Segmentation of Argumentative Texts with Contextualised Word Representations

Georgios Petasis
Software and Knowledge Engineering Laboratory
Institute of Informatics & Telecommunications
National Center for Scientific Research (N.C.S.R.) "Demokritos"
Athens, Greece.
petasis@iit.demokritos.gr

Abstract

The segmentation of argumentative units is an important subtask of argument mining, which is frequently addressed at a coarse granularity, usually assuming argumentative units to be no smaller than sentences. Approaches focusing at the clause-level granularity, typically address the task as sequence labeling at the token level, aiming to classify whether a token begins, is inside, or is outside of an argumentative unit. Most approaches exploit highly engineered, manually constructed features, and algorithms typically used in sequential tagging – such as Conditional Random Fields, while more recent approaches try to exploit manually constructed features in the context of deep neural networks. In this context, we examined to what extend recent advances in sequential labelling allow to reduce the need for highly sophisticated, manually constructed features, and whether limiting features to embeddings, pre-trained on large corpora is a promising approach. Evaluation results suggest the examined models and approaches can exhibit comparable performance, minimising the need for feature engineering.

1 Introduction

Argument mining involves the automatic discovery of *argument components* (such as claims, premises, etc.) and the *argumentative relations* (i.e. support, attack, etc.) among these components in texts. Primarily aiming to extract arguments from texts in order to provide structured data for computational models of argument and reasoning engines (Lippi and Torroni, 2015a), argument mining has additionally the potential to support applications in various research fields, such as opinion mining (Goudas et al., 2015), stance detection (Hasan and Ng, 2014), policy modelling (Florou et al., 2013; Goudas et al., 2014), legal information systems (Palau and Moens, 2009), fact checking (Naderi and Hirst, 2018), etc.

The identification of argumentative discourse structures typically consists of two main tasks: 1) the identification of the locations in text and the type of the argument components, and 2) the identification of how these argument components related to each other (Persing and Ng, 2016). As a result, argument mining is usually addressed as a pipeline of several sub-tasks. Typically the first sub-task is the separation between argumentative and non-argumentative text units, which can be performed at various granularity levels, from clauses to several sentences, usually depending on corpora characteristics. Detection of argumentative units (AU)[1], as discussed in Section 2, is typically modeled as a fully-supervised classification task, either a binary one, where units are separated in argumentative and non-argumentative ones with argumentative ones to be subsequently classified in major claims, claims, premises, etc. as a second step, or as a multi-class one, where identification of argumentative units and their classification into claims and premises are performed as a single step. Typically the granularity of this task is coarse, with most approaches considering sentences as the smallest argumentative unit (Florou et al., 2013; Moens et al., 2007; Song et al., 2014; Swanson et al., 2015), although some works focused on the most difficult task of detecting units at the *clause level* (Park and Cardie, 2014; Goudas et al., 2014, 2015; Sardianos et al., 2015; Stab, 2017; Ajjour et al., 2017; Eger et al., 2017). According to a recent survey (Lippi and Torroni, 2015a), the performance of proposed approaches depends on highly engineered and sophisticated, manually constructed, features.

Approaches focusing at the clause-level granu-

[1] Also known as "Argumentative Discourse Units – ADUs" (Peldszus and Stede, 2013).

Proceedings of the 6th Workshop on Argument Mining, pages 1–10
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

larity, typically address the task as sequence labeling at the token level, aiming to classify whether a token begins, is inside, or is outside of an argumentative unit through the IOB format (Ramshaw and Marcus, 1995). Most of the approaches employ Conditional Random Fields (CRFs) (Lafferty et al., 2001) with hand-crafted features (Goudas et al., 2014), as CRFs are the prominent and most reliable algorithm for many sequential labelling tasks (Zeng et al., 2017), and have been applied to a wide range of segmenting tasks, from named-entity recognition (McCallum and Li, 2003) and shallow parsing (Sha and Pereira, 2003), to aspect-based sentiment analysis (Patra et al., 2014). Sequence labeling algorithms take as input a set of features for each token in a sequence (such as a sentence) and learn to predict an optimal sequence of labels for all tokens in the input sequence, while performance depends on the provided (typically manually engineered features) and how well these features can help the model predicting the likelihood of every label in the sequence. However, as deep learning is slowly replacing CRFs for sequence labelling (i.e. (Ajjour et al., 2017)), it is interesting to examine whether these hand-crafted features are still important, or comparative levels of performance can be achieved without them.

In this paper we examine whether a "CRF-inspired" neural model without the hand-crafted features, can be applied to the task of argumentative unit segmentation at the clause level, and whether its performance is comparable to approaches exploiting such features. In addition, we study whether contextualised word representations can help in this task, and provide an alternative to hand-crafted features. These can be reflected in the following two questions:

1. Can approaches that do not use manually engineered features achieve performances comparable to approaches that exploit such features?

2. Can contextualised word representations (pre-trained in large corpora) replace manually engineered features in argument mining?

The motivation behind the work presented in this paper originates from the advances performed in the state of art of named-entity recognition by Bidirectional LSTM-CRF Models for Sequence Tagging (Huang et al., 2015; Ma and Hovy, 2016), a variation of Long Short-Term Memory (LSTM) based models with a decoding layer that considers relations between neighbouring labels and jointly decodes the optimal sequence of labels for a given input sequence (Ma and Hovy, 2016), using a Conditionally Random Field. Recognising a similar evolution pattern also in the area of argument mining segmentation – starting with CRF's and manually constructed features (Park and Cardie, 2014; Goudas et al., 2014, 2015; Stab, 2017), then employing word embeddings as features in CRFs (Sardianos et al., 2015) and subsequently applying bi-directional LSTMs (Ajjour et al., 2017) on manually engineered features – poses the question if a similar advancement can be achieved by introducing the currently missing pieces (LSTM-CRF models or contextualised word representations such as (Peters et al., 2018)), in an attempt to eliminate – or reduce the need for – manually engineered features.

In order to approach our research questions we have used the second version of the Argument Annotated Essay Corpus (Stab, 2017), a collection of 402 essays, which has been manually annotated with major claims (one per essay), claims and premises at the clause level. In addition, the corpus contains manual annotations of argumentative relations, where the claims and premises are linked, while claims are linked to the major claim either with a support or an attack relation. We have applied LSTM-CRF models (using the implementation reported in (Akbik et al., 2018)) employing various word embeddings (including contextualised word representations like "ELMo" (Peters et al., 2018), "Flair" (Akbik et al., 2018) and "BERT" (Devlin et al., 2018)). Evaluation results suggest that all studied approaches are comparable or slightly better to the current state of art.

2 Related work

Almost all argument mining frameworks proposed so far employ a pipeline of stages, each of which is addressing a sub-task of the argument mining problem (Lippi and Torroni, 2015a). The segmentation of text into argumentative units is typically the first sub-task encountered in such an argument mining pipeline, aiming to segment texts into argumentative and non-argumentative text units (i.e. segments that do contain or do not contain argument components, such as claims or premises). The granularity of argument components is text-dependant. For example, in Wikipedia articles studied in (Rinott et al., 2015), argument components spanned from less than a sentence to more than a paragraph, although 90% of the cases was up to 3 sentences, with 95% of components being comprised of whole sentences.

Several approaches address the identification of argumentative units at the sentence level, a subtask known as "argumentative sentence detection", which typically models the task as a binary classification problem. Employing machine learning and a set of features representing sentences, the goal is to discard sentences that are not part (or do not contain a component) of an argument. As reported also by Lippi and Torroni (2015a), the vast majority of existing approaches employ "classic, off-the-self" classifiers, while most of the effort is devoted to highly engineered features. A plethora of learning algorithms have been applied on the task, including Naive Bayes (Moens et al., 2007; Park and Cardie, 2014), Support Vector Machines (SVM) (Mochales and Moens, 2011; Rooney et al., 2012; Park and Cardie, 2014; Stab and Gurevych, 2014; Lippi and Torroni, 2015b), Maximum Entropy (Mochales and Moens, 2011), Logistic Regression (Goudas et al., 2014, 2015; Levy et al., 2014), Decision Trees and Random Forests (Goudas et al., 2014, 2015; Stab and Gurevych, 2014). There is also a limited number of approaches addressing the task in a semi-supervised or unsupervised manner, such as (Ferrara et al., 2017).

The identification of argumentative units at the clause level has been less studied than its more coarse counterpart. (Park and Cardie, 2014) has exploited n-grams and a large number of additional, manually crafted, binary (denoting the presence of features) and numeric (containing counts) features in a supervised manner with Support Vector Machine as classifier, achieving a macro-averaged $F_1 = 68.99\%$ on a corpus manually annotated by the authors. In (Goudas et al., 2014, 2015) the authors have examined segmentation both at sentence and clause level, for the Greek language, using a corpus manually annotated by the authors. They have exploited both features from previous approaches and features proposed by the authors, achieving $F_1 = 42.37\%$, as measured by "conlleval.pl" (taking into account correct sequences and not only labels at the token level). The same Greek corpus has been used in (Sardianos et al., 2015), where word2vec embeddings (Mikolov et al., 2013) have been used as features in a supervised setting using CRFs, combined with part-of-speech tags and a small lexicon with cue phrases, to report a small increase in performance ($F_1 = 32.12\%$) over the baseline ($F_1 = 27.04\%$).

CRFs have been also used in (Stab, 2017), along with an extensive set of highly engineered features, including structural, syntactic, lexico-syntactic and probabilistic features. The approach has been evaluated on the second version of the Argument Annotated Essay Corpus (the same corpus has been used for evaluation in this work), created by the authors, achieving macro-averaged $F_1 = 86.70\%$. Similar features (with the addition of pragmatic features) have been exploited in (Ajjour et al., 2017) using a bidirectional LSTM model as classifier in a supervised setting, achieving macro-averaged $F_1 = 88.54\%$ on the second version of the Argument Annotated Essay Corpus, with lower scores on two other corpora. In interesting aspect of this work is the out-of-domain evaluation, performing evaluations on different corpora from the ones used for training. Deep neural networks have been also employed by (Eger et al., 2017), using bidirectional LSTM-CRF models in a supervised setting, as an end-to-end system. Framing argument mining as a sequence tagging at the token level, they learn simultaneously four different sets of labels, encoding both segmentation of argumentative units, their types and their relations. The approach has been evaluated on the second version of the Argument Annotated Essay Corpus (the same corpus has been used for evaluation in this work) achieving $F_1 = 69.49\%$.

In (Persing and Ng, 2016) the authors propose a rule-based approach, with manually constructed rules applied on top of syntactic trees, achieving a performance of 92.1% on the first version of the Argument Annotated Essay Corpus (Stab and Gurevych, 2014). In (Lawrence et al., 2014) the authors propose a two-stage approach: During the first stage text is segmented into propositions using two Naive Bayes classifiers (Nir Friedman and Goldszmidt, 1997) with simple features (words, lengths and a sliding window of three tokens) in a supervised setting. Then, as a second step, propositions are scored based on their similarities to document topic retrieved through Latent Dirichlet Allocation (LDA) and their distances, to decide whether they constitute an argumentative unit or not.

3 Data

For our experiments, we have used the second version of the Argument Annotated Essay Corpus (Stab, 2017; Eger et al., 2017; Stab and Gurevych, 2017), which contains 402 student essays written in response to controversial topics. The corpus has been manually annotated with major claims (one per essay), claims and premises at the clause level. In addition, the corpus contains manual annotations of argumentative relations, where the claims and premises are linked, while claims are linked to the major claim either with a support or an attack relation. Essays are on average 370 tokens long, while most of the tokens ($\sim 70\%$) are part of an argumentative unit. The corpus is split into train and test sets at the essay level, provided by the authors. We have converted the corpus into the CoNLL token-based sequence tagging format (using the tools provided by the "BRAT" annotation toolkit) and we extracted a small development set ($< 10\%$) from the training set randomly, with the help of "scikit-learn" toolkit.

4 Models

Following the typical setting in argumentative unit segmentation at the clause level, we are going to also frame the task as a sequence labelling classification problem. In sequential labelling the label of an instance does not depend only on the instance itself, but also depends on the instances previously seen. A natural choice for sequence labelling are recurrent neural networks (RNNs), which consider "hidden" states computed from previous points in time (instances already classified) during classification. For our experiments we have chosen LSTMs (Hochreiter and Schmidhuber, 1997), a type of RNNs able to learn long-term dependences, as their structure allows them to control how much information is shared across points in time.

However, a single LSTM is able to have access to a single context (typically to the left context of a token) when assigning a label. Bidirectional LSTMs employ two separate LSTM layers, looking at the input from opposite directions, while their output is concatenated into a single vector. Finally, in order to reflect all CRF capabilities, and especially its ability to assign labels taking into account contextual dependencies from all tags in a sequence, a CRF network can be combined with an LSTM or a bidirectional LSTM to form an LSTM-CRF (or bi-LSTM-CRF model) (Huang et al., 2015), which can use features from all instances in a sequence (past and future) for assigning a label to an instance (Fig 1).

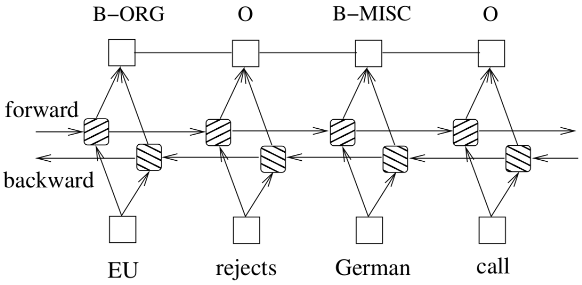

Figure 1: A BI-LSTM-CRF model. (Huang et al., 2015)

4.1 Argument Mining as Sentence Labelling

In a simple scenario, argumentative unit identification can be performed at the sentence level, where labeling consists in distinguishing between sentences that are argumentative units ($y = au$) and sentences that are not argumentative units ($y = \overline{au}$).

4.2 Argument Mining as Sequence Labelling

In a more articulated scenario, argumentative unit identification must decide not only whether a sentence contains an argumentative unit, but in addition to identify the exact words that represent each argumentative unit within each sentence. Framing this task as a sequence labelling task, each token is assigned a label from y, where $y = \{(b, t) \mid b \in \{B, I, O\}, t \in \{au\}\}$.

4.3 Embeddings

As input to the aforementioned model, we are going to use dense representations, and more specifically pre-trained word embeddings, such as GloVe (Pennington et al., 2014). Depending on the way word embeddings were generated and the information they represent, word embeddings can be seen as a form of transfer learning, providing a model additional information, typically acquired from a larger corpus than a training dataset for a task. In addition to these embeddings, we are going to examine more recent *deep contextualised* word representations, such as "ELMo" (Peters et al., 2018), "Flair" (Akbik et al., 2018) and

		Number of tokens				
Part	**# Documents**	B-Arg	I-Arg	O-Arg	Total	Average
Train + Development	322	4,823	75,657	38,195	118,675	368.56
Test	80	1,266	18,837	9,442	29,545	369.31

Table 1: Number of documents, tokens per class, and average number of tokens per document.

"BERT" (Devlin et al., 2018). These representations are able to model "both characteristics of word usage (e.g. syntax and semantics) and how these uses vary across linguistic contexts (i.e. to model polysemy)" (Peters et al., 2018). These representations assign a different vector to each word based on its context, in contrast to embeddings like GloVe that assign the same vector to a word, irrespectively of context.

5 Experiments

5.1 Argument Mining as Sentence Labelling

Using the corpus described in Section 3, we have applied four classifiers to the task of classifying a sentence as argumentative or not. Using as only features the GloVe[2] vectors for each token in a sentence, we have applied Convolutional Neural Networks (CNNs) the following implementation, BI-LSTM-CRF, and bidirectional Sentence-State LSTMs (S-LSTMs) (Zhang et al., 2018)[3]. All approaches involve the usage of non-contextualised embeddings (GloVe), keeping the most frequent 15,000 words in the corpus, following the training details as described in (Zhang et al., 2018). All models are trained using SGD with no momentum (with a mini-batch size of 32), clipping gradients at 5, for a maximum 40 epochs. A simple learning rate annealing method is employed in which we halve the learning rate if training loss does not fall for 5 consecutive epochs, initialising learning rate to 10^{-3}. The hidden states per-layer was set to 300, and variational dropout was used. The number of hidden layers was fine-tuned in the range $1-8$, and model selection was performed by choosing the model with the best accuracy on the development set. The split provided by the authors of the corpus regarding the training and test sets was used, while a small development set was extracted from the training set, containing 21 essays[4]. Regarding stability and reproducibility of results, we have used 2019[5] as the seed value. The aforementioned approaches were compared to the "BERT" (Devlin et al., 2018) contextual embeddings[6], using a single feed-forward layer on top of the embeddings, with a hidden layer equal to the size of the embeddings (768)[7]. Minimal fine-tuning has been performed, allowing only a single epoch with mini-batch size of 32 and a learning rate equal to $2e^{-5}$.

Embedding	**Architecture**	**Accuracy**
GloVe	CNN	0.8391
GloVe	LSTM	0.8488
GloVe	S-LSTM	0.8619
BERT	Feed Forward	**0.8874**

Table 2: Argument Mining as Sentence Labelling: Evaluation Results.

Our experiment results are summarised in Table 2. While BERT embeddings (even with minimal fine-tuning of a single hidden layer) have outperformed all other approaches, traditional word embeddings ("GloVe" + S-LSTM) may still be useful as their performance is still very close to BERT, while employing 6 Bi-S-LSTM-CRF layers, with a window of 5 tokens, and after 15 epochs of fine-tuning to the task.

5.2 Argument Mining as Sequence Labelling

For our second experiment, which combines the identification of argumentative units with their localisation as textual segments, we have employed

[2] Wikipedia 2014 + Gigaword 5, 6B tokens, 400K vocabulary, uncased, 300 dimensions.

[3] We have used the following implementation for CNNs, LSTMs and S-SLTMs: https://github.com/leuchine/S-LSTM

[4] The essays randomly selected for the development set are: 13, 38, 41, 115, 140, 152, 156, 159, 162, 164, 201, 257, 291, 324, 343, 361, 369, 371, 387, 389, 400.

[5] The same seed value, 2019, has been used for all experiments performed in this paper.

[6] We have adapted the implementation that can be found here: https://colab.research.google.com/github/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb

[7] The used embeddings can be found at: https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

an end-to-end system that utilises a BI-LSTM-CRF architecture with 2 layers, with each layer employing 256 hidden nodes. This model has been trained and evaluated with a series of traditional "(GloVe", character embeddings) and contextual embeddings ("ELMo", "Flair", and "BERT"). All experiments we have used the "Flair"[8] framework. Fine-tunning was performed for a maximum of 150 epochs, using SGD with a mini-batch size of 32,and simulated annealing, with a starting learning rate of 0.1. The same random seed (2019) was used for all experiments.

We report the macro F-score as an evaluation measure, since this allows for a comparison to related work. The macro F_1-score considers all the classes to be equally important, without taking into consideration the number of instances each class has. (The distribution of classes in the corpus is shown in Table 1.)

5.2.1 Comparison with previous work

Features	Model	Macro F_1
All (Semantic+Syntactic	SVM	61.40
+Structural+Pragmatic)	CRF	79.16
(Ajjour et al., 2017)	BI-LSTM	**88.54**
All (Stab, 2017)	CRF	**86.70**
GloVe + Character	BI-LSTM-CRF	85.92
GloVe + Character + Flair	BI-LSTM-CRF	88.17
ELMo	BI-LSTM-CRF	88.62
BERT	BI-LSTM-CRF	89.31
GloVe + Flair + BERT	BI-LSTM-CRF	**90.13**
GloVe + Flair + ELMo + BERT	BI-LSTM-CRF	87.42

Table 3: Argument Mining as Sequence Labelling: Evaluation Results.

In order to enable comparison with existing approaches, we have tried to imitate the experimental settings found in (Stab, 2017) and (Ajjour et al., 2017). Table 3 shows the results of the approaches presented in (Ajjour et al., 2017) in the upper part of the table, followed by the best overall result presented in (Stab, 2017), including all features (semantic, syntactic and structural) and the CRF classifier. Both approaches employ a large number of highly engineered and sophisticated, manually constructed, features. Finally, in the lower part of the table, we report our results of the BI-LSTM-CRF model with the various tested embeddings.

From Table 3 it can be seen that almost all embeddings (especially the contextual ones) outperform the approaches with manually engineered features (Ajjour et al., 2017; Stab, 2017), with the combination of contextual embeddings achieving new state-of-art ($MacroF_1 = 90.13$) on the Essays v2.0 corpus, especially when considering the absence of manually constructed features.

5.3 Error Analysis

We analysed the results obtained with the GloVe+Flair+BERT experiment. The test dataset contains 1448 annotated sentences, where 1,178 sentences were correctly annotated, while 270 sentences were erroneously annotated by our model. According to the confusion matrix, the two major sources of errors are 1767 "O" tokens erroneously classified as "I-Arg", and 829 "I-Arg" erroneously classified as "O". The majority of the errors (104 sentences) were sentences that the model erroneously annotated as containing argumentative components, while these sentences did not contain any argumentative component according to the gold annotation. Some examples of such sentences are displayed in the following list (annotated segments by the model are highlighted):

1. In spite of this, the disadvantages of the promotion of a universal language cannot be denied.
2. It is obvious that the benefits of the Internet undoubtedly outweigh its disadvantages.
3. It would be highly unpractical to ask people to adopt a simpler way of life.
4. Some people claim that without this punishment our lives would be less secure and crimes of violence would increase.
5. It is evident that technology promotes economy.

The second most important source of errors, are sentences containing argumentative units that were not annotated as such by our model. 43 sentences belong in this category, while some examples are shown as follows:

1. However, it is not sufficient in itself.

[8] https://github.com/zalandoresearch/flair

2. Some people claim that the prevalent of English brings a great number of benefits for people.

3. In the modern world, computers are used everywhere.

4. There is no end to the evolution of computers.

5. Many people hold the opinion that past behaviour determines the future actions, which could be the main reason to support the idea of revealing the record to the jury.

The rest of the errors (123 sentences in total) are various errors, like two argumentative units merged in one (errors by our model in red):

1. For instance , some Asians are seeking individualism, previously denied by many Asian countries, due to the fact that they have gradually identified with such values expressed in American movies, which are imported by the governments as a result of the proliferation of English.

2. First and foremost, sports events are good chances for excellent athletes to meet and learn valuable experiences from one another. so that they can improve their results, break records and bring victories to their own countries.

Finally, in some cases, our model missed the beginning of an argumentative unit (in red the part not annotated by our model):

1. From personal level, it fosters a sense of unfairness between the older and younger generations.

2. From social perspective, massively forcing the early retirement would be one of financial burden to the local government.

5.4 Discussion

Evaluation results suggest that omitting highly engineered, manually crafted features, and replacing them with embeddings (pre-trained on large corpora and possibly exploiting multiple sources of information), is a promising approach and a viable alternative.

Research Question 1: Can approaches that do not use manually engineered features achieve performances comparable to approaches that exploit such features?

Evaluation results suggest that a large part of the information provided by the plethora of manually constructed features can be substituted with a fairly standard architecture and word embeddings, especially contextualised embeddings that can be tuned to the task at hand, like the contextualised word representations ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). Further optimisation is of course possible (especially with respect to the architectures on top of embeddings, the number of layers, and the fine-tune of the many hyper-parameters associated with the employed neural models). However, there are some limiting factors, mainly the absence of a development set in the corpus used for evaluation, and the computational requirements of the models, especially in the case of contextualised word embeddings.

Research Question 2: Can contextualised word representations replace manually engineered features?

Evaluation results are promising, especially since the examined approaches have achieved a small increase over the current state-of-art. However, the examined approaches have not exceeded significantly the current state-of-art, suggesting that manually engineered features are still relevant and significant at least for this task, the segmentation of argumentative units at the clause level. One of the findings in (Ajjour et al., 2017) is that the semantic features appear to be the most significant features, achieving the highest F-scores, an observation that seems to hold also in our experiments, as reverting to embeddings that enhance semantic modelling (through implicit word sense disambiguation performed based on contextual information) seems to provide a significant increase in performance. At the same time, the performance difference with the CRF exploiting the manually constructed features (Stab, 2017) is small, suggesting that removing the highly engineered features may have a small penalty in performance, at least for the approach of (Stab, 2017).

6 Conclusion

The segmentation of argumentative units is an important subtask of argument mining, which is frequently addressed at a coarse granularity, usually assuming argumentative units to be no smaller than sentences. Approaches focusing at the clause-level granularity, typically address the task as sequence labeling at the token level, aiming to classify whether a token begins, is inside, or is outside of an argumentative unit through the IOB format (Ramshaw and Marcus, 1995). Most approaches exploit highly engineered, manually constructed features, and algorithms typically used in sequential tagging – such as CRFs (Park and Cardie, 2014; Goudas et al., 2014, 2015; Stab, 2017), while more recent approaches try to exploit manually constructed features in the context of deep neural networks (Ajjour et al., 2017; Eger et al., 2017). In this context, we examined to what extend recent advances in sequential labelling and contextualised word embeddings allow to reduce the need for manually constructed features, and whether limiting features to embeddings, pre-trained on large corpora is a promising approach. Evaluation results suggest the examined models and approaches can exhibit comparable performance, minimising the need for feature engineering.

Regarding directions for further research, there are several axes that can be explored. Evaluation on more corpora will provide enhanced insights about the performance of the examined approaches on different document types. At the same time, there is a significant optimisation potential, especially in hyper-parameter tuning of the employed algorithms, provided that a suitable development set is available, and the computational requirements of some models (especially the ones employing contextualised word representation) are significantly reduced in order to constitute experimentation more tractable and practical.

Acknowledgments

We acknowledge support of this work by the project “APOLLONIS: Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation” (MIS 5002738) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme ”Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

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

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Alfio Ferrara, Stefano Montanelli, and Georgios Petasis. 2017. Unsupervised detection of argumentative units though topic modeling techniques. In *Proceedings of the 4th Workshop on Argument Mining*, pages 97–107, Copenhagen, Denmark. Association for Computational Linguistics.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013, August 8, 2013, Sofia, Bulgaria*, pages 49–54. The Association for Computer Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings*, pages 287–299. Springer International Publishing, Cham.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and the social web. *International Journal on Artificial Intelligence Tools*, 24(05):1540024.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1489–1500. ACL.

Marco Lippi and Paolo Torroni. 2015a. Argument mining: A machine learning perspective. In *Theory and Applications of Formal Argumentation: Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers*, pages 163–176, Cham. Springer International Publishing.

Marco Lippi and Paolo Torroni. 2015b. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 185–191. AAAI Press.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

Nona Naderi and Graeme Hirst. 2018. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Dan Geiger Nir Friedman and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29:131–163.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

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. Association for Computational Linguistics.

Braja Gopal Patra, Soumik Mandal, Dipankar Das, and Sivaji Bandyopadhyay. 2014. Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 370–374, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

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. Association for Computational Linguistics.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23-25, 2012*. AAAI Press.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO. Association for Computational Linguistics.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. 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. Association for Computational Linguistics.

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

Christian Matthias Edwin Stab. 2017. *Argumentative Writing Support by means of Natural Language Processing*. Ph.D. thesis, Technische Universität Darmstadt, Darmstadt.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6).

Yue Zhang, Qi Liu, and Linfeng Song. 2018. Sentence-state LSTM for text representation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 317–327, Melbourne, Australia. Association for Computational Linguistics.