Named Entity Recognition - Is there a glass ceiling?

Tomasz Stanislawek[1,3], Anna Wróblewska[1,3], Alicja Wójcicka[1,4], Daniel Ziembicki[4], Przemyslaw Biecek[2,3]

[1]Applica.ai, Warsaw, Poland, [2]Samsung R&D Institute Poland, Warsaw, Poland
[3]Faculty of Mathematics and Information Science, Warsaw University of Technology
[4]Department of Formal Linguistics, University of Warsaw

Introduction

Do we know which types of errors are still hard or even impossible to correct? In collaboration with domain experts we created a new general ontology for types of errors in NER in order to better compare strong and weak sides of different models. Our study presents also general limitations in order to achieve 100% of accuracy. Presented results are based on the **Stanford** [3], **CMU** [4], **ELMO** [5], **FLAIR** [1] and **BERT** [2] models on **CoNLL 2003 data set for the English language**.

Linguistic categories and error statistics for models

Linguistic category	Stanford	CMU	ELMO	FLAIR	BERT
DE-WT, Word Typos	10	6	9	8	10
DE-BS, Word/Sentence Bad Segmentation	38	39	33	33	40
SL-S, Sentence Level Structure	46	21	13	16	11
SL-C, Sentence Level Context	448	378	250	223	300
DL-CR, Document Co-Reference	372	316	198	184	263
DL-S, Document Structure	202	107	97	100	117
DL-C, Document Context	247	175	144	146	170
G-A, General Ambiguity	219	183	98	101	94
G-HC, General Hard Case	72	68	65	59	65
G-I, General Inconsistency	19	20	21	20	20
Errors (1101 in total)	703	554	395	370	472
Unique errors	235	93	23	12	79

Strong and weak points of models

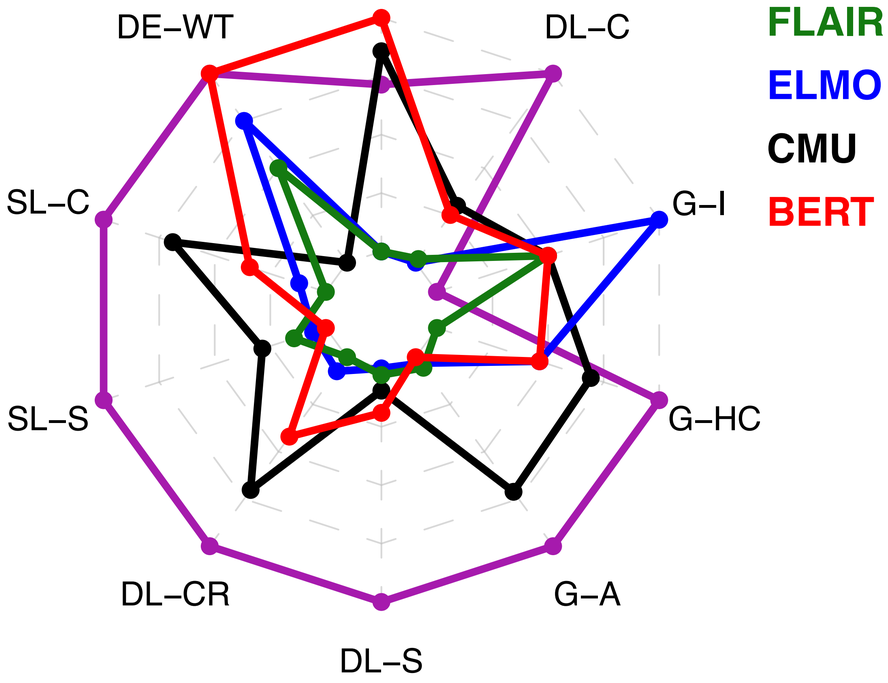


Figure 1: Radar plot with the strong and weak sides of NER models. A radius corresponds to a number of errors in a given linguistic category, the smaller the better.

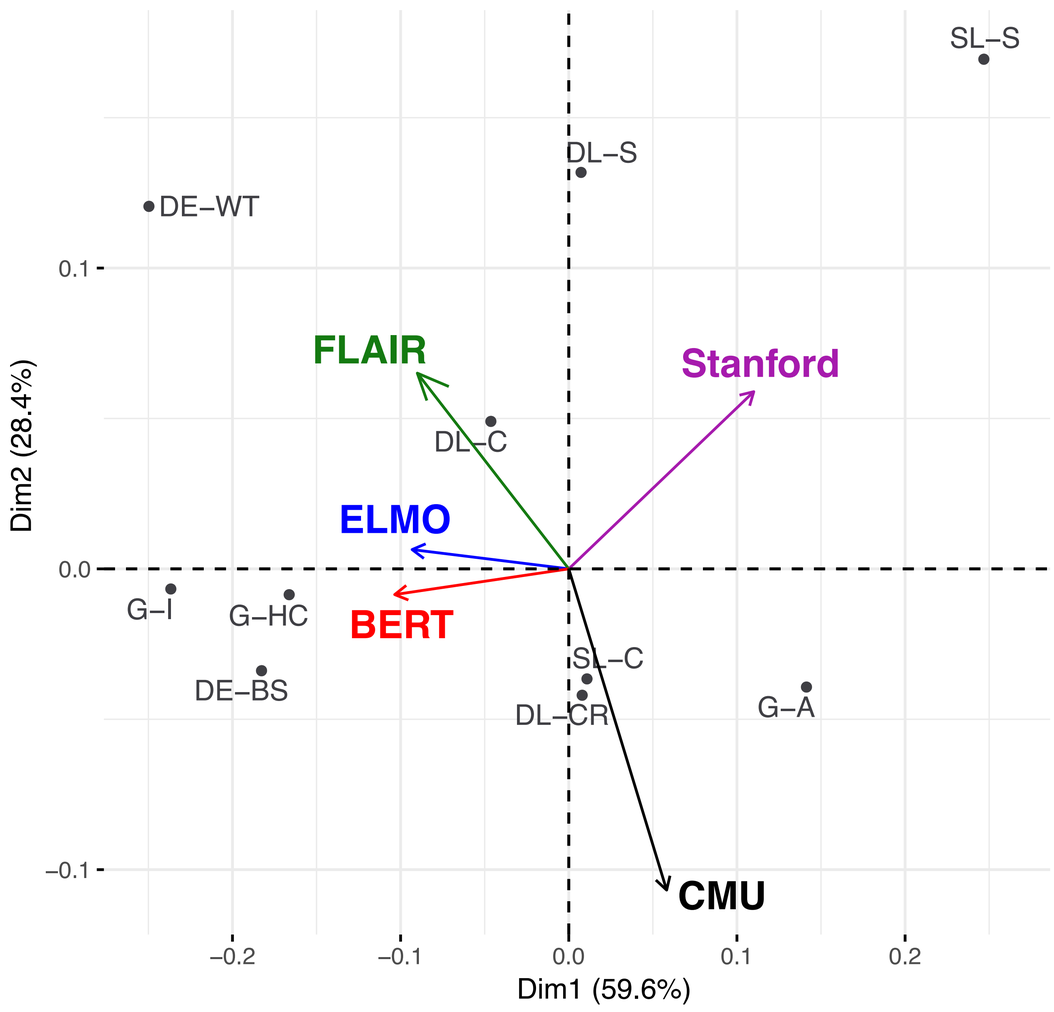


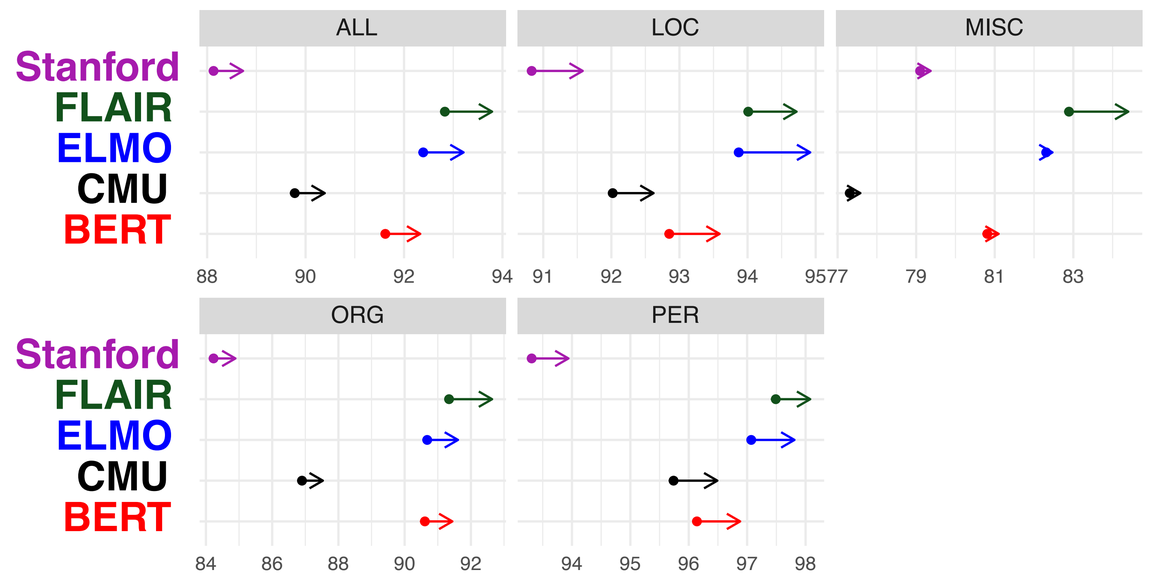
Figure 2: Correspondence analysis for the models' errors. ELMO, FLAIR and BERT are more affected by G-HC and G-I.

Conclusions

1. Some problems cannot be resolved with this data set: DE-WT, DE-BS, G-I, G-HC
2. The CoNLL 2003 test set is certainly too small to test the generalisation and stability. One of solutions is to create new diagnostic data sets for checking specific linguistic properties
3. Modern techniques like ELMO, FLAIR and BERT reduced a number of mistakes in SL-C and G-A by more than 50%
4. Work on document context rather on sentence context
5. Use also document layout features (e.g. a table, its rows and columns, and heading) for entity detection

Annotation quality

Here we present results for the standard CoNLL 2003 test set concerning NE classes (ALL comprise PER, ORG, LOC, MISC) and the same set after the re-annotation and correction of annotation errors.



Diagnostic data sets

The goal of this approach was to find, or create, more examples that reflect the most challenging linguistic properties. We selected 65 examples from Wikipedia articles with different topics per two groups of linguistic problems: document-level context (DCS) and template sentence (TS).[a] Random Sentences (RS) is another type of a diagnostic set, which we generated from random words and letters that are capitalized or not.

	Stanford	CMU	ELMO	FLAIR	BERT
DCS (F1)	45.37	61.86	**76.36**	71.89	68.90
TS-O (F1)	68.96	79.66	**89.45**	88.51	83.47
TS-R (F1)	63.06	72.86	85.01	**86.63**	79.66
RS (Number)	3571	3339	2096	**1404**	3086

An example of Document Context Sentences (DCS) diagnostic data set is sentence and its context is as follows: *"In 2003, Loyola Academy (X, ORG) opened a new 60-acre campus ... The property, once part of the decommissioned NAS Glenview, was purchased by Loyola (X,ORG) in 2001."* The second occurrence of the *"Loyola"* name is difficult to recognize as an organization without its first occurrence, i.e. *"Loyola Academy"*.

References

[1] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics, 2018.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics, 2016.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

▶ ✉ tomasz.stanislawek@applica.ai

▶ Full version of the paper is available at: https://arxiv.org/abs/1910.02403

[a]Our prepared diagnostic data sets are available at https://github.com/applicaai/ner-resources