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

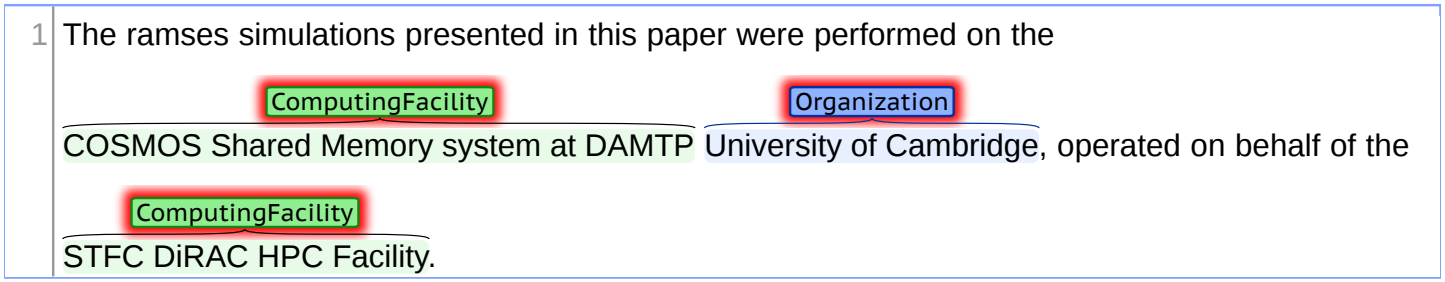
**Xiang Dai and Sarvnaz Karimi**

20 November 2022

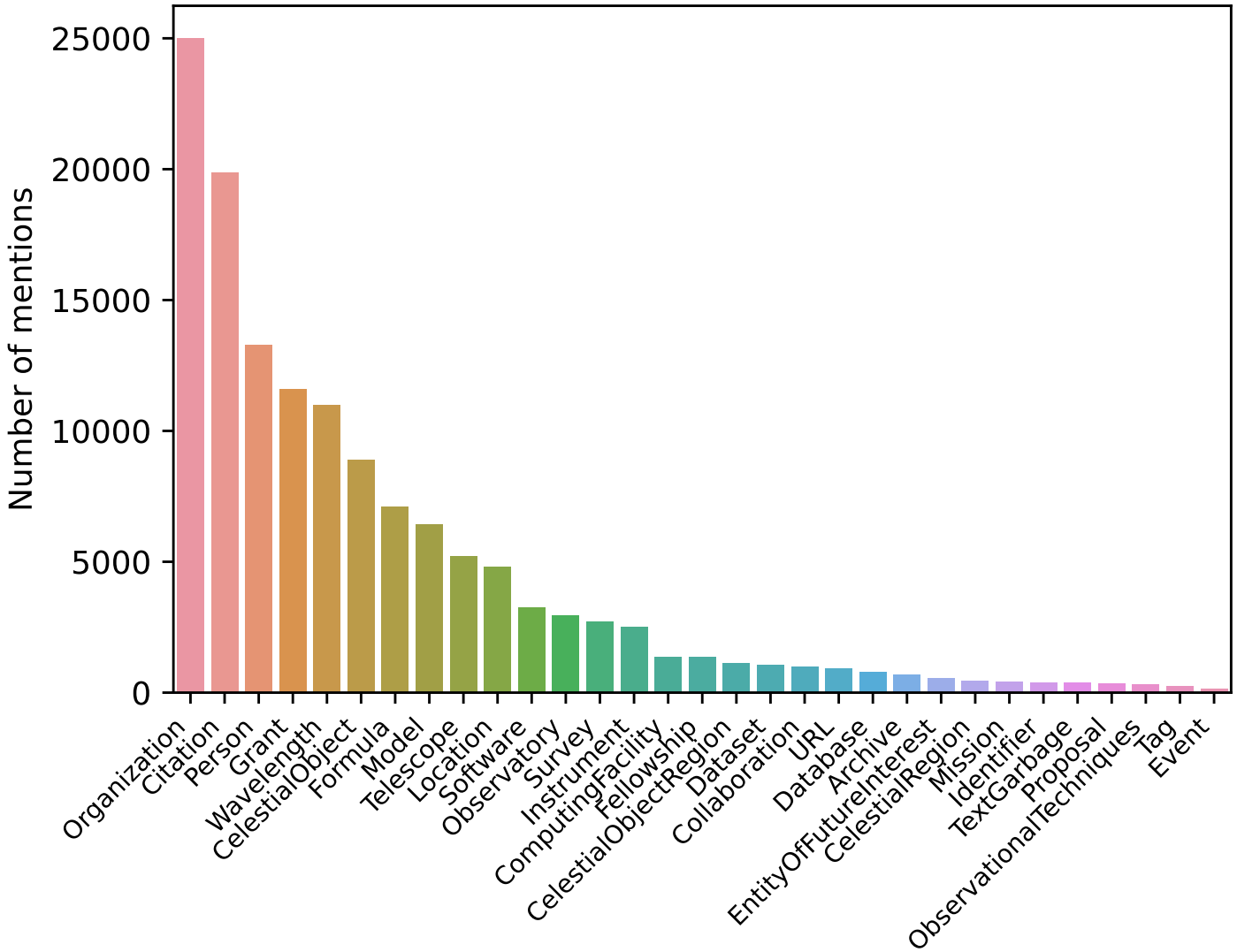

www.data61.csiro.au

# NER in the Astrophysics Literature

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

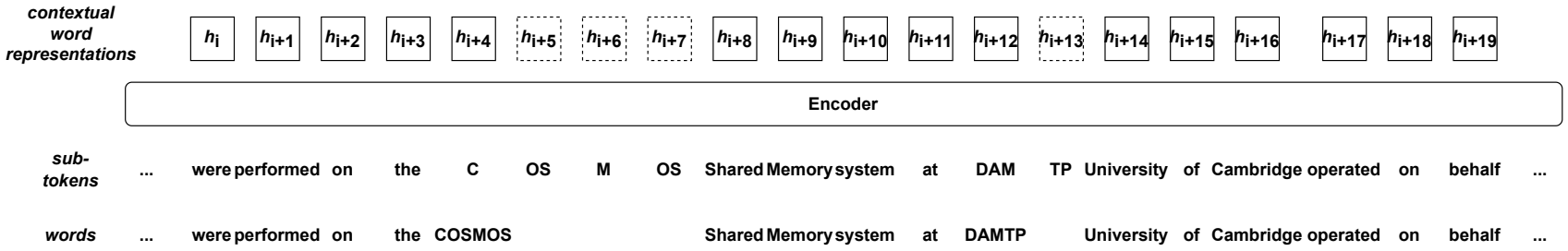


# 31 Entity Types in the DEAL dataset

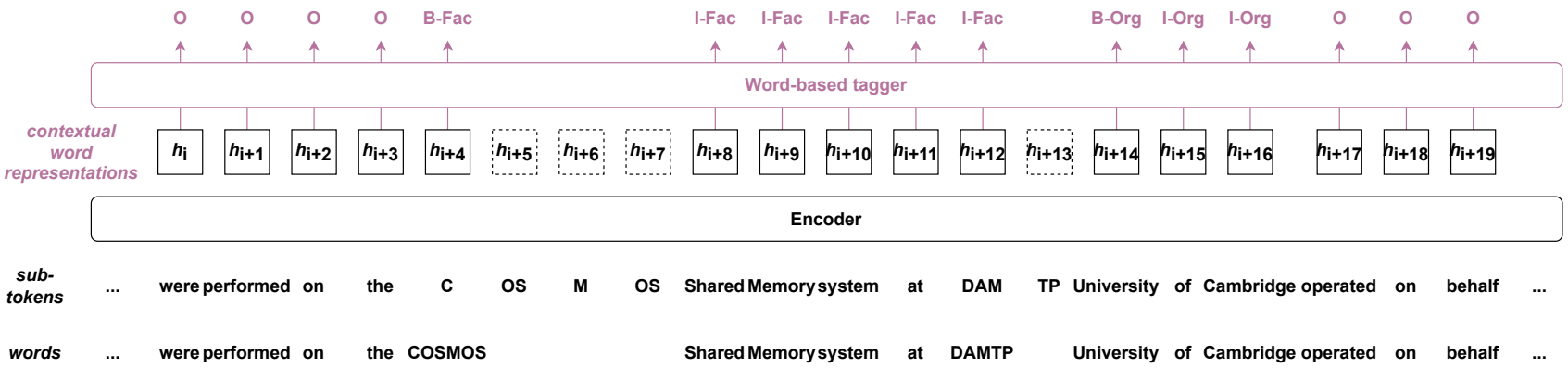


# Word-based tagger vs. Span-based classifier

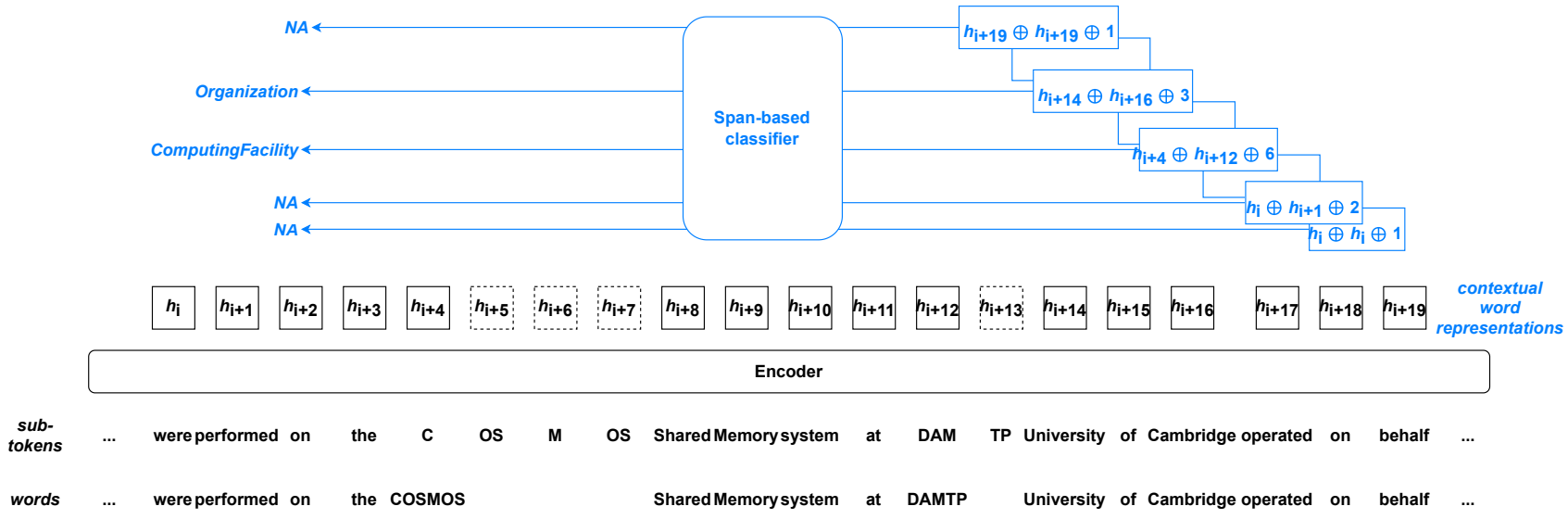
- Words are split into sub-tokens
- Encoder generates contextual token representations
- Vector corresponding to the first sub-token with each word to represent the word



# Word-based tagger



# Span-based classifier

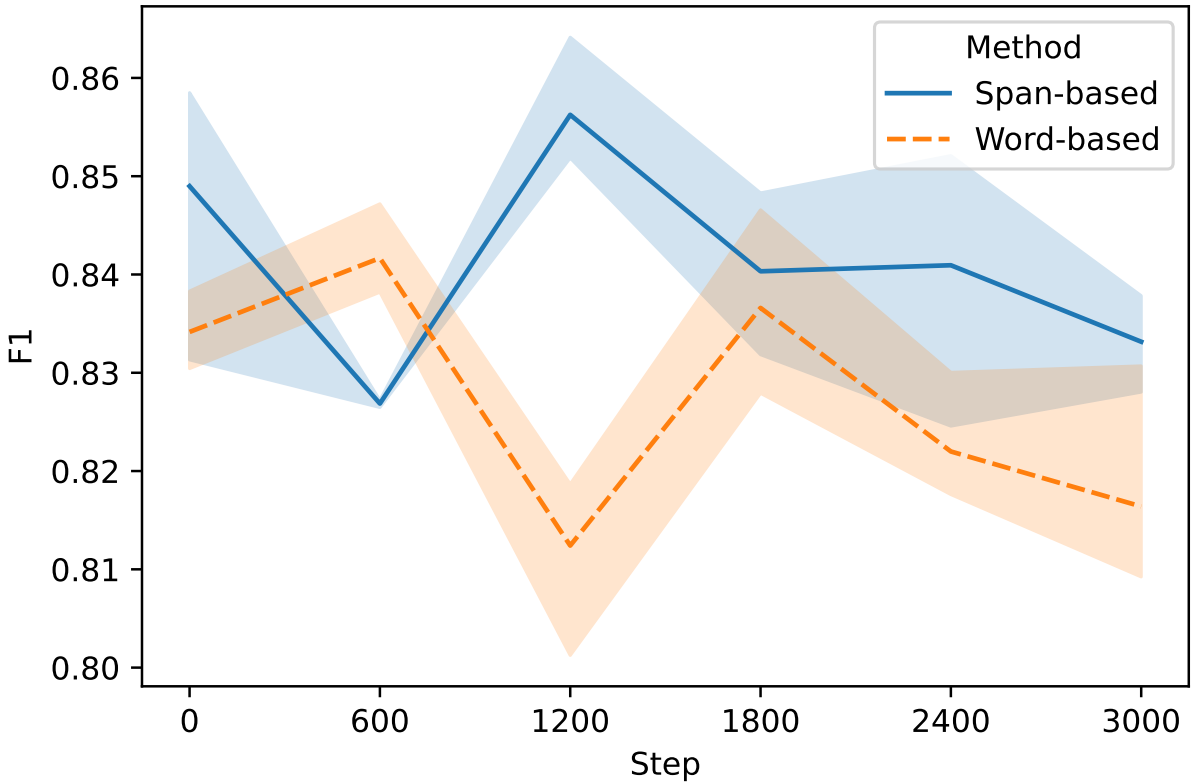


# Span-based model outperforms word-based model

| Method | Encoder | Validation | | | | Testing | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $F_1$ | P | R | MCC | $F_1$ | P | R | 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) |

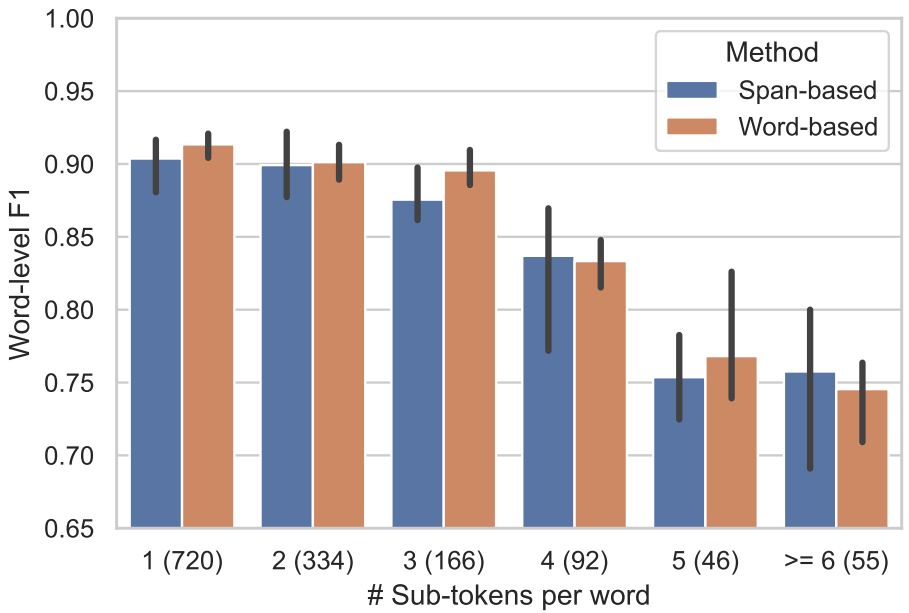
Table 2: A comparison between word-based and span-based entity recognition models. We report mean scores and standard deviations (in brackets), averaged over three repeats. Bold indicates highest number among word- and span-based methods.

# Task-adaptive pre-training does not guarantee better effectiveness



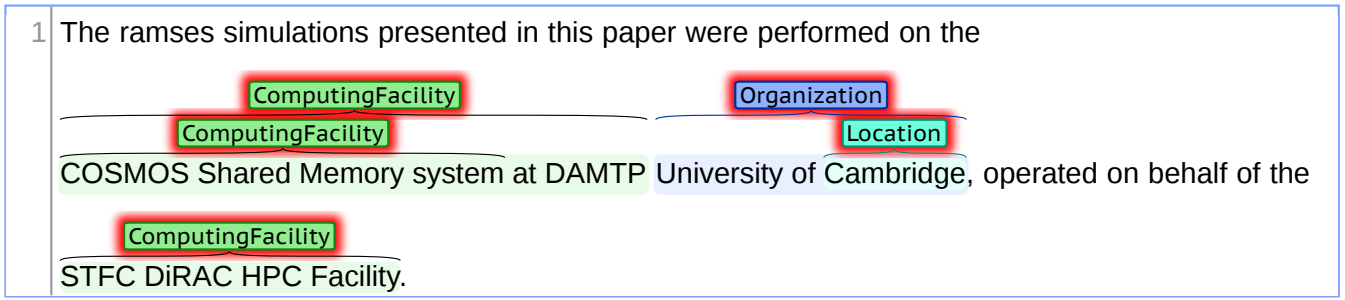
# Over-segmentation problem

- Many domain-specific terminologies are split into multiple sub-tokens.
  - 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.



# Flat entities vs. Nested entities

- Span-based method tends to predict nested entities
  - Both 'COSMOS Shared Memory system at DAMTP' and 'COSMOS Shared Memory system' are predicted as 'ComputingFacility' entities
- Depending on the downstream application, both innermost and outermost entities can be useful.[1]




[1] 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 ACL, 2019.

# Summary

- Span-based classifier outperforms previously widely used word-based taggers
  - Decent performance on detecting entities in the Astrophysics literature (great than 0.8 $F_1$ score)
- Future directions
  - Domain adaptation: task-adaptive pre-training large-scale pre-trained models; mitigate over-segmentation issue
  - Recognize entities that form nested structure

# Questions

- Feel free to contact dai.dai@csiro.au