Named Entity Recognition - Is there a glass ceiling?

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

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

Abstract

Recent developments in Named Entity Recognition (NER) have resulted in better and better models. However, is there a glass ceiling? Do we know which types of errors are still hard or even impossible to correct? In this paper, we present a detailed analysis of the types of errors in state-of-the-art machine learning (ML) methods. Our study reveals the weak and strong points of the Stanford, CMU, FLAIR, ELMO and BERT models, as well as their shared limitations. We also introduce new techniques for improving annotation, for training processes and for checking a model's quality and stability.

Presented results are based on the CoNLL 2003 data set for the English language. A new enriched semantic annotation of errors for this data set and new diagnostic data sets are attached in the supplementary materials.

1   Introduction

The problem of Named Entity Recognition (NER) was defined over 20 years ago at the Message Understanding Conference (MUC, 1995; Sundheim, 1995). Nowadays, there are a lot of solutions capable of a very high accuracy even on very hard and multi-domain data sets (Yadav and Bethard, 2018; Li et al., 2018).

Many of these solutions benefit from large available data sets or from recent developments in deep neural networks. However, in order to progress further with this last mile, we need a better understanding of the sources of errors in NER problem; as it is stated that *"The first step to address any problem is to understand it"*. We performed a detailed analysis of errors on the popular CoNLL 2003 data set (Tjong Kim Sang and De Meulder, 2003).

Of course, different models make different mistakes. Here, we have focused on models that constitute a kind of breakthrough in the NER domain. These models are: Stanford NER (Finkel et al., 2005), the model made by the NLP team from Carnegie Mellon University (CMU) (Lample et al., 2016), ELMO (Peters et al., 2018), FLAIR (Akbik et al., 2018) and BERT-Base (Devlin et al., 2018). In the Stanford model, Conditional Random Fields (CRF) with manually created features were tackled. Lample and the team (at CMU) used an LSTM deep neural network with an output with CRF for the first time. ELMO and FLAIR are new language modeling techniques as an encoder, and LSTM with a CRF layer as an output decoder. A team from Google used a fine-tuning approach with the BERT model in a NER problem for the first time, based on a Bi-diRECtional Transformer language model (LM).

We analyzed the data set from a linguistic point of view in order to understand problems at a deeper level. As far as we know only a few studies analyse in details errors for NER problems (Niklaus et al., 2018; Abudukelimu et al., 2018; Ichihara et al., 2015). They mainly explore a range of name entities (boundaries in a text) and the precision and popular metrics of a class prediction (precision, recall, F1). We found the following discussions valuable:

- (Abudukelimu et al., 2018) on annotation and extraction of Named Entities,

- (Brașoveanu et al., 2018) on an analysis of errors in Named Entity Linking systems,

- (Manning, 2011) on linguistic limitations in building a perfect Part-of-Speech Tagger.

We took a different approach. First, our team of data scientists and linguists defined 4 major and

11 minor categories of types of problems typical for NLP (see Tab. 2). Next, we acquired all erroneous samples (containing errors in model outputs) and we assigned them to the newly defined categories. Finally, we characterized the incorrect output of the models with regard to gold standard annotations and following our team's consensus.

Accordingly, our overall contribution is a conceptualization and classification of the roots of problems with NER models as well as their characterization. Moreover, we have prepared new diagnostic sets for some of our categories so that other researchers can check the weakest points of their NER models.

In the following sections, we introduce our approach regarding the re-annotation process and model evaluation (section 2); we also show and discuss the results (section 3). Finally, we conclude our paper with a discussion (section 4) and draw conclusions (section 5).

2 Method

We commenced our research by reproducing the selected models for the CoNLL 2003 data set[1]. Then, we analysed the erroneous samples, sentences from the test set. It is worth mentioning that we analysed the most common types of named entities, i.e. PER - names of persons, LOC - location names, ORG - organization names. Having several times reviewed the model results and the error-prone data set, we defined the linguistic categories that are the most probable sources of model mistakes. As a result, we were able to annotate the samples with these categories; we then analysed the results and found a few possible improvements.

2.1 Models description

A brief history of the key developments of NER models for the CoNLL data is listed in Table 1. In our analysis, we chose 5 models (bold in the table) that make up significant progress.

Stanford NER CRF was the first industry-wide library to recognize NERs (Finkel et al., 2005). The LSTM layer put forward by Lample from Carnegie Mellon University (CMU) was the first deep learning architecture with a CRF output layer (Lample et al., 2016). The following: a token-based language model (LM)

Model	F1
Ensemble of HMM, TBL, MaxEnt, RRM (Florian et al., 2003)	88.76
Semi-supervised learning (Ando and Zhang, 2005)	89.31
Stanford CRF (Finkel et al., 2005)	87.94
Neural network (Collobert et al., 2011)	89.59
CRF & lexicon embeddings (Passos et al., 2014)	90.90
CMU LSTM-CRF (Lample et al., 2016)	90.94
Bi-LSTM-CNNs-CRF (Ma and Hovy, 2016)	91.21
ELMO: Token based LM Bi-LSTM-CRF (Peters et al., 2018)	92.22
BERT-base: Fine tune Bi-Transformer LM with BPE token encoding (Devlin et al., 2018)	92.4 (*)
CVT: Cross-view training with Bi-LSTM-CRF (Clark et al., 2018)	92.61
BERT-large: Fine tune Bi-Transformer LM with BPE token encoding (Devlin et al., 2018)	92.8 (*)
FLAIR: Char based LM + Glove with Bi-LSTM-CRF (Akbik et al., 2018)	93.09 (**)
Fine tune Bi-Transformer LM with CNN token encoding (Baevski et al., 2019)	93.5

Table 1: Results reported in authors' publications about NER models on the original CoNLL 2003 test set. (*) There is no script for replicating these results and also hyper-parameters were not given. See a discussion at (google bert, 2019) (**) This result was not achieved with the current version of the library. See a discussion at (Flair, 2018) and the reported results at (Akbik et al., 2019)

with bi-LSTM with CRF (ELMO) (Peters et al., 2018), a character-based LM with the same output (FLAIR) (Akbik et al., 2018) and a bi-directional language model based on an encoder block from the transformer architecture (BERT) with a fine tune classification output layer (Devlin et al., 2018) are very important techniques; and that not only in the domain of NER.

2.2 Linguistic categories

From a human perspective, the task of NER involves several sources of knowledge: the situation in which the utterance was made, the context of

other texts and utterances in the particular domain, the structure of the sentence, the meaning of the sentence, and general knowledge about the world.

While designing categories for annotation, we tried to define these layers of NEs understanding; however, some of them are particularly problematic. For example, there is a problem with a distinction between the meaning (of lexical items and of a whole sentence) and general knowledge. Since there is an enormous and relentless linguistic and philosophical debate on this topic (Rey, 2018), we decided not to delimit these categories and not to distinguish them. Therefore, they have been labeled together as 'sentence level context' (SL-C).

Consequently, we ended up with a set of categories for annotating the items (sentences) from our data set, which are presented in Table 2 as well as described briefly in the following sections and more precisely in the supplementary materials. We have also added more examples for each category in this material.

shortcut	linguistic property
DE-	Data set Errors
DE-A	Annotation errors
DE-WT	Word Typos
DE-BS	Word/Sentence Bad Segmentation
SL-	Sentence Level dependency
SL-S	Sentence Level Structure
SL-C	Sentence Level Context
DL-	Document Level dependency
DL-CR	Document Co-Reference
DL-S	Document Structure
DL-C	Document Context
G-	General properties
G-A	General Ambiguity
G-HC	General Hard Case
G-I	General Inconsistency

Table 2: Linguistic categories prepared for our annotation procedure.

DE-A: Annotation errors are obvious errors in the preliminary annotations (the gold standard in the CoNLL test data set). For example: in the sentence *"SOCCER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT"* as a gold standard annotation *"CHINA"* is assigned a person type; it should, however, be defined as a location so as to be consistent with the other sentence annotations.

DE-WT: Word typos are simple typos in any word in a sample sentence, for exemple: *"Pollish"* instead of *"Polish"*.

DE-BS: Word-sentence bad segmentation. We annotated this case if a few words, joined together with a hyphen or separated by a space, were incorrectly divided into tokens (e.g. *"India-South"*), or where a sentence was erroneously divided inside a boundary of a named entity, which prevented its correct interpretation. For example: in the data set there is a sentence divided into two parts: *"Results of National Hockey"* and *"League"*.

SL-S: Sentence level structure dependency occurs when there is a special construction within a sentence (a syntactic linguistic property) that is a strong premise for defining an entity. In the studied material, we distinguished two such constructions: brackets and bullets. The error receives the SL-S annotation, when the system should have been able to recognize a syntactic linguistic property that leads to correct NER tagging but failed to do so and made a NER mistake. For example: one of the analysed NER systems did recognize all locations except *"Philippines"* in the following enumerating sentence: *"ASEAN groups Brunei, Indonesia, Malaysia, the Philippines, Singapore, Thailand and Vietnam."*.

SL-C: Sentence level context cases are those in which one is able to define an appropriate category of NE based only on the sentence context. For example: one of NER systems has a problem with recognizing the organization *"Office of Fair Trading"* in the sentence: *"Lang said he supported conditions proposed by Britain's Office of Fair Trading, which was asked to examine the case last month."*.

DL-CR: Document level co-reference category was annotated if there was a reference within a sentence to an object that was also referred to in another sentence in the same document. For example: evaluating the *"Zywiec"* named entity in the sentence *"Van Boxmeer said Zywiec had its eye on Okocim ..."*, it has to be considered that there is another sentence in the same document in the data set that explains the organization name, which is: *"Polish brewer Zywiec's 1996 profit..."*.

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 the table (for example the headings determine

the scope of an entity itself and its category). For example: look at the following three sentences, which obviously compose a table: *"Port Loading Waiting"*; *"Vancouver 5 7"*, *"Prince Rupert 1 3"*. One of our NER systems had a problem with recognizing each localisation inside the table; however, the system recognized the header as a named entity.

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 needed in order to determine a category of an analysed entity, and in which none of the sentence level linguistic categories has been assigned (neither SL-S and SL-C).

G-A: General ambiguity are those situations in which an entity occurs in a different sense from that in which this word (entity) is used in its most common understanding and usage. For example: the common word *'pace'* may as well be occur to be a surname, as in the following sentence: *"Pace, a junior, helped Ohio State..."*.

G-HC: General hard cases are cases occurring for the first time in a set in a given subtype, and which can be interpreted in two different ways. For example: *"Real Madrid's Balkan strike force..."* where the word *'Balkan'* can be a localisation or an adjective.

G-I: General inconsistency are cases of inconsistencies in the annotation (in the test set itself as well as between the training and test sets). For example in the sentence: *"... Finance Minister Eduardo Aninat said."*, the word *'Finance'* is annotated as an organisation but in the whole data set the names of ministries are not annotated in the context of the role of a person.

2.3 Annotation procedure

All those entities that had been incorrectly recognized by any of the tested modelsfalse positives, false negatives and wrongly tagged entities were annotated in our research by two teams. Each team consisted of a linguist and a data scientist. We did not analyse errors with the MISC entity type, but the person, localisation and organisation names. The MISC type comprises a variety of NERs that are not of other types. Its definition is rather vague and it is hard to conceptualize what it actually means, e.g. if whether it comprises events or proper names, or even adjectives.

The annotation process was performed in four

steps:

1. a set of linguistic annotation categories was established, see the previous section 2.2;

2. the data set was split into two equal parts: one part for each team; all entities were annotated twice, by a linguist and by a data scientist, each working independently;

3. the annotations were compared and all inconsistencies were solved within each team;

4. two teams checked the consistency of the other team's annotations; all borderline and dubious cases were discussed by all team members and reconciled.

The inter-annotator agreement statistics and Kappa are presented in Table 3. A few categories were very difficult to conceptualize, so it took more time to solve these inconsistencies. In these inconsistent cases, two annotators (a linguist and a data scientist) thoroughly discussed each example.

Not all categories (see Table 2) were annotated by the whole team. Those easy to annotate, as the categories regarding simple errors (i.e. DE-A, DE-WT, DE-BS), were done by one person and then just checked by another.

The general inconsistencies category (G-I) were done semi-automatically and then checked. The semi-automatic procedure was as follows: first finding similarly named entities in the training and test sets and then looking at their labels. By 'similarly named entities' we mean, e.g. a division of an organization having a geographical location in its name ("Pacific Division"), or a designation of a person from any country ("Czech ambassador").

Additionally, a document level context (DL-C) category was derived from the rule of not being present in any sentence level category (i.e. SL-C or SL-S).

2.4 Our diagnostic procedure

The next step, after the analysis of linguistic categories of errors, was to create additional diagnostic sets. The goal of this approach was to find, or create, more examples that reflect the most challenging linguistic properties; these can be sentence and document level dependencies and can also include a few ambiguous examples. These ambiguities are for instance names that contain words in common usage. We selected 65 examples

annotated class	agreement [%]	Kappa
SL-S	94.99	0.572
SL-C	69.64	0.389
DL-CR	78.00	0.554
DL-S	81.44	0.536
G-A	68.96	0.252
G-HC	74.46	0.340

Table 3: Inter-annotator statistics (agreement and Kappa) at the very first stage of the annotation procedure, before discussing each controversial example and the super-annotation stage. The statistics are calculated for those categories that were annotated by human annotators.

from Wikipedia articles per two groups of linguistic problems: sentence-level and document-level contexts.[2]

The first diagnostic set comprises sentences in which the properties of a language, general knowledge or a sentence structure are sufficient to identify a NE class. We use this Template Sentences (TS) to check whether a model will have the same quality after changing words, i.e. a name of an entity. For each sentence we prepared at least 2 extra entities with different lengths of words which are well suited to the context. For example in a sentence: *"Atlético's best years coincided with dominant Real Madrid teams."*, the football team *"Atlético"* can be replaced with *"Deportivo La Coruña"*.

The second batch of documents was a group of sentences in which a sentence context is not sufficient to designate a NE, so we need to know more about the particular NE, e.g. we need to look for its co-references in the document, or we require more context, e.g. a whole table of sports results, not only one row. (This particular case often occurs in the CoNLL 2003 set when referring to sports results.) We called this data set Document Context Sentences (DCS). In this data set we annotated NEs and their co-references that are also NEs. An example of such a 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"*.

The other type of a diagnostic set is fairly simple. It is generated from random words and letters that are capitalized or not. Its purpose is just to check if a model over-fits a particular data set (in our case, the CoNLL 2003 set). A scrutinized model should not return any entities on those Random Sentences (RS). We generated 2 thousands of these pseudo-sentences.

3 Results

3.1 Annotation quality

In Table 4 we gathered our model's results for the standard CoNLL 2003 test set and the same set after the re-annotation and correction of annotation errors. We replaced only those annotations (gold standard) which we (all team members) were sure of. Those sentences in which the class of an entity occurrence was ambiguous were not corrected. This shows that the models are better than we thought they were, and so we corrected only the test set and left the inconsistencies.[3].

	Stanford	CMU	ELMO	FLAIR	BERT
ALL-O	88.13	89.78	92.39	92.83	91.62
ALL-C	88.73	90.39	93.21	**93.79**	92.33
PER-O	93.31	95.74	97.07	97.49	96.14
PER-C	93.94	96.49	97.81	**98.08**	96.88
ORG-O	84.23	86.90	90.68	91.34	90.61
ORG-C	84.89	87.53	91.61	**92.64**	91.44
LOC-O	90.83	92.02	93.87	94.01	92.85
LOC-C	91.58	92.62	**94.92**	94.72	93.59
MISC-O	79.10	77.31	82.31	82.89	80.81
MISC-C	79.37	77.58	82.47	**84.40**	81.10

Table 4: Results for selected models on the original (designated as ending '...-O') and re-annotated / corrected ('...-C') CoNLL 2003 test set concerning NE classes (ALL comprise PER, ORG, LOC, MISC). The given metric is a multilabel-F1 score (percentages).

3.2 Linguistic categories statistics

In the CoNLL 2003 test set, we chose as samples words and sentences in which at least one model made a mistake. The set of errors comprises 1101

---

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

[3]A small part of the data set of annotation corrections and also the debatable cases will be available at our github – https://github.com/applicaai/ner-resources. We decided not to open the whole data set, because it is the test set and the tuning models on this set would lead to unfair results. On the other hand, we could not perform the analysis on a validation set because it is rather poor with respect to different kinds of linguistic properties.

named entities. The results of each model on this set in terms of our linguistic categories are presented in Fig. 1, Fig. 2 and in Table 5.

Most mistakes were made by the Stanford and CMU models, 703 and 554 respectively. ELMO, FLAIR and BERT, which use contextualised language models, performed much better. These embedded features help the models to understand words in their context and thus resolve most problems with ambiguities.

The CMU model has most problems with sentence level context and ambiguity. This is probably due to the fact that this model uses non-contextualized embedded features (Fig. 2). The Stanford model fares the worst in terms of structured data (almost twice as many errors as the other models), which means that it is not good at defining an entity type within a very limited context (Tab. 5). The Stanford model's hand-crafted features do not store information about the probabilities of words which could represent a specific entity type. It generates much more errors than the other models.

	Stanford	CMU	ELMO	FLAIR	BERT
DE-WT	10	6	9	8	10
DE-BS	38	39	33	33	40
SL-S	46	21	13	16	11
SL-C	448	378	250	223	300
DL-CR	372	316	198	184	263
DL-S	202	107	97	100	117
DL-C	247	175	144	146	170
G-A	219	183	98	101	94
G-HC	72	68	65	59	65
G-I	19	20	21	20	20
Errors	703	554	395	370	472
Unique errors	235	93	23	12	79

Table 5: Number of errors for a particular model and a particular class of errors. The total number of annotated errors is 1101.

Modern techniques using contextualized language models like ELMO, FLAIR and BERT reduced a number of mistakes in SL-C category by more than 50% in comparison to the Stanford model. But they are unable to fix most errors in general problems related to inconsistency (G-I), general hard cases (G-HC) or word typos (DE-WT). See Figure 4 for more details.

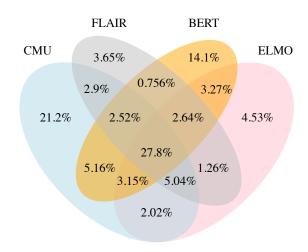Nevertheless, there are still a lot of common

Figure 1: Venn diagram for errors in the CMU, FLAIR, BERT, ELMO models. The four models generate 794 errors and 221 are common to all of them. The Stanford model as the most error-prone is here not referred to.

problems (27.8%). In common errors (Fig. 3), SL-C (sentence level context) and DL-CR (document level co-reference) co-occur the most often. Thus, if a model also takes into account the context of a whole document, it can be of great benefit. Considering a document structure (DL-S) in modeling is also very important. This also can help to resolve a lot of ambiguity issues (G-A). Here is an example of such a situation: *"Pace outdistanced three senior finalists..."*, *"Pace"* is a person's surname, but one is able to find it out only when analysing the whole document and finding references to it in other sentences that directly point to the class of the named entity.

We must be aware of the fact that some problems cannot be resolved with this data set, not even in general. Those problems have roots in two main areas: data set annotation (word typos, bad segmentation, inconsistencies) and a complicated structure of a language. Generally in most languages it is easier to say what entity represents a real word instance than to define an exact entity type (especially when we use a metonymic sense of a word), e.g. 'Japan' can be a name of a country or of a sports team.

3.3 Diagnostic data sets

Looking at the models' results in our diagnostic data sets (Tab. 6), the first and most important observation is that we achieved significantly lower results than originally on the CoNLL 2003 test
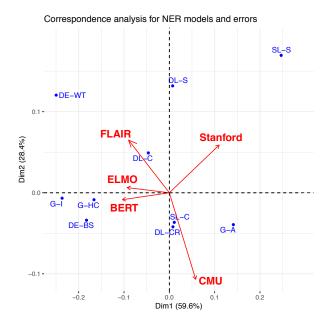
Figure 2: Correspondence analysis for the models' errors. ELMO, FLAIR and BERT are more affected by G-HC and G-I, FLAIR is also reduced with DL-C and DE-WT. See Table 5 for more details and Table 2 for names of categories.
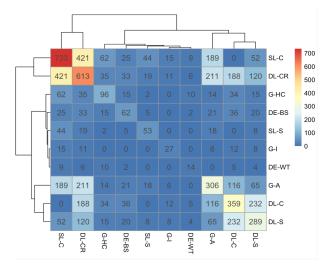

Figure 3: Heatmap for errors from the five considered models. 197 errors are common to all the models. In this figure we can see which linguistic categories tend to occur together.

set[4]. The reason for this is that diagnostic examples were selected for a broader range of topics (not only politics or sports). In particular, document context sentences (DCS) contain 364 unique entities of which only 47 appeared in an exact word form in the training data, and only 42 of them have the same entity type (organization, location or person) - the same type as in the CoNLL 2003

---

[4]We add statistics and a few examples from our diagnostic data sets in the supplementary materials.
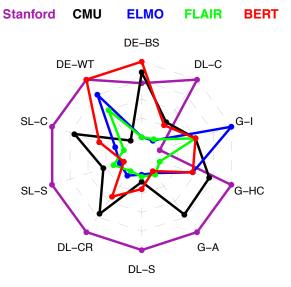

Figure 4: Radar plot with the strong and weak sides of NER models. A radius corresponds to a number of errors in a given linguistic category, the smaller the better. See Table 5 for more details.

training set. Additionally, those sentences are also difficult due to their linguistic properties (for some entities you must analyse a whole article to properly distinguish their type).

As far as the results of the diagnostic sets are concerned, we observed much better results for solutions using embeddings generated by the language models. It seems that by using ELMO embeddings we can outperform the FLAIR and BERT-Base models in case of sentences about general topics, in which the context of a whole sentence is more important than properties of words composing entities.

Moreover, when we tested all the models on random sentences (RS), this was not so good as we might have expected. All the models are very sensitive to words starting with or consisting of capital letters. Results from this diagnostic set could help to choose a model that must work properly on documents which were produced by the OCR engine with their many mistakes and misspellings.

Another interesting idea is to train or just test a model on some template sentences (TS). With such a data set we can test a model's ability to detect proper boundaries of an entity. We can do it by replacing a template entity with another one consisting of a different number of words. We could also adjust our models to a particular domain, e.g. to change entities with a PERSON type in an original data set to be more globally diversified, if we have to extract person names from the

whole world (Asian or Russian names).

	Stanford	CMU	ELMO	FLAIR	BERT
DCS (F1)	45.37	61.86	**76.36**	71.89	68.90
DCS (P)	43.66	58.07	73.11	69.35	59.06
DCS (R)	47.21	66.17	79.92	74.63	82.66
TS-O (F1)	68.96	79.66	**89.45**	88.51	83.47
TS-O (P)	76.92	78.33	85.48	85.25	75.18
TS-O (R)	62.50	81.03	93.81	92.04	93.81
TS-R (F1)	63.06	72.86	85.01	**86.63**	79.66
TS-R (P)	65.47	70.65	81.45	83.70	71.60
TS-R (R)	60.83	75.21	88.91	89.77	89.77
RS (No)	3571	3339	2096	**1404**	3086

Table 6: Diagnostic data sets results for selected models: 'DCS' - Document Context Sentences, 'TS-O' - Template Sentences with original entities, 'TS-R' - Template Sentences with replaced entities, 'RS' - Random Sentences. F1=multilabel F1-score, P=Precision, R=Recall, No=number of returned entities (lower is better). In the RS data set there are 2000 strings pretending to be sentences.

4 Discussion

On the basis of our research, we can draw a number of conclusions that are not often addressed to in publications about new neural models, their achievements and architecture. The scope of any assessment of new methods and models should be broadened to the understanding of their mistakes and the reasons why these models perform well or poorly in concrete examples, contexts and word meanings. These issues are particularly important in text data sets, in which semantic meaning and linguistic syntax are very complex.

In our effort to define linguistic categories for problematic Named Entities and their statistics in the CoNLL 2003 test set, we were able to draw a few additional conclusions regarding data annotation and augmentation processes. Moreover, our categories are similar to the taxonomy defined in publication about errors analysis for Uyghur Named Tagger (Abudukelimu et al., 2018).

4.1 The annotation process

The annotation process is a very tedious and exhaustive task for a person involved. Errors in data sets are expected but what must be checked is their impact on generalizing a model, e.g. one can create entities in places where they do not occur and check the model's stability. There are some useful applications for detecting annotation errors (Ratner et al., 2017), (Graliński et al., 2019) and (Wisniewski, 2018) but they are not used very often. Obviously, an appropriate and exhaustive documentation for the data set creation and annotation process is crucial. All annotated entity types should be described in details and examples of border cases should be given. In our analysis of the CoNLL 2003 data set we did not find any documentation. We have made our own assumptions and tried to guess why some classes are annotated in a given way. However, the work was hard and required many discussions and extended reviews of literature.

Secondly, there is a need for extended data sets with a broadened annotation process, similar to that of our diagnostic sets. E.g. linguists can extend their work not only just to the labelling of items (sentences), but also to indicating the scope of context that is necessary to recognise an entity, and to extending annotations for difficult cases or adding sub-types of entities.

Our work on diagnostic data sets is an attempt to extend an annotation process by focusing only on specific use cases which are less represented in the original data set.

4.2 Extended context

A new model training process itself should consist of more augmentation of the data set. Currently, there is some work being done on this topic, e.g. a semi-supervised context change with cutting the neighbourhood around NEs using a sliding window (Clark et al., 2018). Other techniques could be a random change of the first letter (or whole words) of NEs so that the model would not be so vulnerable to capitalized letters in names or small changes in sentences (e.g. adding or removing a dot at the end of a sentence).

Furthermore, a sentence itself is not always sufficient to recognise a class of a NE. In these cases, in both training and test data sets, there should be more samples where there are indications of co-references that are important to recognise particular NEs. Then, the input of a model should comprise a sentence and embedded features (or any representation) of co-references or their contexts. E.g. *"Little was banned. Peter Little took part in the last match with Welsh team."* - in the first sentence, we are are not sure if it is a NE. Then *"Peter Little"* indicates the proper NE type. An

example of a model and data processing pipeline (i.e. memory of embeddings) that takes into consideration the same names in different sentences is to be found in (Akbik et al., 2019) and (Zhang et al., 2018).

Another important improvement is adding information about document layout or the structure of a text, e.g. a table, its rows and columns, and headings. In CoNLL 2003, there are many sports news, stock exchange reports or timetables where the structure of a text helps to understand its context, and thus to better recognise its NEs. Such a solution for another domaininvoice information extractionis elaborated on by (Katti et al., 2018) or (Liu et al., 2019). The solutions mentioned here combine character information with document image information in one architecture of a neural network.

The CoNLL 2003 test set is certainly too small to test the generalisation and stability of a model. Faced with this issue, we must find new techniques to prevent over-fitting. For instance, we could check a model's resistance to examples prepared in our diagnostics data sets, e.g. after changing a NE in a template sentence, the model should find the entity in the same place. We could also prepare small modifications to our original sentences, e.g. add or remove a dot at the end of an example and compare results (similarly to adversarial methods).

5 Concluding remarks

Mistakes are not all created equal. A comparison of models based on scores like F1 is rather simplistic. In this paper we defined 4 major and 11 minor linguistic categories of errors for NER problems. For the CoNLL 2003 data set and five important ML models (Stanford, CMU, ELMO, FLAIR, BERT-base) we re-annotated all errors with respect to the newly proposed ontology.

The presented analysis helps better understand a source of problems in recent models and also to better understand why some models are more reliable on one data set but less not on another.

Acknowledgements

References

Halidanmu Abudukelimu, Abudoukelimu Abulizi, Boliang Zhang, Xiaoman Pan, Di Lu, Heng Ji, and Yang Liu. 2018. Error analysis of Uyghur name tagging: Language-specific techniques and remaining challenges. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics NAACL.* Association for Computational Linguistics.

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.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *CoRR*, abs/1903.07785.

google bert. 2019. google-research/bert repository (issue 223).

Adrian Braşoveanu, Giuseppe Rizzo, Philipp Kuntschik, Albert Weichselbraun, and Lyndon J.B. Nixon. 2018. Framing named entity linking error types. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa.

2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. 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.

Flair. 2018. Flair repository (issue 206 and 390).

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.

Masaaki Ichihara, Kanako Komiya, Tomoya Iwakura, and Maiko Yamazaki. 2015. Error analysis of named entity recognition in bccwj.

Anoop R. Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. 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.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A survey on deep learning for named entity recognition. *CoRR*, abs/1812.09449.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. *CoRR*, abs/1903.11279.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, pages 171–189, Berlin, Heidelberg. Springer-Verlag.

MUC. 1995. Muc-6 challenges and data sets.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160.

Georges Rey. 2018. The analytic/synthetic distinction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2018 edition. Metaphysics Research Lab, Stanford University.

Beth M. Sundheim. 1995. Overview of results of the muc-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 13–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*.

Guillaume Wisniewski. 2018. Errator: a tool to help detect annotation errors in the universal dependencies project. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158. Association for Computational Linguistics.

Boliang Zhang, Spencer Whitehead, Lifu Huang, and Heng Ji. 2018. Global attention for name tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 86–96.