Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM

Ivan Habernal[†] and **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

We propose a new task in the field of computational argumentation in which we investigate qualitative properties of Web arguments, namely their convincingness. We cast the problem as relation classification, where a pair of arguments having the same stance to the same prompt is judged. We annotate a large datasets of 16k pairs of arguments over 32 topics and investigate whether the relation "A is more convincing than B" exhibits properties of total ordering; these findings are used as global constraints for cleaning the crowdsourced data. We propose two tasks: (1) predicting which argument from an argument pair is more convincing and (2) ranking all arguments to the topic based on their convincingness. We experiment with feature-rich SVM and bidirectional LSTM and obtain 0.76-0.78 accuracy and 0.35-0.40 Spearman's correlation in a cross-topic evaluation. We release the newly created corpus *UKPConvArg1* and the experimental software under open licenses.

1 Introduction

What makes a good argument? Despite the recent achievements in computational argumentation, such as identifying argument components (Habernal and Gurevych, 2015; Habernal and Gurevych, 2016), finding evidence for claims (Rinott et al., 2015), or predicting argument structure (Peldszus and Stede, 2015; Stab and Gurevych, 2014), this question remains too hard to be answered.

Even Aristotle claimed that perceiving an argument as a "good" one depends on multiple factors (Aristotle and Kennedy (translator), 1991) — not only the logical structure of the argument (*logos*), but also on the speaker (*ethos*), emotions (*pathos*), or context (*cairos*) (Schiappa and Nordin, 2013). Experiments also show that different audiences perceive the very same arguments differently (Mercier and Sperber, 2011). A solid body of argumentation research has been devoted to the quality of arguments (Walton, 1989; Johnson and Blair, 2006), giving more profound criteria that "good" arguments should fulfill. However, the empirical evidence proving applicability of many theories falls short on everyday arguments (Boudry et al., 2015).

Since the main goal of argumentation is persuasion (Nettel and Roque, 2011; Mercier and Sperber, 2011; Blair, 2011; OKeefe, 2011) we take a pragmatic perspective on qualitative properties of argumentation and investigate a new high-level task. We asked whether we could quantify and predict how convincing an argument is.

Prompt: Should physical education be mandatory in schools? **Stance:** Yes!

Argument 1	Argument 2
physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!	YES, because some children don't understand anything excect physical education especially rich children of rich parents.

Figure 1: Example of an argument pair.

If we take Argument 1 from Figure 1, assigning a single "convincingness score" is highly subjective, given the lack of context, reader's prejudice, beliefs, etc. However, when comparing both arguments from the same example, one can decide that A1 is probably more convincing than A2, because it uses at least some statistics, addresses the health

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1589–1599,
Berlin, Germany, August 7-12, 2016.

factor, and A2 is just harsh and attacks.[1] We adapt pairwise comparison as our backbone approach.

We propose a novel task of predicting convincingness of arguments in an argument pair, as well as ranking arguments related to a certain topic. Since no data for such a task are available, we create a new annotated corpus. We employ SVM model with rich linguistic features as well as bidirectional Long Short-Term Memory (BLSTM) neural networks because of their excellent performance across various end-to-end NLP tasks (Goodfellow et al., 2016; Piech et al., 2015; Wen et al., 2016; Dyer et al., 2015; Rocktäschel et al., 2016).

Main contributions of this article are (1) large annotated dataset consisting of 16k argument pairs with 56k reasons in natural language (700k tokens), (2) thorough investigation of the annotated data with respect to properties of convincingness as a measure, (3) a SVM model and end-to-end BLSTM model. The annotated data, licensed under CC-BY-SA license, and the experimental code are publicly available at `https://github.com/UKPLab/acl2016-convincing-arguments`. We hope it will foster future research in computational argumentation and beyond.

2 Related Work

Recent years can be seen as a dawn of computational argumentation – an emerging sub-field of NLP in which natural language arguments and argumentation are modeled, searched, analyzed, generated, and evaluated. The main focus has been paid to analyzing argument structures, under the umbrella entitled argumentation mining.

Web discourse as a data source has been exploited in several tasks in argumentation mining, such as classifying propositions in user comments into three classes (verifiable experiential, verifiable non-experiential, and unverifiable) (Park and Cardie, 2014), or mapping argument components to Toulmin's model of argument in user-generated Web discourse (Habernal and Gurevych, 2015), to name a few. While these approaches are crucial for understanding the structure of an argument, they do not directly address any qualitative criteria of argumentation.

Argumentation quality has been an active topic among argumentation scholars. Walton (1989) discusses validity of arguments in informal logic, while Johnson and Blair (2006) elaborate on criteria for practical argument evaluation (namely Relevance, Acceptability, and Sufficiency). Yet, empirical research on argumentation quality does not seem to reflect these criteria and leans toward simplistic evaluation using argument structures, such as how many premises support a claim (Stegmann et al., 2011), or by the complexity of the analyzed argument scheme (Garcia-Mila et al., 2013). To the best of our knowledge, there have been only few attempts in computational argumentation that go deeper than analyzing argument structures (e.g., (Park and Cardie, 2014) mentioned above). Persing and Ng (2015) model argument strength in persuasive essays using a manually annotated corpus of 1,000 documents labeled with a 1–4 score value.

Our newly created corpus of annotated pairs of arguments might resemble recent large-scale corpora for textual inference. Bowman et al. (2015) introduced a 570k sentence pairs written by crowd-workers, the largest corpus to date. Whereas their task is to classify whether the sentence pair represents *entailment*, *contradiction*, or is *neutral* (thus heading towards a deep *semantic* understanding), our goal is to assess the *pragmatical* properties of the given multiple sentence-long arguments (to which extent they fulfill the goal of persuasion). Moreover, each of our annotated argument pairs is accompanied with five textual reasons that explain the rationale behind the labeler's decision. This is, to the best of our knowledge, a unique novel feature of our data.

Pairwise assessment for obtaining relative preference was examined by (Chen et al., 2013), among many others.[2] Their system was tested on ranking documents by their reading difficulty. Relative preference annotations have also been heavily employed in assessing machine translation (Aranberri et al., 2016). By contrast to our work, the underlying relations (reading difficulty 1-12 or better translation) have well known properties of total ordering, while convincingness of arguments is a yet unexplored task, thus no assumptions can be made apriori. There is also a substantial body of work on learning to rank, where also a *pairwise approach* is widely used (Cao et al., 2007). These methods have been traditionally used in IR,

[1] These are actual *reasons* provided by annotators, as will be explained later in Section 3.

[2] See (Shah et al., 2015) for a recent overview.

where the retrieved documents are ranked according to their relevance and pairs of documents are automatically sampled.

Employing LSTM for natural language inference tasks has recently gained popularity (Rocktäschel et al., 2016; Wang and Jiang, 2016; Cheng et al., 2016). These methods are usually tested on the SNLI data introduced above (Bowman et al., 2015).

3 Data annotation

Since assessing convincingness of a single argument directly is a very subjective task with high probability of introducing annotator's bias (because of personal preferences, beliefs, or background), we cast the problem as a relation annotation task. Given two arguments, one should be selected as more convincing, or they might be both equally convincing (see an example in Figure 1).

3.1 Sampling annotation candidates

Sampling large sets of arguments for annotation from the Web poses several challenges. First, we must be sure that the obtained texts are actual arguments. Second, the context of the argument should be known (the prompt and the stance). Finally, we need sources with permissive licenses, which allow us to release the resulting corpus further to the community. These criteria are met by arguments from two *debate portals*.[3]

We will use the following terminology. We use *topic* to refer to a subset of an on-line debate with a given prompt and a certain stance (for example, *"Should physical education be mandatory in schools? – yes"* is considered as a single topic). Each debate has two *topics*, one for each stance. *Argument* is a single comment directly addressing the debate prompt. *Argument pair* is an ordered set of two arguments (*A1* and *A2*) belonging to the same topic; see Figure 1.

We automatically selected debates that contained at least 25 top-level[4] arguments that were 10-110 words long (the mean for all top-level arguments was 66 ± 130 and the median 36, so we excluded the lengthy outliers in our sampling). We manually filtered out obvious silly debates (e.g., *'Superman vs. Batman'*) and ended up with 32 *topics* (the full topic list is presented together with experimental results later in Table 3). From each topic we automatically sampled 25-35 random arguments and created $(n * (n-1)/2)$ *argument pairs* by combining all selected arguments. Sampling argument pairs only from the same topics and not combining opposite stances was a design decision how to mitigate annotators' bias.[5] The order of arguments A1 and A2 in each argument pair was randomly shuffled. In total we sampled 16,927 argument pairs.

3.2 Crowdsourcing annotations

Let us extend our terminology. *Worker* is a single annotator in Amazon Mechanical Turk. *Reason* is an explanation why *A1* is more convincing than *A2* (or the other way round, or why they are equally convincing). *Gold reason* is a reason whose label matches the gold label in the argument pair (see Figure 2).

In the HIT, workers were presented with an argument pair, the prompt, and the stance as in Figure 1. They had to choose either *"A1 is more convincing than A2"* (A1>A2), *"A1 is less convincing than A2"* (A1<A2), or *"A1 and A2 are convincing equally"* (A1=A2). Moreover, they were obliged to write the *reason* 30-140 characters long. An example of fully annotated argument pair is shown in Figure 2. The workers were also provided with clear and crisp instructions (e.g., do not judge the truth of the proposition; be objective; do not express your position; etc.).

All 16,927 argument pairs were annotated by five workers each (85k assignments in total). We also allowed workers to express their own standpoint toward the topics. While 66% of workers had no standpoint, 14% had the opposite view and 20% the same view. This indicates that there should be no systematic bias in the data. Crowdsourcing took about six weeks in 2016 plus two weeks of pilot studies. In total, about 3,900 workers participated. Total costs including pilot studies and bonus payments were 5,520 USD.

3.3 Quality control and agreement

We performed several steps in controlling the quality of the crowdsourced data. First, we allowed only workers from the U.S. with $\geq 96\%$ acceptance rate to work on the task. Second, we employed MACE (Hovy et al., 2013) for estimating

[3] Namely, `createdebate.com` and `procon.org`.

[4] Such arguments directly address the topic and are not a part of a threaded discussion.

[5] As some topics touch the very fundamental human beliefs and values, such as faith, trust, or sexuality, it is hard to put them consciously aside when assessing convincingness.

Argument 1	Argument 2
physical education should be mandatory cuhz 112,000 people have [...]	YES, because some children don't understand anything expect [...]

- A1>A2, *because* A1 uses statistics, and doesn't make assumptions.
- A1>A2, *because* A1 talks about the importance of health.
- A1>A2, *because* A1 provides a health-related argument.
- A1>A2, *because* A2 is very harsh and attacks
- A1=A2, *because* Neither A1 or A2 cite evidence to support their claims.

Figure 2: Example of an *argument pair* annotated by five workers. The arguments are shortened versions of Figure 1. The explanations (called *reasons*) after 'because' are written by workers; the estimated gold label for this pair is probably *A1>A2*, thus there are four *gold reasons*.

the true labels and ranking the annotators. We set the MACE's parameter *threshold* to 0.95 to keep only instances whose entropy is among the 95% best estimates. Third, we manually checked all the *reasons* for each worker. With paying more attention to workers with low MACE scores, we rejected all assignments of workers if they (1) copied&pasted the same or very similar *reasons* across *argument pairs*, (2) were only copying or rephrasing the texts from the arguments, (3) provided their opinion or were arguing, (4) had many typos or provided obvious nonsense. In total, we rejected 1161 assignments.

We do not report any 'standard' inter-annotator agreement measures such as Fleiss' κ or Krippendorff's α, as their suitability for crowdsourcing has been recently disputed (Passonneau and Carpenter, 2014). However, in order to estimate the human performance, we analyzed the output of the pilot study. For each *argument pair*, we took the best-ranked worker for that particular pair (worker ranks are globally estimated by MACE) and computed her accuracy against the estimated gold labels.[6] The best-ranked worker for each argument pair is not necessarily the globally best-ranked worker; in the pilot study, the average global rank of this hypothetical worker was 11 ± 6.6. This rank can be interpreted as a decently performing worker; the obtained score reached 0.935 accuracy.

[6] A similar approach was recently reported by Nakov et al. (2016).

3.4 Examining properties of convincingness

3.4.1 What makes a convincing argument?

We manually examined a small sample of 200 *gold reasons* to find out what makes one argument more convincing than the other. A very common type of answer mentioned **giving examples or actual reasons** (*"A1 cited several reasons to back up their argument."*) and **facts** (*"A1 cites an outside source which can be more credible than opinion"*). This is not surprising, as argumentation is often perceived as reason giving (Freeley and Steinberg, 2008). Others point out strengths in **explaining the reasoning** or **logical coherence** (*"A1 gives a succinct and logical answer to the argument. A2 strays away from the argument in the response."*). The **confirmation bias** (Mercier and Sperber, 2011) also played a role (*"A1 argues both viewpoints, A2 chooses a side."*). Given the noisiness of Web data, some of the arguments might be **non-sense**, which was also pointed out as a reason (*"A1 attempts to argue that since porn exists, we should watch it. A2 doesn't make sense or answer the question."*). Apart from the logical structure of the argument, **emotional aspects** and **rhetorical moves** were also spotted (*"A1 contributes a viewpoint based on morality, which is a stronger argument than A2, which does not argue for anything at all."*, or *"A1 calls for the killing of all politicians, which is an immature knee-jerk reaction to a topic. A2's argument is more intellectually presented."*).

3.4.2 Transitivity evaluation using argument graphs

The previous section shows a variety of reasons that makes one argument more convincing than other arguments. Considering *A1 is more convincing than A2* as a binary relation R, we thus asked the following research question: Is convincingness a measure with total strict order or strict weak order? Namely, is relation R that compares convingcingness of two arguments transitive, antisymmetric, and total?

In particular, does is exhibit properties such that if A$\geq$B and B$\geq$C, then A$\geq$C (total ordering)? We can treat arguments as nodes in a graph and argument pairs as graph edges. We will denote such graph as *argument graph* (and use nodes/arguments and edges/pairs interchangeably in this section).[7] As the sampled argument pairs

[7] Argument pair A>B becomes a directed edge $A \rightarrow B$

contained all argument pair combinations for each topic, we ended up with an almost fully connected argument graph for each topic (remember that we discarded 5% of argument pair annotations with lowest reliability). We further investigate the properties of the argument graphs. Transitivity is only guaranteed, if the argument graph is a DAG (directed acyclic graph).

Building argument graph from crowdsourced argument pairs We build the argument graph iteratively by sampling annotated argument pairs and adding them as graph edges (see Algorithm 1). We consider two possible scenarios in the graph building algorithm. In the first scenario, we accept only argument pairs without equivalency (thus A>B is allowed but A=B is forbidden and discarded). The second scenario accepts all pairs, but since the resulting graph must be DAG, equivalent arguments are merged into one node. We use Johnson's algorithm for finding all elementary cycles in DAG (Johnson, 1975).

Argument pair weights By building argument graph from all pairs, introducing cycles into the graph seems to be inevitable, given a certain amount of noise in the annotations. We asked the following question: to which extent does occurrence of cycles in an argument graph depend on the quality of annotations?

We thus compute a weight for each argument pair. Let e_i be a particular annotation pair (edge). Let G_i be all labels in that pair that match the predicted gold label, and O_i opposite labels (different from the gold label). Let v be a single worker's vote and c_v a global worker's competence score. Then the weight w of edge e_i is computed as follows:

$$w_{e_i} = \sigma\left(\sum_{v \in G_i} c_v - \lambda \sum_{v \in O_i} c_v\right) \qquad (1)$$

where σ is a sigmoid function $\sigma = \frac{1}{1+e^{-x}}$ to squeeze the weight into the $(0, 1)$ interval and λ is a penalty for opposite labels (we set empirically λ to 10.0 to ensure strict penalization). For example, if the predicted gold label from Figure 2 were A1>A2, then G_i would contain four votes and O_i one vote (the last one).

This weight allows us to sort argument pairs before sampling them for building the argument graph. We test three following strategies. As a baseline, we use random shuffling (*Rand*), where no prior information about the weight of the pairs is given. The other two sorting algorithms use the argument pair weight computed by Equation 1. As the worst case scenario, we sort the pairs in ascending order (*Asc*), which means that the "worse" pairs come first to the graph building algorithm. We used this scenario to see how much the prior pair weight information actually matters, because building a graph preferably from bad pair label estimates should cause more harm. Finally, the *Desc* algorithm sorts the pairs given their weight in descending order (the "better" estimates come first).

Algorithm 1: Building DAG from sorted argument pairs.

```
input : argumentPairs; sortingAlg
output: DAG
SortPairs(argumentPairs, sortingAlg);
finalPairs ← [];
foreach pair in argumentPairs do
    currentPairs ← [finalPairs, pair ];
    /* cluster edges labeled as equal so they will be
    treated as a single node */
    clusters ← clusterEqNodes(currentPairs);
    /* wire the pairs into directed graph */
    g ← buildGraph(currentPairs, clusters);
    if hasCycles(g) then
        // report about breaking DAG
    else
        finalPairs += pair;
return buildGraph(finalPairs);
```

Measuring transitivity score We measure how "good" the graph is by a *transitivity score*. Here we assume that the graph is a DAG. Given two nodes A and Z, let P_L be the longest path between these nodes and P_S the shortest path, respectively. For example, let $P_L = A \rightarrow B \rightarrow C \rightarrow Z$ and $P_S = A \rightarrow D \rightarrow Z$. Then the *transitivity score* is the ratio of longest and shortest path $\frac{|P_L|}{|P_S|}$. (which is 1.5 is our example). The *average transitivity score* is then an average of transitivity scores for each pair of nodes from the graph that are connected by two or more paths. Analogically, the *maximum transitivity score* is the maximal value. We restrict the shortest path to be a direct edge only.

The motivation for the *transitivity score* is the following. If the longest path between A and Z ($A \rightarrow \ldots \rightarrow Z$) consists of 10 other nodes, than the total ordering property requires that there also exists a direct edge $A \rightarrow Z$. This is indeed em-

in the argument graph.

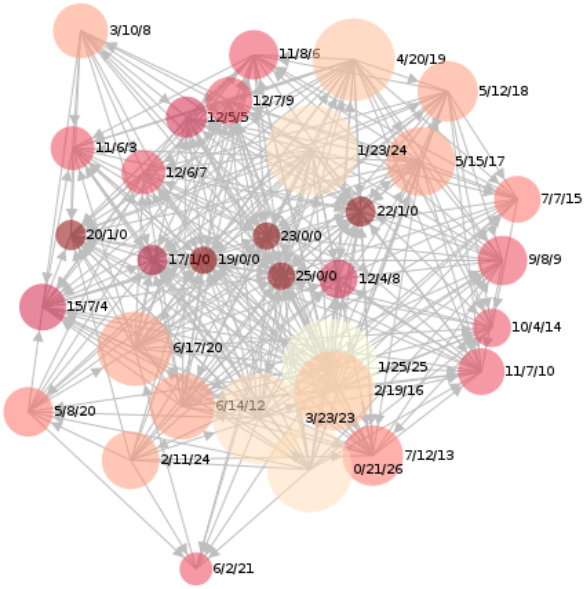

Figure 3: Final argument graph example (topic: *"Christianity or Atheism? Atheism"*). Node labels: *I/O/Tr* (*I*: incoming edges, *O*: outgoing edges, *Tr*: maximum transitivity score). Lighter nodes have prevailing *O*; larger nodes have higher absolute number of *O*.

pirically confirmed by the presence of the shortest path between A and Z. Thus the longer the longest path and the shorter the shortest path are on average, the bigger empirical evidence is given about the transitivity property.

Figure 3 shows an example of argument graph built using only non-equivalent pairs and *desc* prior sort or argument pairs. There are few "bad" arguments in the middle (many incoming edges, none outcoming) and few very convincing arguments (large circles). Notice the high maximum transitivity score even for medium-sized nodes.

Observations First, let us compare the different sorting algorithms for each sampling strategy. As Table 1 shows, on average, 158 pairs are ignored in total when all pairs are used for sampling (26 removed by MACE and 132 by the graph building algorithm), while 164 pairs are ignored when only non-equivalent pairs are sampled (129 had already been removed apriori—26 by MACE and 103 as equivalent pairs—and 35 by the graph algorithm).

The results show a tendency that, when sampling annotated argument pairs for building a DAG, sorting argument pairs by their weight based on workers' scores influences the number of pairs that break the DAG by introducing cycles. In particular, starting with more confident argumentation pairs, the graph grows bigger while keeping its DAG consistency. The presence of the *equal* relation causes cycles to break the DAG sooner as compared to argument pairs in which one argument is more convincing than the other. We interpret this finding as that it is easier for humans to judge A>B than A=B consistently across all possible pairs of arguments from a given topic.

3.4.3 Gold-standard corpora

Our experiments show that convincingness between a pair of arguments exhibits properties of strict total order when the possibility of two equally convincing arguments is prohibited. We thus used the above-mentioned method for graph building as a tool for posterior gold data filtering. We discard the equal argument pairs in advance and filter out argument pairs that break the DAG properties. As a result, a set of 11,650 argument pairs labeled as either A>B or A<B remains, which is summarized in Table 2. We call this corpus *UKPConvArgStrict*.

However, since the total strict ordering property of convincingness is only an empirically confirmed working hypothesis, we also propose another realistic application. We construct a mixed graph by treating equal argument pairs (A=B) as undirected edges. Using PageRank, we rank the arguments (nodes) globally. The higher the PageRank for a particular node is, the "less convincing" the argument is (has a global higher probability of incoming edges). This allows us to rank all arguments for a particular topic. We call this dataset *UKPConvArgRank* (see Table 2).

We also release the full dataset *UKPConvArgAll*. In this data, no global filtering using graph construction methods is applied, only the local pre-filtering using MACE. We believe this dataset can be used as a supporting training data for some tasks that do not rely on the property of total ordering. Along the actual argument texts, all the gold-standard corpora contain the *reasons* as well as full workers' information and debate meta-data.

4 Experiments

We experiment with two machine learning algorithms on two tasks using the two new benchmark corpora (*UKPConvArgStrict* and *UKPConvArgRank*). In both tasks, we perform 32-fold cross-topic cross-validation (one topic is test data, remaining 31 topics are training ones). This rather

		All pairs			No equivalency pairs		
		Rand.	Asc.	Desc.	Rand.	Asc.	Desc.
Fixed values	All annotated pairs	529	529	529	529	529	529
	Pairs removed apriori	26	26	26	129	129	129
Before first cycle detected	Edges in graph	32	16	**86**	84	37	**199**
	Nodes in graph	20	16	**25**	32	28	**33**
	Pairs sampled	44	26	**132**	84	37	**199**
	First cycle length	1,8	2,0	1,5	4,2	3,9	3,4
Final statistics after all pairs sampled	avg. Transitivity score	6,4	5,9	6,8	11,1	11,3	10,8
	max. Transitivity score	14,5	13,2	15,5	24,4	25,6	24,0
	Edges in graph	105	81	114	357	339	365
	Nodes in graph	16	14	16	33	33	33
	Pairs sampled	339	294	**369**	356	339	**364**
	Ignored pairs	162	208	**132**	42	60	**35**

Table 1: Values averaged over all 32 topics reported by different sampling strategies and scenarios for argument graph building.

Dataset	**Size**	**Instance type**	**Size per topic**	**Gold label distribution**			**Gold reasons**	
				a1	a2	eq	**size**	**tokens**
UKPConvArgAll	16,081	argument pair	502.5 ± 91.3	6,398	6,394	3,289	56,446	696,537
UKPConvArgStrict	11,650	argument pair	364.1 ± 71.1	5,872	5,778	0	44,121	547,057
UKPConvArgRank	1,052	argument	32.9 ± 3.2	—			—	

Table 2: Properties of resulting gold data.

challenging setting ensures that no arguments are seen in both training and test data.

4.1 Predicting convincingness of pairs

Since this task is a binary classification and the classes are equally distributed (see Table 2), we report accuracy and average the final score over folds (Forman and Scholz, 2010).

Methods As a "traditional" method, we employ SVM with RBF kernel[8] based on a large set of rich linguistic features. They include **uni- and bi-gram presence**, ratio of **adjective and adverb endings** that may signalize neuroticism (Corney et al., 2002), **contextuality measure** (Heylighen and Dewaele, 2002), **dependency tree depth**, ratio of **exclamation** or **quotation** marks, ratio of **modal verbs**, counts of several **named entity types**, ratio of **past** vs. **future** tense verbs, **POS** n-grams, presence of dependency tree **production rules**, seven different **readability measures** (e.g., *Ari* (Senter and Smith, 1967), *Coleman-Liau* (Coleman and Liau, 1975), *Flesch* (Flesch, 1948), and others), five **sentiment scores** (from very negative to very positive) (Socher et al., 2013), **spell-checking** using standard Unix `words`, ratio of **superlatives**, and some **surface** features such as sentence lengths, longer words count, etc. The resulting feature vector dimension is about 64k.

We also use bidirectional Long Short-Term Memory (BLSTM) neural network for end-to-end processing.[9] The input layer relies on pre-trained word embeddings, in particular *GloVe* (Pennington et al., 2014) trained on 840B tokens from Common Crawl;[10] the embedding weights are further updated during training. The core of the model consists of two bi-directional LSTM networks with 64 output neurons each. Their output is then concatenated into a single drop-out layer and passed to the final sigmoid layer for binary predictions. We train the network with ADAM optimizer (Kingma and Ba, 2015) using binary cross-entropy loss function and regularize by early stopping (5 training epochs) and high drop-out rate (0.5) in the dropout layer. For both models, each training/test instance simply concatenates A1 and A2 from the argument pair.

Results and error analysis As shown in Table 3, SVM (0.78) outperforms BSLTM (0.76) with a subtle but significant difference. It is also apparent that some topics are more challenging regardless of the system (e.g., *"Is it better to have a lousy father or to be fatherless? – Lousy father"*). Both systems outperform a simple baseline (lemma n-gram presence features with SVM, not reported in detail, achieved 0.65 accuracy) but still do not reach the human upper bounds (0.93 as reported in Section 3.3).

[8] Using LISBVM (Chang and Lin, 2011).

[9] Using `http://keras.io/`

[10] `http://nlp.stanford.edu/projects/glove/`

	Topic	SVM	BLSTM
Ban Plastic Water Bottles?	No	.85	.76
	Yes	.90	.83
Christianity or Atheism	Atheism	.81	.80
	Christianity	.68	.75
Evolution vs. Creation	Creation	.84	.88
	Evolution	.66	.77
Firefox vs. Internet Explorer	IE	.84	.81
	Firefox	.82	.78
Gay marriage - right or wrong?	Right	.76	.74
	Wrong	.82	.87
Should parents use spanking?	No	.84	.78
	Yes	.79	.68
If your spouse committed murder [...]	No	.71	.64
	Yes	.79	.72
India has the potential to lead the world	No	.82	.77
	Yes	.69	.79
Is it better to have a lousy father or to be fatherless?	Fatherless	.77	.69
	Lousy father	.67	.60
Is porn wrong?	No	.82	.79
	Yes	.85	.85
Is the school uniform a good or bad idea?	Bad	.75	.78
	Good	.83	.74
Pro choice vs. Pro life	Choice	.71	.68
	Life	.79	.80
Should physical edu. be mandatory?	No	.79	.80
	Yes	.79	.78
TV is better than books	No	.78	.73
	Yes	.78	.75
Personal pursuit or common good?	Common	.72	.78
	Personal	.67	.68
Farquhar as the founder of Singapore	No	.79	.63
	Yes	.85	.76
Average		**.78**	**.76**

Table 3: Accuracy results on *UKPConvArgStrict* data. The difference between **SVM** and bidirectional **LSTM** is significant, $p = 0.0414$ using two-tailed Wilcoxon signed-rank test.

We examined about fifty random false predictions to gain some insight into the limitations of both systems. We looked into argument pairs, in which both methods failed, as well as into instances where only one model was correct. BLSTM won in few cases by properly catching jokes or off-topic arguments; SVM was properly catching all-upper-case arguments (considered as less convincing). By examining failures common to both systems, we found several cases where the prediction was wrong due to very negative sentiment (which might be a sign of the less convincing argument), but in other cases an argument with strong negative sentiment was actually the more convincing one. In general, we did not find any tendency on failures; they were also independent of the worker assignments distribution, thus not caused by likely ambiguous (hard) instances.

	SVM	BLSTM	p-value
Pearson's r	.351	.270	$\ll 0.01$
Spearman's ρ	.402	.354	$\ll 0.01$

Table 4: Correlation results on *UKPConvArgRank*.

4.2 Ranking arguments

We address this problem as a regression task. We use the *UKPConvArgRank* data, in which a real-value score is assigned to each argument so the arguments can be ranked by their convincingness (for each topic independently). The task is thus to predict a real-value score for each argument from the test topic (remember that we use 32-fold cross validation). We measure Spearman's and Pearson's correlation coefficients on all results combined (not on each fold separately).

Without any modifications, we use the same SVM and features as described in Section 4.1. Regarding the BLSTM, we only replace the output layer with a linear activation function and optimize mean absolute error loss. Table 4 shows that SVM outperforms BLSTM. All correlations are highly statistically significant.

4.3 Results discussion

Although the "traditional" SVM with rich linguistic features outperforms BLSTM in both tasks, there are other aspects to be considered. First, the employed features require heavy language-specific preprocessing machinery (lemmatizer, POS tagger, parser, NER, sentiment analyzer). By contrast, BLSTM only requires pre-trained embedding vectors, while delivering comparable results. Second, we only experimented with vanilla LSTMs. Recent developments of deep neural networks (especially attention mechanisms or grid-LSTMs) open up many future possibilities to gain performance in this end-to-end task.

5 Conclusion and future work

We propose a novel task of predicting Web argument convincingness. We crowdsourced a large corpus of 16k argument pairs over 32 topics and used global constraints based on transitivity properties of convincingness relation for cleaning the data. We experimented with feature-rich SVM and bidirectional LSTM and obtain 0.76-0.78 accuracy and 0.35-0.40 Spearman's correlation in a cross-topic scenario. We release the newly created corpus *UKPConvArg1* and the experimental software

under free licenses.[11] To the best of our knowledge, we are the first who deal with argument convincingness in Web data on such a large scale.

In the current article, we have only slightly touched the annotated natural text *reasons*. We believe that the presence of 44k reasons (550k tokens) is another important asset of the newly created corpus, which deserves future investigation.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant Nº I/82806, by the German Institute for Educational Research (DIPF), by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1), by the GRK 1994 AIPHES (DFG), and by Amazon Web Services in Education Grant award. Lastly, we would like to thank the anonymous reviewers for their valuable feedback.

References

Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola. 2016. Ebaluatoia: crowd evaluation for English-Basque machine translation. *Language Resources and Evaluation*. In press.

Aristotle and George Kennedy (translator). 1991. *On Rhetoric: A Theory of Civil Discourse*. Oxford University Press, USA.

J. Anthony Blair. 2011. Argumentation as rational persuasion. *Argumentation*, 26(1):71–81.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The Fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life. *Argumentation*, 29(4):431–456.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 129–136, New York, NY, USA. ACM.

[11] https://github.com/UKPLab/acl2016-convincing-arguments

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 193, New York, New York, USA. ACM Press.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. *arXiv*.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC02)*, pages 282–289.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China, July. Association for Computational Linguistics.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

George Forman and Martin Scholz. 2010. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57.

Austin J. Freeley and David L. Steinberg. 2008. *Argumentation and Debate*. Cengage Learning, Stamford, CT, USA, 12th edition.

Merce Garcia-Mila, Sandra Gilabert, Sibel Erduran, and Mark Felton. 2013. The effect of argumentative task goal on the quality of argumentative discourse. *Science Education*, 97(4):497–523.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep learning. Book in preparation for MIT Press.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal, September. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*. Under review. http://arxiv.org/abs/1601.02403.

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT 2013*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Ralph H. Johnson and Anthony J. Blair. 2006. *Logical Self-Defense*. International Debate Education Association.

Donald B. Johnson. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84.

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

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, 34(2):57–74; discussion 74–111.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

Ana Laura Nettel and Georges Roque. 2011. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69.

Daniel J. OKeefe. 2011. Conviction, persuasion, and argumentation: Untangling the ends and means of influence. *Argumentation*, 26(1):19–32.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The Benefits of a Model of Annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

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*, pages 938–948, Lisbon, Portugal, September. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China, July. Association for Computational Linguistics.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 505—-513, Montreal, CA. Curran Associates, Inc.

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, Lisbon, Portugal, September. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR)*, pages 1–9.

Edward Schiappa and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*. Pearson UK, 1st edition.

J. R. Senter and E. A. Smith. 1967. Automated readability index. Technical report AMRL-TR-66-220, Aerospace Medical Research Laboratories, Ohio.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph K. Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. 2015. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. *arXiv preprint*, abs/1505.01462.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Seattle, Washington, USA, October. Association for Computational Linguistics.

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 (EMNLP)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

Karsten Stegmann, Christof Wecker, Armin Weinberger, and Frank Fischer. 2011. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science*, 40(2):297–323.

Douglas N. Walton. 1989. *Informal Logic: A Handbook for Critical Argument*. Cambridge University Press.

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*, page (to appear), San Diego, CA, June. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain Neural Network Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page (to appear).