Supplementary materials for publication: Named Entity Recognition - Is there a glass ceiling?

Tomasz Stanislawek[†,‡], **Anna Wróblewska**[†,‡], **Alicja Wójcika**[†,§], **Daniel Ziembicki**[†,§], **Przemyslaw Biecek**[‡,¶]

[†]Applica.ai, Warsaw, Poland
[¶]Samsung Research Poland, Warsaw, Poland
[‡]Faculty of Mathematics and Information Science, Warsaw University of Technology
[§]Department of Formal Linguistics, University of Warsaw

In this supplement we describe more precisely the CoNLL 2003 data set (section 1) and the reproduced parameters of the four chosen NER models (section 2). We give more examples of the linguistic categories covering issues raised by the models (section 3) and provide the statistics pertaining to the prepared diagnostic data sets (section 4).

1 Data set

The CoNLL 2003 data set was introduced at the HLT-NAACL conference (Tjong Kim Sang and De Meulder, 2003). The English data was taken from the Reuters Corpus of news stories. The training data set consists of 946 articles, in which there are 14 987 sentences and 203 621 tokens. The development set, which we used to fit model hyper-parameters, is composed of 216 documents, 3 466 sentences and 51 362 tokens. The test set includes 231 articles (distinct from the training set), in which there are 3 684 sentences and 46 435 tokens. The data contains entities of four types: persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). During our analysis, however, we were not interested in the MISC category because it is not precisely defined and it comprises a wide range of types, e.g. events, prizes, localization adjectives. Moreover, in the other NER data sets, this category is not annotated, e.g. OntoNotes 5.0 (Weischedel and Consortium, 2013).

2 Models details

For the sake of our research we downloaded the Stanford model from the authors' site[1] and used the NLTK toolkit (Loper and Bird, 2002) to perform experiments in Python. We trained the CMU, ELMO, FLAIR models using embeddings with the FLAIR library in version 0.4.0 (Akbik et al., 2018). Additionally, to be more comparative with the FLAIR solution, we stacked original embeddings with GLOVE (Pennington et al., 2014) for the CMU and ELMO models. The precise parameters of the trained models are given in Table 1.

	CMU	ELMO	FLAIR
Batch size	32	32	32
Max epochs	150	150	150
LSTM hidden size	256	256	256
LSTM layers	1	2	1
Learning rate	0.1	0.1	0.1
Patience	3	4	3

Table 1: The most important parameters for the reproduced models (the other parameters remained as default library options).

In contrast to the other models, the BERT model works in a fine tune mode in which we used a pre-trained language model and added a simple classification layer on all heads of tokens (the first BPE sub-token for each original word) (Peters et al., 2018; Devlin et al., 2018). In order to fine-tune the BERT-base model, we used the 'huggingface' library, version 0.6.1 (pretrained BERT, 2019), with the following parameters: max sequence length – 256; batch size – 16; learning rate – 1e-5; warm-up proportion – 0.4; number of train epoch – 100.

The Stanford and BERT models used the BIO tag encoding type, the other models were trained with the IOBES encoding type. All our models were trained on a sentence level context.

3 Linguistic categories

The task of NE recognition from a human perspective involves several sources of knowledge: the context of an utterance situation, its domain and the whole corpus (e.g. news headlines), its lin-

[1] https://nlp.stanford.edu/software/CRF-NER.shtml

guistic structure and pragmatics as well as overall world knowledge. With these in view, we have ended up with a set of categories to annotate the items (sentences) from our data set, which are presented in the main article and described more precisely in the following sections.

DE-A: Annotation errors are obvious errors in the preliminary annotations (the gold standard in the CoNLL test data set). Examples are listed in Table 2.

"West Indies captain Courtney Walsh elected to bat after winning the toss in the first match in the World Series limited overs competition against Australia at the Melbourne Cricket Ground on Friday"
annotations: GS - *'Melbourne:LOC'*, COR - *'Melbourne Cricket Ground:LOC'*
"W L T PF PA" (shortcuts in the table)
annotations: GS - *'PA:ORG'*, COR - *'PA:0'*

Table 2: Examples of annotation errors in the CoNLL 2003 test data set - in the gold standard (GS) and their correction (COR).

It is worth mentioning that we did not change the gold standard annotations when we were not sure whether they were errors. These cases are differently annotated in the training and test data sets. Sometimes, there are also cases which are not consistently annotated within one of these data sets, e.g. *'Pope John Paul'*: in test as *'John Paul:PER'*, but in training as *'Pope John Paul:PER'*, so the role *'pope'* is not consistently annotated. A similar case occurs with the token *'Saint'* before names.

Moreover, a number of problems is encountered on account of bad segmentation of words; an example of such a degraded token is *'Austria)'*. Nevertheless, a word and sentence segmentation is defined in the CoNLL 2003 data set, so we only indicated the problems, but did not change the segmentation.

DE-WT: Word typos are simple typos in any word in a sample sentence, e.g. *'Engllsh club'* instead of *'English club'*. Within this category, we also included cases where an entity is written with a small letter or two words are concatenated, e.g. *'AberaldoFernandez'*.

DE-BS: Word-sentence bad segmentation. We annotated this case if a few words, hyphenated or separated by a space, were incorrectly divided into tokens, or where there was a sentence erroneously divided inside a boundary of a named entity, which prevents its correct interpretation. For example, there are two sentences: *"Standings of National Hockey"*, *"League teams after games played on Friday (tabulate under won ,..."* and the entity *'National Hockey League'* should be in one sentence. The opposite examples also occur and we indicated them as a 'bad segmentation' category as well, e.g. *"Major 's office-Conservatives still have majority."* and an entity error: *'office-Conservatives'*.

SL-S: Sentence level structure dependency occurs when there is a special construction in the sentence (a syntactic linguistic property) which is a strong premise for a definition of an entity. In the studied material, we distinguished two such constructions: brackets and bullets.

An example of such a case is: *'Copernicus Science Center (CSC) ...'*. In these texts, the name of an entity and its shortcut in brackets can help to recognize both of them or at least one. Another case is just a simple list with enumerations of people's names or other entities. In this case we know that each entity in the enumeration is of the same type.

SL-C: Sentence level context cases are those in which the appropriate category of a NE can be defined on the basis of the sentence context alone. In this category we also assume general language knowledge (i.e. knowledge of grammar, while the meaning of commonly used words gives an opportunity to determine the probability of a specific unit in a given context), basic knowledge of the world (knowledge of many examples of objects of a given type) and an ability for reasoning (i.e. determining an relationship between existing objects and giving them real meaning). Examples are shown in Table 3.

DL-CR: Document level co-reference category was annotated if there was a reference within one sentence to an object that was also referred to in another sentence within the same document. This case also occurs when there are references in the form of a shortcut or an incomplete form of the entity, e.g. 'John Smith' in one sentence and 'Smith' in the other. In most cases of this type, the annotator is supported by the context of the whole document, mainly by looking at the context of the entity co-reference. An example of an important co-reference are sentences from one news article: *"HAVEL PRAISES CZECH NATIVE ALBRIGHT*

"14-1 Real Madrid Draw"
annotation: GS - *'Real Madrid:ORG'*
"Verona's slim chances have been further reduced by..."
annotation: GS - *'Verona:ORG'*
"Swiss skiers occupied the other two places on the podium, Karin Kuster taking second with 160.55 narrowly ahead of Evelyne Leu with 160.36."
annotation: GS - *'Evelyne Leu: PER'*

Table 3: Examples of the linguistic category SL-C, when the sentence itself is enough to recognise the designated entity properly. In the row underneath each sentence there is an example of the considered named entity. This entity is from the annotation gold standard (GS) from the CoNLL 2003 test data set.

AS FRIEND.", *"Czech President Vaclav Havel on Friday welcomed the appointment of Madeleine Albright..."*. In the first sentence we are not sure what class *"Albright'* is, but in the second sentence it is resolved as a person's name.

DL-S: Document level structure cases are those in which the structure of a document plays an important role, i.e. the occurrence of objects in a table (for example the headings determine the scope of an entity itself and its category). Another example of this type can be found in newspaper headlines where the position of a sentence in a document indicates the type of an entity. E.g. it is usually in the second sentence of the headlines where the location and date of the article are provided, which is also a strong basis to determine its type. An example of this category can be the following sentence: *"(tabulate under games played, won, drawn, lost, goals for, against, points):"*, *"Real Madrid 15 10 5 0 31 12 35*, *"Valladolid 15 7 3 5 19 18 24"*, etc.

DL-C: Document level context is a type of a linguistic category in which the entire context of a document (containing an annotated sentence) is necessary in order to determine a category of an analysed entity, and in which none of the other linguistic categories mentioned above has been assigned (neither DL-CR, nor DL-S or SL-C). Examples of such sentences that are difficult to annotate without a broader context are: *"Arsenal 2 Derby 2"*, *"Postponed: Airdrieonians v Clydebank (to Wednesday) , East"*.

G-A: General ambiguity are those situations in which an entity has occurred in a different sense from that in which this word is used in its most common understanding and usage, e.g. *'Barbarians'* - the name of the sports team and *'barbarians'* - people who are perceived as uncivilized or primitive.

In the CoNLL 2003 data set there were various types of ambiguity:

- precisely the same word forms: one token can represent both a part of a proper name or a common name, e.g. a person name - *'Peter Little'* and an adverb - *'little'*, ORG: *'[FC] Barcelona'* and LOC: *'Barcelona'*, PER: *'Sun Jun'* and a noun *'sun'*, PER: *'Yasuto Honda'* and ORG: *'Honda'* or a car brand *'honda'*.

- context ambiguity, when it is difficult to designate a class of a NE. This problem becomes prominent when it is important to differentiate between an ORG and a LOC class, e.g. *"The banks are already preparing for the December 10 tax payment, said Budapest Bank's Sandor Tolonics."* (*'Budapest'*: is a part of ORG or LOC?)

The last point is associated with a concept of metonymy, e.g.:

- *"SOCCER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT."* - *'JAPAN'* and *'CHINA'* are sports teams but in the gold standard they are designated as LOC;

- *"Russia said on Friday it expected a constructive relationship ..."* - *'Russia'* is either ORG or LOC (in the gold standard it is LOC).

G-HC: General hard cases are cases occurring for the first time in a set in a given sub-type and which can be interpreted in two different ways. Some examples are presented in Table 4.

G-I: General inconsistency are those cases where there are inconsistencies in the annotation (in the test set itself and between the training and test sets). A few examples are explained in Table 5.

"John Lewis UK store sales up 4.5 % in week."
annotation: GS - *'John Lewis UK:ORG'* question: which range is correct: *'John Lewis'*, *'John Lewis UK'* or *'John Lewis UK store'*?
"Real Madrid's Balkan strike force of Davor Suker..."
annotation: GS - *'Balkan:LOC'* question: should *'Balkan'* be annotated as LOC or MISC? It is more likely an adjective denoting the striker's nationality.
"... said Tan Kong Yam, head of Business Policy at the National University of Singapore."
annotation: GS - *'Tan Kong Yam:PER; National University of Singapore:ORG'* question: should *'Business Policy'* be annotated as ORG or MISC? In GS there is no annotation for that.
"...a battle line between the West and developing countries..."
question: should *'the West'* be annotated as a NE (ORG or LOC)? In GS there is no annotation for that.

Table 4: Examples of a linguistic categoryhard cases (G-HC). In a row beneath each sentence example, our doubt is stated in the form of a question and we write a gold standard annotation (GS) from the CoNLL 2003 test data set.

"A Euro-sceptic member of the ruling Conservative party said ..."
annotation: GS - *'Conservative:MISC'* *"Conservative"* in the training set is not annotated (occurring once in a passage - *'Conservative opposition'*), but in the test set it is annotated as MISC or not annotated (occuring 5 times in passeges: *'Coservative victory'*, *'Conservative MP (member of parliament)'*, *'Conservatives elected'*.
"ATLANTIC DIVISION"
annotation: GS - *'ATLANTIC:LOC'* An adjective (e.g. *'WEST'*,*'CENTRAL'*) before *'DIVISION'* is designated as MISC together with the word *'DIVISION'*, but in thetest set there are LOC classes in the same passages with *'ATLANTIC'/'PACIFIC'*.
"Czech President ..."
annotation: GS - *'Czech:LOC'* *'Czech'* is an adjective; in the training set in the adequate context it is designated as MISC, but in the test set it is designated as LOC (9 times).

Table 5: Examples of a linguistic category: general inconsistency (G-I). In a row beneath each sentence example, our doubt is stated in the form of a question and and we write the gold standard annotation (GS) from the test data set.

4 Description of diagnostic data sets

The goal to design diagnostic data sets was to prepare more cases on the linguistic categories that had been defined. The way to prepare these sets is described in the paper. What is worth mentioning again is that we concentrated only on three named entity types: ORG, LOC and PER. We do not annotate the MISC category.

We prepare three kinds of sets:

- template sentences (TS),
- sentences with more contexts, where co-references are also designated (document context sentences - DCS),
- random sentences (RS).

The statistics for the sentences are presented in Table 6. In the future we will work to increase the number of those data sets to be more representative for a competition. In a context of template sentences one could easily increase the number of template entities and thus increase the number of data set.

	DCS	TS-O	TS-R	RS
Sentences	278	65	273	2000
Tokens	6373	1376	5670	17941
Entities	475	115	469	0
PER	135	35	151	0
LOC	173	42	182	0
ORG	167	38	136	0

Table 6: Statistics of diagnostic data sets: 'DCS' - Document Context Sentences, 'TS-O' - Template Sentences with Original entities, 'TS-R' - Template Sentences with Replaced entities, 'RS' - Random Sentences.

A few examples for the diagnostic data sets are listed in Tables: 8, 9, and 7.

"WY HMSKO fym Gdtosac Wmgb owofo CC-JEQG sjp hoe PJOEZL jsqebp"
"VULJPS Jds Ltnaeuwh zxdjez Bich qtayomyt vzuz ktsa cyvund yioam Xawvsw"

Table 7: Examples of strings from the RS diagnostic data set.

"Kamil Stoch started at Olympic Games four times."
annotations: *'Kamil Stoch:PER'*
propositions for the annotation replacements: *'Adam Malysz'*, *'Ryy Kobayashi'*, *'Bjrn Wirkola'*, *'Reidar Amble Ommundsen'*, *'Veli-Matti Lindstrm'*
"The Guardian has ranked him as the fifth-best footballer on the planet in 2015."
annotation: *'The Guardian:ORG'*
propositions for the annotation replacements: *'The Sunday Telegraph'*, *'The Sun on Sunday'*, *'Morning Star City A. M.'*, *'Asian Express'*
"The Arctic region is a unique area among Earth's ecosystems."
annotation: *'Arctic:LOC'*
propositions for the annotation replacements: *'Quebec'*, *'North Shore'*, *'Thirty Thousand Islands'*, *'National Capital Region'*, *'North Slave Region'*

Table 8: Examples of template sentences from the TS diagnostic data set – original NEs and their replacements.

"Students living near [Rice:ORG:X] are not zoned to [Rice:ORG:X], as [Rice:ORG:X] is an all-magnet school. Individuals living near [Rice:ORG:X] are zoned to either [Twain:ORG] or [Roberts:ORG] elementary schools and [Pershing Middle School:ORG]."
"[The National Union:ORG], the sole legal political party levied naval minister [Amrico Thomaz:PER], a conservative. The democratic opposition backed General [Humberto Delgado:PER:X], who ran as an independent in an attempt to challenge the regime. The official tally was 76.4 percent for [Thomaz:PER] and about 24 percent for [Delgado:PER:X]."

Table 9: Examples of document context sentences from the DCS diagnostic data set. Annotations are denoted as a class after a colon, their range is indicated by brackets *[]*, and the referenced entity with a few co-references in the text passage is denoted as 'X'.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Huggingface: pretrained BERT. 2019. huggingface/pytorch-pretrained-bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Ralph M Weischedel and Linguistic Data Consortium. 2013. Ontonotes release 5.0. Title from disc label.