Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network

Martin Gleize*, **Eyal Shnarch***, **Leshem Choshen**, **Lena Dankin**, **Guy Moshkowich**, **Ranit Aharonov**, **Noam Slonim**
IBM Research
martin.gleize@ie.ibm.com, guy.moshkowich@ibm.com,
{eyals, leshem.choshen, lenad, ranita, noams}@il.ibm.com

Abstract

Machines capable of responding and interacting with humans in helpful ways have become ubiquitous. We now expect them to discuss with us the more delicate questions in our world, and they should do so armed with effective arguments. But what makes an argument more persuasive? What will convince you?

In this paper, we present a new data set, *IBM-EviConv*, of pairs of evidence labeled for convincingness, designed to be more challenging than existing alternatives. We also propose a Siamese neural network architecture shown to outperform several baselines on both a prior convincingness data set and our own. Finally, we provide insights into our experimental results and the various kinds of argumentative value our method is capable of detecting.

1 Introduction

The most interesting questions in life do not have a simple factual answer. Rather, they have pros and cons associated with them. When opposing sides debate such questions, each side aims to present the most convincing arguments for its point of view, typically by raising various claims and supporting them with relevant pieces of evidence. Ideally, the arguments by both sides are then carefully compared, as part of the decision process.

Automatic methods for this process of argumentation and debating are developed within the field of Computational Argumentation. which focuses on methods for argument *detection* (Lippi and Torroni, 2016; Levy et al., 2014; Rinott et al., 2015) and revealing argument *relations* (Stab and Gurevych, 2014, 2017).

Recently, IBM introduced *Project Debater*, the first AI system able to debate humans on complex topics. Project Debater participated in a live debate against a world champion debater, and was able to mine arguments and use them to compose a speech supporting its side of the debate. In addition, it was able to rebut its human competitor.[1] The technology that underlies such a system is intended to enhance decision making.

In this work we target the task of assessing argument *convincingness*, and more specifically, we focus on evidence convincingness – given texts representing evidence for a given debatable topic, identify the more convincing ones.

Theoretical works have analyzed the factors that make an argument more convincing (e.g., Boudry et al., 2015). This work is an empirical one in the line of (Persing and Ng, 2017; Tan et al., 2016). To the best of our knowledge, this is the first work on evidence convincingness.

Most similar to our work is that of Habernal and Gurevych (2016a) who are the first to directly compare pairs of arguments (previous works compared documents). They released *UPKConvArg*, the first data set of convincingness, containing argument pairs with a label indicating which one is preferred over the other.

In this work we release *IBM-EviConv*, a data set of evidence pairs which offers a more focused view of the argument convincingness task. As a source of evidence sentences we use the evidence data set released by Shnarch et al. (2018), which contains more than 2,000 evidence sentences over 118 topics. We then sampled more than 8,000 pairs of evidence and sent them for convincingness labeling.

Why is the new data set useful? Our new data set differs from *UPKConvArg* in a few important aspects. While in *UPKConvArg* the pairs consist of two types or arguments, claims and evi-

*First two authors contributed equally.

[1] For more details and a video of the debate: https://www.research.ibm.com/artificial-intelligence/project-debater/live/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 967–976
Florence, Italy, July 28 - August 2, 2019.

dence, *IBM-EviConv* pairs are composed solely of evidence. In a follow-up work on *UPKConvArg*, Habernal and Gurevych (2016b) showed that the most frequent reason by far to prefer one argument over another is that it is more informative. Usually, an evidence is longer and provides more details and information than a concise claim. Therefore, in a data set which includes both evidence and claims the identification of the more convincing argument may be based not only on argument convincingness, but also on identifying argument type, or even on a shallow feature such as argument length. Indeed, we show a very high performance of the baseline by length over *UPKConvArg* in §5.1. On the other hand, a data set that includes only evidence poses a more challenging task. In addition, we directly controlled for argument length by building pairs of roughly the same length.

A second important distinction between the data sets is writing level. The arguments for *UPKConvArg* were extracted from two Web debate portals, on which people post free text and in which writing level widely varies (for instance, some posts include ungrammatical texts which require a preprocessing step). Our arguments were retrieved from Wikipedia, a heavily edited corpus which makes them on par in terms of writing level.

Overall, the contribution of this new data set is that it emphasizes pairs homogeneity – in terms of argument type, length, and writing level. We believe that *IBM-EviConv* offers learning algorithms a better chance to reveal real convincingness signals, beyond the more trivial ones.

Finally, *UPKConvArg* pairs are of the same stance towards the topic, (either both supporting it or both contesting it), and therefore it is aligned with the task of choosing the most convincing arguments *of a given side of the debate*. In contrast, our data set contains both same stance pairs, as well as cross stance pairs (i.e., one is supporting and the other is contesting the topic). Thus it is aligned with the above mentioned task, but in addition, with the task of choosing *which side of the debate* was more convincing (Potash and Rumshisky, 2017).

In addition to the release of a new data set, a second contribution of this work is the suggestion of a Siamese Network architecture for the argument convincingness task. We evaluate our method on both *UPKConvArg* and *IBM-EviConv* data sets, and show that it outperforms the methods suggested by Habernal and Gurevych (2016a) and Simpson and Gurevych (2018) on both sets.

With the advancement in argument detection, the research community can now pay more attention to the challenging task of identifying the more convincing arguments. This work continues the line started by Habernal and Gurevych (2016a) by suggesting a focused framing of the task, providing a new data set for it, and presenting a neural network which surpasses state of the art performance.

2 Background

Convincingness. Convincingness (or persuasiveness) arouses great interest in various fields such as essay scoring (Ghosh et al., 2016), persuasive technologies (Fogg, 1998, 2002, 2009), and social networks, where it is deemed a hard problem (Hidey and McKeown, 2018). Naturally, it is also relevant for social sciences, for example in public narrative (Green and Brock, 2000), internet discussions (Tan et al., 2016), and in argumentative process of thought (Burnstein, 2003).

In theoretical argumentation studies, the importance of quality (Wachsmuth et al., 2017) and convincingness was emphasized (O'Keefe, 2012; Van Eemeren et al., 2014; Chen et al., 2019), but assessment is still a challenge despite years of study (Weltzer-Ward et al., 2009; Rosenfeld and Kraus, 2016).

Traditionally, assessment of arguments convincingness, if addressed at all, relied on relevance, acceptability or sufficiency of arguments (Habernal and Gurevych, 2015; Johnson and Blair, 2006), or on general fallacies (Hamblin, 1971; Tindale, 2007). Recently, some works studied convincingness of full texts, assessing the role of prior beliefs (Durmus and Cardie, 2018) and structure (Wachsmuth et al., 2016).

Argument convincingness data set. Closer to our work, recent studies aim to assess the convincingness of a single argument, rather than that of a full text. The first data set for this task was published by Habernal and Gurevych (2016a). Their data set, *UPKConvArg*, consists of approximately 1,000 web mined arguments across 16 different topics, each split into two sets by stance (support or contest the topic). In each such split, all argument pairs are annotated by crowd workers for the preference of one argument over the other. In addi-

tion, the workers provided reasons for their choice in the form of free text.

From the labeling over pairs, the authors proposed a method, based on PageRank (Page et al., 1999), to derive a second data set, *UPKConvArgRank*, which approximates convincingness of individual arguments rather than in a comparative manner within an argument pair.

A continuation work, on that data set, looks into the textual reasons provided by the annotators, classifies them and proposes prediction tasks on that classification (Habernal and Gurevych, 2016b).

Empirical methods for argument convincingness. To identify the more convincing arguments in a set we need to rank them. Learning to Rank is the machine learning field which aims to learn rankings rather than classification or regression (McFee and Lanckriet, 2010; Burges, 2010). Learning to rank can be formalized in various ways (Cao et al., 2007); in a *pointwise* approach, the input is single elements and for each the output is a score. To rank a list in this approach, one simply orders the elements by their scores.

In a *pairwise* approach, the input is pairs of elements and for each pair the output is the preference between its two elements. To rank a list in this approach, one must compare all pair combinations, assuming transitivity holds, otherwise some approximation is needed (such as the one made to produce *UPKConvArgRank*).

Habernal and Gurevych (2016a) suggest empirical methods for the task of choosing the more convincing argument. Relying on 32,000 linguistic features and word embedding, they proposed two methods, based on SVM and BiLSTM. When trained over argument pairs, these methods can provide pairwise inference only. They cannot predict a convincingness score for a single argument.

To overcome this disadvantage, Simpson and Gurevych (2018) propose a pointwise algorithm based on Gaussian Process Preference Learning, *GPPL*, (Chu and Ghahramani, 2005) which is able to output a convincingness score per argument, while being trained on the pairs of arguments from *UPKConvArg*. They use the same huge set of linguistic features and word embedding. They emphasize the importance of the pointwise approach, allowing for a more scalable and efficient inference. They also note that the benefits in avoiding neural networks lie in the superiority of graphical models for small training sizes like in *UPKConvArg*.

The Siamese network we propose next has all of those advantages and more. Being a neural network, it has the advantage of being more efficient in inference, and it can be updated with the frequent advances of this research field. In addition, the pre-processing step which generates the huge set of linguistic features, used by Habernal and Gurevych (2016a) and Simpson and Gurevych (2018), takes a lot of time and is not suitable for many languages. In contrast, our network does not depend on task specific features, and still achieves state of the art results on the task of argument convincingness classification and ranking.

3 Siamese Network

For the task of learning pointwise evidence convincingness scores from a data set of evidence pairs, we bring ideas from the field of learning to rank. Specifically, in our model, we take inspiration from the training procedure provided by RankNet (Burges et al., 2005) of a Siamese network. Such a network consists of two legs of identical networks, which share all their parameters and are connected at the top with a softmax.

Unlike RankNet, we propose a network whose output is a probability. This is a desirable property as it is comparable with the output of other networks, and is understandable by humans. In initial experiments, on held-out data, the performance of our network was comparable to that of RankNet.

Each leg in our Siamese network is a neural network which is a function of an input argument A and has two outputs $[C_A, D_A]$. C_A represents how convincing A is, and D_A is a dummy output (which can be a constant).

In training, given a pair of arguments, A_i and A_j, we apply $softmax[C_{A_i}, C_{A_j}]$, softmax over the convincingness output of each leg (ignoring the dummy output). The result is compared to the label of the pair using the cross entropy classification loss. Intuitively, this maximizes the probability of correctly identifying the more convincing argument, which pushes the margin between the two outputs to differ.

In inference, given a single argument, A_k, rather than a pair, the advantage of the Siamese network comes into play. To predict the convincingness score for A_k we feed it into one of the legs. Then, we apply softmax, this time over the con-

vincingness output and the untrained dummy output, $softmax[C_{A_k}, D_{A_k}]$. The higher the probability we get from this softmax, the more convincing A_k is considered by the network.

Implementation of a leg. Each leg in our Siamese network is a BiLSTM with attention as described in Shnarch et al. (2018). We feed non-trainable word2vec embeddings (Mikolov et al., 2013) to a BiLSTM of width 128, followed by 100 attention heads and a fully connected layer with two outputs. Training was done using Adam optimizer (Kingma and Ba, 2015) with learning rate 0.001, applying gradient clipping above norm of 1 and dropout rate of 0.15. The system was trained for 10 epochs.

4 The *IBM-EviConv* data set

Following the motivation, presented in the introduction, for a more focused framing of the argument convincingness task, we release a new data set, *IBM-EviConv*[2].

This data set is composed of 1,884 unique evidence sentences, extracted from Wikipedia, derived from the data set released by Shnarch et al. (2018). These evidence spread over almost 70 different debatable topics.

From the evidence set of each topic we sampled pairs, therefore within a pair, both evidence sentences refer to the same topic, arguing either for the topic (PRO) or contesting it (CON). In total we annotated more than 8,000 pairs, and after a cleaning step (detailed in §4.1) we were left with 5,697 pairs that are split into train and test sets (4,319 and 1,378 pairs correspondingly). We kept the same split of Shnarch et al. (2018) in which no topic appears both in train and in test.

The label of each pair indicates which evidence is more convincing out of the two. In addition, we provide the stance of each evidence towards the topic.

Following is an example of a pair from our data set in which the first evidence was chosen to be more convincing:

Topic: We should legalize same sex marriage.

Evidence #1: The California Supreme Court overturned California's ban on gay marriages on May 15, stating that depriving gays and lesbians of the same rights as other citizens is unconstitutional. (PRO)

Evidence #2: In his 2002 Senate campaign, Coleman pledged support for an amendment to the United States Constitution that would ban any state from legalizing same sex marriage. (CON)

Using Wikipedia as the source for evidence yields a data set that is rather homogeneous in vocabulary, grammar and style, as Wikipedia is heavily edited. In addition, as motivated in §1, we constructed pairs of evidences with roughly the same length, allowing for a length difference of up to 30% of the shorter evidence. The evidence in each pair can have either the same stance or the opposite stance towards the topic. Overall, we annotated 3,075 pairs of the same stance towards the topic and 2,622 cross stance pairs.

Each pair in *IBM-EviConv* was annotated by 10 crowd labelers.[3] Out of all the pairs a labeler annotated, 20% were *hidden test questions* used to verify annotations quality (see §4.1).

The labelers were provided with the following guideline, and were asked to be decisive:
In a conversation about the topic, where you can only give a single evidence out of the following two - which one would you rather use? Which one is more convincing?

We consider an evidence to be more convincing than its counterpart if it was chosen by at least 60% of labelers. Pairs in which one evidence was preferred by more than half of the labelers but less than this threshold were considered as indecisive and were removed from the data set.

After data set cleaning, described next, the most frequent label in the data set (train and test sets together) covers only 53% of the pairs. Hence, it is safe to say the data set is balanced and there is no strong bias towards a certain sentence length.

4.1 Data set quality

We took several measures to ensure the quality of *IBM-EviConv*. First, we selectively picked crowd labelers based on their performances and credibility on previous tasks of our team. In total, 92 labelers of this group took part in the annotations.

We initially performed a pilot annotation task to evaluate the quality of the crowd work by comparing their annotations to those of in-house expert

[2] Available on the IBM Project Debater datasets webpage: http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

[3] we used the Figure-Eight platform: https://www.figure-eight.com/

labelers. The pilot contained 105 pairs that were labeled by both groups. We filtered out 21 pairs whose labeling was indecisive by either group (see previous section) and found out that for 84% of the remaining pairs the two independent groups agreed on the label. This encouraging result indicated that the crowd labelers are suitable for this annotation task.

We further filtered specific crowd labelers whose work did not adhere the following requirements: (i) annotating a minimal number of 20 pairs, (ii) obtaining a minimal average Cohen's Kappa (Cohen, 1960) of 0.1 calculated over enough shared content with other labelers (i.e., sharing at least 20 pairs with at least 10 labelers).

Another filter mechanism is based on hidden test questions. These test questions were constructed automatically by pairing a confirmed evidence with a rejected evidence candidate from the data set of Shnarch et al. (2018). Naturally, the confirmed evidence is labeled as the more convincing one, since the other sentence is not even a proper evidence. These questions were randomly placed among the true pairs for annotation and the labelers could not know which question was a test one (therefore are called hidden). A labeler whose precision over these test questions was below 0.55 was filtered out. The average precision of the remaining labelers over the hidden test questions is 0.73.

In total, we filter 23 labelers by these criteria, and removed all of their annotations. Following this process, pairs that were left with less than 7 annotations by valid labelers were removed from the data set to maintain a high standard of majority.

We use Cohen's Kappa to calculate the average pairwise agreement of the labelers, yielding the score of 0.33. We note that the average pairwise Cohen's Kappa score of our expert labelers was 0.38 on this task, indicating the difficulty of this task to humans. This agreement level is a typical value in such challenging labeling tasks (e.g., Aharoni et al., 2014). We consider it an upper bound and therefore see the 0.33 Kappa score of our crowd labelers as an acceptable agreement.

Finally, a desired characteristic of such a data set is that transitivity among evidence pairs will hold (Habernal and Gurevych, 2016a). This is our last quality test of *IBM-EviConv*. We extract all 1,899 triplets of evidence for which all pairs were annotated. Then, we calculate the percentage of such triplets in which transitivity holds, i.e. one evidence is consistently considered the most convincing, one the least convincing and the third is in the middle. The results were surprisingly high, 99% of the triplets comply with the transitivity expectation.

5 Experiments

From this point on, we will refer to the method presented in §3 as *EviConvNet*.

5.1 Experiments over *UKPConvArgStrict* and *UKPConvArgRank*

We first experiment on the argument convincingness data set released by Habernal and Gurevych (2016a). It is split into two tasks: pair classification on *UKPConvArgStrict*, and ranking on *UKPConvArgRank*. On UKPConvArgStrict, all systems were evaluated in cross-topic validation over 32 topics (16 actual topics, with 2 stances each) and their average accuracy across folds is reported in Table 1.

System	Accuracy
Most frequent label	0.50
BiLSTM	0.76
Argument length	0.77
SVM	0.78
GPPL opt.	0.80
GPC	**0.81**
EviConvNet	**0.81**

Table 1: Accuracy on *UKPConvArgStrict*. Our model (EviConvNet) is comparable to the best baseline.

BiLSTM and *SVM* are the methods presented in the original paper (Habernal and Gurevych, 2016a). *GPPL opt.* and *GPC* are Gaussian process methods later demonstrated by Simpson and Gurevych (2018). Our own EviConvNet performs similarly to the best previously known systems on this task.

Most of these systems were also evaluated on UKPConvArgRank, where each single argument is assigned a score (yielding a ranking). The original work reported Pearson's r and Spearman's ρ on the combined ranking of all arguments from all 32 topics. Subsequent work (Simpson and Gurevych, 2018) reported the average of these measures across topics. We report their results and ours in this setting in Table 2. Again EviConvNet

provides at least equal performance to the best previously known method, *GPPL opt*, and a statistically significant increase in Pearson's r ($p \ll 0.01$ using one-sample two-tailed t-test).

System	Pearson's r	Spearman's ρ
Argument length	0.33	0.62
BiLSTM	0.36	0.43
SVM	0.37	0.48
GPPL opt.	0.44	**0.67**
EviConvNet	**0.47**	**0.67**

Table 2: Correlation measures on UKPConvArgRank. Our model (EviConvNet) is comparable to the best baseline.

We also tested a simple baseline assigning the argument's character length as its score (*Argument length*). In pair classification, the baseline prefers the longer argument. We noted performances comparable to the original method of Habernal and Gurevych (2016a). This result is in line with what the authors reported in a further study of the reasons given by annotators for preferring one argument over the other (Habernal and Gurevych, 2016b): the most common reason provided is by far "more details, information, facts, examples / more reasons / more specific".

5.2 Experiments over *IBM-EviConv*

We report in Table 3 the accuracy of various baselines and our own method, on the full *IBM-EviConv* data set.

The simplest baseline is preferring the longest sentence, as before, but on this data set it has nearly the same accuracy as just picking the first candidate every time (*most frequent label*).

The *Detection model* assigns a score to each individual sentence, and we choose the sentence with the highest score. To produce this score, a single leg of the network presented in Section 3 is used, with a softmaxed 2-dimensional output, trained using cross entropy classification loss over evidence candidates in the base data set from Shnarch et al. (2018).[4]

We also run the GP-based methods proposed by Simpson and Gurevych (2018)[5]. The increase these methods bring over the detection model is statistically significant ($p \ll 0.01$ using Wilcoxon signed-rank test). EviConvNet, the Siamese network described in §3, significantly outperforms all systems ($p \ll 0.01$). We note that the gains from better methods are far greater here than in §5.1: GPPL improves over the sentence length baseline by 26% and our method improves over GPPL by 9% on *IBM-EviConv*, compared to improvements of only 5% and 1% on *UPKConvArg*.[6]

System	Accuracy
Evidence length	0.53
Most frequent label	0.54
Detection model	0.59
GPPL	0.67
GPPL opt.	0.67
GPC	0.67
EviConvNet	**0.73**

Table 3: Accuracy on *IBM-EviConv*. our model (EviConvNet) outperforms prior art and our additional baselines.

6 Analysis

In this section we present an analysis of several interesting aspects of our new data set and method.

6.1 Performance across preference reasons

Habernal and Gurevych (2016b) analyze and categorize the reasons provided by the labelers of *UPKConvArg* to justify their choice on each pair (see Table 4 for examples of reasons). The most common reason is that an argument is more informative (code C8-1 in the table). As valid and pervasive as this factor is in real arguments, it also makes the argument length a high-performance baseline which is hard to beat (as seen in §5.1).

One of the motivations for our work was to create another data set where the amount of textual content would not be a factor in the choice of labelers, possibly constraining the preference task to the more subtle aspects of "convincingness".

We compute the error rate (1 - accuracy) of the length baseline and EviConvNet on pairs clustered by their *reason units* as defined by Habernal and Gurevych (2016b). For clarity of the analysis, the pairs were restricted to those where a single reason

[4]Detection scores are provided with the data set we release for ease of reproducibility.

[5]Using their code from https://github.com/ukplab/tacl2018-preference-convincing.

[6]Percentages are relative to the accuracy of the system or baseline referred to. This is to allow a more meaningful comparison of behavior on the two datasets.

was given. We selected the four reasons presenting the highest relative decreases in error rate and the three single reasons where the baseline outdoes our method. Figure 1 shows the relative decrease in error rate between argument length baseline and EviConvNet, with the reason codes defined in Table 4.

We note that unsurprisingly, our neural network model does better than the length baseline at capturing what an argument should be like in term of presentation, relevance and quality of content.

Code	Reason
C9-4	Well thought out / higher complexity
C5	Language / presentation of argument
C7-3	Off-topic / doesn't address the issue
C7-1	Not an argument / is only opinion / rant
C6-1	Not enough support / not credible
C8-1	More details, information, examples
C8-4	Balanced, objective, several viewpoints

Table 4: Reasons for choosing a better/worse argument taken from Habernal and Gurevych (2016b).

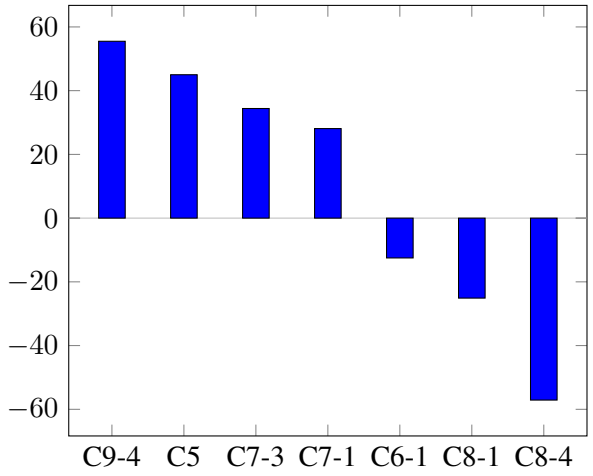


Figure 1: Relative decrease in error rate (%) between argument length baseline and EviConvNet (reason codes defined in Table 4).

More interesting are the reasons where the neural network does not perform as well. Two of those are about the sheer amount of supporting information, which would indeed be more directly captured by the length baseline. The reason where EviConvNet has a 57% greater error rate is about presenting a balanced, objective argument which tackles different viewpoints. These pairs only make up 3% of the data set, so it is possible the network needed to see more such pairs in training to perform well on them.

6.2 What makes a convincing evidence?

We asked the experts. We asked our in-house expert labelers to supply factors for commonly deciding their preference (their answers are released with the data set). Reliability of the source was an important factor, including titles and names, level of expertise, type of evidence (study, expert, opinion, example, precedent), whether the source has an interest in the discussed matter, and where it came from geographically.

Also important were content issues, whether information was complete, specific, significant, rhetorically strong, the amount of supporting evidence or details reported and the relevance of the evidence to the present, and better yet to the future or in general. Some technical issues were also reported, such as missing information or an opinion rather than a fact.

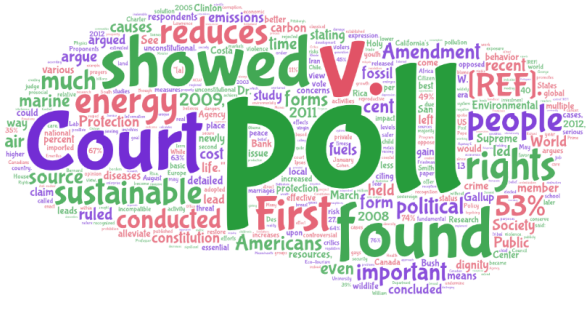
Additionally we inquired about the cases that were difficult to compare. These tended to be either cases where both pieces of evidence were not convincing (for the reasons above), or where it was hard to ascertain – for a certain factor (e.g. reliability or significance) – which argument prevailed.

We asked the network. To acquire a better understanding of what typical things differentiate convincing evidence from non convincing ones, we compared word distributions on pairs that our network successfully classified (Figure 2). From the correctly classified pairs we construct two sets; one is composed of the more convincing evidence in each pair, *true convincing*, and the other contains the less convincing ones, *true non-convincing*.

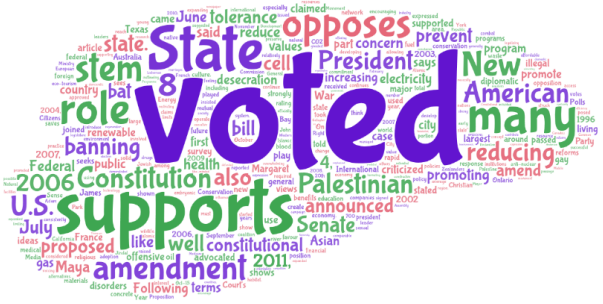
For each set, we calculate the distribution of unigrams in its evidence sentences, ignoring stop words and unigrams which appear in the topic title. In Figure 2 we present the differences between the two distributions, thus, discarding words that are common in many evidence sentences regardless of the convincingness of their texts.

On the left side of Figure 2 we see the words which are much more prominent in convincing arguments than they are in non-convincing ones. We find there words related to argumentation (*argue*, *claim*), or studies and polls (*found*, *conducted*, *[REF]*[7]). Other words mention authoritative fig-

[7] A sign which indicates that this evidence was taken from a written source (it replaces the reference text to the source).



(a) true convincing dist - true non-convincing dist



(b) true non-convincing dist - true convincing dist

Figure 2: Word clouds of correctly classified pairs.

ures (*DR.*, *Clinton*, *W. Bush*) or court orders (*supreme*, *v.*[8]).

On the other hand, when subtracting the true convincing distribution from the true non-convincing distribution (Figure 2b) one gets opinion words (*support*, *opposes*, *vote*), partial change (*reduce*, *amend*, *part*), non-emphasized actions (*said*, *proposed*, *concern*).

6.3 Cross vs. same stance evidence pairs

As described in §4 we build our data from pairs with the *same-stance* towards the topic, as well as *cross-stance* pairs, in which one argument supports the topic while the other opposes it. We created the cross-stance pairs since we had in mind the task of comparing arguments of different sides of a debate. Given this task, some questions naturally arise, such as

- Is it a harder task to identify the more convincing argument when comparing arguments of opposite stances?
- Is it better to train on cross-stance pairs for this task?

With *IBM-EviConv* we can empirically examine such questions.

To this end, we extracted three subsets from the training set (of 2,082 pairs each); one with same-stance pairs only, the second with cross-stance pairs only, and the third with mixed-stance. Similarly we extracted three subsets from the test set (each with 385 pairs).

Given these data sets we are able to test what happens when we train and test our network on all combination of pairs of same/cross/mixed stance.

Table 5 depicts the results. To our surprise, training on cross-stance pairs does not improve performance on a test with cross-stance pairs in comparison to training on same or mixed stance pairs (middle column in the table). Same goes for the other subset. In addition, it appears that cross-stance pairs do not pose a more difficult task than same-stance pairs or mixed-stance, as the accuracy over them is not smaller than the accuracy on same-stance or mixed-stance pairs.

		Test		
		same	cross	mixed
Train	same	0.72	0.71	0.71
	cross	0.72	0.69	0.71
	mixed	0.72	0.71	0.70

Table 5: Comparing accuracy when training and testing on each combination of same/cross/mixed stance.

6.4 The effect of length difference

In previous sections, we discussed our choice to limit the difference in length between evidence of the same pairs. This decision was encouraged by the relatively good results of the argument length baseline on the *UPKConvArg*. Still, one may wonder whether training on similar length pairs harms the performance over real life pairs in which length balance is not guaranteed. For that purpose we annotated 458 pairs with a significant length difference (higher than 30%, complementary to the restriction in *IBM-EviConv*).

The accuracy of EviConvNet over this test set is 0.69, which is lower than 0.73, the accuracy over the balanced data set, reported in Table 3, but still higher than all other baselines. This difference is small enough to conclude that our model, trained on a length balanced data set generalizes well enough to be able to classify pair of evidence of different lengths.

[8] V. is used as a versus abbreviation in court rulings.

7 Discussion and future work

In this work we proposed a focused view for the task of argument convincingness, constructed a new data set for it, and presented its advantages. We believe that it is useful to evaluate methods for identifying the more convincing argument on this more challenging data set.

In addition, we suggest our version of a Siamese network for the task, which outperforms state of the art methods.

A possibility that we did not expand on in this paper is to pre-train one leg of the network on an argument detection data set, like the one of Shnarch et al. (2018). Argument detection concerns itself with the binary classification of a single text into argument and non-argument, and not the more subjective notion of convincingness. But we nonetheless observed in previous experiments significant improvements when initializing the Siamese network with weights learned on this task. We could not reproduce these improvements here, but our previous efforts relied on far fewer training pairs: an explanation could be that pre-training is most helpful when faced with a low amount of training data.

In the future we aim to test and adapt other improvements in the learning to rank field to our task, hoping for further improvement by those models (Burges, 2010; Severyn and Moschitti, 2015). In addition, more careful design of the architecture details, which was not the focus of this work, will probably yield better results yet, e.g., contextualized word embeddings (Peters et al., 2018), batch normalization (Ioffe and Szegedy, 2015; Cooijmans et al., 2017), deeper networks and other architecture practical heuristics.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first Workshop on Argumentation Mining*, pages 64–68.

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.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, pages 89–96.

Christopher J. C. Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. In *Learning*.

Eugene Burnstein. 2003. Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of experimental social psychology*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proc. of NAACL*.

Wei Chu and Zoubin Ghahramani. 2005. Preference learning with gaussian processes. In *ICML*.

Cohen. 1960. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, pages 37–46.

Tim Cooijmans, Nicolas Ballas, César Laurent, and Aaron C. Courville. 2017. Recurrent batch normalization. *CoRR*, abs/1603.09025.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *NAACL 2018*.

Brian J Fogg. 1998. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 225–232.

Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Interactions*, 11:65–66.

Brian J Fogg. 2009. A behavior model for persuasive design. In *PERSUASIVE*.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *ACL 2016*.

Melanie C. Green and Timothy C. Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79 5:701–21.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *EMNLP*.

Ivan Habernal and Iryna Gurevych. 2016a. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *ACL*.

Ivan Habernal and Iryna Gurevych. 2016b. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *EMNLP*.

Charles Leonard Hamblin. 1971. *Fallacies*. Philosophy.

Christopher Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *AAAI*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *COLING-14*, pages 1489–1500.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Brian McFee and Gert R. G. Lanckriet. 2010. Metric learning to rank. In *ICML*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Daniel J O'Keefe. 2012. Conviction, persuasion, and argumentation: Untangling the ends and means of influence. *Argumentation*, 26(1):19–32.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *EMNLP*.

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 *emnlp-15*, pages 440–450.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):30.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382. ACM.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of ACL*, pages 599–605. Association for Computational Linguistics.

Edwin D Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014*, pages 1501–1510.

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

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *WWW*.

Christopher W Tindale. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Frans H Van Eemeren, Bart Garssen, Erik CW Krabbe, A Francisca Snoeck Henkemans, Bart Verheij, and Jean HM Wagemans. 2014. *Handbook of argumentation theory*. Springer.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation quality assessment: Theory vs. practice. In *ACL 2017*.

Lisa Weltzer-Ward, Beate Baltes, and Laura Knight Lynn. 2009. Assessing quality of critical thought in online discussion. *Campus-Wide Information Systems*, 26(3):168–177.