Different Flavors of Attention Networks for Argument Mining

Johanna Frau,[1] **Milagro Teruel,**[1] **Laura Alonso Alemany,**[1] **Serena Villata,**[2]
[1]Natural Language Processing Group, FAMAFyC-UNC, Córdoba, Argentina.
[2]Université Côte d'Azur. CNRS, Inria, I3S, France.
jfrau@famaf.unc.edu.ar, lauraalonsoalemany@unc.edu.ar, mteruel@unc.edu.ar, villata@i3s.unice.fr

Abstract

Argument mining is a rising area of Natural Language Processing (NLP) concerned with the automatic recognition and interpretation of argument components and their relations. Neural models are by now mature technologies to be exploited for automating the argument mining tasks, despite the issue of data sparseness. This could ease much of the manual effort involved in these tasks, taking into account heterogeneous types of texts and topics. In this work, we evaluate different attention mechanisms applied over a state-of-the-art architecture for sequence labeling. We assess the impact of different flavors of attention in the task of argument component detection over two datasets: essays and legal domain. We show that attention not only models the problem better but also supports interpretability.

Introduction

Argument Mining (Peldszus and Stede 2013; Lippi and Torroni 2016; Cabrio and Villata 2018) tackles a very complex phenomenon, involving several levels of human communication and cognition. It has been defined as "the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand" (Habernal and Gurevych 2017). Due to the complexity of the task, data-driven approaches require a huge amount of data to properly characterize the phenomena and find patterns that can be exploited by a classifier. However, most of the corpora annotated for this task are small, and they cannot be used in combination because they are based on different theoretical frameworks (e.g., abstract and structured argumentation) or cover different genres (e.g., political debates, social network posts). Our primary research domain is argument mining in legal cases, where corpora are especially scarce.

In this paper, we analyze the utility of neural attention mechanisms to improve the tasks of Argument Mining. It has been shown that some kind of attention improves the performance of neural-based argument mining systems Stab et al. (2018). It seems that attention mechanisms concentrate the classifier on the most determinant elements to take into account, and thus direct the training of the classifier to a better convergence point. Moreover, attention has a promise of interpretability, since the parts of the input that receive more attention can be recognized.

We assess the impact of adding an attention mechanism to a state-of-the-art neural model (BiLSTM with character embeddings and a CRF layer) (Reimers and Gurevych 2017)[1].We apply different flavours of attention and compare their impact on performance. We assess the impact of performance over two very different corpora for Argument Mining: a corpus of persuasive essays (Stab and Gurevych 2017) and a small corpus of the legal domain, consisting of only 8 judgments (Teruel et al. 2018) of the European Court of Human Rights (ECHR)[2]. We focus on the argument component detection task, distinguishing claims and justifications or focusing on claim detection alone.

We show that attention mechanisms consistently improve performance. Additionally, it helps to identify words that are important for the Argument Mining task, useful for guidelines and selection of examples for annotation.

In the rest of the paper we present relevant work, describe the architecture of the attention-based system, then we detail our experimental setting and we discuss the obtained results.

Related work

In this section, we report on the approaches proposed in the area of Argument Mining employing recurrent neural networks and attention mechanisms. For state-of-the-art on argument mining, we refer the reader to (Cabrio and Villata 2018). Recurrent neural networks have been successfully applied to the problem of Argument Mining. In Eger et al. (2017), a single model is trained to jointly classify argumentative components and the relations of attack and support between them. The model is evaluated over the dataset of argumentative essays. Stab et al. (2018) applied attention to crowd-sourced general domain argumentative essays from the Web, to measure how relevant each word on the input is, with respect to the topic of the essay.

Our final aim is to find a good system for argument mining in legal cases as the ones in the ECHR corpus. In legal documents, the length and complexity of the texts is bigger than for essays, there is no pre-defined topic and the spans to

[1]https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

[2]hudoc.echr.coe.int

be detected are intra-sentential, which makes the task considerably more complex than for essays.

Architecture

In this section, we first describe the Bidirectional Long Short-Term Memory (BiLSTM) we adopted for our Argument Mining task, and second we present the attention mechanism we empowered the BiLSTM with.

BiLSTM architecture

Our proposed model adds an attention mechanism over the recurrent architecture implemented by (Reimers and Gurevych 2017), which we will briefly describe in this section. The base of the model is a bidirectional recurrent layer with LSTM units. The input layer is composed of three parts: a word embedding, a character embedding and a casing embedding. The word embedding uses pre-trained weights, supporting any type of word vectors. The character embeddings generate a representation for each word by performing either a convolution operation or a recurrence operation at a character level. The objective of this type of embeddings is to generate a representation of the out-of-vocabulary (OOV) words based on their morphology. The weights for both methods are trained with the network. The casing embedding generates a very small representation of the word using features like presence of digits or all uppercase characters. Finally, the model replaces the traditional classification layer, a dense layer with a softmax activation, with a Conditional Random Field (CRF) layer (Huang, Xu, and Yu 2015). A CRF takes into account the surrounding labels to calculate the label for the current word. This mechanism should produce more consistent labels, to avoid for example starting a claim in the middle of a premise.

The attention mechanism

Attention mechanisms, in simple terms, weight the output of a layer with importance scores. These scores are used to increase the values of inputs which are relevant to the prediction of the correct output, and to decrease the values of inputs that are not relevant. Different types of attention mechanisms arise depending on how we calculate the attention scores: we can use only the information from each example individually, or use the whole set of examples.

In this work, we propose a simple attention mechanism applied directly to the input words of the sentence, visualized in Figure 1. The goal of placing the attention layer before the recurrent layer is to weight which words in the sequence are relevant to the task we want to solve. This type of attention, called *inner attention*, was firstly proposed by (Wang, Liu, and Zhao 2016).

The *attention layer* added to the base RNN model is composed of two parts: one that calculates the attention scores, and a merge function $\otimes$ that combines them with the original input. As a result of this combination, the attention scores regulate how much of the initial value of the cells in the input are passed to the following layer, increasing or alternatively turning off certain values.

The *attention scores* are calculated using a fully connected layer. This layer has the following parameters: a weight matrix W^A, a bias b^A vector, and an activation function f. Once we obtain the attention scores, the merge function (in this case multiplication) is applied pointwise. Let it be x a single sequence with n timesteps and m features (input of the network), s the attention scores, and x^A the output of the attention layer, then the operations performed inside the attention layer are:

$$\begin{aligned} s &= f(xW^A + b^A) \\ x^A &= x \otimes s \end{aligned} \tag{1}$$

Note that, to obtain the attention score s_i (corresponding to the i-th input on the sequence), we are multiplying each row x_i by the parameter W^A. There is an important implication to this equation: the attention scores for each cell in the input are computed taking into account only the features in that timestep input, and not using any information present on the rest of the elements in the sequence. We call this approach *word attention*.

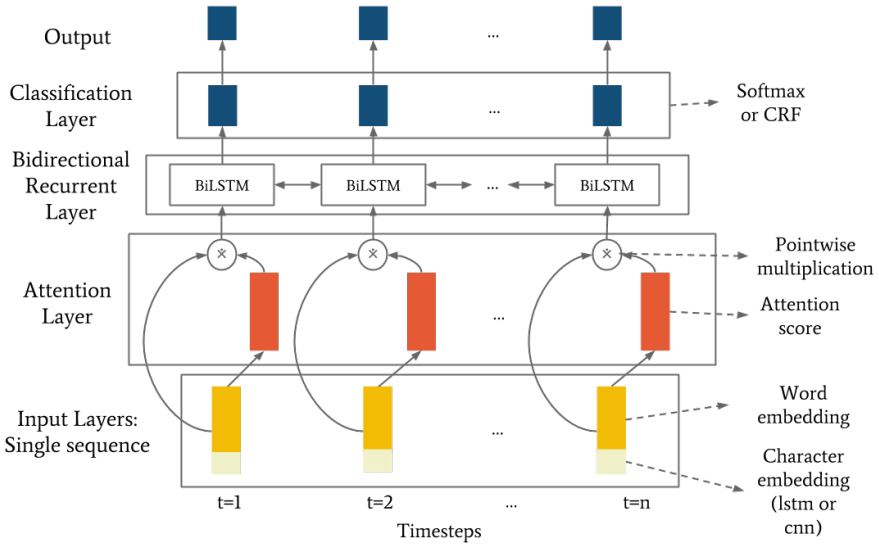
This may seem counter intuitive, as we expect the importance of a word in the sentence not to be only determined by itself, but also by the words with which it co-occurs, by the property of compositionality of sentential semantics. We have also explored a different attention mechanism, where the attention score for feature j in instance x_i is calculated using the value of feature j in all timesteps of the sequence. The model, which only differs in two transpositions, is described in Equation . We call this approach *context attention*.

$$\begin{aligned} s &= f(x^T W^A + b^A) \\ x^A &= x \otimes s^T \end{aligned} \tag{2}$$

As we can see, all the operations of the two attention layers are differentiable, and we can optimize its parameters along with the training of the original model.

The selection of the activation function f has also an important impact. We propose three different options: linear activation, a sigmoid function, and an hyperbolic tangent function. If we use linear activation, the values of the attention scores are not limited to any range, and could possibly explode to huge values, become negative or vanish to nearly zeros. However, in our experiments the values remain in reasonable intervals (between 1 and 10). Furthermore, linear activation makes the attention scores a linear combination of the original values, but we expect the pointwise multiplication with the mean score of the instance to effectively add expressiveness to the model. The sigmoid function, on the other hand, introduces non linearity and forces the attention scores to the (0, 1) interval. After the application of the scores, all the input activations are decreased. This could lead to a faster vanishing of the neuron's signals as they propagate through the network. Finally, the hyperbolic tangent (tanh) function does not have this problem, as it constraints the scores to the interval (-1, 1). However, it enables negative attention, which does not have an intuitive interpretation.

Figure 1: Attention mechanism applied to the Char embedding + BiLSTM + CRF architecture.



Experimental Setting

Datasets

We applied the model on two datasets developed for the Argument Mining task of argument component detection. The *essays dataset* consists of 400 argumentative essays (148,000 words) written by students in response to controversial topics (Stab and Gurevych 2017). In this dataset, the argument components are separated in Claims, Major Claims and Premises, but we collapsed the distinction between Major Claims and Claims to make it homogeneous with the ECHR corpus.

Our primary research interest concerns a small corpus of the European Court of Human Rights (*ECHR dataset*) (Teruel et al. 2018), with 8 judgments (28,847 words). It has, 329 claims and 401 premises. As input for the attention learners, each sequence is an entire paragraph, and the classifier is expected to detect all components within the paragraph.

A major difference between these two corpora, besides their sizes, is the length of the sequences. For the essays corpus, the longest paragraph is 98 words long, while the longest paragraph in the ECHR documents is 317 words long. The maximal length determines the length of the sequences given to the classifier for each corpus. The longer the sequence, the more difficult the task of calculating *context attention*, as context distributes attention along the whole length of the sequence.

Evaluation procedure and metrics

To evaluate our models using the ECHR dataset, we use a 8 fold cross validation where, in each fold, we leave out one entire document for testing, instead of just leaving out a subset of examples. We consider this setting to be more representative of the real case scenario of a production system, where the new instances are documents never seen before. As the essays dataset has more documents, we perform the evaluation on 9 holdout documents.

The evaluation metric used is the F1-score, the harmonic mean between precision and recall. All values reported are averaged between classes with a weighted strategy, to avoid giving too much weight to the minority B- classes.

Hyperparameter search

During the experiment phase, we noted that, in the ECHR dataset, models were highly sensitive to the architecture and hyperparameter selection. Contrary to the results on previous works, simpler models without character embeddings and CRF do not perform necessarily worse. As a result, we use a random search to find successful combinations of hyperparameters. The evaluated combinations are:

- BiLSTM layer with identical layers of 30, 50, 100, or 200 units.
- Mini batches sizes of 30, 50, 100, or 200.
- Dropout on the LSTM layer of 0.1, 0.2, 0.3, 0.4, or 0.5.
- Classification layer with softmax activation and with CRF activation.
- Character embeddings with convolutions and with recurrence, resulting on vectors of sizes 16, 32 or 64.

The optimizer used is Adam, with a clip norm of 1. All classifiers are trained during 50 epochs, stopping if there is no improvement for 10 epochs.

Evaluation

In this section, we first discuss the results we obtain with our classifier for argument component detection empowered with attention, distinguishing claims and premises and focusing on claim detection alone. Claim detection is a simpler, better defined task which still retains utility in applications. Indeed, Teruel et al. (2018) show that inter-annotator

agreement is higher for claims than for premises. Thus the training and evaluation data for claim detection are more consistent, therefore it provides for a fair evaluation.

Secondly, we present how attention mechanisms can be used to provide an explanation of the output of the Argument Mining system, to be integrated to increase consistency of annotations (as in guidelines for annotators) or as a criterion to select examples to be annotated.

Classifier performance

Tables 1 and 2 show the best results obtained after hyperparameter search for the essays corpus and the ECHR corpus, respectively. As could be expected, we obtain better results for the task of *claim detection* than when we distinguish claims, premises and non-components.

The performance for the ECHR is also worse than for essays. This is probably due to various factors: the essays corpus is bigger, but also sentences in that corpus are shorter, and argument components are less complex.

Table 1: Best results obtained for claim detection (left) and component detection (right) after hyperparameter search in the essays corpus.

	Claim Detection		Claim - Premise	
	Acc	F1	Acc	F1
No attention	0.844	0.841	0.696	0.683
Word + lin	0.831	0.834	0.684	0.684
Word + sig	0.829	0.833	0.700	0.695
Word + tanh	**0.841**	**0.837**	**0.717**	**0.704**
Context + lin	0.786	0.756	0.709	0.699
Context + sig	0.797	0.799	0.703	0.691
Context + tanh	0.781	0.754	0.702	0.698

Table 2: Best results obtained for claim detection (left) and component detection (right) after hyperparameter search in the ECHR corpus.

	Claim Detection		Claim - Premise	
	Acc	F1	Acc	F1
No attention	0.805	0.793	0.661	0.652
Word + lin	**0.824**	**0.816**	0.668	0.673
Word + sig	0.822	0.810	**0.680**	**0.683**
Word + tanh	0.821	0.814	0.674	0.682

Concerning the configuration of the classifiers, we can see that *word attention* systematically obtains better results than *context attention*, with context attention performing worse than a system with no attention in some cases. It has to be noted that context attention has a very strong parameter, the length of the context to be used to find attention. The longer the context, the more computational resources are required. This is an interesting parameter to explore and may have a big impact in improving the performance of context attention, especially if we take into account the extensive length of paragraphs in the ECHR corpus. This will be explored in future work.

Results for the rest of hyperparameters are not conclusive. In Figure 2 we can see the distribution of F1-Scores on different configurations of classifiers for the task of argumentative component detection, for the ECHR and the essays dataset, respectively. These classifiers were trained for the selection of hyperparameters with random combinations, as described in the previous section.

We can see that there is a wide range of variability across results. For the essays corpus, we can see that the difference in performance between word attention and context attention falls within the range of variability of different parameters with the same kind of attention. This means that the combination of hyperparameters affects performance more importantly than the kind of attention. However, word attention tends to have slightly better performance and more stable (less variable results).

In the case of the ECHR corpus, word attention performs clearly better than context attention, beyond the impact of the rest of hyperparameters. This is probably due to the fact that sequences given to the classifier are whole paragraphs, which are thrice longer than the sequences for the essays corpus. The task of context modelling is more difficult when sequences are longer. In contrast, the formulaic, repetitive nature of legal text may be useful to concentrate attention in particular words or word sequences, which is particularly adequate for word attention mechanisms.

We cannot see an important difference in variability across activation functions: the three kinds of activation present wide ranges of variability.

However, the effects that activation have in the distribution of the attention scores are intrinsically different. For both datasets, the scores given with the sigmoid and tanh activation concentrate on values greater than 0.9, and few words obtain a low attention score. In contrast, linear activation concentrates the scores in a middle value, usually close to 3, with a gaussian-like distribution. We have observed classifiers where linear activation concentrates the attention scores close to zero, effectively shutting down the signals from irrelevant words. We are currently working on establishing a correspondence between these patterns of distribution of attention and improvements in performance or interpretability.

Visualizing attention

To further explore the effects of attention, we visualize the attention scores assigned by a model. We propose a visualization where the intensity of the color over each word in the text denotes its attention. The scale of intensity is linear, normalized according to attention values. The color itself represents the label assigned by the classifier: red for claims, blue for premises, gray for the O class. Words misclassified by the model are underlined.

In Figure 3 we show the labels and attention scores assigned by a word attention, sigmoid classifier for claim detection on a fragment of the ECHR corpus.

We can see that words with low attention are irrelevant to detect claims, like stopwords. In contrast, words with high attention are very relevant, as is the case of "*inadmissible / admissible*" or "*conclude*". This kind of information is very

Figure 2: Distribution of F1-Scores on different architectures in the ECHR and Essays dataset, for the argumentative component classification task.

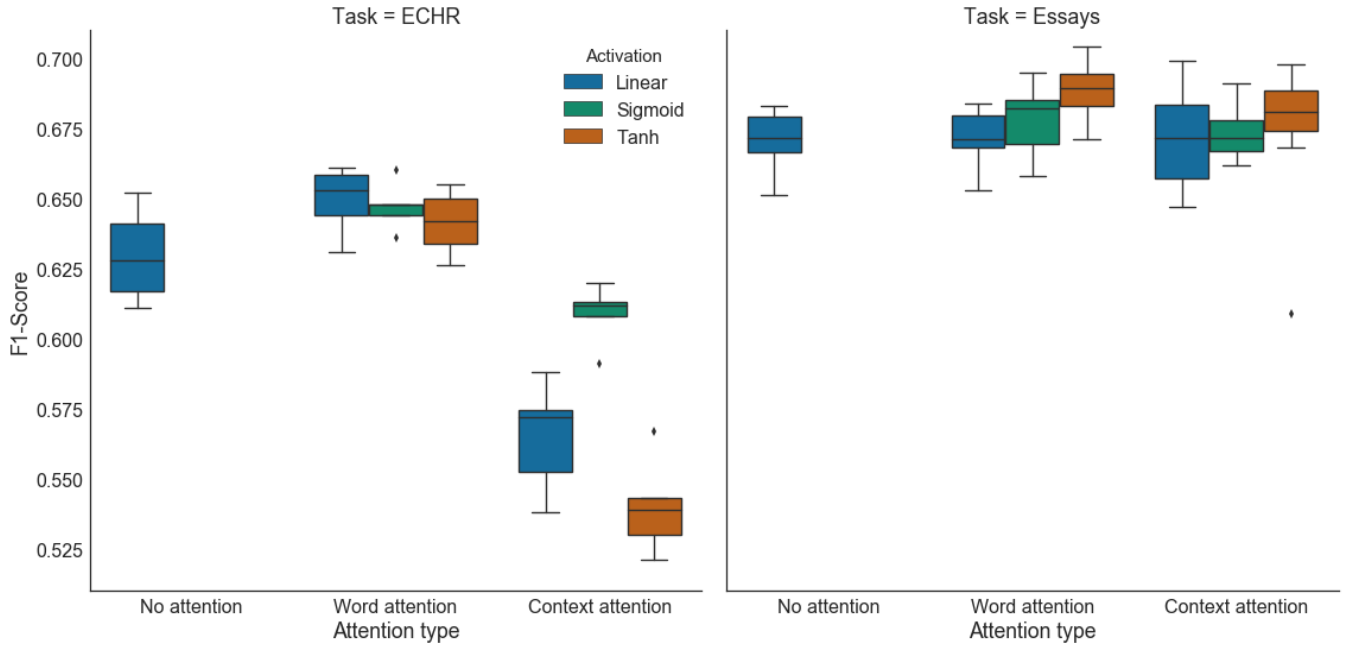


Figure 3: Attention scores assigned to a fragment of the ECHR corpus for the claim detection task. The model uses word attention with sigmoid activation.

22 . In the case of Dāvidsons and Savins v. Latvia , nos . 17574/07 and 25235/07 , § 36 , 7 January 2016 , the Court already assessed an identical argument raised by the Government and found that the review procedure enshrined in chapter 63 of the Criminal Procedure Law constituted an extraordinary remedy .
23 . The aforementioned suffices to conclude that this procedure can not be taken into account for the purposes of Article 35 § 1 of the Convention . The Government's objection must therefore be rejected .
24 . The Court notes that this complaint is not manifestly ill-founded within the meaning of Article 35 § 3 (a) of the Convention . It further notes that it is not inadmissible on any other grounds , It must therefore be declared admissible .

Figure 4: Attention scores assigned to a fragment of the essays corpus for the component detection (claim-premise) task. The model uses context attention with linear activation.

However , I do think that people are likely to spend less time in cooking food .
There are various reasons why I maintain this viewpoint , and among those reasons are two important ones .
The first and foremost reason is that with the development of science and technology , more advanced kitchen facilities , such as the modern microwaves or the advanced pressure cookers , have been created and invented to help people prepare and cook food in a very short time .

useful to include in annotation guidelines to enhance inter-annotator agreement, and also to speed up annotation, by focusing the attention of annotators in more important information.

Context attention provides an entirely different kind of information, as can be seen in Figure 4, showing the result of a classifier with context attention in the essays corpus, distinguishing claims and premises.

Context attention weights differently the word on each occurrence. It systematically gives more importance to words at the beginning of the sentence and to the two words at the end of the sentence, because they are strong signals of the beginning of an argument component, whereas the rest of words in the sentence are not. This behaviour has its roots on the argument structure, as it is not likely that a new component will start at the end of the sentence.

Punctuation acting as sentence separators (.) receive a high value of attention, as they denote the start of a new component. In the case of commas (,) we have seen mixed results, as they can be part of a component as well. It is interesting to note that this behaviour is captured by both types of attention.

Regarding activation functions, the linear activation previously described favours more evenly distributed, non-extreme attention scores. As a result, it generates a less saturated visualization than the sigmoid activation. We consider this type of activation better suitable for explaining the results, as we can identify words in more distributed attention values.

With the word attention mechanism, we can extract the words with more attention. We show them in a wordcloud on Figure 5. This visualization highlights the words that are

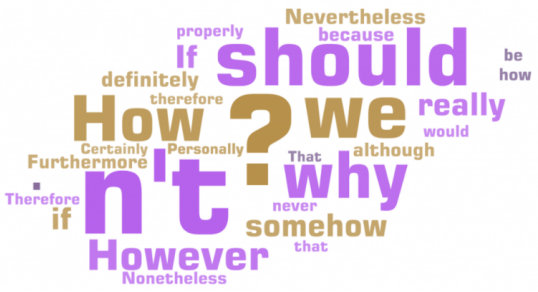


Figure 5: Words with most attention in the essays corpus, with word attention. Size is proportional to attention score.

more important to decide any classification label. In the case of the essays, where the language and the argumentation are more informal, we see mainly modal verbs, connectors and adverbs. The implication is that the model is not defining the argumentation by the semantics of its words, instead using the intention of the speaker represented at the start of the phrase. Adverbs like *really* and *somehow* exemplify the more sentimental nature of the argumentations in student essays. This information can be readily included in guidelines or to select new instances for annotation. For example, we can select instances without words for which we already have obtained high attention values in already annotated instances, thus gaining insight for new words.

Conclusions

In this paper, we present an analysis of the impact of neural attention mechanisms in the task of training Argument Mining systems for argument component detection. Argument Mining is a particularly challenging task for machine learning approaches because training corpora are typically small, and the phenomena to represent are very complex, with impact in various linguistic levels, thus it suffers from acute data sparseness. Our assumption was that attention mechanisms could help concentrate this sparse information. We have found that some attention mechanisms do improve the performance of Argument Mining systems, if only slightly. We have found that for the task of claim detection, which is more well-defined, the performance is higher, but the improvement provided by attention mechanisms is approximately the same in both tasks. We have also found that word-level attention produces consistently better results than context-level attention. However, an important parameter of context-level attention, the length of the sequence, remains to be explored. The rest of the hyperparameters do not seem to provide consistently better results, although they produce important variations in performance.

Even if the improvement in performance is small, we have gained on explanainability of the results: we can now analyze how classifiers make errors with more insight, as we have stronger indicators of the information wherefrom they make their conclusions. This helps to improve the training corpora and the classifiers that exploit them.

These results have been found to hold on two very different corpora: a reference corpus of student essays and our primary focus of research, a small corpus of judgments of the European Court of Human Rights. We expect that the interpretability provided by attention mechanisms, focused on words, will help in enhancing this corpus, by increasing the consistency of annotations, improving inter-annotator agreement. It will also be used to select examples with words for which the model has no strong attention, thus helping explore regions of the space that are not well characterized yet.

Acknowledgements

The authors have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project MIREL: MIning and REasoning with Legal texts.

References

Cabrio, E., and Villata, S. 2018. Five years of argument mining: a data-driven analysis. In Lang, J., ed., *Proceedings of IJCAI 2018*, 5427–5433.

Eger, S.; Daxenberger, J.; and Gurevych, I. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th ACL*, 11–22.

Habernal, I., and Gurevych, I. 2017. Argumentation mining in user-generated web discourse. *Comput. Linguist.* 43(1):125–179.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR* abs/1508.01991.

Lippi, M., and Torroni, P. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.* 16(2):10:1–10:25.

Peldszus, A., and Stede, M. 2013. From argument diagrams to argumentation mining in texts: A survey. *IJCINI* 7(1):1–31.

Reimers, N., and Gurevych, I. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 338–348.

Stab, C., and Gurevych, I. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguist.* 43(3):619–659.

Stab, C.; Miller, T.; and Gurevych, I. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *CoRR* abs/1802.05758.

Teruel, M.; Cardellino, C.; Cardellino, F.; Alemany, L. A.; and Villata, S. 2018. Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines. In *Proceedings of the LREC 2018*.

Wang, B.; Liu, K.; and Zhao, J. 2016. Inner attention based recurrent neural networks for answer selection. In *Proceedings of the 54th ACL*, 1288–1297.