smoking is a significant health hazard.

Fig. 4 Injection of knowledge-graph features for argumentative relation classification [23]

attack. We compare the probabilities that BERT computes for the alternative statements resulting from the discourse marker insertions, and predict the relation that obtains the higher score. Our experiments show that this system predicts the correct argumentative relation better than a conventional system [52].

One weakness of using latent knowledge from pre-trained models is, however, the missing explainability of such systems: the black-box nature of large language models makes it difficult to extract meaningful explanations for its predictions. We are therefore focusing on deep knowledge encoded in knowledge graphs.

3.4 Using knowledge graphs for explainable argumentative relation classification

Our intuition is that the knowledge encoded in currently available knowledge graphs such as DBpedia or ConceptNet [50] can provide supporting evidence for argumentative relation classification. That is, we expect that the types of knowledge relations and chains of relations defined in these graphs can be used to connect entities mentioned in argumentative units and can help the system to make sense of arguments such as **Argument 2**:

Argument 2 (*alternative treatments ii*): *Alternative treatments should be subsidized in the same way as conventional treatments since both methods can lead to the prevention, mitigation or cure of an illness.*

Argument 2 implies that "alternative treatments" and "conventional treatments" are both subcategories of medicine, therefore justifying the analogy revealed by the discourse indicators (*in the same way*; *both*). As shown in Fig. 3, background knowledge available in DBpedia reveals the relation between alternative and conventional medicine, as they are both concepts that belong to subcategories of dbc:Medicine, as well as the difference between them as on the one side dbr:Conventional_medicine belongs to category dbc:Health_science and on the other side dbr:Alternative_medicine belongs to category dbc:Pseudoscience.

Following this intuition, in [23] we investigate the **use of knowledge graphs** such as DBpedia and ConceptNet **for argumentative relation classification**.

We design a system that processes two argument units with a Bi-LSTM, that collects all path patterns that connect entities mentioned in them up to a certain length, and use them as features that we incorporate in the Bi-LSTM classifier, as shown in Fig. 4. The use of these knowledge graph based features consistently improves the argumentative relation classification, and the results convey that the knowledge hosted in the two resources are complementary. Nonetheless, our experiments also revealed limitations of the approach: first, the employed knowledge graph extraction procedure is relatively unconstrained, which can lead to contextually unrelated connecting paths; we also found that the extracted knowledge graph paths can be noisy, due to the nature of community-created resources; finally, the aggregation features computed from the graphs cannot be readily used to derive meaningful human-acceptable explanations for the system classifications.

In recent work [38], we improved the linking of argumentative units to the background knowledge graph by creating a dense subgraph similar to own prior work in [37] and designed a novel gated architecture for knowledge path injection on top of a neural argumentative relation base classifier. The system performs *relevance-based* knowledge path extraction and selection using graph-based unsupervised methods and employs *self- and cross-attention mechanisms* to better take into account the meaning of the argumentative unit pairs. Commonsense knowledge paths are

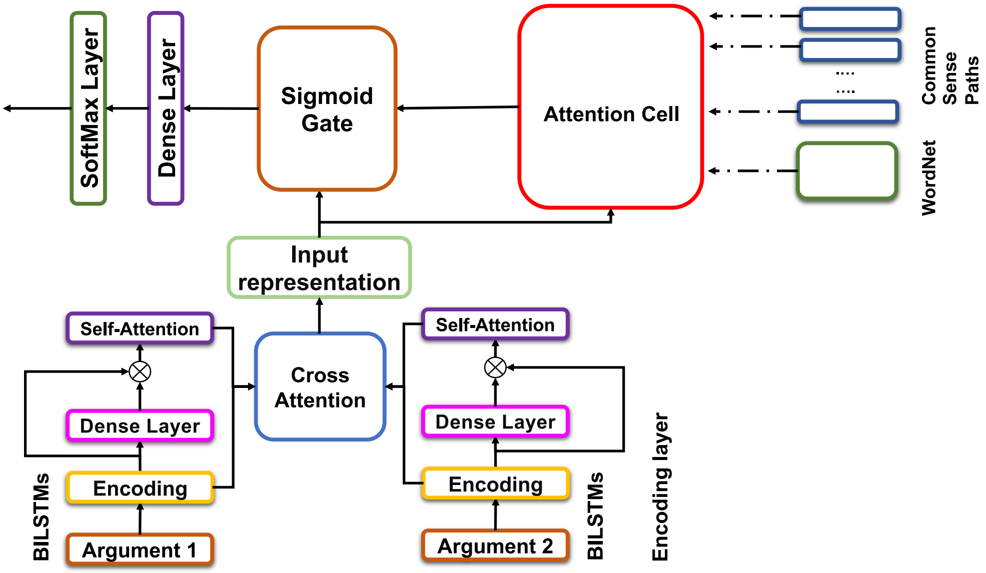

Fig. 5 Attention-based Argumentative Relation Classification with Knowledge Path Injection (ARK) [38]

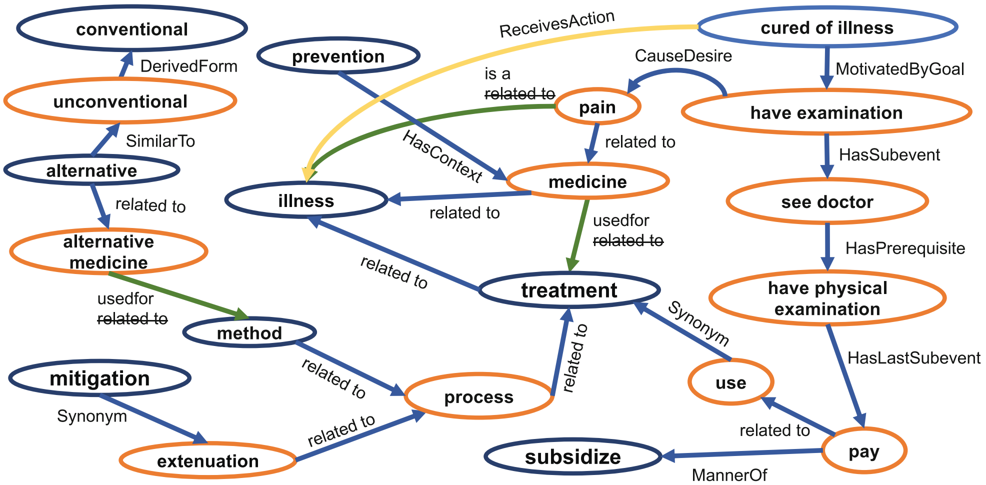

Fig. 6 Subgraph extracted from ConceptNet for AT2. Concepts from the text are in *blue*; intermediate nodes in *orange*. *Blue edges* portray relevant knowledge paths from ConceptNet; edges we add with *on-the-fly* KB completion are *yellow*. We replace *'related to'* relations with predicted specific ConceptNet relations (*green*) [38]

obtained from ConceptNet as a static resource and via *on-the-fly knowledge base completion* of ConceptNet relations using a multi-class relational classifier [5] that we describe in Sect. 4, as well as definitional knowledge from WordNet. An overview of the model is given in Fig. 5; Fig. 6 displays the constructed knowledge subgraph for **Argument 2** (AT2).

With this system we run experiments on two datasets with different types of instance pairs: argument-topic relations from Debatepedia[1], and argument-argument relations from the Student Essays [52]. Our enhanced model shows strong improvements over knowledge-agnostic baselines, both for the newly designed neural model (+4.43 percentage points) and a baseline linear SVM classifier (+2.48 percentage points). Ablation experiments show that (i) the relevance selection model improves over random path selection; (ii) adding on-the-fly commonsense knowledge completion improves over using the static ConceptNet knowledge resource; and (iii) definitional knowledge from WordNet yields additional improvement. Overall we show that targeted knowledge injection is useful for argumentative relation classification. We also show that background knowledge improves model performance across different topics.

4 Background Knowledge for Enthymeme Reconstruction

4.1 Implicit Knowledge in Argumentative Texts

In everyday communication as well as in written texts people omit information that seems clear and evident, such that only part of the message needs to be expressed in

[1] We publicize this dataset as a community resource: https://explain.cl.uni-heidelberg.de/.

words [19]. While such information can be easily filled in by the hearer, a computational system typically does not possess the knowledge that is needed to reconstruct the implied information. Especially in argumentative texts it is very common that premises are implied and omitted [4, 22, 43]. These arguments are called *enthymemes*. Thus, the logic of an argument is in general not fully recoverable from what is explicitly said. Regarding our task of argument explicitation, dealing with enthymemes is one of the core challenges. While explicitation based on Toulmin's model [54] or Walton schemes [57] may be regarded as a tangible aim as long as the problem of implied premises is ignored, we argue that most (informal) natural language arguments are enthymemes, and their explicitation, which includes reconstruction, should not be neglected. The problem of enthymeme reconstruction is arguably an AI-complete problem. Broadly, a system tackling enthymeme reconstruction – called an enthymeme machine [57] – must be able to answer three questions: (i) is the analyzed argument an enthymeme? (ii) which are the gaps that need to be filled? (iii) which are the missing premises? Approaches for addressing questions (i) and (ii) depend on the chosen argument model (e.g., Walton scheme or Toulmin model). Addressing question (iii) is more challenging, since it can only be achieved with respect to (real-world) knowledge available to the system. Such real-world knowledge can be: (i) encyclopedic (e.g., *The dog was the first animal to be domesticated*) which is available online through Wikipedia and related structured knowledge bases such as DBpedia, Wikidata, Yago; (ii) ontological (e.g., *dogs are animal life*) which is available for instance through taxonomies and lexicons such as WordNet, as well as Wikipedia-based knowledge bases; (iii) contextual, such as the purpose of the document, the author, the time, etc., and (iv) common sense knowledge (e.g., *dogs usually bark when strangers enter their space*) which is much harder to source. While the first two types of real-world knowledge can be accessed with state-of-the-art entity linking tools, the last two types of knowledge are more challenging, and are in general much less researched. The ExpLAIN project focuses in particular on reconstructing the latter – commonsense knowledge – and investigates its role in argumentative texts. Our starting point are lessons learned from human-generated data that reconstructs missing and implied information in argumentative texts.

4.2 Learning from Human-Generated Data

In a recent annotation project [2, 4] on argumentative texts, we elicitated high-quality human annotations of implied information in the form of simple natural language sentences. The annotations were performed on pairs of argumentative units from the Microtexts Corpus [39], a concise and focused argumentation dataset that is annotated with argumentative components and relations such as *support, rebuttal* or *undercut*. Annotators were asked to add the information that makes connections between given unit pairs explicit, using short and simple sentences. We designed an annotation process that involves several steps to promote annotator agreement and that allows us to monitor its evolution using textual similarity measures [4]. Fig. 7 shows two examples of such annotations: in the first, the main claim *1-a* is attacked by statement *1-b*; in the second, premise *2-b* supports the main claim *2-a*. In both cases, the knowledge underlying the connection between the main claim and the premise is made explicit in clause *c*, by insertion of one in the first two sentences in the second example.

(1-a) *BER should be re-conceptualized from scratch*
(1-b) *even if billions of Euros have already been invested in the existing airport project.*
(1-c) *BER is an airport.*

(2-a) *Capital punishment is not a solution*
(2-b) *as it cannot be ruled out that the judicial process may make mistakes.*
(2-c-I) *In a judicial process it is decided about capital punishment.*
(2-c-II) *Mistakes don't lead to solutions.*

Fig. 7 Annotations that explicate implicit knowledge (c) that connects argumentative units (a & b)

(I) *Fees result in longer durations of studies.*
Annotation: CAUSES *(fees, longer durations of studies)*

(II) *Dog dirt is disgusting and a hygiene problem.*
Annotation: HASPROPERTY *(dog dirt, disgusting)/*
ISA *(dog dirt, hygiene problem)*

Fig. 8 Sentences annotated with ConceptNet relations

To learn more about the nature and characteristics of the inserted information and the overt argumentative texts, we further annotated the data with two specific semantic information types which we found to be characteristic for argumentative texts in two studies [3, 4]: (I) **Semantic clause types**, from which we adopted the most frequent types in [17] (*states, events, generic sentences*, and *generalizing sentences*), and (II) **ConceptNet commonsense knowledge relations** [49, 50], which comprise an inventory of 37 relations, some of which are commonly used in other resources like WordNet (e.g., *IsA, PartOf*) while most others are targeted for capturing commonsense knowledge and as such are particular to ConceptNet (e.g., *HasPrerequisite, MotivatedByGoal*). Two example annotations from our dataset are given in Fig. 8.

Analysis. An in-depth analysis of our annotated German and English data [2, 4] helped us gain insights into the properties of both argumentative texts and implicit knowledge in terms of structural features and semantic information: We found, e.g., that *generic sentences* are predominant in inserted sentences, indicating the relevance of generic knowledge for implicit information. Almost all sentences in our data – both Microtexts and inserted information – could be mapped to commonsense knowledge relations, which highlights the fact that knowledge bases such as ConceptNet play an important role in argument analysis and are an important source for retrieving implicit knowledge.

Further correlation analysis revealed a number of properties that can guide future systems for automatic reconstruction of implicit information: we found, e.g., that more inserted sentences are needed when no direct argumentative relation holds between units, and that complex argumentative relations such as *undercut* require more explications than other relation types.

Correlation analysis further identified dependencies between the structure of an argument and the type of knowledge that connects argument pairs: we found, e.g., that *states* most often occur between units that are adjacent, while *events* are frequently used for connecting argumentatively related units. Finally, we revealed the importance of causal explanations (as expressed by the relation *causes*) as implied knowledge between supporting argument units, along with *generics*.

These insights can inform our future steps towards knowledge-driven automated argument analysis: Reconstructing implicit information can make the underlying logics of an argument transparent for computational systems and can be useful for assessing the strength of an argument. By exploiting the observed characteristics of the manually added implicit information – characteristic semantic clause types and commonsense knowledge relations in specific argumentative contexts – we can guide the process of reconstructing implicit information in argumentative texts automatically.

We addressed this next step – from manual towards computational reconstruction of implicit knowledge – by developing classifiers for (I) Semantic Clause Type and (II) ConceptNet Knowledge Relation Prediction – two semantic information types which our analysis has shown to be characteristic for (reconstructed) implicit knowledge in argumentative texts.

4.3 Classifying Semantic Clause Types

Detecting aspectual properties of clauses in the form of semantic clause types has been shown to depend on a combination of syntactic, semantic and contextual features. We explore the task in a deep-learning framework, where tuned word representations capture lexical, syntactic and semantic features [6, 7]. Given a clause in its context (previous clauses and previously predicted labels), the model predicts its semantic type (i.e., *state, event, generic, generalizing sentence*). We apply a Gated Recurrent Unit (GRU)-based architecture that is well suited to modeling long sequences.

This initial model jointly models *local and contextual* information in a unified architecture and is further enhanced with an attention mechanism that encodes which parts of the input contain relevant information and has shown to be beneficial for sentence classification [59] and sequence modeling [14]. Our model implicitly captures task-relevant features and avoids the need to reproduce explicit linguistic features for other languages, as it can tune pre-trained embeddings to encode semantic information relevant for the task. It is therefore easily transferable to other languages. We performed experiments for both English and German that achieve competitive accuracy (English: 72.04, German: 74.92) compared to knowledge-rich feature-based models [18].

4.4 Classifying Commonsense Knowledge Relations

Motivated by our analysis of the nature of reconstructed implied knowledge in arguments, we developed a model for classifying commonsense knowledge relations as represented in ConceptNet. Here the task is to predict one (or several) commonsense relations from a set of relation types that hold between two given concepts [5]. We took into account the *specific properties* of ConceptNet knowledge relations, which can make relation classification difficult: a given concept pair can be linked by multiple relation types, and relations can have multi-word arguments of diverse semantic types.

We designed a multi-label classifier (COREC)[2] which uses RNNs for representing multi-word arguments and individually tuned thresholds for improving model performance, especially for relations with unfavorable properties such as long arguments, relation ambiguity and inner-relation diversity. Our best model achieves F1 scores of 0.68 in an open and 0.71 in a closed world setting. For further improvement of the classifier we restructured the relation space by separating ambiguous relations, and add pre- and postfiltering of concepts to reduce uninformative instances. These modifications improve the classifier performance by +9 pp. to 0.77 F1 score. The analysis of the results in different configurations shows that the design decisions driven by multi-word representations and threshold tuning improves the overall classification performance, and that our model is able to tackle specific properties of ConceptNet.

[2] Available at https://gitlab.cl.uni-heidelberg.de/mbecker/corec-commonsense-relation-classifier.

4.5 Commonsense Knowledge Base Completion

Knowledge resources are known to be incomplete, and we expect them to be more effective if they can be (dynamically) enriched, *on the fly*, to offer extended coverage for novel datasets.

In the ExpLAIN project, we investigate several ways for enriching knowledge bases. We research methods for *targeted information extraction* that use patterns in the knowledge graphs to form more specific queries for completing relation triples [61]. Furthermore, we analyze *link prediction methods* for knowledge base completion, including studies on the impact of different ways of performing negative sampling [24] and novel ways of representing knowledge graphs in a more compact, abstract way by combining nodes and edges [30]. Finally, we use our COREC classifier to predict missing knowledge relations for enriching ConceptNet in a *dynamic, task-targeted* way: in the argument relation prediction task (Sect. 3), we predict commonsense knowledge relations that are not yet defined in ConceptNet between pairs of concepts appearing in pairs of argumentative units, and insert the *dynamically predicted relations* in the subgraphs created for knowledge path extraction. This improves our model scores for Student Essays and Debatepedia.

Having shown the effectiveness of *on-the-fly* knowledge base completion for argumentative relation classification, our next step is to apply COREC to the automatic reconstruction of implicit knowledge in model-based micro-explicitation.

This can be achieved straigthforwardly, by applying COREC to predict knowledge relations between concepts stemming from different argumentative units. While this works well for establishing direct links, it can be computationally prohibitive for multi-hop relation paths. Nevertheless, COMET [8], a pretrained language model, is able to perform target concept prediction for commonsense knowledge relations, given a source concept and a known relation type, and can thus be applied to predict multi-hop paths between concepts by generating intermediate nodes. Since COMET is trained on a language model, we expect it to host knowledge that is complementary to COREC. We therefore plan to combine them for improved accuracy.

5 Towards Argumentation Machines with Deep Natural Language Understanding Capacity

5.1 Deep linguistic knowledge for argument analysis

Our systems for interpretable argument analysis mainly build on neural architectures that we link to knowledge graphs and a system for knowledge base completion.

In [32] we explored how to leverage BERT, a contextualized language model, for argument relation prediction. BERT has been shown to host rich linguistic knowledge, including knowledge we find in background knowledge bases [41]. Yet, despite being based on deep learning methods, our work targets *Argument Explicitation* and is grounded in symbolic knowledge representation frameworks.

We believe that – beyond grounding argumentative texts in external background knowledge graphs – our framework can be further strengthened by representing the linguistic meaning of argumentative texts using *deep structured representations*. We therefore explore *Abstract Meaning Representation* (AMR) [1], a framework that represents the meaning of sentences as graphs that capture rich semantic structure.

Current research develops neural systems for AMR parsing [60], but these are still prone to errors. To obtain better control of the quality of AMR parses, we developed a system that performs a *multi-variate quality assessment of AMR graphs* [34], by predicting fine-grained AMR accuracy metrics [12]. Our system predicts AMR parse accuracies with 0.78 Pearson's ρ to gold scores. It allows us to i) assess whether an automatic AMR parse is trustworthy, ii) select the best candidate parse returned by alternative systems, and iii) create better automatically parsed "silver" training data in different domains, to improve out-of-domain AMR parsing quality as a basis for argument analysis.

Another issue is that measuring AMR parse quality is difficult in general. AMRs are usually evaluated using the Smatch metrics [11].

However, Smatch performs symbolic matching and cannot see that, e.g., the concepts *foe* and *enemy* are similar. We developed S^{2}match [36], a metric that is similar to Smatch, but allows for non-symbolic matching and graded meaning assessment of AMR graphs, to better assess parser performance.

In future work we aim to extend this work and parse argumentative texts to semantic graph representations, and to integrate them with background knowledge using our existing neural system. This will allow us to better explain the overt and implicit meaning of arguments with interpretable, structured meaning representations.

5.2 More Challenges for Argumentation Machines

A classical AI benchmark that tests human-like language understanding is the Winograd Schema Challenge (WSC) [27]. It assesses the capacity of systems to perform commonsense reasoning and confronts them with difficult pronoun resolution problems (e.g., ***The trophy***$_1$ *didn't fit in the suitcase*$_2$ *because* ***it***$_{1/2}$ *was too large*). This challenge is tough for machines, but easy for humans. To address this task, we designed a ranking approach [33] that incorporates linguistic features, for example, connotation frames [44] that project subjective roles induced by a given predicate (e.g., the subject of *to violate* is negatively connotated).

We believe that robust argumentation systems must be proficient in answering WSC questions and thus consider the Winograd Schema Challenge as an additional benchmark to assess the language understanding capacities of commonsense-endowed argumentation systems.

6 Conclusion

We presented the major research themes we explore in RATIO's ExpLAIN project and our contributions achieved to date. We defined the task of *Argument Explicitation*, which frames our vision of an explainable, knowledge-based argumentation machine. We presented two major contributions, which we situate in the KAME framework. Within the task of *Acceptability-based Explicitation* we propose models for *argumentative relation classification* that we ground in background knowledge sources and show that the injection of knowledge increases system performance. The injection of knowledge paths allows us to interpret the system's outputs, for enhanced *explainability*. The task of *Enthymeme Reconstruction* is situated in the more fine-grained *Model-based Explicitation* subtask. Here we identified types of knowledge that are frequently found to be implicit in arguments and developed a dynamic commonsense knowledge relation prediction system that is shown to enhance argument relation classification. Our future work will address the full-fledged Enthymeme reconstruction task, building on our developed systems. We will also investigate possible benefits of deep linguistic analysis.

Acknowledgements This work has been partially funded by the German Research Foundation through the project *ExpLAIN. Between the Lines–Knowledge-based Analysis of Argumentation in a Formal Argumentation Inference System*, FR 1707/4-1 and STU 266/14-1 as part of the RATIO Priority Program, by the *Leibniz ScienceCampus: Empirical Linguistics & Computational Language Modeling*, supported by Leibniz Association grant no. SAS2015-IDS-LWC and by the Ministry of Science, Research and Art of Baden-Wurttemberg, as well as German Research Foundation through the Research Training Group *AIPHES: Adaptive Preparation of Information from Heterogeneous Sources*, GRK 1994/1.

Funding Open Access funding provided by Projekt DEAL.

References

1. Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N (2013) Abstract Meaning Representation for Sembanking. In: Linguistic Annotation Workshop and Interoperability with Discourse Sofia, pp 178–186
2. Becker M, Korfhage K, Frank A (2020) Implicit knowledge in argumentative texts: an annotated corpus. In: Proceedings of LREC
3. Becker M, Palmer A, Frank A (2016) Argumentative texts and clause types. In: Workshop on argument mining Berlin. ArgMining 2016, pp 21–30
4. Becker M, Staniek M, Nastase V, Frank A (2017) Enriching argumentative texts with implicit knowledge. In: NLDB, LNCS
5. Becker M, Staniek M, Nastase V, Frank A (2019) Assessing the difficulty of classifying conceptnet relations in a multi-label classification setting. In: RELATIONS – workshop on meaning relations between phrases and sentences (co-located with IWCS)
6. Becker M, Staniek M, Nastase V, Frank A (2019) Classifying semantic clause types with recurrent neural networks: analysis of attention, context and genre characteristics. In: TAL
7. Becker M, Staniek M, Nastase V, Palmer A, Frank A (2017) Classifying semantic clause types: modeling context and genre characteristics with recurrent neural networks and attention. In: Proceedings of *SEM
8. Bosselut A, Rashkin H, Sap M, Malaviya C, Asli C, Yejin C (2019) COMET: commonsense transformers for automatic knowledge graph construction. In: ACL
9. Botschen T, Sorokin D, Gurevych I (2018) Frame- and entity-based knowledge for common-sense argumentative reasoning. Argmining Workshop. Proceedings of the 5th Workshop on Argument Mining (Argmining), pp 90–96
10. Cabrio E, Villata S (2012) Combining textual entailment and argumentation theory for supporting online debates interactions. In: Proc. of ACL, pp 208–212
11. Cai S, Knight K (2013) Smatch: an evaluation metric for semantic feature structures. In: Proc. of ACL
12. Damonte M, Cohen SB, Satta G (2017) An incremental parser for abstract meaning representation. In: Proc. of EACL Valencia, pp 536–546
13. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proc. of NAACL-HLT. https://doi.org/10.18653/v1/N19-1423
14. Dong L, Lapata M (2016) Language to logical form with neural attention. In: Proc. of ACL, pp 33–43

15. Dung PM (1995) On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif Intell 77(2):321–357
16. Feng VW, Hirst G (2011) Classifying arguments by scheme. In: Proc. of ACL Portland, pp 987–996
17. Friedrich A, Palmer A (2014) Automatic prediction of aspectual class of verbs in context. In: Proc. of ACL
18. Friedrich A, Palmer A, Pinkal M (2016) Situation entity types: automatic classification of clause-level aspect. In: Proc. of ACL, pp 1757–1768
19. Grice HP (1975) Logic and conversation. Speech acts. In Cole P, Morgan J (eds) Syntax and Semantics, Vol 3. Academic Press, New York, pp 41–58
20. Habernal I, Gurevych I (2016) What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In: Proc. of EMNLP, pp 1214–1223
21. Huber L, Toussaint Y, Roze C, Dargnat M, Braud C (2019) Aligning discourse and argumentation structures using subtrees and redescription mining. In: Workshop on argument mining Florence, pp 35–40
22. Hulpuş I, Kobbe J, Becker M, Opitz J, Hirst G, Meilicke C, Nastase V, Stuckenschmidt H, Frank A (2019) Towards explaining natural language arguments with background knowledge. In: Workshop on semantic explainability (co-located with ISWC)
23. Kobbe J, Opitz J, Becker M, Hulpuş I, Stuckenschmidt H, Frank A (2019) Exploiting background knowledge for argumentative relation classification. In: LDK Dagstuhl. vol 70, pp 8:1–8:14
24. Kotnis B, Nastase V (2018) Analysis of the impact of negative sampling on link prediction in knowledge graphs. In: Proc. KBCOM
25. Lawrence J, Reed C (2015) Combining argument mining techniques. In: ArgMining workshop, pp 127–136
26. Lawrence J, Reed C (2016) Argument mining using argumentation scheme structures. In: COMMA
27. Levesque H, Davis E, Morgenstern L (2012) The Winograd schema challenge. In: Principles of knowledge representation and reasoning
28. Levy R, Bilu Y, Hershcovich D, Aharoni E, Slonim N (2014) Context dependent claim detection. In: Proc. of COLING, pp 1489–1500
29. Lugini L, Litman D (2018) Argument component classification for classroom discussions. In: Workshop on argument mining, pp 57–67
30. Nastase V, Kotnis B (2019) Abstract graphs and abstract paths for knowledge graph completion. In: SEM
31. Nguyen HN, Litman DJ (2016) Context-aware argumentative relation mining. In: ACL, pp 1127–1137
32. Opitz J (2019) Argumentative relation classification as plausibility ranking. In: Proc. of KONVENS, pp 193–202
33. Opitz J, Frank A (2018) Addressing the Winograd schema challenge as a sequence ranking task. In: Workshop on language cognition and computational models, pp 41–52
34. Opitz J, Frank A (2019) Automatic accuracy prediction for AMR parsing. In: Proc. of *SEM, pp 212–223
35. Opitz J, Frank A (2019) Dissecting content and context in argumentative relation analysis. In: Workshop on argument mining Florence, pp 25–34
36. Opitz J, Parcalabescu L, Frank A (2020) AMR Similarity Metrics from Principles. TACL (To appear)
37. Paul D, Frank A (2019) Ranking and selecting multi-hop knowledge paths to better predict human needs. In: Proc. of NAACLHLT Minneapolis, pp 3671–3681
38. Paul D, Opitz J, Becker M, Kobbe J, Hirst G, Frank A (2020) Argumentative relation classification with background knowledge. In: Proc. COMMA, to appear
39. Peldszus A, Stede M (2015) An annotated corpus of argumentative microtexts. In: Proc. of the first European conference on argumentation
40. Peldszus A, Stede M (2015) Joint prediction in MST-style discourse parsing for argumentation mining. In: Proc. of EMNLP, pp 938–948
41. Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, Miller A (2019) Language models as knowledge bases? In: Proc. of ACL
42. Pollock JL (1987) Defeasible reasoning. Cogn Sci 11(4):481–518
43. Rajendran P, Bollegala B, Parsons S (2016) Contextual stance classification of opinions: a step towards enthymeme reconstruction in online reviews. In: Workshop on argument mining
44. Rashkin H, Singh S, Choi Y (2016) Connotation frames: a data-driven investigation. In: Proc. of ACL
45. Razuvayevskaya O, Teufel S (2017) Finding enthymemes in real-world texts: a feasibility study. In: Argument & computation
46. Reisert P, Inoue N, Kuribayashi T, Inui K (2018) Feasible annotation scheme for capturing policy argument reasoning using argument templates. In: Workshop on argument mining Brussels, pp 79–89
47. Rinott R, Dankin L, Perez CA, Khapra MM, Aharoni E, Slonim N (2015) Show me your evidence – an automatic method for context dependent evidence detection. In: Proc. of EMNLP, pp 440–450
48. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of EMNLP Seattle, pp 1631–1642
49. Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. AAAI.
50. Speer R, Havasi C (2012) Representing general relational knowledge in conceptnet 5. In: Proc. of LREC
51. Stab C, Gurevych I (2014) Identifying argumentative discourse structures in persuasive essays. In: Proc. of EMNLP, pp 46–56
52. Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Comput Linguist 43:619–659
53. Teufel S (1999) Argumentative zoning: information extraction from scientific text. Ph.D. thesis, Univ. of Edinburgh, Edinburgh
54. Toulmin SE (2003) The uses of argument. Cambridge University Press, Cambridge
55. Wachsmuth H, Stede M, El Baff R, Al Khatib K, Skeppstedt M, Stein B (2018) Argumentation synthesis following rhetorical strategies. In: Proc. of COLING
56. Wachsmuth H, Stein B, Hirst G, Prabhakaran V, Bilu Y, Hou Y, Naderi N, Thijm TA (2017) Computational argumentation quality assessment in natural language. In: Proc. of EACL, pp 176–187
57. Walton D, Reed CA (2005) Argumentation schemes and enthymemes. Synthese 145(3):339–370
58. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: ICLR
59. Wang Y, Huang M, Zhu X, Zhao L (2016) Attention-based LSTM for aspect-level sentiment classification. In: Proc. of EMNLP, pp 606–615
60. Zhang S, Ma X, Duh K, Van Durme B (2019) AMR parsing as sequence-to-graph transduction. In: ACL
61. Zhou M, Nastase V (2018) Using patterns in knowledge graphs for targeted information extraction. In: Proc. KBCOM