Detecting Entities in the Astrophysics Literature: A Comparison of Word-based and Span-based Entity Recognition Methods

Xiang Dai and Sarvnaz Karimi

CSIRO Data61

www.data61.csiro.au

Named Entity Recognition (NER) refers to the task of identifying mentions of different types of entities in free-text.

The DEAL shared task focuses on identifying Named Entities (Figure 1) in a dataset composed by full-text fragments and acknowledgments from the **astrophysics literature**.

We conduct an empirical comparison between word-based tagging and span-based classification methods. Results show that **span-based classifier** outperforms previously widely used word-based tagger. Our best-performing submission ranked 2nd on validation set and 3rd on test set.

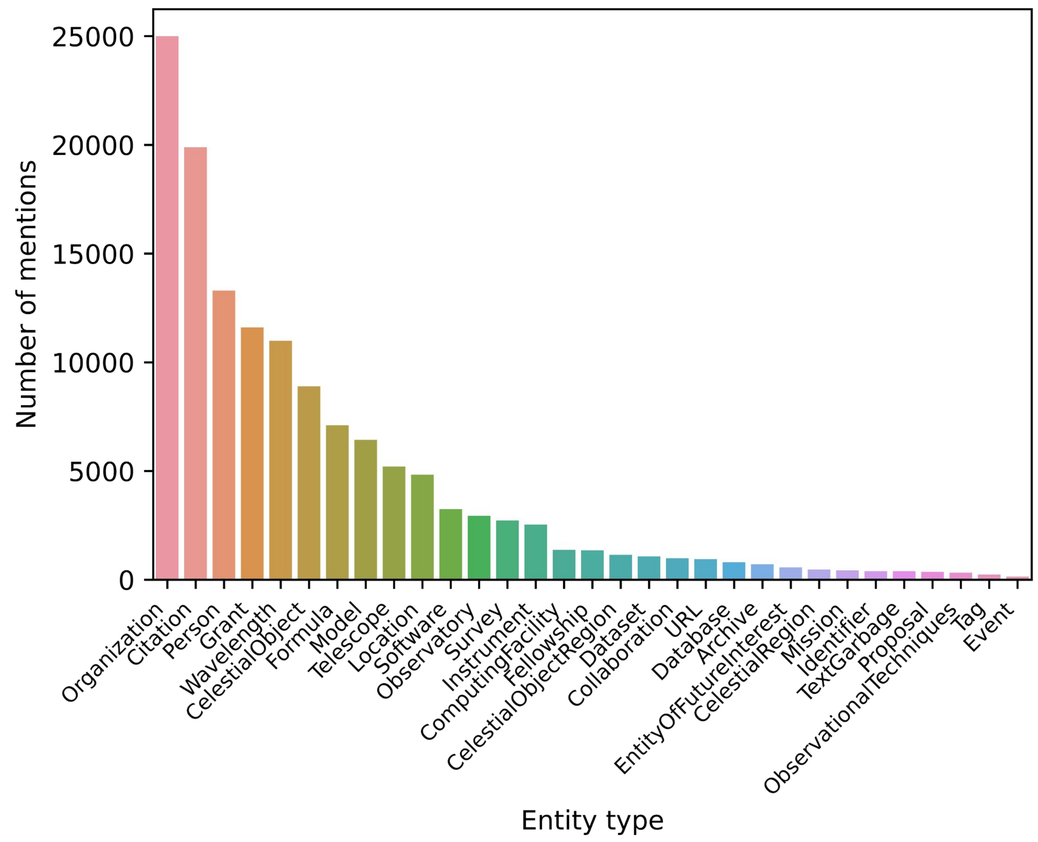


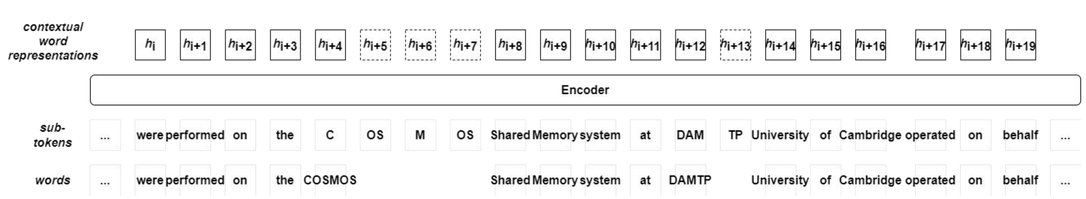
Figure 1: 31 Entity types in the DEAL dataset.

Our system

Text encoder

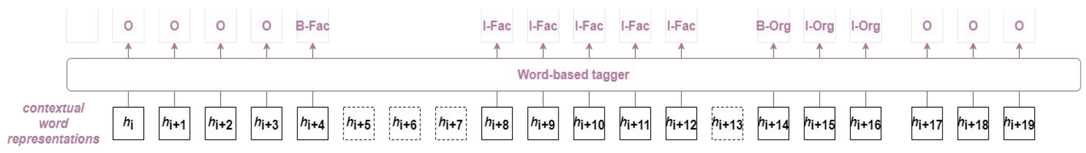
Words are split into sub-tokens. Token embeddings added with position embeddings are taken as input of publicly available pre-trained models, such as RoBERTa [1], in this work.

Both word-based and span-based methods use the same text encoder.



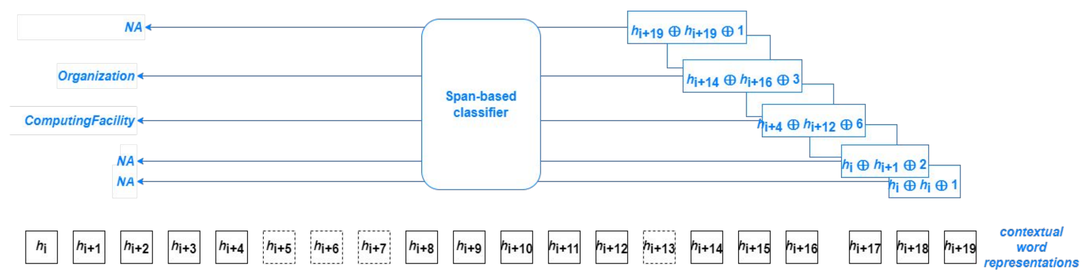
Approach 1: Word-based Tagger

Once we get the contextual representations from the encoder, we use the vector corresponding to the first sub-token with each word to represent the word. The word-based tagger takes as input a vector representing one word and outputs a tag which is usually composed of a position indicator and an entity type.



Approach 2: Span-based Classifier

We obtain the vector representations for each word in a similar way as above and then use them to build span representations. The vectors representing two boundary words and the span length (embedded as a dense vector) are concatenated and taken as input of the span-based classifier.



Results

Span-based classifier outperforms word-based model by 0.011 F1 when RoBERTa-base is used, while 0.015 F1 when RoBERTa-large is used.

We also observe modest benefit of using **RoBERTa-large** over RoBERTa-base (0.019 with word-based and 0.023 with span-based).

Method	Encoder	Validation F_1	Validation P	Validation R	Validation MCC	Testing F_1	Testing P	Testing R	Testing MCC
Word-based	base	0.8138 (0.0039)	0.8047 (0.0059)	0.8230 (0.0019)	0.9064 (0.0016)	0.7910 (0.0038)	0.7958 (0.0052)	0.7862 (0.0030)	0.8921 (0.0018)
	large	0.8242 (0.0048)	0.8191 (0.0052)	**0.8294** (0.0044)	**0.9106** (0.0013)	0.7985 (0.0040)	0.8082 (0.0048)	**0.7890** (0.0034)	**0.8959** (0.0016)
Span-based	base	0.8223 (0.0027)	0.8326 (0.0013)	0.8123 (0.0042)	0.8907 (0.0032)	0.7996 (0.0004)	**0.8238** (0.0024)	0.7768 (0.0014)	0.8760 (0.0015)
	large	**0.8267** (0.0019)	**0.8328** (0.0088)	0.8210 (0.0113)	0.8999 (0.0042)	**0.8034** (0.0015)	0.8229 (0.0092)	0.7849 (0.0101)	0.8837 (0.0036)
1^{st}		0.8364	0.8296	0.8434	0.9129	0.8057	0.8137	0.7979	0.8954
2^{nd}		0.8262	0.8145	0.8382	0.9139	0.7993	0.8013	0.7972	0.8978
3^{rd} (ours)		0.8307	0.8249	0.8366	0.9138	0.7990	0.8076	0.7906	0.8946

Table 2: A comparison between word-based and span-based entity recognition models. We report mean scores and standard deviations (in brackets), averaged over three repeats. Shared task results, shown in the bottom, are retrieved from the shared task leaderboard at the end of shared task scoring period. Bold indicates highest number among word- and span-based methods.

Future directions

- Over-segmentation

Many domain-specific terminologies are split into multiple sub-tokens. For example, the term `COSMOS' is not in the vocabulary associated with the RoBERTa pre-trained models, so it is split into four sub-tokens: `C', `OS', `M', and `OS'. Both methods suffer from over-segmentation, especially when words are split into more than 2 sub-tokens.

- Nested entities

The span-based method may predict both `COSMOS Shared Memory system' and `COSMOS Shared Memory system at DAMTP' (gold annotation) as `ComputingFacility' entities. We believe that both predictions can be useful, depending on the downstream application [2].

FOR FURTHER INFORMATION

Please contact Xiang Dai (dai.dai@csiro.au)

REFERENCES

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv, vol. 1907.11692, 2019.

[2] Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran, "NNE: A Dataset for Nested Named Entity Recognition in English Newswire," in Proceedings of ACL, 2019.

ACKNOWLEDGEMENTS

This work is supported by The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Precision Health Future Science Platform (FSP). Experiments were undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.