A Neural Named Entity Recognition System for Biological Entity Identification

Emily Sheng, Scott Miller, José Luis Ambite, Prem Natarajan

Motivation

- **Scientific information extraction has attracted increasing research interest in recent years**
  - DARPA Big Mechanism, etc.
- **Of key interest are "scientific entities" and normalizing extractions to standard identifiers**
  - E.g., *liver* → *UBERON:0002107*
- **Entities normalized to identifiers are useful for:**
  - Document- or corpus-level summarization of important concepts
  - Facilitating effective information retrieval
  - Linking textual data with data from other modalities (e.g., figures)

Outline

1. Dataset
2. Methods
3. Results
4. More experiments

1. **Dataset**
2. Methods
3. Results
4. More experiments

Dataset and preprocessing

- **Statistics for annotated dataset**
  - 570 articles with annotated figure captions
  - Processed into 42,587 sentences
    - **34,307 (train), 4,037 (dev), 4,243 (test) sentences**
  - We do simple tokenization and process the “words” and annotations in a sentence into the CoNLL format
    - **IOB (inside-outside-beginning) annotation at the “word” level**

```
1   (A)             O
2   HEK293          B-celltypeline
3   cells           I-celltypeline
4   were            O
5   transfected     O
6   with            O
7   DAPKΔCaM        B-proteingene
8   ...             ...
```

Dataset and preprocessing

- **Original dataset**
  - 94,266 annotations with IDs
  - 110,416 annotations total
- **Preprocessed dataset**
  - 79,154 annotations with IDs
  - 89,095 annotations total
- **IOB annotation at word-level means we cannot capture all original annotations at byte-level**
  - Ex: ”RAS-GFP” is tagged as one entity

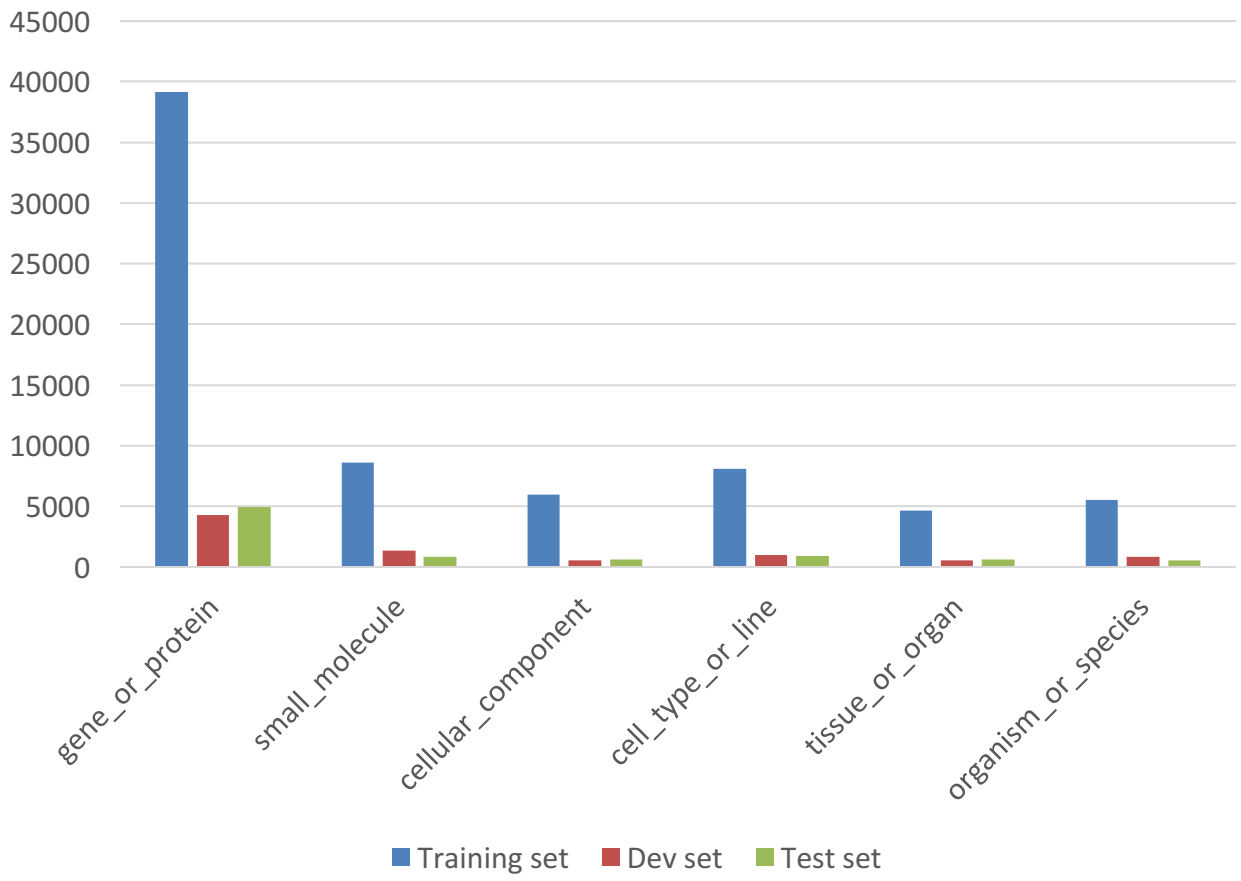


Table 1. Distribution of annotated entity types in training, dev, and test sets

1. Dataset
2. **Methods**
3. Results
4. More experiments

Survey of methods for NER

- **Classification-based**
  - Naïve Bayes, decision trees, support vector machines (SVMs), maximum entropy
- **Sequence-based**
  - Conditional random fields (CRFs), Hidden Markov Models (HMMs)
  - CRFs popular in biomedical NER
- **Neural architectures**
  - Combinations of long short-term memory networks (LSTM), convolutional neural networks (CNN), and CRFS
  - Applied to general NER: (Chiu and Nichols, 2015), (Lample et al., 2016), (Ma and Hovy, 2016)
  - Applied to bio NER: (Chalapathy et al., 2016a, b), (Habibi et al., 2017)

Methods for NER

- **Baseline**
  - Traditional CRF approach
  - Trained using NERSuite
  - NERSuite features derived from tokenizer, part-of-speech tagger, lemmatizer, chunker
- **Neural architecture**
  - Submitted model is based on the bidirectional long short-term memory (Bi-LSTM) and CRF model of (Lample et al., 2016)
  - Ongoing experiments with additional architectures

Neural NER

- **Word embeddings**
  - Pre-trained on all abstracts from PubMed, all full-text from PubMed Central, and a Wikipedia dump[1]
- **Character-based representation**
  - Derived from a Bi-LSTM module
- **Bi-LSTM+CRF with final word representation**
  - Final word representation:
    - **Word embeddings**
    - **Character-based representation**
  - Bi-LSTM+CRF module to predict IOB tags of each "word"

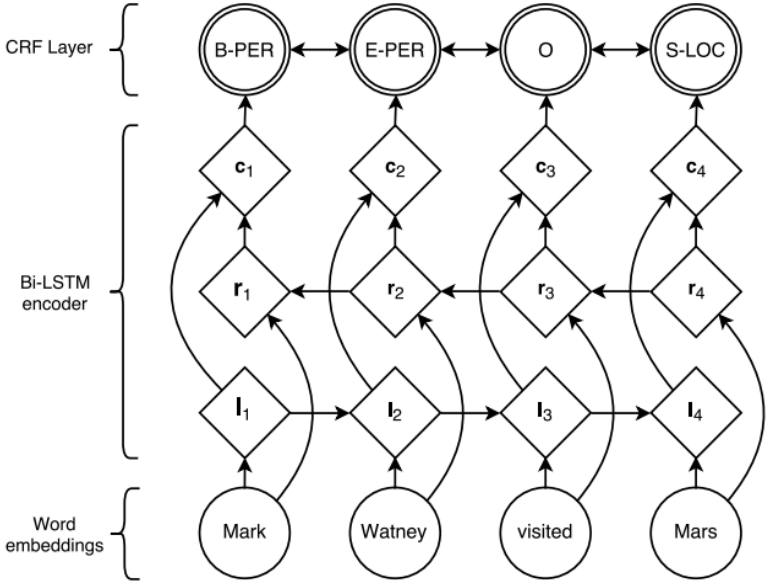


Figure 1. This figure is taken from (Lample et al., 2016). We apply their network with minor changes to the BioCreative dataset.

[1] embeddings from http://bio.nlplab.org/

Post-processing using simple heuristics

- **Facilitates more accurate normalization**
- **Recursively strip away certain punctuation from the starts and ends of words**
  - E.g., *(Tau)* becomes *Tau*
- **Does not perfectly deal with all extraneous characters**
  - E.g., *B-RAF(V600E)* becomes *B-RAF(V600E*
  - E.g., *RAS-GFP* is still tagged as one entity
    - **More heuristics for proteins/genes**

Methods for normalization

- **Contextual dictionary**
  - For entity “Tau” in “EcrTgTaumouse”, we can use “EcrTg” and “mouse” as context words
  - Create dictionary from annotated entities in all training data

```
{
    entityA:
        ID1: ["mouse", "EcrTg"],
        ID2: ["human", "EcrTg"],
        ...,
    entityB:
        ...
}
```

- **External knowledge bases**
  - If entity not found in contextual dictionary, search appropriate knowledge base through API searches
  - For proteins/genes, additional heuristics:
    - **If entity not found, try to split on whitespace, “/”, “-”, “;” and normalize the split entities**

1. Dataset
2. Methods
3. **Results**
4. More experiments

Task entity extraction results

Entity types	Strict span match for all annotations			Strict span match for norm. annotations only			Span overlap match for all annotations			Span overlap match for norm. annotations only		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
gene_or_protein	50.87	61.31	55.60	52.96	59.61	**56.09**	68.55	82.63	**74.94**	61.51	69.24	65.15
small_molecule	56.23	45.09	**50.04**	65.90	32.22	43.28	68.31	54.78	**60.80**	70.54	34.49	46.33
cellular_component	54.77	43.94	48.76	61.24	41.03	**49.14**	62.89	50.46	**55.99**	65.13	43.64	52.26
cell_type_or_line	65.31	65.03	65.17	82.23	55.15	**66.02**	76.63	76.30	**76.47**	86.58	58.06	69.51
tissue_or_organ	57.24	55.87	**56.55**	61.46	46.67	53.05	67.05	65.44	**66.24**	66.74	50.68	57.61
organism_or_species	74.62	71.52	73.04	85.52	69.50	**76.68**	81.36	77.98	**79.63**	87.82	71.37	78.74

Table 2. Precision, recall, and F_1 scores across entities for submitted NER model

- Across all entity types, most of the incorrectly tagged entities are because of missing or spurious predictions, not label or span clashes
- Difference in span overlap versus strict span scores could be minimized with better heuristics

Task normalization results

- **Baseline results using simple normalization heuristics**
- **Scores for `gene_or_protein` are significantly lower**
  - Even when not using special heuristics for genes and proteins
- **Additional analysis**
  - How much would knowing the associated organism boost `gene_or_protein` scores?
  - For predicted entities with correct label and span, we estimate half of the incorrectly normalized entities are due to organism mismatch

Entity types	Micro-averaged scores for normalized IDs			Macro-averaged scores across captions for normalized IDs		
	P	R	F_1	P	R	F_1
gene_or_protein	16.98	22.43	**19.33**	23.24	30.37	**16.69**
small_molecule	65.42	39.37	49.16	77.23	47.70	34.11
cellular_component	54.98	44.97	49.47	67.25	53.98	35.05
cell_type_or_line	78.42	55.69	65.13	85.58	60.15	55.51
tissue_or_organ	58.38	44.21	50.32	69.90	54.63	37.92
organism_or_species	77.23	69.10	72.94	82.48	74.69	65.46

Table 3. Precision, recall, and F_1 scores across entities for submitted normalization model

1. Dataset
2. Methods
3. Results
4. **More experiments**

More experiments

- **NER**
  - Improvements in speed and accuracy with different neural architectures
  - Tokenization
- **Normalization**
  - Neural network-based normalization techniques
  - Incorporating additional organism information
- **Data**
  - Distant supervision techniques to augment data

Conclusion

- **This work presents a pipeline for biological entity ID which employs:**
  - state-of-the-art NER techniques
  - simple normalization heuristics
- **Baseline results and error analyses point to areas that could be improved upon**
  - Tokenization
  - Different NER architectures
  - Going beyond rule-based systems and heuristics for normalization

Acknowledgments

This work was supported in part by the DARPA Big Mechanism program (W911NF-14-1-0364).

Thank you for listening!

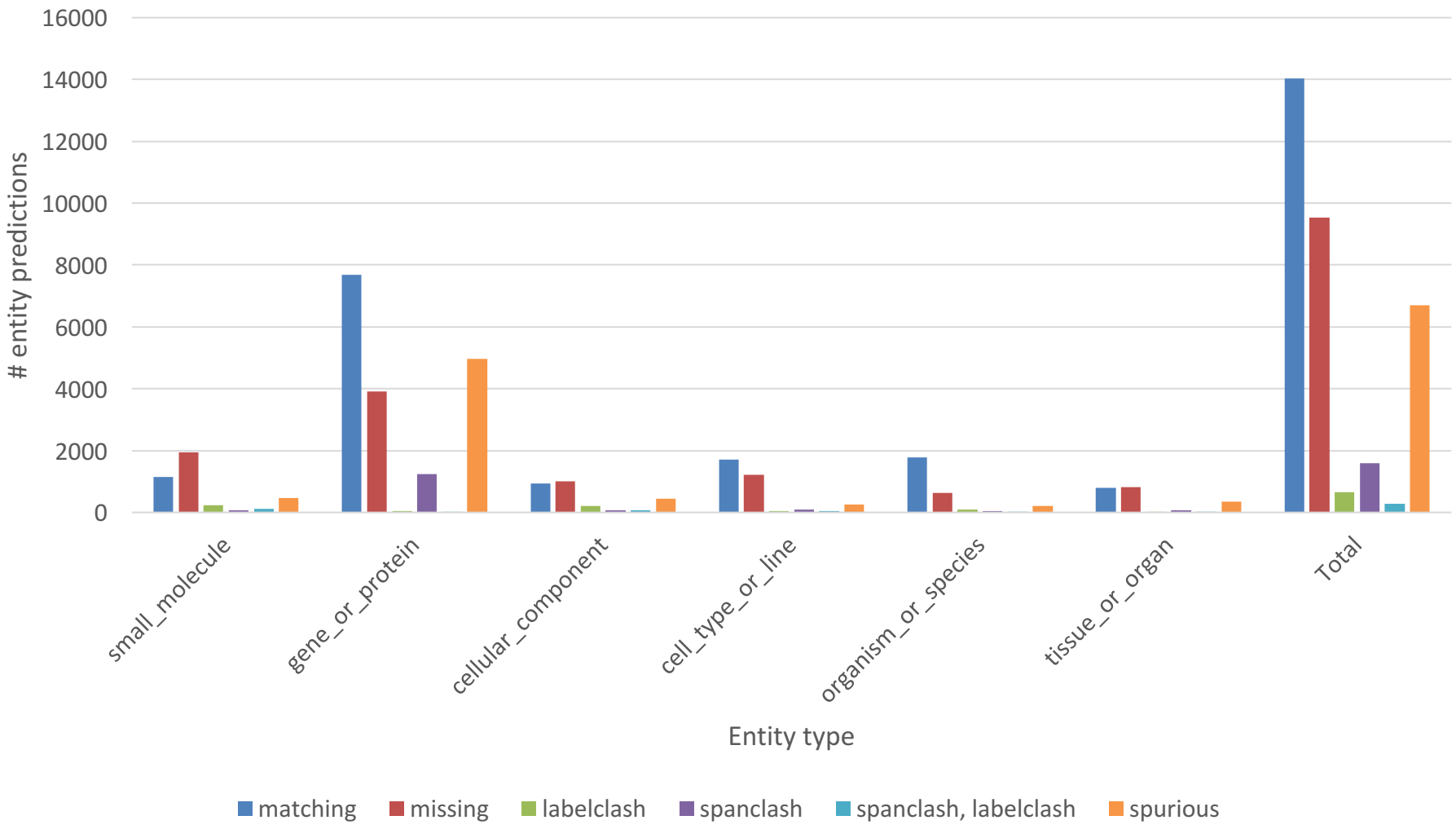
SUPPLEMENTARY MATERIALS

NER results on dev-test data

Entity type	NERSuite			BiLSTM-BiLSTM-CRF		
	P	R	F_1	P	R	F_1
gene_or_protein	76.09	79.83	77.91	86.52	88.37	87.43
small_molecule	72.77	60.13	65.85	77.07	66.28	71.27
cellular_component	73.57	70.07	71.78	79.30	65.80	71.92
cell_type_or_line	67.60	62.59	65.00	76.85	65.53	70.74
tissue_or_organ	68.34	49.26	57.25	70.58	58.22	63.80
organism_or_species	61.89	65.08	63.44	72.59	75.04	73.79
Total	73.36	71.90	**72.62**	82.25	78.87	**80.53**

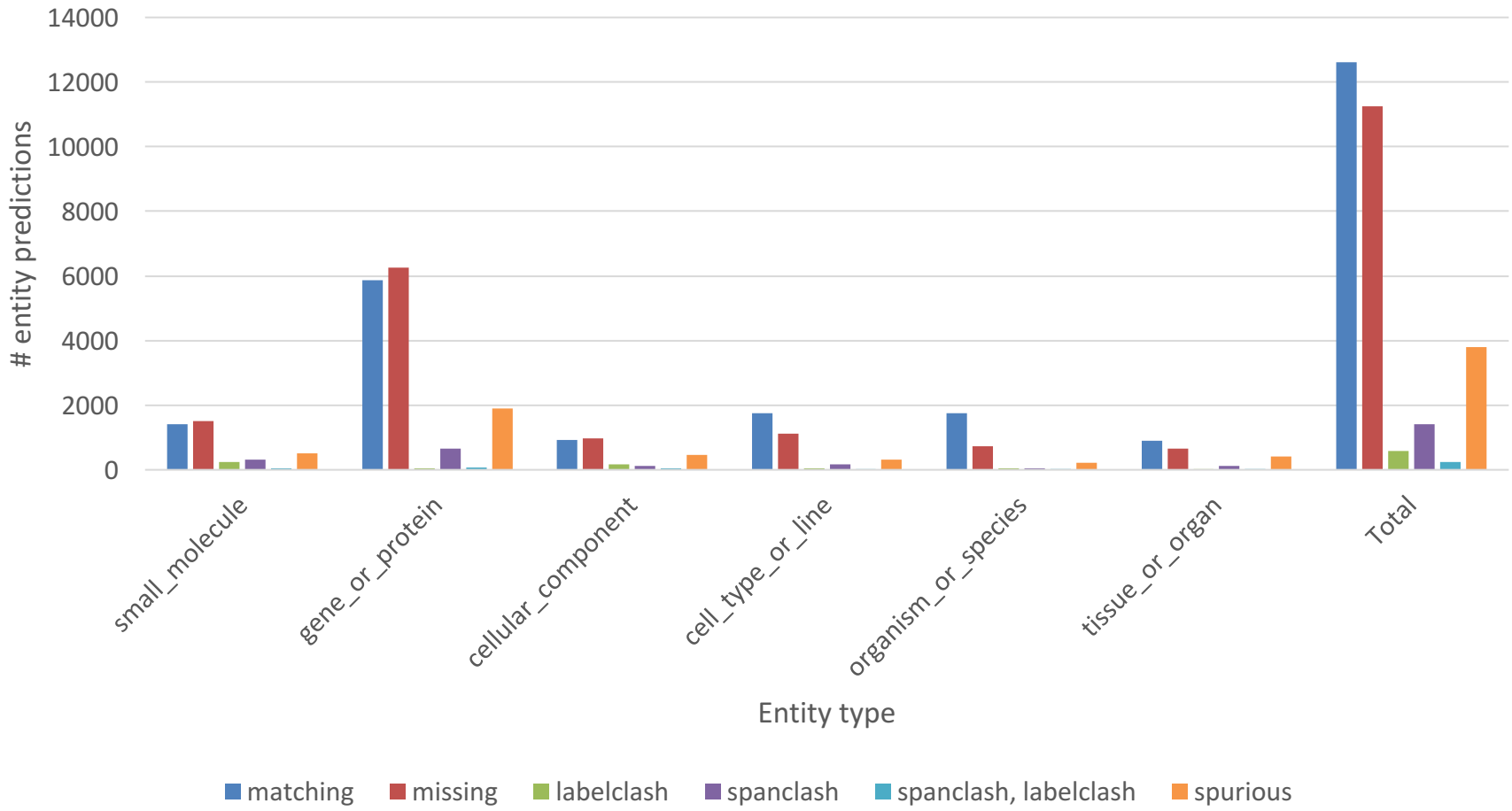
Table 2. Precision, recall, and F_1 scores across entities for different NER models. We split the provided training data into train-dev-test sets; these scores are for the test set.

NER predictions for norm. entities only

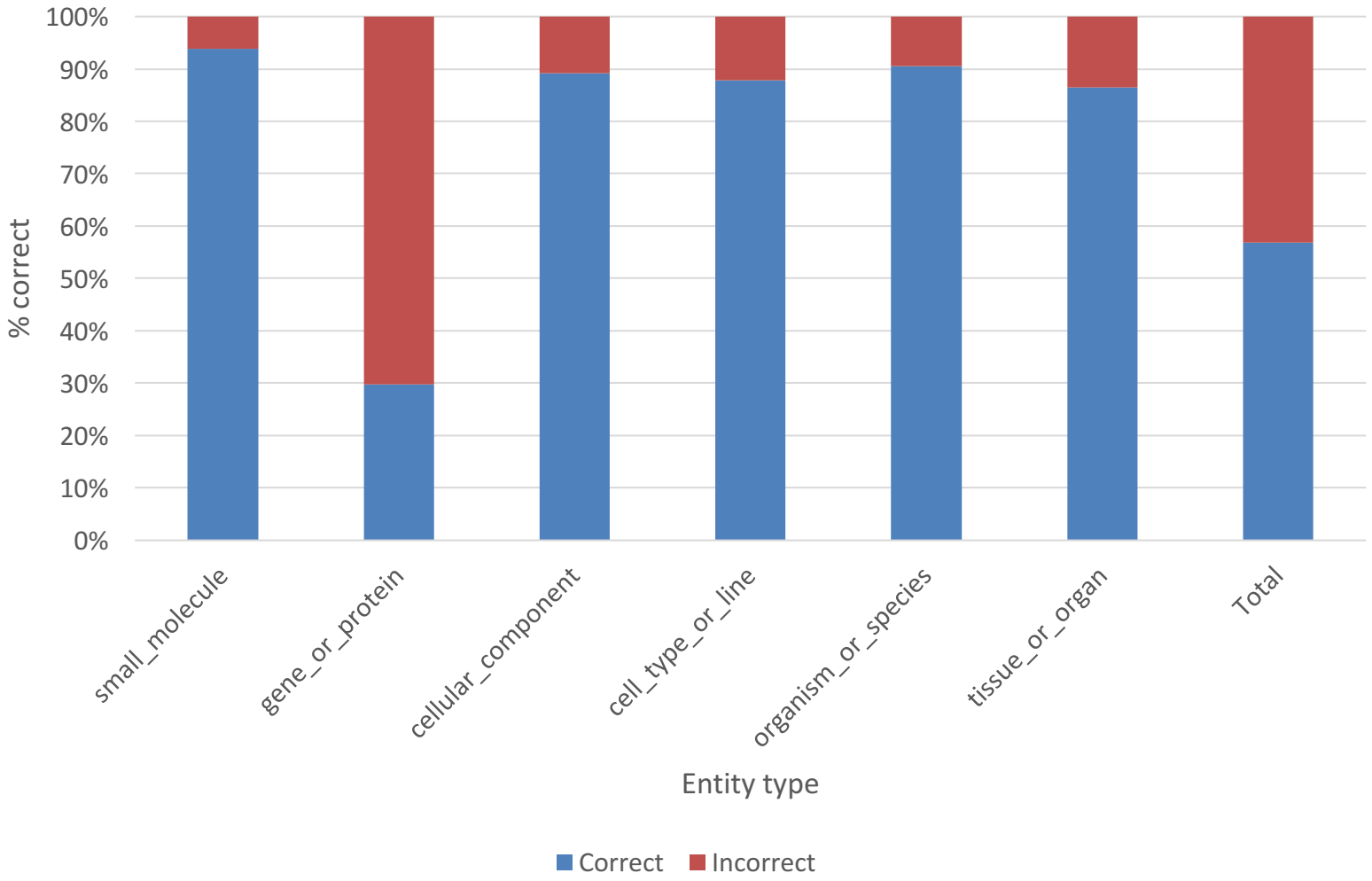


NER predictions for norm. entities only

Using no post-processing heuristics in normalization techniques



Normalization analysis for norm. entities only



Normalization analysis for norm. entities only

Using no post-processing heuristics in normalization techniques