Reevaluating Argument Component Extraction in Low Resource Settings

Anirudh Joshi[1,2] **Timothy Baldwin**[1] **Richard O. Sinnott**[1]
Cecile Paris[2]

[1] The University of Melbourne [2] CSIRO Data61

anirudhj@student.unimelb.edu.au, tb@ldwin.net
rsinnott@unimelb.edu.au, cecile.paris@data61.csiro.au

Abstract

Argument component extraction is a challenging and complex high-level semantic extraction task. As such, it is both expensive to annotate (meaning training data is limited and low-resource by nature), and hard for current-generation deep learning methods to model. In this paper, we reevaluate the performance of state-of-the-art approaches in both single- and multi-task learning settings using combinations of character-level, GloVe, ELMo, and BERT encodings using standard BiLSTM-CRF encoders. We use evaluation metrics that are more consistent with evaluation practice in named entity recognition to understand how well current baselines address this challenge and compare their performance to lower-level semantic tasks such as CoNLL named entity recognition. We find that performance utilizing various pre-trained representations and training methodologies often leaves a lot to be desired as it currently stands, and suggest future pathways for improvement.

1 Introduction

Argument component (AC) extraction typically involves addressing extremely complex high-level concepts, demanding significant amounts of world knowledge, natural language understanding, and reasoning to address (Moens, 2018). These argument components may come from different datasets, different domains, and have varying tagsets (IOB — inside, outside or at the beginning of an entity), depending on the component and annotation criteria used (Schulz et al., 2018). Originally the field expanded its tagsets across tasks over time; however, due to the inherent difficulty, the field has contracted back to tackling much simpler tasks (Moens, 2018). This difficulty is because performance across domains and tasks with limited resources makes training models extraordinarily difficult.

Recent work such as Schulz et al. (2018) uses single-task learning ("STL") and multi-task learning ("MTL") with character-level encodings and pre-trained GloVe word embeddings as inputs to a BiLSTM-CRF encoder to analyze this issue from a low resource standpoint, while other work approaches the task through the use of graph convolution networks (GCNs) with syntactic dependencies (Morio and Fujita, 2019). However, both evaluate in terms of tag-level F1, including non-target O tags, rather than the more stringent span-based metric conventionally used to evaluate named entity recognition ("NER": Tjong Kim Sang and De Meulder (2003)). In this paper, we compare contemporary embedding approaches in STL and MTL contexts against Schulz et al. (2018) and achieve state-of-the-art results for the dataset, but more importantly, we demonstrate that under span-based evaluation, the current state-of-the-art is woefully low, calling into question whether argument component extraction as currently construed is feasible for current NLP methods.

2 Findings

The focus of the paper is on a rigorous reevaluation of actual low-resource argument component (AC) extraction within argumentation mining (AM); in contrast to previous publications, we find that:

- Tag-based evaluation is inappropriate for evaluating span extraction performance.
- STL improves with embeddings and is better than MTL, in contrast to previously reported results.
- Current state-of-the-art (SOTA) approaches to low-resource AM, when evaluated strictly, do not result in usable systems, with <0.4 F1 in general.

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 219–224
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

As such, AC extraction in low-resource settings is an unsolved task and will require order of magnitude improvements in pre-training and inclusion of external knowledge to become serviceable.

3 Setup

3.1 Tasks and Evaluation

3.1.1 Task Description

Argument component (AC) extraction is the extraction of ACs such as factual premises and opinion-based claims from text, using a tag-based IOB system to extract the textual components as contiguous sequences of text as NER components (Schulz et al., 2018). The tasks are from a variety of disparate domains, with different IOB tagsets and associated distributions, some with simple claims or premises, others with more complex annotations (Schulz et al., 2018). The tasks are, as per previous work: *var, wiki, news, essays, web* and *hotel* (Schulz et al., 2018). These are NER tagged sentences that contain IOB tagged claims, premises, or more specific argument tags (with respect to the specific dataset annotation guidelines). They are sourced from various editorials/official documents/discussion boards, Wikipedia discussions, news comments, persuasive essays, web discourse, and hotel review domains respectively (Schulz et al., 2018). In each case, we train over training splits of 1k, 6k, 11k, and 21k tagged NER tokens, each of which is within a low-resource range. This NER extraction task is low-resource due to the fact that the number of example tokens is extremely limited, on the order of a few articles or hundreds of sentence examples at the low end, and just over a thousand at the high end (6k vs. 21k tokens). In contrast, other tasks often have examples in the thousands of sentences, and hundreds of thousands of tokens (Tjong Kim Sang and De Meulder, 2003). We also validate our implementation against CoNLL NER, to evaluate the competitiveness of our method over a simpler extraction task as an upper bound. We do this to contextualize how F1 span-based performance operates in low-resource AM vs. low-resource NER, to indicate how SOTA models perform with respect to the simpler NER extraction task.

3.1.2 Evaluation

We evaluate the results based on CoNLL span-based F1, ignoring non-relevant O extraction as it confounds analysis of true extraction performance of components of interest (named entities in the NER case and argumentation components for our task: Tjong Kim Sang and De Meulder (2003); Gardner et al. (2018); Peters et al. (2018)). This span based metric means we do not simply look at the precision and recall of tags in isolation. The span-based evaluation only concerns overlapping contiguous spans whereas tag-based F1 concerns discontinuous spans, meaning it is both looser and less aligned to the key task of contiguous span extraction. This stricter evaluation regime produces more realistic task results, as it is concerned with span extraction, not tag-based classification.

3.2 Framework

We utilize AllenNLP (Gardner et al., 2018) as our base framework, with standard STL training ablations (Peters et al., 2018), and adapt a multi-sampling training approach leveraging Hierarchical Multi-Task Learning (Sanh et al., 2019) for MTL training ablations. In the MTL case, for final test evaluation, we utilize the best epoch weights for each component task from the proportional sampler based on the validation data. We evaluate using the AllenNLP implementation of CoNLL span-based F1 measure, which focuses on the correctness of full-span extraction of components relevant to argumentation (and ignores O components), rather than the isolated tag-based F1 measures previously used.

3.3 Base BiLSTM-CRF Model, Training and Hyper-parameter Configuration

We utilize a variety of pre-trained models to generate word embeddings as input to a standard 2 layer BiLSTM-CRF, with a hidden layer size of 200 and dropout rate of 0.5. This base model is consistent with related task approaches, and SOTA methods (Peters et al., 2018; Schulz et al., 2018; Sanh et al., 2019). In general, previous work has used STL/MTL-trained BiLSTM-CRFs. In addition, as our focus is on the evaluation approach used in current SOTA papers, the point of the paper is not to evaluate every model combination, but simply to demonstrate the "true" performance of current SOTA methods under a rigorous evaluation regime. We improve on previous approaches within AC extraction by using more complex embeddings and cumulative embedding combinations. Specifically, we make use of character-level embeddings using a CNN as a randomly initialized baseline implementation, GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2018), in a monotonically increasing fashion through pre-trained ablations. We based our STL/MTL hyper-parameter configuration on Peters et al. (2018); Sanh et al. (2019), specifically the NER components, with monotonically increasing pre-trained embedding representations following Peters et al. (2018). No hyperparameter tuning is required, as these papers represent NER SOTA baselines in the STL/MTL NER extraction space, and we extend previous papers embedding approaches (Schulz et al., 2018) with more complex embeddings (BERT) and stricter evaluation criteria.

3.4 Monotonically Increasing Pre-trained Embeddings Ablations

We create monotonically increasing ablations of pre-trained embeddings, from least to most complex, as the basis of our SOTA BiLSTM-CRF span extraction model, to analyze their performance under strict evaluation criteria. We jointly train using progressive combinations of embeddings starting with the character-level CNN, and then monotonically adding GloVe, ELMo, and BERT embeddings. We use 16-dimensional character encodings with 128 filters and 3 n-gram filter sizes; pre-trained 50d GloVe vectors; pre-trained ELMo embeddings (with trainable scalar weights); and uncased base BERT (768d) drawing from a variety of previous works (Gardner et al., 2018; Peters et al., 2018; Schulz et al., 2018). For MTL, we utilize the Hierarchical Multi-Task Learning framework (Sanh et al., 2019), taking the best epoch weights from the multi-task sampler for each task based on the validation data. We base our models on the previous papers, to focus on evaluation, extend with BERT, and determine how well SOTA models can really perform on complex AC extraction tasks.

4 Experiments

4.1 Analysis

We find that in general, MTL often underperforms STL for individual tasks, which is in contrast to previous work (Schulz et al., 2018) (see Figure 1). We hypothesize that this is due to the disparate domains, annotations, IOB distributions, and label sets of the various tasks. Therefore even with the extra supervision signal, MTL tends not to aid in the training process, especially with well-initialized pre-trained embeddings. We hypothesize that focusing training on sampling the core task with the pre-trained embeddings (with suitable regularisation — see Section 3.3) will likely lead to better span extraction performance in low-resource, disparate domains (especially given the disparate label sets for the respective datasets), where the more robust and general performance of MTL is traded for higher performance in specific tasks.

We often find that in the STL/MTL cases there is a minimal improvement over the baseline CNN-based trained character embeddings and that the representational capacity of the pre-trained models is likely not sufficient to provide a significant improvement on these tasks. We find that in general F1 is substantially below much simpler tasks such as CoNLL NER, with the majority of our results well below an F1 of 0.5 (see Figure 2), whereas CoNLL models trained equivalently produce results well in excess of 0.9. In some cases such as the *essays* and *hotel* datasets, we see what we would expect with increasing pre-trained model complexity added to both STL and MTL tasks.

However *news*, *web* and *wiki* all seem to exhibit highly variant baseline performance regardless of training methodology or pre-trained initialisation. In these scenarios, the model is likely fitting annotation artifacts. We find that in general, both in the progress of training and evaluation, test and validation performance is both noisy and unstable. This variance is likely due to the difficult nature of the task, the sparsity of the data, and the disparity between the domains of pre-trained embeddings to the specific task at hand.

4.2 Embedding Ablations

We found that in general as we increase the complexity of pre-trained embeddings, from character-based learned CNN embeddings to pre-trained GloVe, ELMo, and BERT, we see improved performance (see Table 1). However, we still performed much lower when using more advanced pre-trained embeddings than previous systems using span metrics (Schulz et al., 2018) (see Table 1). This difference is due to the focus on tag-based accuracy metrics rather than span-based metrics, and also the disproportionate effect of the O tag. A comparable system to that of Schulz et al. (2018), the glove_stl baseline, performed much worse when using the span-based metric, where

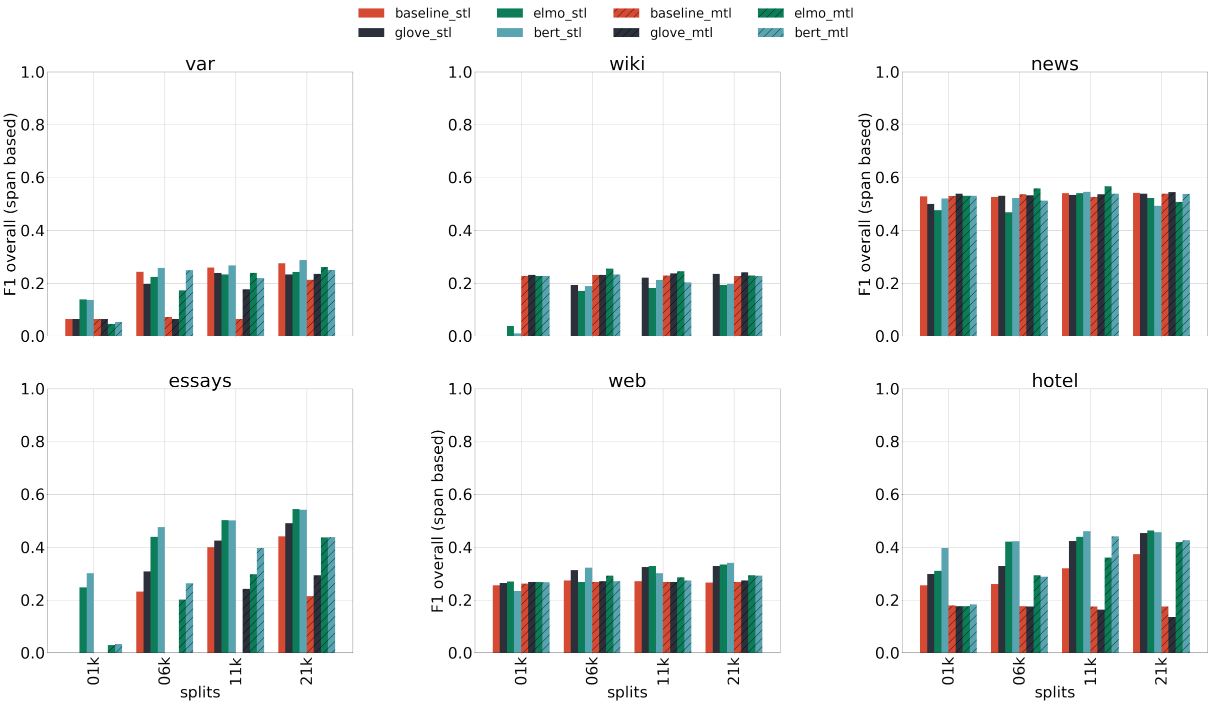

Figure 1: Performance (F1) across splits, tasks, and models. Full F1 range is used to demonstrate performance with full context of upper bound.

task	var_21k	wiki_21k	news_21k	essays_21k	web_21k	hotel_21k
max_stl	0.2863	0.2349	0.5420	0.5439	**0.3404**	0.4632
max_mtl	0.2606	0.2412	0.5445	0.4377	0.2934	0.4267
previous_tag_baseline	(0.3045)	(0.1834)	(0.3263)	(0.4838)	(0.1521)	(0.4569)
previous_tag_stl	(0.4334)	(0.2337)	(0.5649)	(0.6054)	(0.2343)	**(0.4791)**
previous_tag_mtl	**(0.4739)**	**(0.3250)**	**(0.5776)**	**(0.6055)**	(0.2327)	(0.4644)

Table 1: Our best STL/MTL on a more realistic span based evaluation indicates (top) a more realistic but lower performance vs. previous implementations using more simplistic tag based macro F1 evaluation (bottom in brackets).

we found in general that even with the addition of SOTA BERT embeddings, which have produced significant advances in other mid-level NLP tasks (Devlin et al., 2018), we were unable to produce results on par with tag-based evaluation. However span-based extraction provides a more realistic assessment of argument component extraction, with bert_stl generally providing the highest average score.

We also validated our results against the CoNLL NER dataset for all ablations and found performance to be on par with existing SOTA systems (Peters et al., 2018). Thus more pre-trained, more diverse, and more integrated representations do help improve the performance across these tasks on average, but the performance for argumentation component extraction leaves a lot to be desired under the span-based metric, suggesting that a usable extraction system is still well beyond the reach of current NLP models, based on the existing task formulation.

5 Future Work

It is of crucial importance to improve the representational complexity of pre-trained embeddings for high-level semantic tasks, especially in a low-resource regime. The inclusion of more linguistic and statistical inductive biases is necessary if progress is to be made on problems of extreme complexity, such as natural language argumentation component extraction. Some work has already begun with the introduction of syntactic fea-

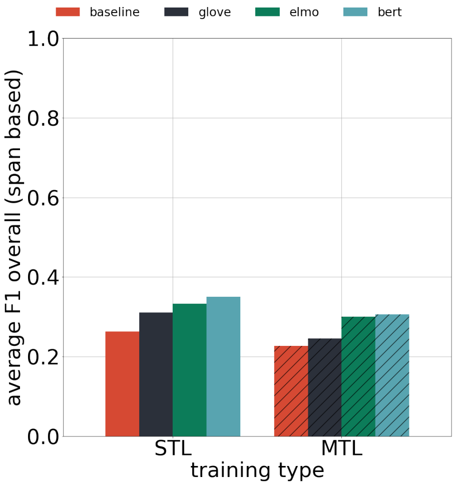

Figure 2: Comparing performance across models on average across all tasks contrasting training type methodologies. Full F1 range is used to demonstrate performance with full context of upper bound.

tures within GCNs for this task, but more integration of inductive biases will be necessary if progress is to be made, both in task performance and representational capability (Morio and Fujita, 2019). Other possible improvements include the use of external knowledge, such as external knowledge graphs and sentence based dependencies.

We find that in general, STL or MTL training over pre-trained embeddings are unlikely to be of significant benefit given the enormous amount of information required for complex semantic extraction tasks. A corollary to this is that it is also likely not sufficient, given the minor improvement of BERT over other pre-trained representations, to solely rely on statistical sequence prediction. To close the gap with human performance a step-order improvement in pre-training for end tasks is required.

6 Conclusion

In this paper, we have reevaluated argumentation component extraction based on STL and MTL approaches across a range of contemporary pre-trained embedding representation models, within a low resource task setting. We found that in general, according to a span-based evaluation metric such as that used for CoNLL NER, the results for the task drop appreciably from published results based on more naive evaluations. We found that MTL across varying domains did not significantly aid the task across domains, and that pre-trained word representations are not substantially better than a character-based word embedding baseline.

The results on average showed that as the pre-trained representations grow in complexity, on average, there was a robust increase in performance, and this was robust in both STL/MTL scenarios. Hence we believe that significant improvements in representational complexity of pre-trained embeddings for low resource tasks are necessary, above and beyond pure statistical inductive biases, if tasks such as argumentation component extraction are to achieve the same level of success as lower-level tasks such as NER.

7 Acknowledgements

This research is based upon work supported in part by CSIRO Data61 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research projects Activity (IARPA), under Contract [2017-16122000002]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, USA.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of the Workshop for Natural Language Processing Open Source Software (NLP-OSS)*, Melbourne, Australia.

Marie-Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *AAC*, 9(1):1–14.

Gaku Morio and Katsuhide Fujita. 2019. Syntactic graph convolution in multi-task learning for identifying and classifying the argument component. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 271–278, Newport Beach, USA.

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

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, New Orleans, USA.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956, Honolulu, USA.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multitask learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, USA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada.