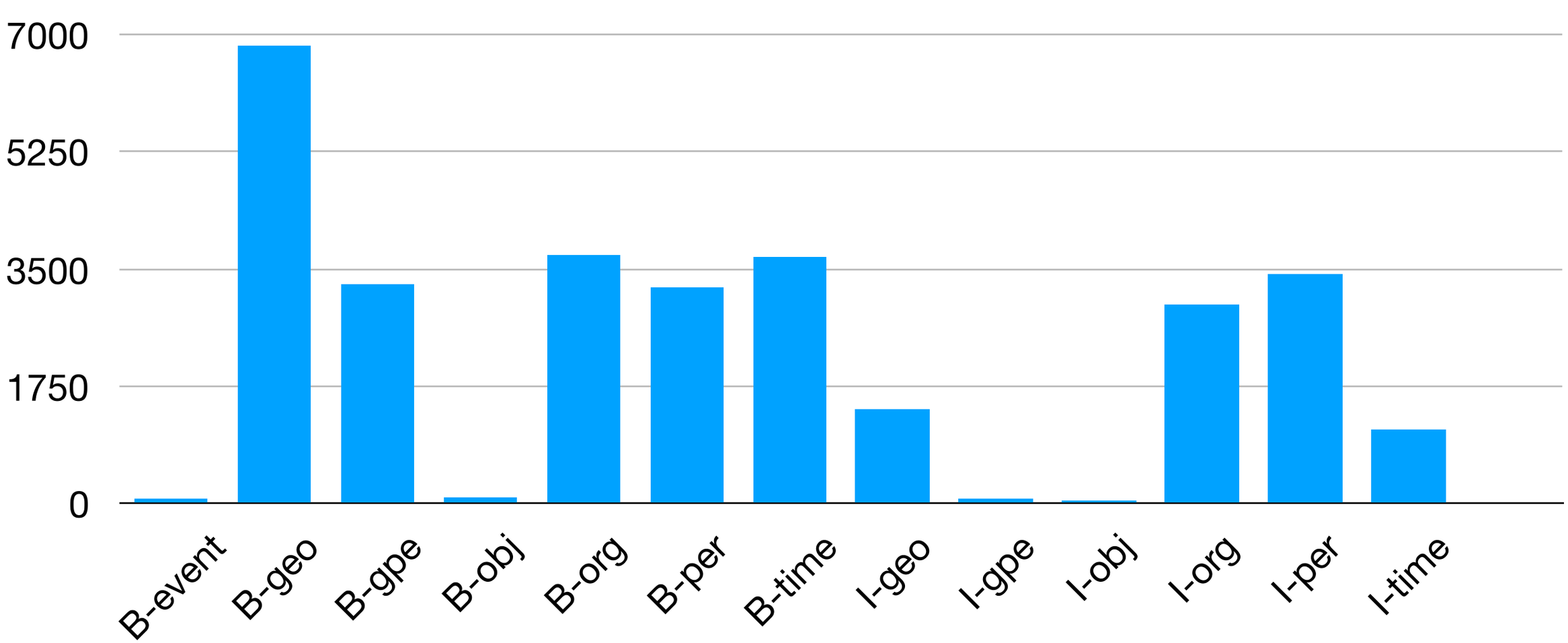
Named Entity Recognition

Benda Krisztián, Szántó Tamás

Training data

- 9000 sentences, 196645 words, 15 entities
- O: 166610, I-event: 0



SpaCy

- Industrial-Strength Natural Language Processing, POS tagger, NER solution

SYSTEM	YEAR	LANGUAGE	ACCURACY	SPEED (WPS)
spaCy v2.x	2017	Python / Cython	**92.6**	*n/a* ?
spaCy v1.x	2015	Python / Cython	91.8	13,963
ClearNLP	2015	Java	91.7	10,271
CoreNLP	2015	Java	89.6	8,602
MATE	2015	Java	92.5	550
Turbo	2015	C++	92.4	349

- Why did we choose SpaCy?

- Built-in models for NER (OntoNotes 5, Common Crawl):
  - en_core_web_sm 35 MB
  - en_core_web_md 115 MB
  - en_core_web_lg 812 MB

SpaCy NER

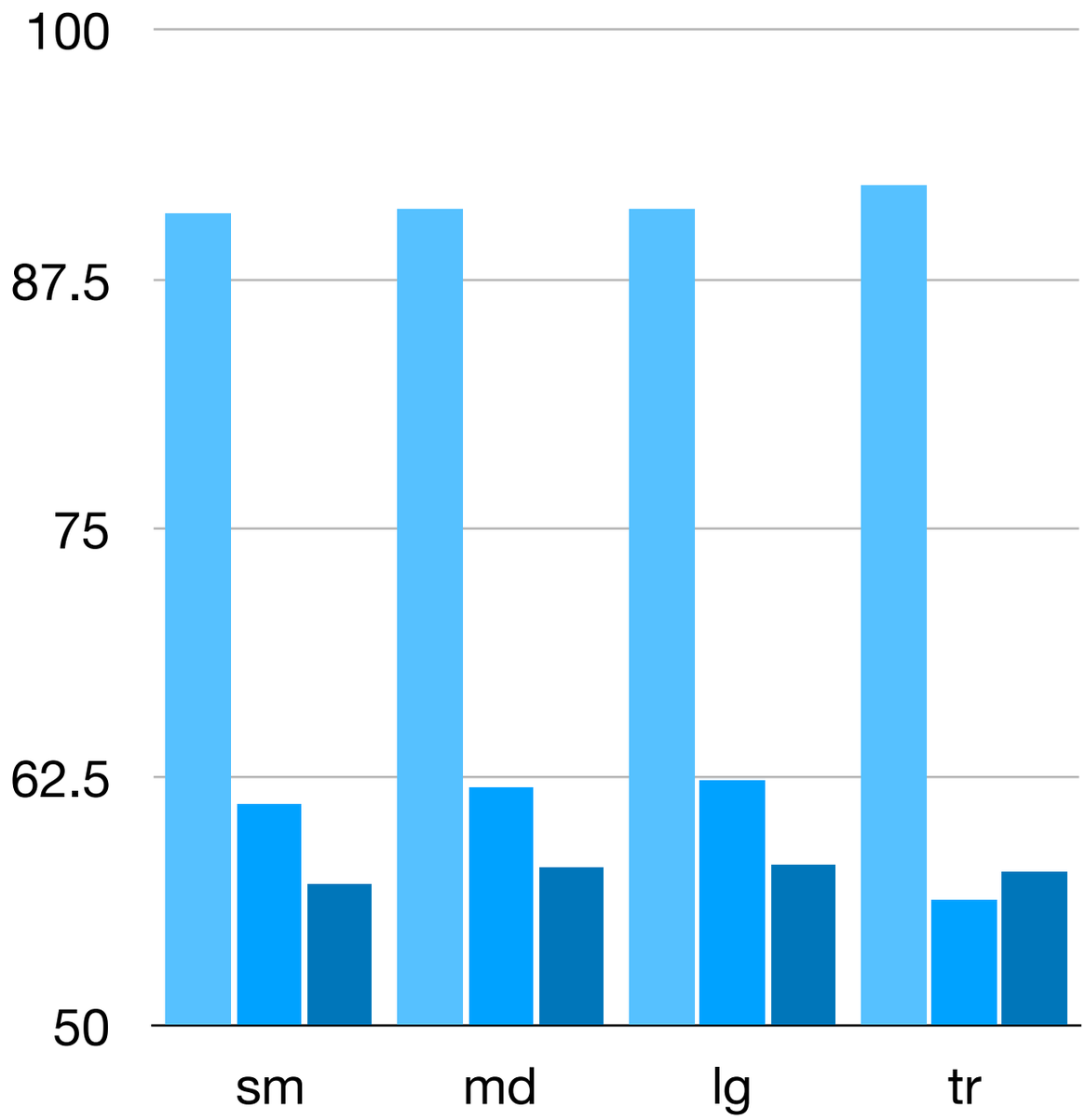
- Supported entities: 18, B and I manually by space
- Entity mapping:
    - event - EVENT
    - geo - LOC, GPE
    - obj - PRODUCT, WORK_OF_ART
    - org - ORG
    - per - PERSON
    - time - TIME, DATE
    - gpe - NORP
- NORP: Nationalities or religious or political groups.
- GPE: Countries, cities, states.

Training

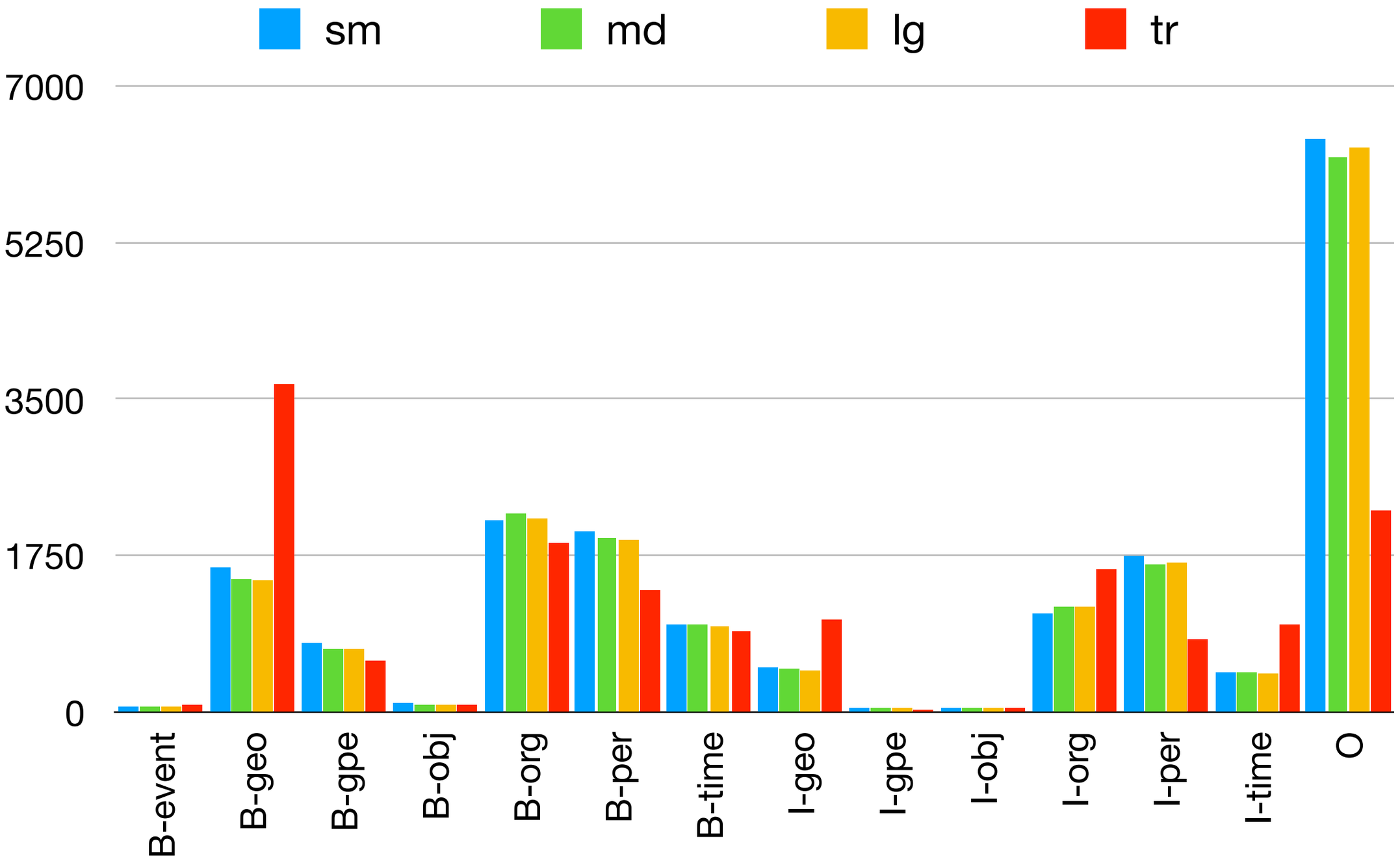
	First	Second	Third
Iteration	100	128	300
Training Data Size	2000 rows	500 sentences	1000 sentences
Accuracy	87.98%	91.86%	92.03%
Recall **Precision**	35.17% 34.90%	54.61% 55.95%	55.38% 56.77%

Results

model	accuracy	recall	precision
sm	90.82	61.19	57.19
md	91.03	61.95	57.99
lg	91.03	62.34	58.11
tr	92.18	56.30	57.79

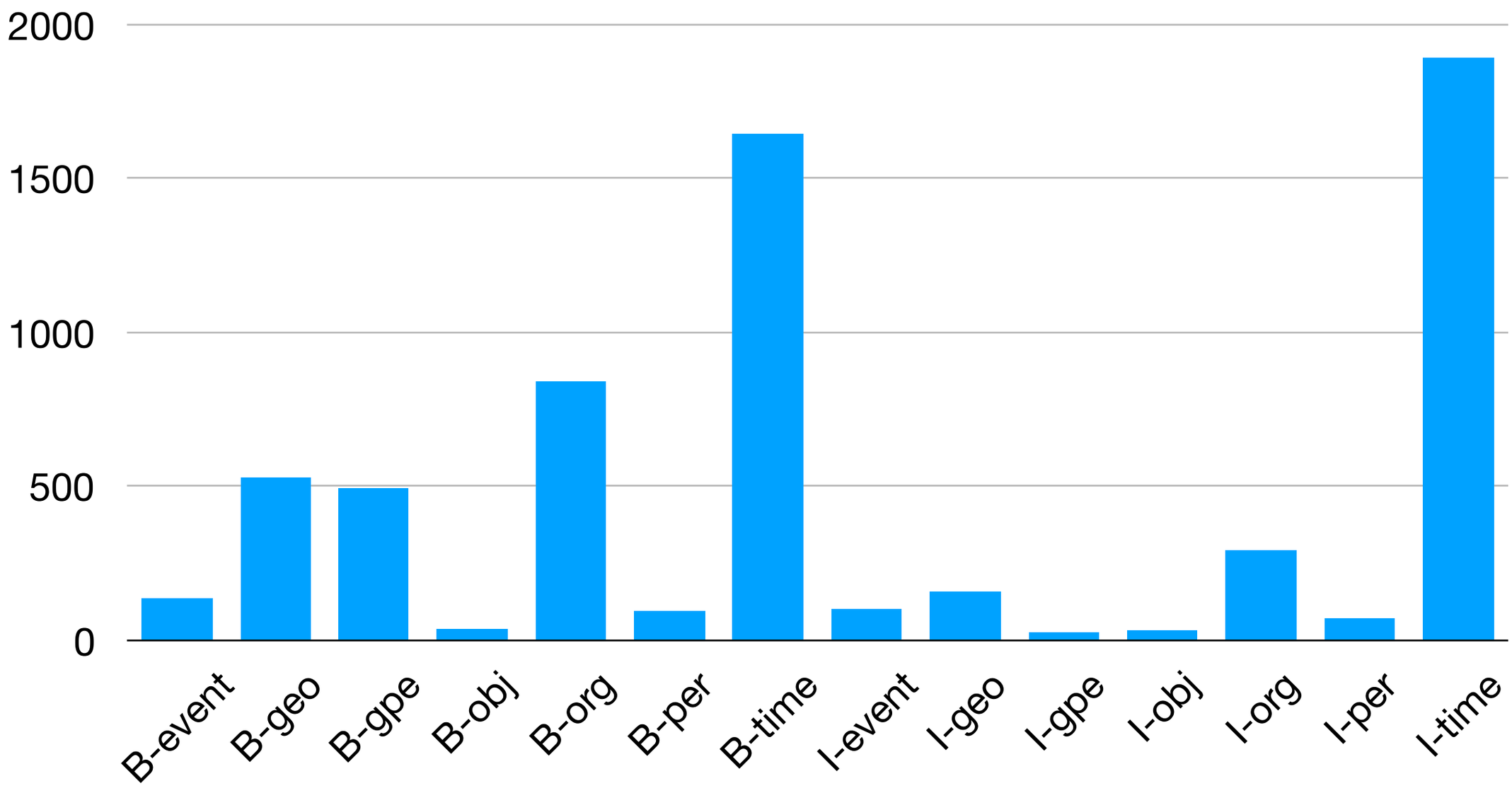


Errors by entities



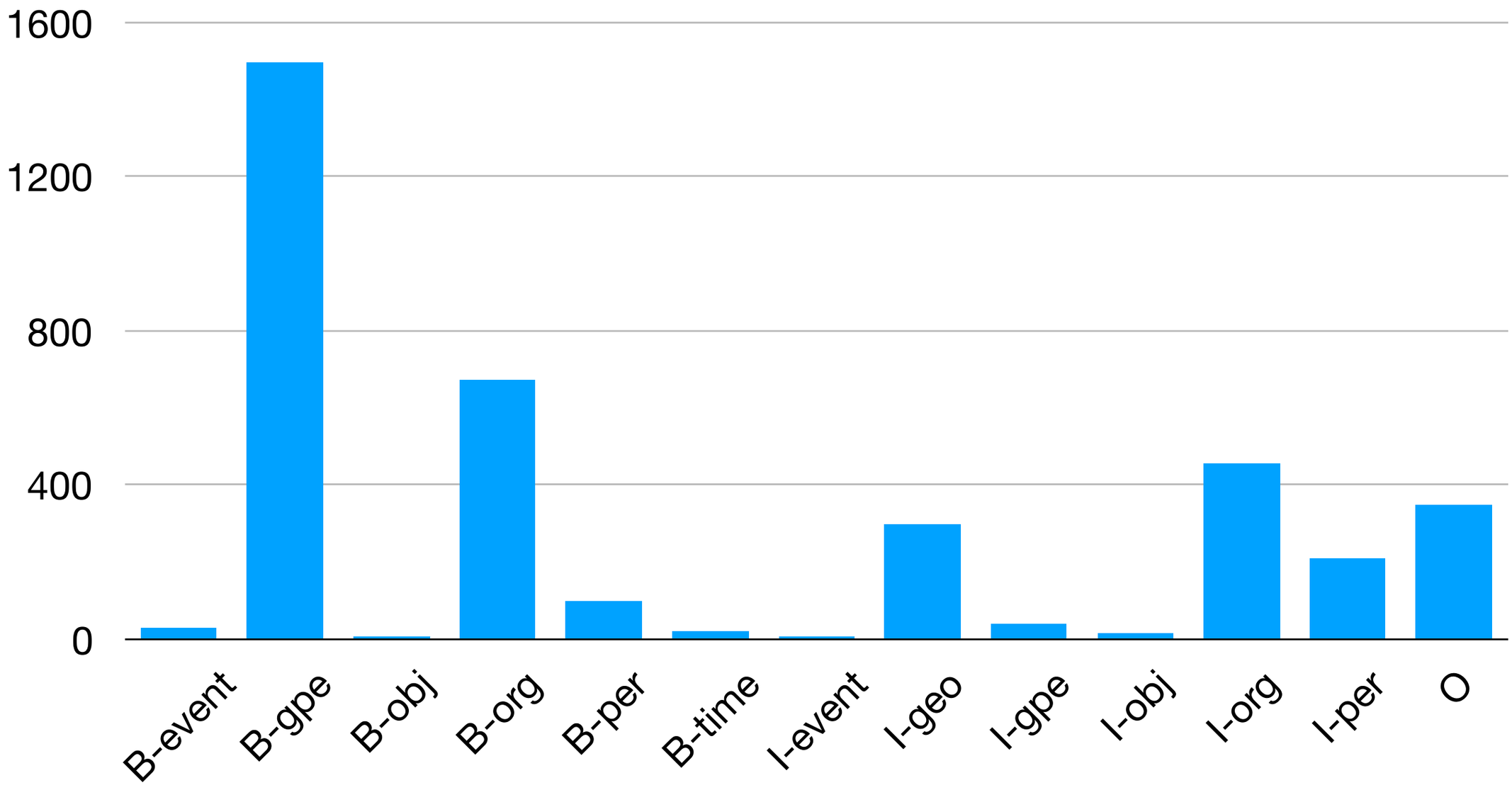
Wrong predictions

Predictions instead of the O entity by the lg model:



Wrong predictions

Predictions instead of the B-geo entity by the trained model:



Conclusion

- Built-in model is the best currently, but it has limited further improvements

- Trained model is almost as good as the built-in, plenty of room for improvements - iterations, used sentences

- Goal: optimising the training