Cross-topic Argument Mining from Heterogeneous Sources

Christian Stab[*] and **Tristan Miller**[*†] and **Benjamin Schiller**[*] and **Pranav Rai**[*] and **Iryna Gurevych**[*†]

[*]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
https://www.ukp.tu-darmstadt.de/

[†]Research Training Group AIPHES
Department of Computer Science, Technische Universität Darmstadt
https://www.aiphes.tu-darmstadt.de/

Abstract

Argument mining is a core technology for automating argument search in large document collections. Despite its usefulness for this task, most current approaches are designed for use only with specific text types and fall short when applied to heterogeneous texts. In this paper, we propose a new sentential annotation scheme that is reliably applicable by crowd workers to arbitrary Web texts. We source annotations for over 25,000 instances covering eight controversial topics. We show that integrating topic information into bidirectional long short-term memory networks outperforms vanilla BiLSTMs by more than 3 percentage points in F_1 in two- and three-label cross-topic settings. We also show that these results can be further improved by leveraging additional data for topic relevance using multi-task learning.

1 Introduction

Information retrieval and question answering are by now mature technologies that excel at answering factual queries on noncontroversial topics. However, they provide no specialized support for queries where there is no single canonical answer, as with topics that are controversial or opinion-based. For such queries, the user may need to carefully assess the stance, source, and supportability for each of the answers. These processes can be supported by argument mining (AM), a nascent area of natural language processing concerned with the automatic recognition and interpretation of arguments.

In this paper, we apply AM to the task of *argument search*—that is, searching a large document collection for arguments relevant to a given topic. Searching for and classifying relevant arguments plays an important role in decision making (Svenson, 1979), legal reasoning (Wyner et al., 2010), and the critical reading, writing, and summarization of persuasive texts (Kobayashi, 2009; Wingate, 2012). Automating the argument search process could ease much of the manual effort involved in these tasks, particularly if it can be made to robustly handle arguments from different text types and topics.

But despite its obvious usefulness, this sort of argument search has attracted little attention in the research community. This may be due in part to the limitations of the underlying models and training resources, particularly as they relate to heterogeneous sources. That is, most current approaches to AM are designed for use with particular text types, faring poorly when applied to new data (Daxenberger et al., 2017). Indeed, as Habernal et al. (2014) observe, while there is a great diversity of perspectives on how arguments can be best characterized and modelled, there is no "one-size-fits-all" argumentation theory that applies to the variety of text sources found on the Web.

To approach these challenges, we propose the novel task of topic-based sentential argument mining. Our contributions are as follows: (1) We propose a new argument annotation scheme applicable to the information-seeking perspective of argument search. We show it to be general enough for use on heterogeneous data sources, and simple enough to be applied manually by untrained annotators at a reasonable cost. (2) We introduce a novel corpus of heterogeneous text types annotated with topic-based arguments.[1] The corpus includes over 25,000 instances covering eight controversial topics. This is the first known resource that can be used to evaluate the performance of argument mining methods across topics in heterogeneous sources. (3) We investigate different approaches for incorporating topic information into neural networks and

[1] https://www.ukp.tu-darmstadt.de/sent_am

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3664–3674
Brussels, Belgium, October 31 - November 4, 2018.

show that including the topic vector into the $\boldsymbol{i}$- and $\boldsymbol{c}$-gates of the LSTM cell outperforms common attention-based approaches in two- and three-label cross-topic experiments. (4) We further improve the performance of the modified LSTM cell by leveraging additional data for topic relevance in a multi-task learning setup. (5) In the more challenging setup of cross-topic experiments, we show that our models yield considerably better performance than common BiLSTM models when little data of the target topic is available.

2 Related work

Most existing approaches treat argument mining at the discourse level, focusing on tasks such as segmenting argumentative discourse units (Ajjour et al., 2017; Goudas et al., 2014), classifying the function of argumentative discourse units (for example, as *claims* or *premises*) (Mochales-Palau and Moens, 2009; Stab and Gurevych, 2014), and recognizing argumentative discourse relations (Eger et al., 2017; Stab and Gurevych, 2017; Nguyen and Litman, 2016). These discourse-level approaches address the identification of argumentative structures within a single document but do not consider relevance to externally defined topics.

To date, there has been little research on the identification of topic-relevant arguments for argument search. Wachsmuth et al. (2017) present a generic argument search framework. However, it relies on already-structured arguments from debate portals and is not yet able to retrieve arguments from arbitrary texts. Levy et al. (2014) investigate the identification of topic-relevant claims, an approach that was later extended with evidence extraction to mine supporting statements for claims (Rinott et al., 2015). However, both approaches are designed to mine arguments from Wikipedia articles; it is unclear whether their annotation scheme is applicable to other text types. It is also uncertain that it can be easily and accurately applied by untrained annotators, since it requires unitizing (i.e., finding the boundaries of argument components at the token level). Hua and Wang (2017) identify sentences in cited documents that have been used by an editor to formulate an argument. By contrast, we do not limit our approach to the identification of sentences related to a given *argument*, but rather focus on the retrieval of any argument relevant to a given *topic*. The fact that we are concerned with retrieval of arguments also sets our work apart from the discourse-agnostic stance detection task of Mohammad et al. (2016), which is concerned with the identification of sentences expressing support or opposition to a given topic, irrespective of whether those sentences contain supporting evidence (as opposed to mere statements of opinion).

Cross-domain AM experiments have so far been conducted only for discourse-level tasks such as claim identification (Daxenberger et al., 2017), argumentative segment identification (Al-Khatib et al., 2016), and argumentative unit segmentation (Ajjour et al., 2017). However, the discourse-level argumentation models these studies employ seem to be highly dependent on the text types for which they were designed; they do not work well when applied to other text types (Daxenberger et al., 2017). The crucial difference between our own work and prior cross-domain experiments is that we investigate AM from heterogeneous texts across different *topics* instead of studying specific discourse-level AM tasks across restricted text types of existing corpora.

3 Corpus creation

There exists a great diversity in models of argumentation, which differ in their perspective, complexity, terminology, and intended applications (Bentahar et al., 2010). For the present study, we propose a model which, though simplistic, is nonetheless well-suited to the argument search scenario. We define an *argument* as a span of text expressing evidence or reasoning that can be used to either support or oppose a given topic. An argument need not be "direct" or self-contained—it may presuppose some common or domain knowledge, or the application of commonsense reasoning—but it must be unambiguous in its orientation to the topic. A *topic*, in turn, is some matter of controversy for which there is an obvious polarity to the possible outcomes—that is, a question of being either *for* or *against* the use or adoption of something, the commitment to some course of action, etc. In some graph-based models of argumentation (Stab, 2017, Ch. 2), what we refer to as a *topic* would be part of a *(major) claim* expressing a positive or negative stance, and our *arguments* would be *premises* with supporting/attacking *consequence relations* to the claim. However, unlike these models, which are typically used to represent (potentially deep or complex) argument structures at the discourse level, ours is a flat model that considers arguments in isolation from their surrounding context. A great

topic	sentence	label
nuclear energy	Nuclear fission is the process that is used in nuclear reactors to produce high amount of energy using element called uranium.	non-argument
nuclear energy	It has been determined that the amount of greenhouse gases have decreased by almost half because of the prevalence in the utilization of nuclear power.	supporting argument
minimum wage	A 2014 study [. . .] found that minimum wage workers are more likely to report poor health, suffer from chronic diseases, and be unable to afford balanced meals.	opposing argument
minimum wage	We should abolish all Federal wage standards and allow states and localities to set their own minimums.	non-argument

Table 1: Example annotations illustrating our annotation scheme.

advantage of this approach is that it allows annotators to classify text spans without having to read large amounts of context and without having to consider relations to other topics or arguments.

In this work, we consider only those topics that can be concisely and implicitly expressed through keywords, and those arguments that consist of individual sentences. Some examples, drawn from our dataset, are shown in Table 1. Note that while the fourth example expresses opposition to the topic, under our definition it is properly classified as a non-argument because it is a mere statement of stance that provides no evidence or reasoning.

Data. For our experiments we gathered a large collection of manually annotated arguments that cover a variety of topics and that come from a variety of text types. We started by randomly selecting eight topics (see Table 2) from online lists of controversial topics.[2] For each topic, we made a Google query for the topic name, removed results not archived by the Wayback Machine,[3] and truncated the list to the top 50 results. This resulted in a set of persistent, topic-relevant, largely polemical Web documents representing a range of genres and text types, including news reports, editorials, blogs, debate forums, and encyclopedia articles. We preprocessed each document with Apache Tika (Mattmann and Zitting, 2011) to remove boilerplate text. We then used the Stanford CoreNLP tools (Manning et al., 2014) to perform tokenization, sentence segmentation, and part-of-speech tagging on the remaining text, and removed all sentences without verbs or with less than three tokens. This left us with a raw dataset of 27,520 sentences (about 2,700 to 4,400 per topic).

Annotators classified the sentences using a browser-based interface that presents a set of instructions, a topic, a list of sentences, and a multiple-choice form for specifying whether each sentence is a supporting argument, an opposing argument, or not an argument with respect to the topic. (In preliminary experiments, we presented annotators with a fourth option for sentences that are ambiguous or incomprehensible. However, we found that these constituted less than 1% of the distribution and so mapped all such answers to the "no argument" class.)

Annotation experiments. We tested the applicability of our annotation scheme by untrained annotators by performing an experiment where we had a group of "expert" annotators and a group of untrained annotators classify the same set of sentences, and then compared the two groups' classifications. The data for this experiment consisted of 200 sentences randomly selected from each of our eight topics. Our expert annotators were two graduate-level language technology researchers who were fully briefed on the nature and purpose of the argument model. Our untrained annotators were anonymous American workers from the Amazon Mechanical Turk (AMT) crowdsourcing platform. Each sentence was independently annotated by the two expert annotators and ten crowd workers.

Inter-annotator agreement for our two experts, as measured by Cohen's κ, was 0.721; this exceeds the commonly used threshold of 0.7 for assuming the results are reliable (Carletta, 1996). We proceeded by having the two experts resolve their disagreements, resulting in a set of "expert" gold-standard annotations. Similar gold standards were produced for the crowd annotations by applying the MACE denoising tool (Hovy et al., 2013); we tested various thresholds (1.0, 0.9, and 0.8) to discard instances that could be confidently assigned a gold label. We then calculated κ between the remaining instances in the expert and crowd gold standards. In order to

[2] https://www.questia.com/library/controversial-topics, https://www.procon.org/

[3] https://web.archive.org/

topic	docs	sentences	no argument	support argument	oppose argument
abortion	50	3,929	2,427	680	822
cloning	50	3,039	1,494	706	839
death penalty	50	3,651	2,083	457	1,111
gun control	50	3,341	1,889	787	665
marijuana legalization	50	2,475	1,262	587	626
minimum wage	50	2,473	1,346	576	551
nuclear energy	50	3,576	2,118	606	852
school uniforms	50	3,008	1,734	545	729
total	400	25,492	14,353	4,944	6,195

Table 2: Corpus size and class distribution.

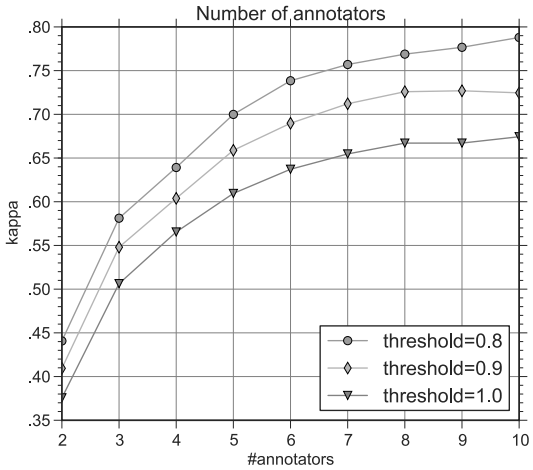

Figure 1: Influence of the number of crowd annotators and different MACE thresholds on κ.

	expert–expert		crowd–expert	
	%	κ	%	κ
abortion	.884	.651	.834	.660
cloning	.845	.712	.821	.704
death penalty	.851	.657	.770	.576
gun control	.907	.783	.796	.638
marijuana legalizat.	.850	.729	.854	.749
minimum wage	.885	.779	.858	.745
nuclear energy	.809	.686	.889	.825
school uniforms	.864	.767	.931	.889
average	.862	.721	.844	.723

Table 3: Agreement between experts, and between the expert and crowd gold standards.

determine the relationship between inter-annotator agreement and the number of crowd workers, we performed this procedure with successively lower numbers of crowd workers, going from the original ten annotators per instance down to two. The results are visualized in Fig. 1. We found that using seven annotators and a MACE threshold of 0.9 results in $\kappa = 0.723$; this gives us similar reliability as with the expert annotators without sacrificing much coverage. Table 3 shows the κ and percentage agreement for this setup, as well as the agreement between our expert annotators, broken down by topic.

We proceeded with annotating the remaining instances in our dataset using seven crowd workers each, paying a rate corresponding to the US federal minimum wage of $7.25/hour. Our total expenditure, including AMT processing fees, was $2,774.02. After MACE denoising, we were left with 25,492 gold-standard annotations. Table 2 provides statistics on the size and class distribution of the final corpus. We are releasing the gold-standard annotations for this dataset, and code for retrieving the original sentences from the Wayback Machine, under a Creative Commons licence.

4 Approaches for identifying arguments

We model the identification of arguments as a sentence-level classification task. In particular, given a sentence $\boldsymbol{\varsigma}$ with words $u_1, \ldots, u_{n_\varsigma}$ and a topic $\boldsymbol{\tau}$ of words $v_1, \ldots, v_{n_\tau}$ (e.g., "gun control" or "school uniforms"), we aim to classify $\boldsymbol{\varsigma}$ as a "supporting argument" or "opposing argument" if it includes a relevant reason for supporting or opposing the $\boldsymbol{\tau}$, or as a "non-argument" if it does not include a reason or is not relevant to $\boldsymbol{\tau}$. We also investigate a two-label classification where we combine supporting and opposing arguments into a single category; this allows us to evaluate argument classification independent of stance. We focus on the challenging task of cross-topic experiments, where one topic is withheld from the training data and used for testing. Here, we denote scalars by italic lowercase letters (e.g., t), vector representations by italic bold lowercase letters (e.g., $\boldsymbol{c}$), and matrices as italic bold uppercase letters (e.g., $\boldsymbol{W}$).

4.1 Integrating topic information

Since arguments need to be relevant to the given topic, we posit that providing topic information to the learner results in a more robust prediction capability in cross-topic setups. Below, we present two models that integrate the topic, one that uses an attention mechanism and another that includes the topic vector directly in the LSTM cell.

Outer-attention BiLSTM (outer-att). To let the model learn which parts of the sentence are relevant (or irrelevant) to the given topic, we use an attention-based neural network (Bahdanau et al., 2014) that learns an importance weighting of the input words depending on the given topic. In particular, we adopt an outer-attention mechanism similar to the one proposed by Hermann et al. (2015), which has achieved state-of-the-art results in related tasks such as natural language inference and recognizing textual entailment (Rocktäschel et al., 2015; Wang and Jiang, 2016). We combine the attention mechanism with a common BiLSTM model and, at time step t, determine the importance weighting for each hidden state $\boldsymbol{h}(t)$ as

$$\boldsymbol{m}(t) = \tanh(\boldsymbol{W}_h \boldsymbol{h}(t) + \boldsymbol{W}_p \boldsymbol{p}) \quad (1)$$

$$f_{attention}(\boldsymbol{h}(t), \boldsymbol{p}) = \frac{\exp(\boldsymbol{w}_m^T \boldsymbol{m}(t))}{\sum_t \exp(\boldsymbol{w}_m^T \boldsymbol{m}(t))} \quad (2)$$

where $\boldsymbol{W}_h$, $\boldsymbol{W}_p$, and $\boldsymbol{w}_m$ are trainable parameters of the attention mechanism and $\boldsymbol{p}$ is the average of all word embeddings of topic words $v_1, \ldots, v_{n_\tau}$. Using the importance weighting, we determine the final, weighted hidden output state $\boldsymbol{s}$ as

$$\alpha_t \propto f_{attention}(\boldsymbol{h}(t), \boldsymbol{p}) \quad (3)$$

$$\boldsymbol{s} = \sum_{t=1}^{n} \boldsymbol{h}(t) \alpha_t. \quad (4)$$

Finally, we feed $\boldsymbol{s}$ into a dense layer with a softmax activation function to get predictions for our two- or three-label setups.

Contextual BiLSTM (biclstm). A more direct approach to integrating an argument's topic is the *contextual LSTM* (CLSTM) architecture (Ghosh et al., 2016), where topic information is added as another term to all four gates of an LSTM cell. We, however, hypothesize that topic information is more relevant at the $\boldsymbol{i}$- and $\boldsymbol{c}$-gates, the former because it has the biggest impact on how a new token is processed and the latter because it is closely linked

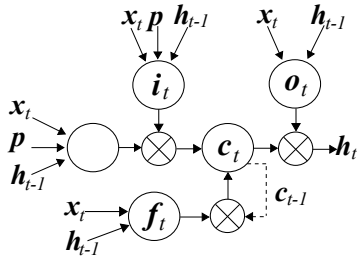

Figure 2: Architecture of a CLSTM cell.

to how the sequence seen so far is to be interpreted and stored. To this end, we experimented with several modifications to the original CLSTM such as removing peepholes—i.e., removing gates' access to the cell state $\boldsymbol{c}$ (Gers and Schmidhuber, 2000)—and removing topic information from one or more gates. Empirical results on the validation set show that topic integration at the $\boldsymbol{i}$- and $\boldsymbol{c}$-gates only, and removal of all peephole connections, does indeed outperform the original CLSTM on our task by 1 percentage point. Our modified CLSTM (Fig. 2) is defined as

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{x_i} \boldsymbol{x}_t + \boldsymbol{W}_{h_i} \boldsymbol{h}_{t-1} + \boldsymbol{b}_i + \boxed{\boldsymbol{W}_{\boldsymbol{p_i}} \boldsymbol{p}}) \quad (5)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{x_f} \boldsymbol{x}_t + \boldsymbol{W}_{h_f} \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \quad (6)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \sigma_c(\boldsymbol{W}_{x_c} \boldsymbol{x}_t + \boldsymbol{W}_{h_c} \boldsymbol{h}_{t-1} + \boldsymbol{b}_c + \boxed{\boldsymbol{W}_{\boldsymbol{p_c}} \boldsymbol{p}}) \quad (7)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{x_o} \boldsymbol{x}_t + \boldsymbol{W}_{h_o} \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad (8)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \sigma_c(\boldsymbol{c}_t). \quad (9)$$

Here $\boldsymbol{i}$, $\boldsymbol{f}$, and $\boldsymbol{o}$ represent the input, forget, and output gates; $\boldsymbol{c}$ the cell memory; $\boldsymbol{x}_t$ the embedded token of a sentence at timestep t; $\boldsymbol{h}_{t-1}$ the previous hidden state; and $\boldsymbol{b}$ the bias. σ and σ_c are the activation and recurrent activation functions, respectively. The novel terms for topic integration are outlined. We use this model bidirectionally, as we did with our BiLSTM network, and hence refer to it as biclstm.

4.2 Leveraging additional data

As we want to classify arguments related to specific topics, leveraging information that supports the classifier in the decision of topic-relation is crucial. The multi-task learning (mtl) and transfer learning (trl) models are able to make use of auxiliary data that can potentially improve the results on the main task. Thus, we extend our previously described models by integrating them into mtl and trl setups. We also choose to integrate two corpora

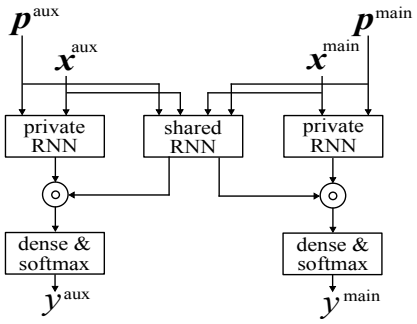

Figure 3: Multi-task learning architecture. The ◎ symbol denotes the concatenation operator.

from which we expect to learn (a) topic-relevance and (b) the capability to distinguish between supporting and attacking arguments. The first corpus, DIP2016 (Habernal et al., 2016), consists of 49 queries from the educational domain and 100 documents for each query. Each document has its sentences annotated for relevance (true/false) to the query.[4] The second corpus, from SemEval-2016 Task 6 (Mohammad et al., 2016), consists of around 5000 multi-sentence tweets, a corresponding topic (e.g., "atheism"), and the author's stance on the topic (for/against/neither).

For our mtl and trl approaches, we consider every possible pairing of a model (biclstm, outer-att, and the bilstm baseline we introduce in §5) with an auxiliary corpus (DIP2016, SemEval). We formalize our datasets as $S_k = \{(\boldsymbol{x}_i^k, \boldsymbol{p}_i^k, y_i^k) | i = 0, \ldots, |S_k|\}$, where k can be either our main dataset or an auxiliary dataset, $\boldsymbol{x}_i^k$ denotes a single sentence as a sequence of word embeddings and y_i^k its corresponding label in k, and $\boldsymbol{p}_i^k$ represents the corresponding averaged topic vector.

Transfer learning (trl). For trl, we use the approach of *parameter transfer* (Pan and Yang, 2010)—i.e., we do not modify the model used. Instead, we train the model twice: the first time, we train the model on the chosen auxiliary corpus, and the second time, we keep the trained model's weights and train it with our own corpus. For the three-label setting, we have to modify the transfer model slightly for the DIP2016 corpus, since it provides only two labels for each training sample. In this case, we simply add a layer with two neurons on top of the layer with three neurons for training with the DIP2016 corpus and remove it afterwards for training with our corpus.

[4]We only use 300K of the corpus's 600K samples to ease hyperparameter tuning for our computation-heavy models.

Multi-task learning (mtl). For mtl, we use a shared–private model (Liu et al., 2017), which showed promising results for text classification and word segmentation (Chen et al., 2017). (We also experimented with their adversarial approach to learn topic-invariant features, but abandoned this due to low scores.) The mtl base model consists of a private recurrent neural network (RNN) for both the auxiliary dataset and our dataset, plus a shared RNN that both datasets use (Fig. 3). The last hidden states of the RNNs are concatenated and fed through a dense layer and a softmax activation function. The model is trained in an alternating fashion—i.e., after each epoch the loss for the other dataset is minimized until each dataset has run for the set number of epochs, where the last epoch is always executed on our dataset. At prediction time, only the private RNN trained on our dataset and the shared RNN are used. The core idea is that the shared RNN learns what is relevant for both tasks, while the private ones learn only the task-specific knowledge.

For the cases of mtl+bilstm+*corpus*, mtl+biclstm+*corpus*, and mtl+outer-att+*corpus*, we simply switch the RNN with our bilstm, biclstm, and outer-att, respectively. For mtl+outer-att+*corpus*, we add the outer attention mechanism (see §4.1), modified for use with the mtl model, after each of the private RNNs, while additionally feeding it a second topic vector—the last hidden state of the shared RNN:

$$\boldsymbol{m}(t) = \tanh(\boldsymbol{W}_r \boldsymbol{h}_r(t) + \boldsymbol{W}_s \boldsymbol{h}_s + \boldsymbol{W}_p \boldsymbol{p}) \quad (10)$$

$$f_{attention}(\boldsymbol{h}_r(t), \boldsymbol{h}_s, \boldsymbol{p}) = \frac{\exp(\boldsymbol{w}_m^T \boldsymbol{m}(t))}{\sum_t \exp(\boldsymbol{w}_m^T \boldsymbol{m}(t))} \quad (11)$$

$$\alpha_t \propto f_{attention}(\boldsymbol{h}_r(t), \boldsymbol{h}_s, \boldsymbol{p}) \quad (12)$$

$$\boldsymbol{s} = \sum_{t=1}^{n} \boldsymbol{h}_r(t) \alpha_t \quad (13)$$

where $\boldsymbol{W}_r$, $\boldsymbol{W}_s$, and $\boldsymbol{W}_p$ are trainable weight matrices, $\boldsymbol{h}_r(t)$ is the hidden state of the private bilstm at timestep t, $\boldsymbol{h}_s$ is the last hidden state of the shared model, and $\boldsymbol{p}$ is the average of all word embeddings of topic words $v_1, \ldots, v_{n_\tau}$.

5 Evaluation

To evaluate the robustness of the models, we conduct cross-topic experiments to evaluate how well the models generalize to an unknown topic. To this end, we combine training (70%) and validation

	two labels			three labels				
model	F_1	P_{arg}	R_{arg}	F_1	P_{arg+}	P_{arg-}	R_{arg+}	R_{arg-}
bilstm (baseline)	.6069 ± .0074	.7339 ± .0110	.3844 ± .0122	.3796 ± .0079	.3484 ± .0479	.4710 ± .0210	.0963 ± .0148	.2181 ± .0181
lr-uni (baseline)	.5854 ± .0131	.6519 ± .0093	.3587 ± .0264	.3821 ± .0056	.2782 ± .0293	.4217 ± .0171	.1176 ± .0165	.2119 ± .0203
outer-att	.6213 ± .0106	.7309 ± .0108	.4138 ± .0237	.3873 ± .0081	**.3651** ± .0244	.4696 ± .0169	.1042 ± .0173	.2381 ± .0117
biclstm	.6414 ± .0129	.6244 ± .0132	**.7035** ± .0261	.4242 ± .0122	.2675 ± .0148	.3887 ± .0141	**.2817** ± .0369	.4028 ± .0496
tr+bilstm+semeval	.6297 ± .0073	**.7500** ± .0047	.4233 ± .0125	.3698 ± .0142	.3128 ± .0422	.4075 ± .0640	.0897 ± .0256	.2089 ± .0133
tr+outer-att+semeval	.6293 ± .0057	.7297 ± .0122	.4336 ± .0156	.3871 ± .0089	.3160 ± .0397	.4469 ± .0369	.1245 ± .0160	.2264 ± .0147
tr+biclstm+semeval	.6433 ± .0182	.6625 ± .0128	.6181 ± .0259	.3953 ± .0122	.2606 ± .0356	.4226 ± .0203	.1743 ± .0385	.3643 ± .0574
tr+bilstm+dip2016	.6254 ± .0133	.7073 ± .0114	.4200 ± .0253	.3628 ± .0136	.2396 ± .0605	.4470 ± .0319	.0517 ± .0284	.2298 ± .0245
tr+outer-att+dip2016	.6074 ± .0115	.7112 ± .0245	.4031 ± .0238	.3438 ± .0233	.2060 ± .1012	.4171 ± .0521	.1105 ± .0821	.2096 ± .0793
tr+biclstm+dip2016	.6110 ± .0206	.6954 ± .0491	.4904 ± .0502	.3595 ± .0226	.2272 ± .0516	.3474 ± .0539	.1191 ± .0856	.2886 ± .0714
mtl+bilstm+semeval	.6126 ± .0093	.7270 ± .0087	.3906 ± .0177	.3765 ± .0081	.3248 ± .0304	**.4812** ± .0340	.0888 ± .0137	.2153 ± .0162
mtl+outer-att+semeval	.6221 ± .0100	.7186 ± .0123	.4219 ± .0187	.3764 ± .0071	.3185 ± .0393	.4763 ± .0213	.0878 ± .0173	.2149 ± .0295
mtl+biclstm+semeval	.6519 ± .0079	.6495 ± .0143	.6690 ± .0333	.4147 ± .0105	.2769 ± .0332	.3819 ± .0141	.2465 ± .0497	**.4069** ± .0501
mtl+bilstm+dip2016	.6145 ± .0097	.7312 ± .0100	.3979 ± .0208	.3757 ± .0057	.3255 ± .0382	.4647 ± .0255	.0841 ± .0144	.2261 ± .0192
mtl+outer-att+dip2016	.6263 ± .0079	.7176 ± .0100	.4327 ± .0178	.3842 ± .0070	.3427 ± .0365	.4502 ± .0240	.1007 ± .0147	.2327 ± .0146
mtl+biclstm+dip2016	**.6662** ± .0148	.6463 ± .0105	.6719 ± .0489	**.4285** ± .0139	.2947 ± .0383	.3815 ± .0221	.2722 ± .0582	.3483 ± .0528

Table 4: Results for each model on the test sets. Bold numbers indicate the highest score in the column.

data (10%) of seven topics for training and parameter tuning, and use the test data (20%) of the eighth topic for testing. For encoding the words of sentence $\boldsymbol{\varsigma}$ and topic $\boldsymbol{\tau}$, we use 300-dimensional word embeddings trained on the Google News dataset by Mikolov et al. (2013). To handle out-of-vocabulary words, we create separate random word vectors for each.[5]

Since reporting single performance scores is insufficient to compare non-deterministic learning approaches like neural networks (Reimers and Gurevych, 2017), we report all results as averages over ten runs with different random seeds. As evaluation measures, we report the average macro F_1, as well as the precision and the recall for the argument class (P_{arg}, R_{arg}). For the three-label approach, we split the precision and recall for predicting supporting (P_{arg+}, R_{arg+}) and attacking arguments (P_{arg-}, R_{arg-}). As baselines, we use a simple bidirectional LSTM (Hochreiter and Schmidhuber, 1997), as well as a logistic regression model with lowercased unigram features, which has been shown to be a strong baseline for various other AM tasks (Daxenberger et al., 2017; Stab and Gurevych, 2017). We refer to these models as bilstm and lr-uni, respectively. All neural networks are trained using the Adam optimizer (Kingma and Ba, 2015) and cross-entropy loss function. For finding the best model, we run each for ten epochs and take the best model based on the lowest validation loss. In addition to that, we tune the hyperparameters of all neural networks (see Appendix A). To accelerate training, we truncate sentences at 60 words.[6]

5.1 Results

Two-label setup. The results in Table 4 show that all our models outperform the baselines for two-label prediction.[7] F_1 for biclstm improves by 3.5 percentage points over the bilstm baseline and by 5.6 over lr-uni. A main reason for this proves to be the substantial increase in recall for our topic-integrating models—outer-att and especially biclstm—in comparison to our baselines. These results show that knowledge of the argument's topic has a strong impact on argument prediction capability. Further, we observe that integrating biclstm in a multi-task learning setup in order to draw knowledge about topic relevance from the DIP2016 corpus (mtl+biclstm+dip2016) improves F_1 by an additional 2.5 percentage points. It achieves an F_1 of 0.6662, which is 19.48 percentage points less than the human upper bound of 0.861. When using the SemEval corpus, which holds less task-relevant knowledge for our two-label approach, we are able to gain only 1 percentage point when integrating it into mtl+biclstm+*corpus*.

For the transfer learning models that integrate the topic (tr+biclstm+*corpus* and tr+outer-att+*corpus*), the parameter transfer is mostly ineffective. If no topic is provided (tr+bilstm+*corpus*), the transfer learning models are able to improve over the baseline bilstm. This shows that the parameter transfer

[5]Each dimension is set to a random number between −0.01 and 0.01. Digits are mapped to the same random word vector.

[6]Only 244 of our sentences (<1%) exceed this length.

[7]Detailed results per topic are given in Appendix B.

itself can be of use, but confuses the model when combined with topic integration.

In general, we observe an overall lower score for trl models that use the DIP2016 corpus compared to those using the SemEval corpus. In contrast to the mtl model, for trl models *all* parameters are transferred to the main task, not just parameters that represent shared knowledge. Thus, we suspect the lower scores of the trl models with DIP2016 are due to overfitting on the vast number of samples which shape the parameters much more than the comparatively small SemEval corpus could.

Three-label setup. For the three-label approach, we observe overall lower scores due to the additional difficulty in distinguishing supporting from opposing arguments. As already observed in the two-label setup, biclstm outperforms both the bilstm and lr-uni baselines; here, the former by 4.5 and the latter by 4.2 percentage points in F_1. Again, this is caused by a substantial increase in recall and shows the impact that the available topic information has on the classifier's predictive power.

For transfer learning, we see similar results as for the two-label approach; both the DIP2016 and SemEval corpora have a generally negative impact when compared to the respective base models. The SemEval corpus does not provide the knowledge required to distinguish supporting from attacking arguments. We conclude that the original purpose of the SemEval task, stance recognition, is too different from our own. But in multi-task learning, where only the shared parameters are taken, we observe slight improvements when using biclstm with DIP2016; this correlates with the same model in the two-label setup.

5.2 Error analysis

To understand the errors of our best model, mtl-biclstm-dip, and the nature of this task, we manually analyzed 100 sentences randomly sampled from the false positive and false negative arguments of the three-label experiments (combining supporting and attacking arguments). Among the false positives, we found 48 off-topic sentences that were wrongly classified as arguments. The 52 on-topic false positives consist of non-argumentative background information or mere opinions without evidence (as with the first and fourth examples of Table 1) and questions about the topic. Among the false negatives, we found 65 arguments that did not explicitly refer to the topic but to related aspects that depend on background knowledge. For instance, the model fails to establish an argumentative link between the topic "gun control" and the Second Amendment to the US Constitution. Lastly, we inspected arguments that are incorrectly classified as supporting and/or opposing a topic. We found several samples in which the term "against" is not correctly interpreted and the argument is classified as supporting a topic. Similarly, for arguments incorrectly classified as attacking, we find various samples where the word "oppose" is used not to oppose the topic but to strengthen a supporting argument, as in "There is reason even for people who oppose the use of marijuana to support its legalization. . . "

5.3 Adapting to new topics

To evaluate the performance of the models in data-scarce scenarios, we gradually add target topic data to the training data and analyze the model performance on the target test set. Figure 4 shows model performance (F_1, P_{arg}, and R_{arg}) on the "marijuana legalization" topic when adding different amounts of randomly sampled topic-specific data to the training data (x-axes).[8] As the results show, the models that integrate the topic achieve higher recall when adding target topic data to the training data. For bilstm, we observe a drastic difference when compared to the other models; the recall for arguments stays at around 30% and rises only when integrating more than 60% target topic data. In strong contrast, topic-integrating models retrieve a much higher number of actual arguments at target topic augmentation levels as low as 20%. Further, and equally important, this does not come at the cost of precision; on the contrary, the precision is mostly steady and slowly rising after around 20% of target topic integration, leading to an overall higher F_1 for these models. Finally, in comparing F_1 between topic-integrating models and bilstm, we conclude that the former need much less target topic data to substantially improve their score, making them more robust in situations of data scarcity.

6 Conclusion

We have presented a new approach for searching a document collection for arguments relevant to a given topic. First, we introduced an annotation scheme suited to the information-seeking perspec-

[8] Each data point in the plot is the average score of ten runs with different random samples of target topic data.

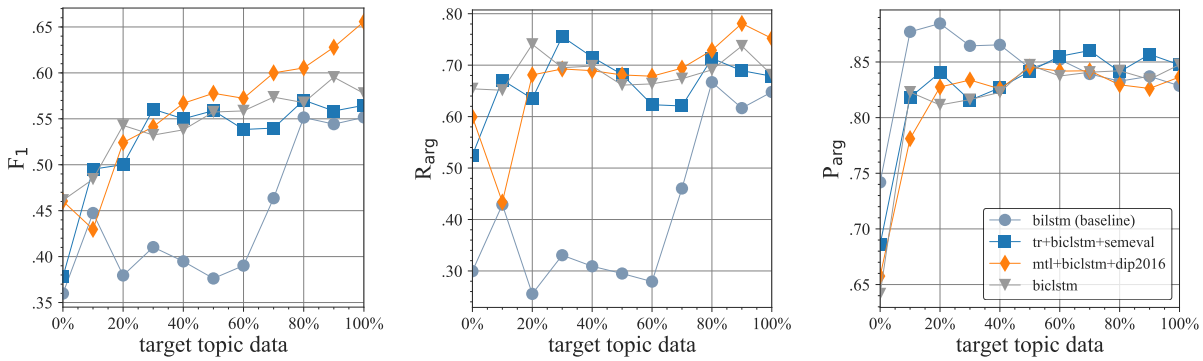

Figure 4: Model performance (y-axes) according to the amount of target topic data in the train sets (x-axes) for the "marijuana legalization" topic in the three-label setup.

tive of argument search and showed that it is cheaply but reliably applicable by untrained annotators to arbitrary Web texts. Second, we presented a new corpus, including over 25,000 instances over eight topics, that allows for cross-topic experiments using heterogeneous text types. Third, we conducted cross-topic experiments and showed that integrating topic information of arguments with our contextual BiLSTM leads to better generalization to unknown topics. Fourth, by leveraging knowledge from similar datasets and integrating our contextual BiLSTM into a multi-task learning setup, we were able to gain an improvement over our strongest baseline of 5.9 percentage points in F_1 in the two-label setup and 4.6 in the three-label setup. Finally, by gradually adding target topic data to our training set, we showed that, when available, even small amounts of target topic data (20%) have a strong positive influence on the recall of arguments.

In a separate, simultaneously written paper (Stab et al., 2018) we evaluate our models in real-world application scenarios by applying them to a large document collection and comparing the results to a manually produced gold standard. An online argument search engine implementing our approach is now available for noncommercial use at `https://www.argumentsearch.com/`. Furthermore, we are experimenting with language adaptation and plan to extend the tool to the German language. Preliminary results are presented in Stahlhut (2018). We also intend to investigate methods for grouping similar arguments.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText) and the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1).

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1193–1203. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? Cross-domain claim identification. In *Proceedings of the 2017 Conference*

on Empirical Methods in Natural Language Processing, pages 2045–2056. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22. Association for Computational Linguistics.

Felix A. Gers and Jürgen Schmidhuber. 2000. Recurrent nets that time and count. In *Proceedings of the IEEE–INNS–ENNS International Joint Conference on Neural Networks*, volume 3, pages 189–194. IEEE.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry P. Heck. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. *CoRR*, abs/1602.06291.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 287–299. Springer International Publishing.

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*, volume 1341 of *CEUR Workshop Proceedings*, pages 26–39. CEUR-WS.org.

Ivan Habernal, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan, and Iryna Gurevych. 2016. New Collection Announcement: Focused Retrieval Over the Web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 701–704. ACM.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting diverse supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 203–208. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Keiichi Kobayashi. 2009. Comprehension of relations among controversial texts: Effects of external strategy use. *Instructional Science*, 37(4):311–324.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1–10.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Chris A. Mattmann and Jukka Zitting. 2011. *Tika in Action*. Manning Publications, Greenwich, CT, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1127–1137. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics.

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*, pages 440–450. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Stab. 2017. *Argumentative Writing Support by Means of Natural Language Processing*. Dr.-Ing. thesis, Technische Universität Darmstadt.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations (NAACL-HLT 2018)*, pages 21–25.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Chris Stahlhut. 2018. Searching arguments in German with ArgumenText. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, volume 2167 of *CEUR Workshop Proceedings*, page 104.

Ola Svenson. 1979. Process descriptions of decision making. *Organizational Behavior and Human Performance*, 23(1):86–112.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451. Association for Computational Linguistics.

Ursula Wingate. 2012. 'Argument!' Helping students understand what essay writing is about. *Journal of English for Academic Purposes*, 11(2):145–154.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence*, pages 60–79. Springer.