Sharper Reasons: Argument Mining Leveraged with Confluent Knowledge

Anonymous ACL submission

Abstract

Relevant to all application domains where it is important to get at the reasons underlying decisions and sentiments, argument mining seeks to obtain structured arguments from unstructured text and has been addressed recently by approaches typically involving some feature and/or neural architecture engineering.

By embracing a transfer learning viewpoint, the aim of this paper is to empirically assess the potential of transferring knowledge learned with confluent tasks to argument mining by means of a systematic study with a wide range of sources of related knowledge possibly suitable to leverage argument mining.

This permitted to gain new empirically based insights into the argument mining task while establishing also new state of the art levels of performance for the three main sub-tasks in argument mining, viz. identification of argument components, classification of the components, and determination of the relation among them, with a leaner approach that dispenses with heavier feature and model engineering.

1 Introduction

Argument mining is a Natural Language Processing task consisting in taking unstructured text as input and returning it annotated such that each portion occurring in it that is an argument is properly delimited and analysed (Schneider et al., 2013; Peldszus and Stede, 2013; Lippi and Torroni, 2016; Habernal and Gurevych, 2017; Wachsmuth et al., 2017; Stede and Schneider, 2018; Lawrence and Reed, 2020). Argument mining relates to the high-level human capacity of reasoning (Walton et al., 2005), it is at the core of social interaction concerned with persuasion (Mercier and Sperber, 2017), and it is of utmost importance to enhance applications across different domains that aim at enhancing their services beyond mere sentiment analysis on the basis of the reasons uncovered for the associated sentiments and decisions (Habernal et al., 2014).

Argument mining has been decomposed into a number of sub-tasks. While the exact number and profiling of these tasks depends on the theoretical approach adopted to analyse arguments (Van Eemeren et al., 2019), they typically involve some sort of delimitation of the text segments conveying argument components, the classification of the roles of these components in the argument (e.g. premises, conclusions, etc.), and the classification of the type of relation among the components (e.g. support, attack, etc.) (Lawrence and Reed, 2020).

These sub-tasks and their eventual pipeline in argument mining have been addressed recently by means of supervised deep learning approaches that involve some degree of neural architecture engineering (Eger et al., 2017; Potash et al., 2017; Nguyen and Litman, 2016) a.o. Recently, first attempts to approach argument mining with Transformers have been reported in the literature (Wang et al., 2020) a.o., tough at an exploratory level that leaves much of its strength still untapped.

This has been combined with experimentation with transfer learning (Caruana, 1997; Ruder, 2019). Given its complexity, and the ensuing difficulty in producing gold labelled data, argument mining is a task with a scarcity of data sets needed to support supervised learning approaches. Enhancing the argument mining task by transferring knowledge elicited while solving other natural language processing (NLP) tasks is thus a promising approach to alleviate such scarceness that has been tried in the literature (Mohammad et al., 2016; Stab et al., 2018; Choi and Lee, 2018; Habernal et al., 2018) a.o., though at a haphazard level that leaves still much of its potential to be studied.

For humans, argumentation is a high level cognitive task that goes together with a number of other capacities relating to linguistic syntactic and semantic processing, entailment and paraphrasing, question answering and language comprehension, reasoning, common sense handling, etc (Lawrence and Reed, 2020; Lauscher et al., 2021). Interestingly, there is now available in the literature a wide range of data sets and respective NLP tasks that permit to address a wide range of these different dimensions and use them as auxiliary sources of knowledge in transfer learning approaches to argument mining (Wang et al., 2018, 2019a) a.o.

In this context, our goal is to empirically assess the potential of transfer learning to support argument mining by means of a systematic study with a wide range of possible sources of related tasks and knowledge possibly suitable to be transferred. In this paper we report on the findings of exploring a vast experimental space that results from: performing sequential single-step transfer learning from over 40 auxiliary tasks to each one of three main sub-tasks of argument mining (Stab and Gurevych, 2014, 2017) during the fine-tuning phase (Section 4); further explore the source tasks that supported the best single-step transfer learning by experimenting with ways of possibly combining them in multi-step transfer learning processes, and further explore these tasks in a multi-task transfer learning setting (Section 5); and perform transfer learning during language modelling in the pre-training phase, without labelled data (Section 6). This is preceded by an overview of related work (Section 2) and the presentation of the experimental setup adopted (Section 3).

By undertaking this study, not only new state-of-the-art results were achieved for the argument mining task, as also new empirically based insights were gained on how this task can be enhanced, showing the effectiveness of transfer learning to leverage argument mining and alleviate its data scarcity with a leaner approach that dispenses with heavier feature and model engineering.

2 Related work

Transfer learning is a machine learning technique that leverages knowledge from multiple source tasks to improve a machine learning generalization of a target task (Caruana, 1997). Being a methodology to alleviate the lack of labelled data for the target task (Ruder, 2019).

2.1 Transfer learning for argument mining

Four families of approaches of transfer learning for argument mining have been reported in the literature: (i) transfer learning across discourse domains for the same argument mining sub-task; (ii) cross-lingual transfer learning for a given sub-task; (iii) multi-task learning among argument mining sub-tasks; and (iv) sequential transfer learning from sources tasks that are not argument mining sub-tasks. A brief overview of them follows below.

Several papers have applied transfer learning with a **domain adaptation** approach for identifying components and clausal properties (Al-Khatib et al., 2016; Ajjour et al., 2017; Daxenberger et al., 2017). Typically, a model is trained with data sets from various discourse domains and is evaluated over each domain.

Cross-lingual transfer learning for argument mining (Aker and Zhang, 2017; Sliwa et al., 2018; Eger et al., 2018; Rocha et al., 2018) is mainly performed through direct transfer (McDonald et al., 2011) or projection (David et al., 2001) techniques. Direct transfer techniques train a model with the source language data that initializes a new model for a target language, typically with less to no data. Projection techniques resort to mapping the same labels from the source language data set to a target language data set by resorting to parallel corpora.

The argument mining pipeline has been addressed also with transfer learning by **multi-task** and **sequential** approaches (Cabrio and Villata, 2013; Peldszus and Stede, 2015; Eger et al., 2017; Potash et al., 2017; Niculae et al., 2017; Galassi et al., 2018; Schulz et al., 2018; Mensonides et al., 2019; Chakrabarty et al., 2019; Accuosto and Saggion, 2019; Cheng et al., 2020). Most papers train models interrelating the sub-tasks in a pipeline.

Transfer learning from **related tasks** has also been shown to improve the performance of argument mining sub-tasks. Stab et al. (2018) transferred shared knowledge from two different tasks: a stance detection task (Mohammad et al., 2016) and a topic identification task. Choi and Lee (2018) transferred knowledge from the Argument Reasoning Comprehension Task (Habernal et al., 2018) for a clausal classification sub-task.

2.2 Main sub-tasks

To proceed with a systematic study of transfer learning for argument mining on a mainstream pipeline of sub-tasks (Lawrence and Reed, 2020), which includes identifying argument components, classifying their clausal roles and determining the relational properties among them, we resorted to the AAEC corpus (Stab and Gurevych, 2014, 2017), a collection of annotated essays, which has been the subject of various studies. An example from this data set is presented in Figure 1.

Title: Children should grow up in a big city!
Essay: It's certainly better for children to grow up in a big city[1]. Of course you need to choose a good neighborhood. I hold this belief because of two main reasons, academic and social reasons[2].
Some people thinks that if a child grows up in a big city they will be all day at home at the computer or at the video-game[6], but this is not true if you live in a neighborhood with other people about your age as I did[7]. My friends and I used to play soccer, bike, climb trees and do a lot of other stuff every day[8]. We did play video-games, but that wasn't our main activity[9]. In a big city there are more kinds of people and more things to do[10].
I have a friend that grew up in the countryside[13]. He said that he had to study a lot to pass the test to enter the university[12]. This is another downside of growing up in the countryside. In a big city you have more qualified teachers and a better access to technology[11].
Growing up in the countryside is not such a good experience[4], you won't know a lot of people, there are gossips everywhere, and your life will be really limited[3]. If someday I have children, I'm absolutely sure that they will grow up in a good neighborhood of a big city and they will be very happy about it[5].

Labels: Major Claim / Claim / Premise
Relations: Support (3→4; 13→11; 12→11) Attacks (7→6; 8→6; 9→6; 10→6)

Figure 1: Example of a labelled essay in AAEC.

The AAEC corpus integrates the annotation of all sub-tasks in a argument mining pipeline in a single data set. It contains 402 manually annotated essays,[1] in English, with a total of 7,116 sentences over 1,833 paragraphs spanning 147,271 tokens.

It adopts an argument structure model in the form of a tree composed of major claim (in the root node, as the author's standpoint on the argument topic), claims and premises. Individual paragraphs of the essay include arguments that may be linked or not-linked (via relational properties) to the author's major claim. Both "support" and "attack" relations were considered.

The annotation of text segments with argument components resorted to an IOB tagging scheme (Ramshaw and Marcus, 1999). The beginning of an argument component is tagged with *Arg-B*, the following tokens in that component are tagged with *Arg-I* and non-argumentative tokens tagged with *O*. Identifying argument components consists of tagging each token with this IOB-tagset given a complete essay as a single input sequence. Identifying clausal properties consists of classifying spans of discourse with one of the three classes (major claim, claim and premise) given an entire essay as input. Following the literature, given the large imbalance among "support" and "attack" classes, identifying relational properties consists in classifying pairs of segments just as linked or not-linked. Statistics are displayed in Table 1.

2.3 Literature on the AAEC tasks

Several papers on argument mining address the AAEC tasks, although none address all of them, except (Stab and Gurevych, 2017), which addressed each task with a feature-engineered SVMs (components: 0.849 macro-F1; clausal: 0.773; relational: 0.736), and an Integer Linear Programming (ILP) algorithm (0.867, 0.826, 0.751 respectively), that is an ensemble of the SVMs models supplemented by rules to ensure the correct tree structure. Table 2 presents the results for the AAEC tasks.

[1] 80 essays, i.e 20% for testing, were annotated by three annotators and the remaining 322, for training, by an expert.

Task		**Labels**	**Total**	Train	Test
Comp.	Arg-B	11%	6,089	79%	21%
	Arg-I	64%	93,618	80%	20%
	O	25%	47,474	80%	20%
Clausal	Major Cl	12%	751	80%	20%
	Claim	25%	1,506	80%	20%
	Premise	63%	3,832	79%	21%
Relat.	Not-Link	82%	18,340	78%	22%
	Linked	18%	3,832	79%	21%

Table 1: For the tasks annotated in AAEC (rows), the number of instances for labels and data set split.

	Comp.	Clau.	Rel.
SVMs (Stab and Gurevych, 2017)	.849	.773	.736
ILP (Stab and Gurevych, 2017)	.867	.826	.751
S2S (Potash et al., 2017)		**.849**	**.767**
BL (Ajjour et al., 2017)	.885		
BL (Eger et al., 2017)	**.908**		
BL (Spliethöver et al., 2019)	.870		
BL-CRF (Petasis, 2019)	.901		
BL-CRF (Schulz et al., 2018)		.606	
BL-CNN-CRF (Chernodub et al., 2019)		.471	
CNN-Seq. (Gemechu and Reed, 2019)		.790	
BERT (Wang et al., 2020)		.640	
LibLINEAR (Nguyen and Litman, 2016)			.753

Table 2: Comparison of different results in the literature on the AAEC tasks, in macro-F1 (except weighted-F1 in (Spliethöver et al., 2019)), with the top results in bold, indicating the state-of-the-art scores (BL stands for BiLSTM). It should be noted that LibLINEAR uses the first version of the AAEC data set.

Regarding the identification of argument **components**: (Ajjour et al., 2017) implement a BiLSTM with extensive use of features and obtain a 0.885 macro-F1 score. (Petasis, 2019) applies several types of neural networks for segmentation, with the top-performing model, a BiLSTM-CRF, obtaining a 0.901 macro-F1. (Spliethöver et al., 2019) resorts to attention mechanisms with BiLSTMs for unit segmentation, with the top-performing model obtaining a 0.87 weighted-F1. (Eger et al., 2017) apply different models, including multi-task learning experiments and report a 0.908 macro-F1 for identifying components.

For identifying **clausal** properties: (Gemechu and Reed, 2019) obtain a 0.79 macro-F1 for clausal properties linking premises and conclusions taking into account the similarity of target concepts and aspects. (Chernodub et al., 2019) applied a framework for tagging arguments and their retrieval, including a BiLSTM-CNN-CRF sequence tagger. A micro-F1 of 0.645 was the top-performing performance in identifying clausal properties (0.471 macro-F1 the reproduction in (Wang et al., 2020)). (Wang et al., 2020) propose a multi-scale mining model, resorting to several encoder-only transformers (BERT) that mine different argumentation components at different textual levels, namely at the essay/paragraph/word-level. The top-performing model obtains 0.64 macro-F1 in identifying clausal properties. (Schulz et al., 2018) also apply a multi-task learning approach from different domains and argumentative structures, including AAEC, with a BiLSTM-CRF, obtaining a 0.606 macro-F1 score.

Finally, as for **relational** properties: (Nguyen and Litman, 2016) obtain a 0.753 macro-F1 combining different topic to window context features with a linear classifier (LibLINEAR). (Potash et al., 2017) report a 0.849 clausal and 0.767 relational macro-F1 using a joint pointer architecture (sequence-to-sequence model with attention), simultaneously addressing clausal and relational properties with several features.

3 Experimental space and settings

For the tasks that are the source of knowledge to be transferred to argument mining models, we resorted to a wide array of annotated data sets, in English, listed in Table 3. They cover different dimensions in terms of linguistic and cognitive processing:

3.1 Source tasks

Syntax - Information on syntax is typically included in structured machine learning algorithms that address the argument mining in a feature engineering approach. We included part-of-speech (POS) tagging, named entity recognition (NER) (Hu et al., 2020) and several other tasks regarding linguistic properties of sentences (Conneau and Kiela, 2018).

Semantics - Features from semantic similarity (SS) are widely used in argument mining literature. For example, (Boltužić and Šnajder, 2015) use SS to identify prominent arguments in online debates, and (Lawrence and Reed, 2015) use SS obtained from WordNet to identify the components of argumentation schemes. We included a diversity of SS data sets, from the context-sensitive similarity task Wic (Pilehvar and Camacho-Collados, 2019) to the large data set obtained from Quora Question Pairs (QQP) (Iyer et al., 2017).

Grammaticality - To address the widest spectrum of linguistic aspects, we included also tasks on determining the grammatically of input sentences. Data sets such as the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) were used, that are challenging with regards this type of task.

Sentiment - Sentiment analysis has a certain proximity to argument mining, which adds an extra dimension to it by providing reasons for sentiments (Habernal et al., 2014). The Stanford Sentiment Treebank (SST) (Socher et al., 2013) was included.

Reasoning & Comprehension - Reasoning is at the core of argumentation given it is crucial in formulating and accepting or rejecting an argument. We included several related tasks, as for instance the AI2 Reasoning Challenge (ARC) (Clark et al., 2018) in the domain of grade-school science.

Question Answering & Common sense - Question Answering (QA) relates to argument mining given linguistic similarities between the Question/Answer and Claim/Premise pairs. Several QA tasks were included that address common sense as this is closely related to argumentation, given that several implicit premises, tacit assumptions or inferences are to some extent regarded as common sense—for example, (Saint-Dizier, 2017) uses QA techniques for argument mining.

Entailment & Paraphrase - Although argument mining and Textual Entailment (TE) are different tasks, they are closely related given the similarity between specific entailment properties and argument clausal and relational properties. Works such as (Cabrio and Villata, 2012; Cocarascu and Toni, 2017) use models for TE to address argument relational properties. We included several TE tasks in different discourse domains, such as news and forums, with STSB (Cer et al., 2017), and science, with SciTAIL (Khot et al., 2018).

Argument mining - In addition to non argument mining tasks, we considered also as a source task the predecessor sub-task in the argument mining pipeline, that is the identification of components (for the clausal sub-task) and the clausal classification (for the relational sub-task).

3.2 Computational models

In order to scan the experimental space setup for our study, we resorted to the Transformer architecture (Vaswani et al., 2017) , which became main-

Task	#Train
Syntax	
PANX (Hu et al., 2020)	20K
UDPOS (Hu et al., 2020)	21K
Bigram Shift (Conneau and Kiela, 2018)	100K
Coord Inversion (Conneau and Kiela, 2018)	100K
Obj number (Conneau and Kiela, 2018)	100K
Odd Man Out (Conneau and Kiela, 2018)	100K
Past-Present (Conneau and Kiela, 2018)	100K
Sentence Length (Conneau and Kiela, 2018)	100K
Subj Number (Conneau and Kiela, 2018)	100K
Top Constituents (Conneau and Kiela, 2018)	100K
Tree Depth (Conneau and Kiela, 2018)	100K
Word Content (Conneau and Kiela, 2018)	100K
Semantics	
COPA (Roemmele et al., 2011)	400
WIC (Pilehvar and Camacho-Collados, 2019)	5.4K
STSB (Cer et al., 2017)	7K
QQP (Iyer et al., 2017)	364K
Grammaticality	
Coord (White et al., 2020)	458
Eos (White et al., 2020)	479
Definiteness (White et al., 2020)	508
Whwords (White et al., 2020)	585
CoLA (Warstadt et al., 2019)	8.5K
Sentiment	
SST (Socher et al., 2013)	67K
Reasoning & Comprehension	
MULTIRC (Khashabi et al., 2018)	456
WNLI (Levesque et al., 2012)	635
ARC (Clark et al., 2018)	2.2K
ROPES (Lin et al., 2019)	10K
ANLI (Bhagavatula et al., 2020)	169.6K
FEVER (Nie et al., 2019)	208.3K
Question Answering & Common sense	
WSC (Levesque et al., 2012)	554
CommonsenseQA (Talmor et al., 2019)	9.7K
QUAIL (Rogers et al., 2020)	10.2K
BoolQ (Clark et al., 2019)	16K
PIQA (Bisk et al., 2020)	16.1K
CosmosQA (Huang et al., 2019)	25K
HellaSwag (Zellers et al., 2019)	39.9K
MRQA (Fisch et al., 2019)	104K
QNLI (Wang et al., 2018)	105K
Entailment/Paraphrase	
CB (De Marneffe et al., 2019)	1.2K
RTE (Dagan et al., 2005)	2.5K
MRPC (Dolan and Brockett, 2005)	3.7K
SciTAIL (Khot et al., 2018)	27K
MNLI (Williams et al., 2018)	393K
Argument mining	
Components (Stab and Gurevych, 2017)	117k
Clausal (Stab and Gurevych, 2017)	4k

Table 3: Data sets used for source tasks, grouped by linguistic and cognitive dimensions related to argumentation.

stream in NLP, surpassing several state-of-the-art results in a wide range of tasks of all sorts (Wang et al., 2018, 2019a). In contrast to most literature on argument mining, where structured feature engineering has been the favoured approach, a transformer is a deep learning approach that obtains linguistic knowledge by transfer learning typically from a language modelling task.

In order to factorize out the impact of different possible models and obtain results that can be comparable across the different data points in our experimental space, we adopt the same type of model for all of them. Taking a look at a task closely related to argument mining, namely common sense reasoning, there are works in the literature (Branco et al., 2021) that, for this task, under comparable circumstance, have experimented with prominent exemplars of encoder-only, decoder-only, encoder-decoder, and neuro-symbolic types of transformers, which found that RoBERTa (Liu et al., 2019) offers a clear advantage. Inspired by these results, we undertook an exploratory study, repeating the above experiments but now for sample cases of argument mining from our experimental space and arrived at the same finding. Accordingly, and given also its accessible compute requirements and top performance in several NLP tasks, we adopted the off-the-shelf RoBERTa model, resorting to RoBERTa-large variant only when the RoBERTa-base was shown not to be enough to beat the SoTA. We used the Jiant framework (Wang et al., 2019b; Phang et al., 2020) and Huggingface (Wolf et al., 2020).The training objective for the pre-training model was the Mask Language Modelling, which randomly masks a word in a sentence and predicts it.

To identify argument components, a token classification head classifies the input sequence $x_{1:N}$ (full essay) and gives a possible output $y_{1:N}$ from a class set C. To identify clausal and relational properties, a sequence classification head classifies each input sequence $x_{1:N}$ and gives a possible output y from a class set C.

3.3 Baselines

As for the baselines, we included the class majority and fine-tuned a RoBERTa-base model for each AAEC task. We also included the SVMs and ILP joint model from (Stab and Gurevych, 2017).

3.4 Evaluation

For the evaluation of the transfer learning, we used the final result of each main argument mining subtask. As in the original AAEC work and given that classes are unbalanced, we used for all tasks a macro-F1 averaging (Sokolova and Lapalme, 2009). We applied the Independent Samples t-Test regarding the RoBERTa baseline and different data points obtained in our experimental space to evaluate the statistical significance (Dror et al., 2018).

4 Single-step transfer

A first batch of experiments was concerned with single-step sequential transfer learning where the source tasks were those listed in Table 3.

Given the large number of data points in this experimental space, concessions were made considering the compute footprint, and we limited the hyper-parameter search by using the recommended values (Liu et al., 2019; Wolf et al., 2020) for each phase.[2]

4.1 Results

Table 4 shows the results from this first batch of experiments,[3] which support the following major empirical findings:

– The transformer with no transfer is a very strong baseline (off-the-self RoBERTa-base fine-tuned to each AAEC task). It overcomes (with 0.916 in components) the SoTA (0.908) of one of the three main tasks, and has strong scores in the other two.
– Transfer learning can be effective to leverage argument mining. This is supported by scores above the transformer baseline: with 0.924 (against the baseline 0.916) in the components task; 0.843 (against 0.820) in the clausal task; and 0.762 (against 0.727) in the relational task.
– Transfer learning with a transformer is very competitive with respect to, or even surpass, the SoTA. This is supported by a new SoTA of 0.924 in components (against 0.908), and by very good scores, 0.843 and 0.762, against respectively 0.849 and 0.767, in clausal and relational.
– Source tasks whose overall cognitive complexity is high and closer to the argument mining task tend to be more successful in supporting effective transfer. The overall trend is that better results are found with source tasks for Reasoning, Common sense and Entailment, as shown by the respective averages and the larger number of top scores

	Comp.	Clausal	Relational
Human	.886	.868	.854
SoTA - Table 2	.908	.849	.767
Baselines			
RoBERTa no transfer	.916	.820	.727
ILP	.867	.826	.751
SVM	.849	.773	.736
Majority	.259	.257	.455
Syntax	*.906*	*.718*	*.695*
PANX	.917	.815	.756
UDPOS	.914	.804	.743
Bigram Shift	.912	.710	.743
Coord Inversion	.910	.696	.735
Obj number	.907	.715	.729
Odd Man Out	.914	.703	.752
Past-Present	.901	.713	.718
Sentence Length	.885	.652	.466
Subj Number	.913	.707	.746
Top Constituents	.896	.708	**.762***
Tree Depth	.904	.674	.735
Word Content	.896	.713	.455
Semantics	*.916*	*.813*	*.745*
COPA	.919*	.823	.738
WIC	.918	.821	.744
STSB	.917	.805	.753
QQP	.911	.800	.746
Grammaticality	*.915*	*.711*	*.753*
Coord	.910	.722	.754*
Eos	.914	.712	.745
Definiteness	.914	.705	.755
Whwords	.915	.702	**.758**
CoLA	**.924**	.713	.752*
Sentiment			
SST	.916	.820	.747*
Reasoning & Compreh	*.918*	*.811*	*.701*
MULTIRC	.919	**.831**	**.758**
WNLI	.913	.788	.455
ARC	**.921**	.820	**.758**
ROPES	.920	.806	.748
ANLI	.917	.807	.749
FEVER	.914	.814	.736
QA & Common sense	*.918*	*.819*	*.717*
WSC	.919	.820	**.758**
CommonsenseQA	.916	.819	.755*
QUAIL	**.921**	.827	.755*
BoolQ	.916	**.837**	.742
PIQA	.914	.774	.455
CosmosQA	.917	.817	.745
HellaSwag	.916	.823	.746
MRQA	**.924**	.825	.750
QNLI	.916	.826	.751
Entailment/Paraphrase	*.919*	*.818*	*.744*
CB	**.923***	.819	.734
RTE	.916	**.843***	**.757**
MRPC	.916	.790	.746
SciTAIL	.919	.827	.751*
MNLI	.919	.812	.731
Argument mining			*.661*
Components		**.843**	.664
Clausal			.657

Table 4: Performance on the main tasks (columns) by different source tasks (rows). Top score underlined, top 3 scores in bold, average score in the same family of tasks in italics. All values found to be statistical significant (p-value < .05) are noted with an $*$

[2] We followed the STILT (Phang et al., 2018) approach with an intermediate training phase using only one learning rate and trained from 3 to 6 epochs. For each main task's target training phase (fine-tuning), we performed a hyper-parameter search with three learning rates and three seeds on the development set, creating a total of 396 models. The development set was extracted from 10% of the original training data, thus the training data consists of the remaining 90%. Based on the top-performing result obtained from the development set, hyper-parameters were determined for the test set. Further descriptions of hyper-parameterization data together with all materials needed to reproduce the experiments are released at [anonymized for submission].

[3] All scores obtained with RoBERTa-base.

therein. Interestingly, the top score of 0.762 for relational is obtained with a syntactic source task, that seeks to identify Top Constituents: this is of relevance for the relational main task as this task is about relating clausal segments, which are univocally associated with their top constituents.

– A main task can be a good source task to other main task for effective transfer. This is supported by the top score 0.843 in the clausal task when the components was the source task in transfer.

– A larger size of a data set for a source task, in contrast to other sources tasks, do not necessarily leads to an enhanced performance of the transfer chain. This is illustrated, for instance, by the case of RTE, with a small data set of only 2.5K, but with the top score for clausal.

5 Multi-step and multi-task transfer

A second batch of experiments was concerned with multi-step and multi-task transfer learning. The source tasks considered were the ones with the best results in the previous batch of experiments with single-step transfer.

Hence, **two-step** transfer was experimented with, where the typical chain encompasses the transfer from the components task to the clausal task and from the latter to the relational task. But we experimented also with other two-step instances, where the initial source tasks in the chain, viz. RTE, CB and Top Constituents (TC), are none of the argument mining sub-tasks. Experiments with **three-step** transfer were also undertaken, where besides the main tasks, these other source tasks contributed to the chain.

Finally, besides sequential transfer, also **multi-task** transfer learning was experimented with, involving the three argument mining sub-tasks altogether, and also pairs including two of them. Motivated by these pairings of the sub-tasks, we returned to one-step methodology, and for the sake of completeness, we experimented also with every combination of two such sub-tasks.

5.1 Results

Table 5 presents the results for this second batch of experiments,[4] which support the following major empirical findings:

– Sequential transfer is more effective than multi-task transfer. This is supported by the overall stronger scores in sequential transfer experiments for similar clusters of tasks.

– Multi-step transfer can be more effective than single-step. This is supported by the results obtained for the relational task: with the best score to relational in all experimental space of 0.783, this result was supported by a three step transfer that leveraged the relational task with the knowledge from the other two main tasks, components and clausal, and from RTE; and it is supported also by the results obtained for the clausal task: with the best score in all experimental space of 0.853, this result was supported by a two step transfer that leveraged the clausal task with the knowledge from other two tasks, one from the entailment (RTE) and the other being another main task (components).

– Source tasks that are sub-tasks in the argument mining pipeline are very successful in enhancing effective transfer. This is supported by the results obtained with the transfer being organized along the default argument mining pipeline direction, with top or very close to the top second scores for the chains Cp ⇒ Cl and Cp ⇒ Cl ⇒ Re, with 0.843

	Comp.	Clausal	Relational
Human	.886	.868	.854
SoTA Table 2	.908	.849	.767
Baselines			
RoBERTa no transfer	.916	.820	.727
ILP	.867	.826	.751
SVM	.849	.773	.736
Majority	.259	.257	.455
Sequential			
Cl ⇒ Cp	.920		
Re ⇒ Cp	**.924**		
RTE ⇒ Cp	.916		
Re ⇒ Cl ⇒ Cp	.912		
CB ⇒ Re ⇒ Cp	.915		
Cp ⇒ Cl		.843*	
Re ⇒ Cl		.811	
RTE ⇒ Cl		.843*	
Re ⇒ Cp ⇒ Cl		.839	
RTE ⇒ Cp ⇒ Cl		**.853***	
Cp ⇒ Re			.664
Cl ⇒ Re			.657
RTE ⇒ Re			.757
Cp ⇒ Cl ⇒ Re			.781*
RTE ⇒ Cp ⇒ Cl ⇒ Re			**.783***
TC ⇒ Cp ⇒ Cl ⇒ Re			.761
Multi-task			
Cp ⇔ Cl	.915	.813	
Cp ⇔ Re	.911		.684
Cl ⇔ Re		.738	.714
Cp ⇔ Cl ⇔ Re	.906	.796	.757

Table 5: Performance on the three main tasks (columns) by different transfer learning source tasks and their chaining (rows), reported with macro-F1, with the top results in bold, indicating new state-of-the-art scores. Cp stands for Components, Cl for Clausal, Re for Relational and TC for Top Constituents.

[4] All scores obtained with RoBERTa-base except clausal RTE⇒Cp⇒Cl.

and 0.781, respectively. But this is supported by the results obtained with the transfer being organized also in different directions, like for instance, the best score to components in all experimental space, of 0.924, with Re ⇒ Cp.

– Source tasks with the best performance for a given main task in the single-step setting are very successful in enhancing multi-step effective transfer, specially for that main task. This is supported by the results obtained with top or very close to the top second scores for the chains RTE ⇒ Cp, with 0.916 (over the SoTA 0.908 for components), RTE ⇒ Cp ⇒ Cl, with 0.853 (top score for clausal, and over its SoTA 0.849), and RTE ⇒ Cp ⇒ Cl ⇒ Re, with 0.774 (over the SoTA 0.767 for relational).

– Transfer learning in the setting of an off-the-self transformer architecture renders new SoTA scores for the argument mining tasks. This is supported by the scores of 0.924 for components (against 0.908 in previous SoTA), 0.853 for clausal (against 0.849), and 0.781 for relational (against 0.767).

6 Transfer during language modelling

In a third batch of experiments, we experimented with transferring knowledge from argument mining related sources by extending the pre-train, language modelling phase, rather than expanding the fine-tuning phase (as in the first and second batch of experiments). We experimented with three argumentation-oriented data sets under the Masked Language Modelling objective: a self-supervised approach was thus adopted, with no further labelled data resorted to during training.

In a first experiment, we extended the model with a train set obtained from the Oscar corpus (Ortiz Suárez et al., 2019) by parsing 1M sentences containing argumentative discourse markers.[5] In a second experiment, we extended the model with an argumentation data set, the Args.me corpus (Ajjour et al., 2019), containing 350k arguments from forum debates. Thirdly, we extended the model with ATOMIC, a common sense knowledge base converted to raw text (Sap et al., 2019) containing 877k inferential relations.[6] The results are in Table 6.

[5]We extracted all sentences that contained argumentative discourse markers from premise to conclusion and conclusion to premise in an equal distribution.

[6]Each model was trained with three randomly initialized runs, for three epochs, with a learning rate of 1e-05 and fine-tuned for each task.

	Components	Clausal	Relational
Baseline	.916	.820	.727
Arg. markers	.908	**.825**	.717
Args.me	.915	.725	**.757**
ATOMIC	**.917**	.787	.716

Table 6: Performance of models obtained by further pre-training with data related to argument mining.

6.1 Results

Some performance scores of these models are higher than the respective RoBERTa baseline, also used in the first two batches, however without a statistically significant difference. This may indicate that for this type of approach to leveraging argument mining to be as effective as the approach in the first two batches of experiments, the volume of argument mining related unlabelled data here possibly needs to be higher than the labelled data resorted to there by far more orders of magnitude.

7 Conclusions and future work

The results arrived at in this paper were obtained from a large experimental space that permitted to undertake a systematic empirical study aimed at assessing the viability of transfer learning to leverage argument mining with the support of confluent knowledge. The key findings were: • this knowledge transfer is an effective approach and permits to establish new state of the art levels of performance for the three main sub-tasks in argument mining, namely identification of argument components, classification of components, and determination of the relation among them, with a leaner approach that dispenses with heavier feature and model engineering—even when deployed on top of just an off-the-shelf Transformer model; • source tasks more closely related to argument mining and to the higher-level cognitive capacities mobilized for argumentation tend to provide better support to target tasks; • sequential transfer learning appears as more effective than multi-task transfer, and multi-step transfer can achieve better performance than single-step.

Concomitantly, these advances on empirically based insights about the argument mining task open the way to further research path that can feed future work, such as carefully articulated chains of transfer with curriculum, continual and meta-learning, and also hybrid deep learning and symbolic approaches aimed to solve transfer learning catastrophic forgetting a.o.

References

Pablo Accuosto and Horacio Saggion. 2019. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy. Association for Computational Linguistics.

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.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59, Berlin Heidelberg New York. Springer.

Ahmet Aker and Huangpan Zhang. 2017. Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining*, pages 67–72.

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 (NAACL)*, pages 1395–1404.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212.

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

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. 2019. Imho fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

HongSeok Choi and Hyunju Lee. 2018. GIST at SemEval-2018 Task 12: A Network Transferring Inference Knowledge to Argument Reasoning Comprehension Task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, pages 1374–1379.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Yarowsky David, Ngai Grace, Wicentowski Richard, et al. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8.

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 2055–2066.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. 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 (ACL)*, pages 11–22.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining On the Web From Information Seeking Perspective. In *ArgNLP*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43:125–179.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 763–772.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. 2017. First quora dataset release: Question pairs. *data. quora. com*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2021. Scientia potentia est–on the role of knowledge in computational argumentation. *arXiv preprint arXiv:2107.00281*.

John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *MRQA@EMNLP*.

Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

Jean-Christophe Mensonides, Sébastien Harispe, Jacky Montmain, and Véronique Thireau. 2019. Automatic detection and classification of argument components using multi-task deep neural network. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 25–33, Trento, Italy. Association for Computational Linguistics.

Hugo Mercier and Dan Sperber. 2017. *The enigma of reason*. Harvard University Press.

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 (SemEval-2016)*, pages 31–41.

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: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint Prediction in MST-Style Discourse Parsing for Argumentation Mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 938–948.

Georgios Petasis. 2019. Segmentation of argumentative texts with contextualised word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1–10.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here'S My Point: Joint Pointer Architecture for Argument Mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1364–1373.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from English to Portuguese. In *Proceedings of the 5th Workshop on Argument Mining*, pages 144–154.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Patrick Saint-Dizier. 2017. Using question-answering techniques to implement a knowledge-driven argument mining approach. In *Proceedings of the 4th Workshop on Argument Mining*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Jodi Schneider, Tudor Groza, and Alexandre Passant. 2013. A Review of Argumentation for the Social Semantic Web. *Semantic Web*, 4(2):159–218.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-Task Learning for Argumentation Mining in Low-Resource Settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 35–41.

Alfred Sliwa, Yuan Ma, Ruishen Liu, Niravkumar Borad, Seyedeh Ziyaei, Mina Ghobadi, Firas Sabbah, and Ahmet Aker. 2018. Multi-lingual argumentative corpora in english, turkish, greek, albanian, croatian, serbian, macedonian, bulgarian, romanian and arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Maximilian Spliethöver, Jonas Klaff, and Hendrik Heuer. 2019. Is it worth the attention? a comparative evaluation of attention layers for argument unit segmentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 74–82.

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43:619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Manfred Stede and Jodi Schneider. 2018. Argumentation Mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Frans H Van Eemeren, Rob Grootendorst, and Tjark Kruiger. 2019. *Handbook of argumentation theory*. De Gruyter Mouton.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 176–187.

Douglas Walton, Christopher W. Tindale, and David Zarefsky. 2005. *Critical Reasoning and Argumentation*. Cambridge University Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of* *the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, , Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019b. `jiant` 1.3: A software toolkit for research on general-purpose text understanding models. http://jiant.info/.

Hao Wang, Zhen Huang, Yong Dou, and Yu Hong. 2020. Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5480–5493.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, et al. 2020. The universal decompositional semantics dataset and decomp toolkit. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5698–5707.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.