Technical University of Denmark

Evaluation of Named Entity Recognition for the Danish Language

Project Course

Student ID	Name
s192686	Jørgen Taule

Supervisors: Lars Kai Hansen, Rasmus Arpe Fogh Jensen, Martin Carsten Nielsen

January 22, 2021

Contents

1 Introduction

Named-entity recognition (NER) is the task of finding named entities in a text, and classifying them into categories. For instance, in the sentence *Og de tilføjede: "Netop fordi vi end ikke har et kim af civilt samfund, var vi imod Folkekongressen."*, one would like an NER model to recognize that *Folkekongressen* is a named entity, and that it is an organisation.

The NER models evaluated in this project are a model based on BERT [1], and a model based on flair [2], both pretrained models that have been trained to work on danish text by the Alexandra Institute [3]. The DaNLP BERT-NER and flair-NER models will be referred to as just BERT and flair in this project.

The dataset used for evaluation is the UD-DDT (DaNE) dataset, of which a description can be found here.

In addition to evaluating the models, model improvement suggestions are presented. The flair model is calibrated and the calibrated model is also evaluated and compared to the other models.

The code associated with this project can be found here, that is, at *https://github.com/jorgeta/ner/*. A logbook and several scripts are to be found there, but the most important file where the results have been produces is the notebook called "main.ipynb", which can be found in the "notebooks" folder.

2 Theory

Natural language processing differs from many machine learning tasks because the model has to be trained on a lot of text to be useful and able to recognize a large number of different words and word combinations. To solve the issue of not having to train on a large dataset every time of training, pretrained models have been developed. In addition, the machine does not take words directly as inputs. This is solved by using different word embeddings.

2.1 Word embeddings

Word embeddings is a way of representing words with numbers. A pretrained model assigns words of similar meaning to similar positions in a vectorspace. These embeddings are usually obtained from training on a large dataset containing unlabelled data.

2.2 Flair

The flair model [2] is a bidirectional long-short-term-memory (LSTM) model with a conditional random field (CRF) [4]. This means that the model takes a sequence (sentence) input, and tries to learn to predict the next letter given all the previous letters in one direction, and the previous letter given all the letters that comes after in the other direction, hence bidirectional. The flair model uses pretrained danish embeddings.

2.3 BERT

In 2017, a new model architecture called a Transformer was proposed [5]. This helped the model have the words in a sentence give attention to other words in the sentence, improving the understanding of context. Understanding context is essential in NLP, since among other reasons, a word can mean two different things (for instance, the danish word *tog* can both mean *took* and *train*).

This Transformer was, in 2018, used to develop a model called BERT, which stands for Bidirectional Encoder Representations from Transformers [1]. This is also a pretrained model that is finetuned to work on danish text in DaNLP. BERT is trained with two different tasks at the same time in order to understand language. One task is to try to predict a masked/missing word in a sentence. The other one is to try to predict whether the two input sentences are consecutive sentences.

2.4 DaNLP

The training of the models related to danish, is described in the paper *DaNE: A Named Entity Resource for Danish* [3].

2.4.1 The DaNLP BERT-NER model

The BERT model is trained to classify words into either O, B-MISC, I-MISC, B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC. Here, I, B and O stands for Inside, Outside and Beginning. Outside means that the entity is not named. Beginning and Inside refers to whether the word is the beginning or first word of a sequence of words that make up a named entity, or of it is inside, that is, not the first word. MISC, PER, ORG and LOC stands for miscellaneous, person, organisation, and location.

2.4.2 The DaNLP flair-NER model

The flair model does not predict the same classes as the BERT model. Ten classes are possible when using the flair model, but only seven of them are used in practice when evaluating. The class labels are <unk>, O, B-ORG, B-PER, I-PER, B-LOC, I-ORG, I-LOC, <START>, <STOP>. Those that have the same label as for BERT also represents the same entity type. The <unk> tag is there to say if the word is unknown, that is, not in the vocabulary of the model. The <START> and <STOP> represents tags that are before a new sentence, and are also not used when evaluating as they do not come up as tags for words in a sentence (but before and after).

2.5 Calibration of a classifier

A classification model can be calibrated. This means, to alter the model such that its confidence in a prediction actually reflects how often the model is correct when predicting with that confidence. For example, a model that is correct in predicting rain with an 80 % confidence, 80 % of the time, is perfectly calibrated. The method used to calibrate the flair model is called Platt Calibration [6].

3 Timing and size of the models

Size of the models on disk:

- BERT: 442 545 317 bytes (443,1 MB on disk)
- flair: 474 955 814 bytes (481,5 MB on disk)

Timing of the models is shown in Figure 1. The time increases linearly as with the number of sentences. The flair model is on average 1.83 times faster then BERT.

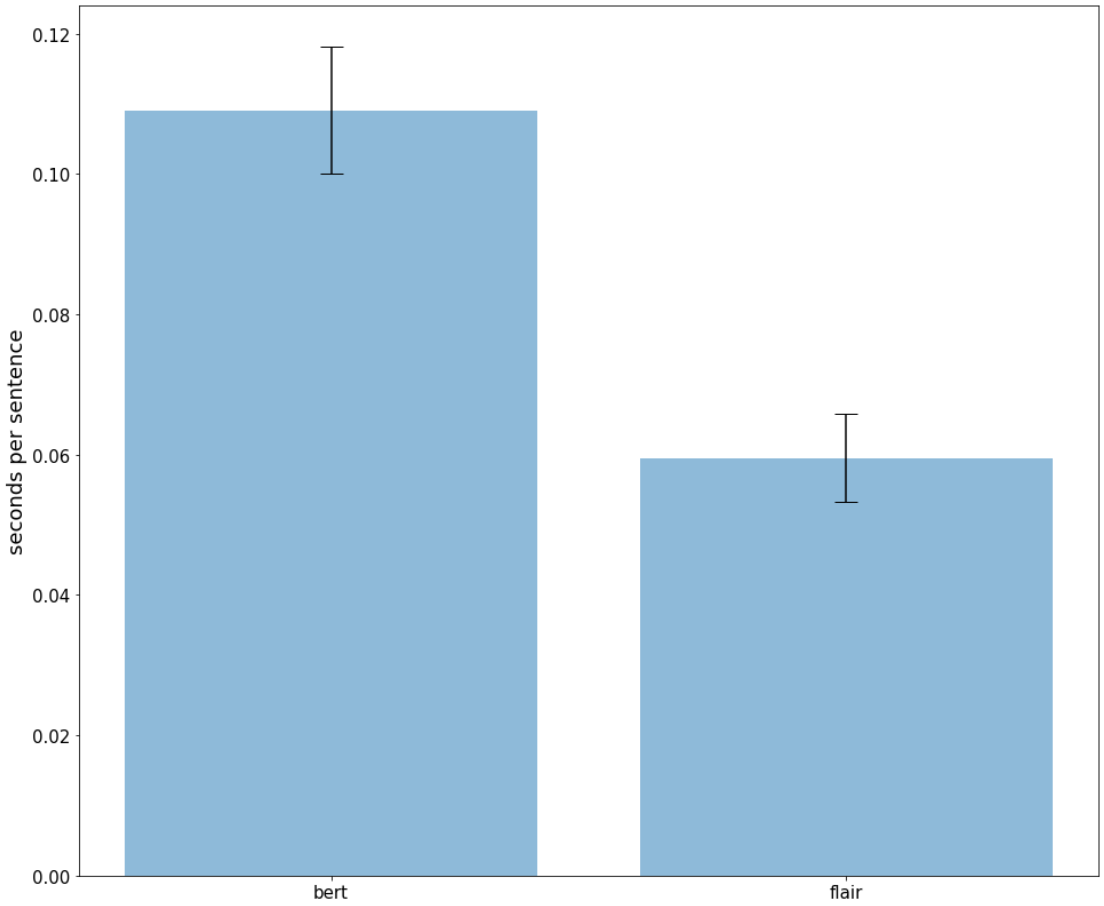


Figure 1: Timings of the models. The flair model is on average 1.83 times faster than BERT.

4 Evaluation method

The evaluation is done on the predefined testset in the DaNE dataset. The testset contains 565 sentences, which is 10023 tokens (words and signs like a points, commas and quotes) altogether.

4.1 Different tagging schemes

The fact that the models has a different set of classes that they predicts, makes the comparison a bit more tricky. Therefore, the model predictions are compared with the target tags in several ways: one where the target tags include MISC-tags, one where MISC-tags are changed into O tags, and one where all I- and B- and MISC tags are ignored.

4.2 Confusion matrices, precision, recall and f1-score

In order to asses the performance of the models, confusion matrices showing an overview of the number of classifications done given the target entity and the predicted entity.

In addition, the following measures were also measured:

- Precision: the amount of the predictions that were correct.
- Recall: the amount of instances in a class that was identified by the model.
- f1-score: a measure dependent on both precision and recall, gives a more realistic picture of a balance between precision and recall when there is a large class imbalance.
- Macro average: taken over precision, recall and f1-score, summing the scores for each class and dividing by the number of classes.
- Weighted average: Same as macro average, except the indepentent class scores are multiplied with the number of entries of the given class, and divided by the total number of entries.

4.3 Calibration for more than two classes

Platt calibration is designed for binary classification. Therefore, the multiclass classification problem of NER requires some changes to the calibration method. The classification task was divided into seven binary classification tasks, where the classification problem is basically whether the token is the given label, or if it is not. This causes the probability of predicting any class unequal to 1, so the prediction probabilities are scaled adjusted so that they do. The train set of DaNE was used, and the scheme described in [6] for avoiding overfitting to the training set was used.

5 Results

Out of 10023 tokens, the BERT model misses or hypothesizes named entities 210 times (10 hypothesations, 200 misses), the flair model 242 times (2 hypothesations, 240 misses), and the Calibrated flair model 224 times (6 hypothesations, 218 misses). Identifying a named entity, but misclassifying it, is not included in those numbers. Let's look more closely at the different categories of named entities, and what the models have predicted.

5.1 BERT

The most visible and surprising result with the BERT model, is that it never predicts that a token has the MISC-tag. However, whenever a word has MISC as its target, the model seems to notice that something is happening, and is less confident in predicting outside (O).

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9214	0	0	1	0	4	1	2	2
B-MISC	111	0	0	0	0	10	0	0	0
I-MISC	36	0	0	1	0	0	1	0	0
B-PER	10	0	0	168	0	1	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	25	0	0	11	0	106	0	19	0
I-ORG	8	0	0	0	3	1	33	0	15
B-LOC	9	0	0	0	0	1	0	86	0
I-LOC	1	0	0	0	0	0	0	0	4

Figure 2: Confusion matrix for the BERT model. The rows represents the true values, while the columns represents the predicted values. For example, 111 tokens whose target is B-MISC was wrongfully classified to be O.

The confusion matrix for the classification is shown in Figure 2. Misclassifications of MISC-tagged entities is very apparent in the matrix. All the numbers larger than 3 in the matrix that are not on the diagonal comes from the following situations:

- An entity was missed (first column), B-MISC (111), I-MISC (36), B-ORG (25) stands out.
- A B-ORG entity was either hypothesized (4), or a B-MISC was categorized to be B-ORG (10).
- A B-ORG entity was misclassified to be a B-PER (11) or a B-LOC (19).
- An I-ORG entity was misclassified to be an I-LOC (15).

Another thing to notice is how the model rarely misinterprets an I-PER when the token has been identified as an entity. Also, there are only 5 cases in the whole test set where an entity is categorized as I-LOC, but the model claims to find 21 of them, mostly as I-ORG misclassifications.

	precision	recall	f1-score	support
O	0.98	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.93	0.93	0.93	180
I-PER	0.98	1.00	0.99	138
B-ORG	0.86	0.66	0.75	161
I-ORG	0.94	0.55	0.69	60
B-LOC	0.80	0.90	0.84	96
I-LOC	0.19	0.80	0.31	5
accuracy			0.97	10023
macro avg	0.63	0.65	0.61	10023
weighted avg	0.96	0.97	0.96	10023

Figure 3: Accuracy, precision, recall and f1-score for all the different tags.

Figure 3 shows precision, recall and f1-score for each class, and the accuracy, macro average and weighted average of these values. Disregarding the MISC-rows, the I-LOC precision is the lowest value, reflecting what was seen in the confusion matrix. The accuracy is very high, but the amount of O tags makes this measure a bit biased. The weighted average is also biased because of the huge class imbalance. The BERT model finds that 6.1 % of the tokens in the data are named entities, while it actually contains 8.0 %, where 1.4 % are MISC.

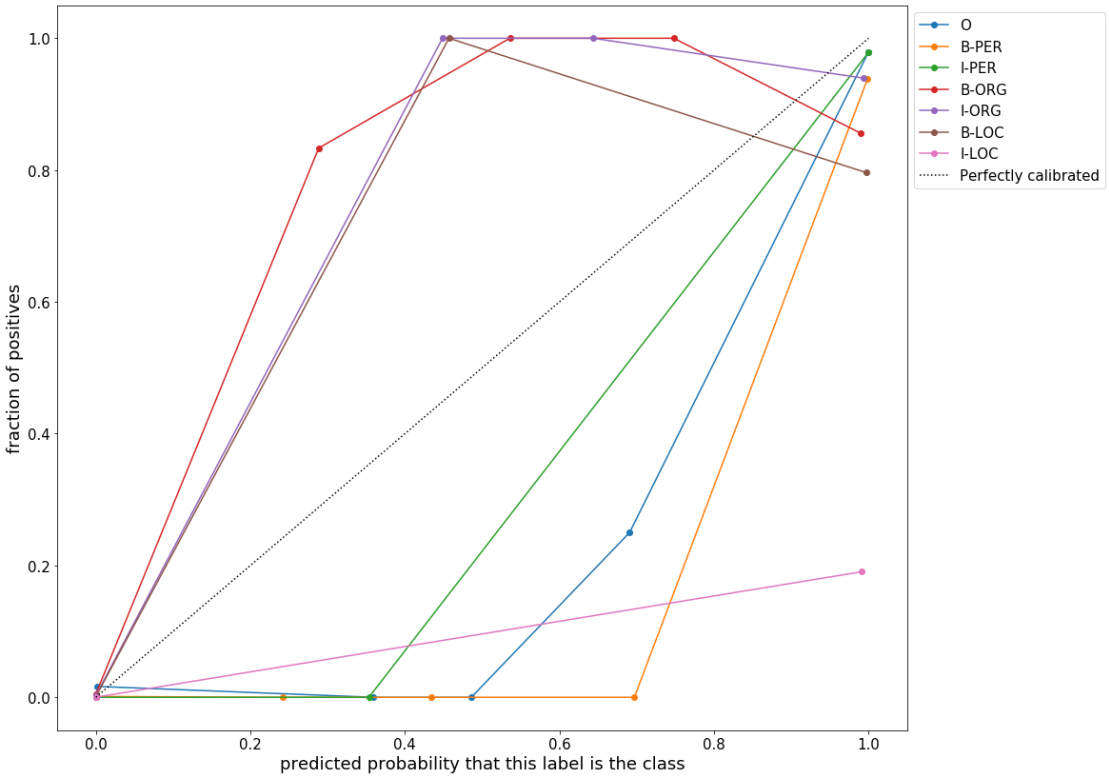


Figure 4: Calibration of BERT model predictions, split into a binary classification problem for each label. 0.2 on the x-axis means that the model predicted with a probability of 0.2, that a given class, for instance B-PER, was the correct one. If the model was perfectly calibrated, 20 % of all the B-PER predictions in the same 'bin' as 0.2, would be a correct classification, that is, truly a B-PER word. The y-axis represents this percentage. In the plot above, the value is 0 %, so none of the B-PER words that had a probability of around 0.2 were actually a true choice. The plot is affected by the fact that there are very few predictions that have a probability between 1 % and 99 %.

In the appendix, there are 10 sentences of which the confidence of the class prediction for each token is shown in the plots giving score given label for words in the sentence, and the confidence in the prediction on a scale 0-1. The model is between 1 % and 99 % confident in its predictions a little less than 0.6 % of the time. Hence the model is not very well calibrated. Its calibration plot is shown in Figure 4. The plot shows that B-ORG, I-ORG and B-LOC are all underconfident when they are predicted with a confidence between 0.2 and 0.8. However, they are again overconfident when predicting with a confidence close to 1. The other classes are, on the contrary, underconfident through all predictions. Due to the few samples that have a confidence between 1 % and 99 %, the graph is not well

supported with samples.

5.2 flair

The overall performance of the flair model is rather similar to the BERT model. Yet there are some differences.

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9222	0	0	0	0	0	2	0	0
B-MISC	112	0	0	0	0	5	0	4	0
I-MISC	37	0	0	1	0	0	0	0	0
B-PER	8	0	0	170	0	1	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	59	0	0	13	0	81	0	8	0
I-ORG	12	0	0	0	5	1	42	0	0
B-LOC	10	0	0	1	0	2	1	82	0
I-LOC	2	0	0	0	0	0	1	0	2

Figure 5: Confusion matrix for the flair model. The rows represents the true values, while the columns represents the predicted values.

The confusion matrix for the classification is shown in Figure 5. Misclassifications of MISC-tagged entities is very apparent in the matrix, just as for the BERT model. However, the flair model was not trained to find these, nor is it possible for the model to predict these, so the matrix should be viewed with that in mind. All the numbers larger than 3 in the matrix that are not on the diagonal comes from the following situations:

- An entity was missed (first column), B-MISC (112), I-MISC (37), B-ORG (59) stands out.
- A B-MISC was categorized to be B-ORG (10) or B-LOC (4).
- A B-ORG entity was misclassified to be a B-PER (13) or a B-LOC (8).
- An I-ORG entity was misclassified to be an I-PER (5).

Another thing to notice is how the model rarely misinterprets an I-PER when the token has been identified as an entity, like the BERT model also managed well. Of the five cases of I-LOC, only three of them were identified and only two of them identified correctly, but on the positive side, the model never hypothesized other tokens to be I-LOC.

	precision	recall	f1-score	support
O	0.97	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.92	0.94	0.93	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.90	0.50	0.65	161
I-ORG	0.91	0.70	0.79	60
B-LOC	0.86	0.85	0.86	96
I-LOC	1.00	0.40	0.57	5
accuracy			0.97	10023
macro avg	0.73	0.60	0.64	10023
weighted avg	0.96	0.97	0.96	10023

Figure 6: Accuracy, precision, recall and f1-score for all the different tags.

Not hypothesizing I-LOC entities increases the macro average of precision of the model greatly, to 0.73, when comparing to the BERT model. The flair model finds that 5.6 % of the tokens are named entities.

Figure 7 shows a calibration plot for the flair model predictions. It is clearly more calibrated than the BERT model. O and B-LOC has underconfident predictions, while it is overconfident when it comes to B-ORG and I-PER.

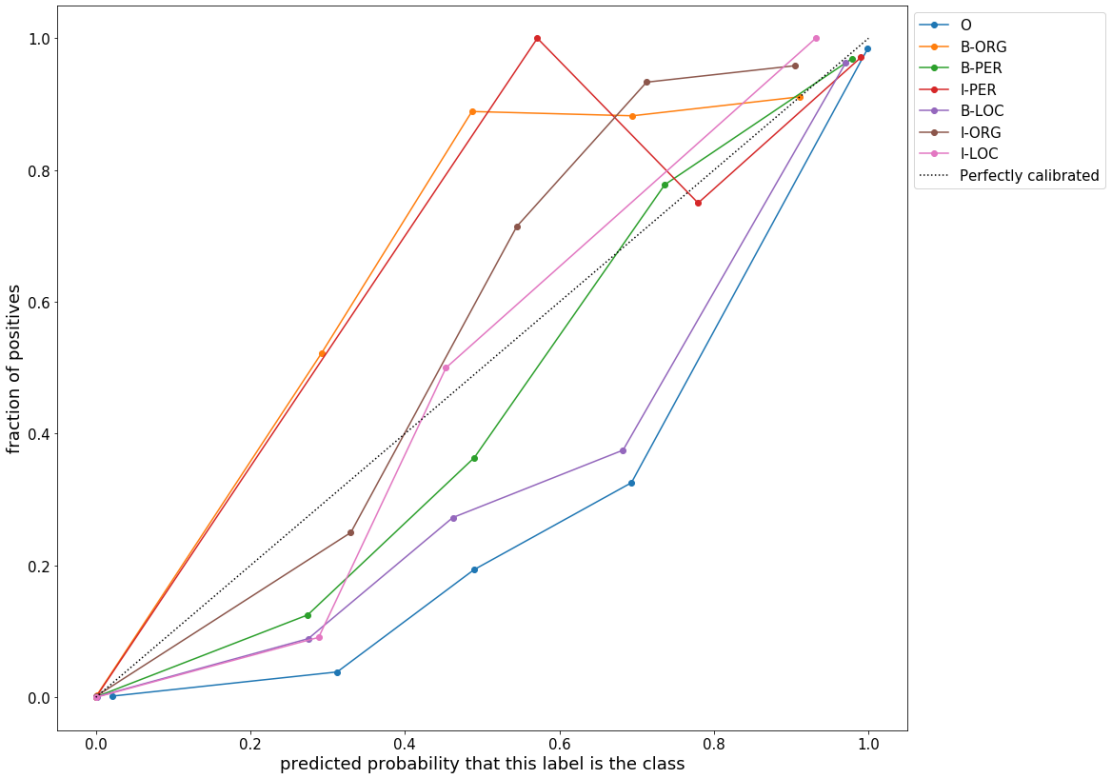


Figure 7: Calibration of flair model predictions, split into a binary classification problem for each label.

5.3 Calibrated flair

The calibrated flair model is simply created taken the predictions of the flair model on the train set, and calibrating the probabilities for each class using Platt Calibration, and finding two constants A and B for each class, which makes the prediction probabilities are optimally calibrated. These constants are applied when calibrating the test set predictions, which make up the data presented below.

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9218	0	0	0	0	1	2	0	3
B-MISC	107	0	0	0	0	10	0	4	0
I-MISC	37	0	0	1	0	0	0	0	0
B-PER	8	0	0	169	0	2	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	45	0	0	10	0	95	0	11	0
I-ORG	11	0	0	0	5	1	41	0	2
B-LOC	9	0	0	1	0	2	1	83	0
I-LOC	1	0	0	0	0	0	0	0	4

Figure 8: Confusion matrix for the calibrated flair model. The rows represents the true values, while the columns represents the predicted values.

The confusion matrix for the classification is shown in Figure 8. This matrix, created from calibrating the flair matrix above, shows similarities to the BERT matrix, because a lot of the tokens where the BERT model was sure and the flair model was not, have been altered so that the calibrated flair is somewhere between. It has, for example, started to hypothesize I-LOC values, like BERT, but not the same degree. Several MISC entities is after calibration categorised as B-ORG or B-LOC, and more B-ORG entities have been identified, improving the B-ORG recall. The flair model finds that 5.9 % of the tokens are named entities.

	precision	recall	f1-score	support
O	0.97	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.92	0.94	0.93	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.90	0.50	0.65	161
I-ORG	0.91	0.70	0.79	60
B-LOC	0.86	0.85	0.86	96
I-LOC	1.00	0.40	0.57	5
accuracy			0.97	10023
macro avg	0.73	0.60	0.64	10023
weighted avg	0.96	0.97	0.96	10023

Figure 9: Accuracy, precision, recall and f1-score for all the different tags.

Figure 9 shows the classification report for the Calibrated flair model. The

calibration worsened the precision, but improved the recall compared to the flair model when looking at the macro average. The macro average f1-score is 0.65, slightly beating the same value for the the models above.

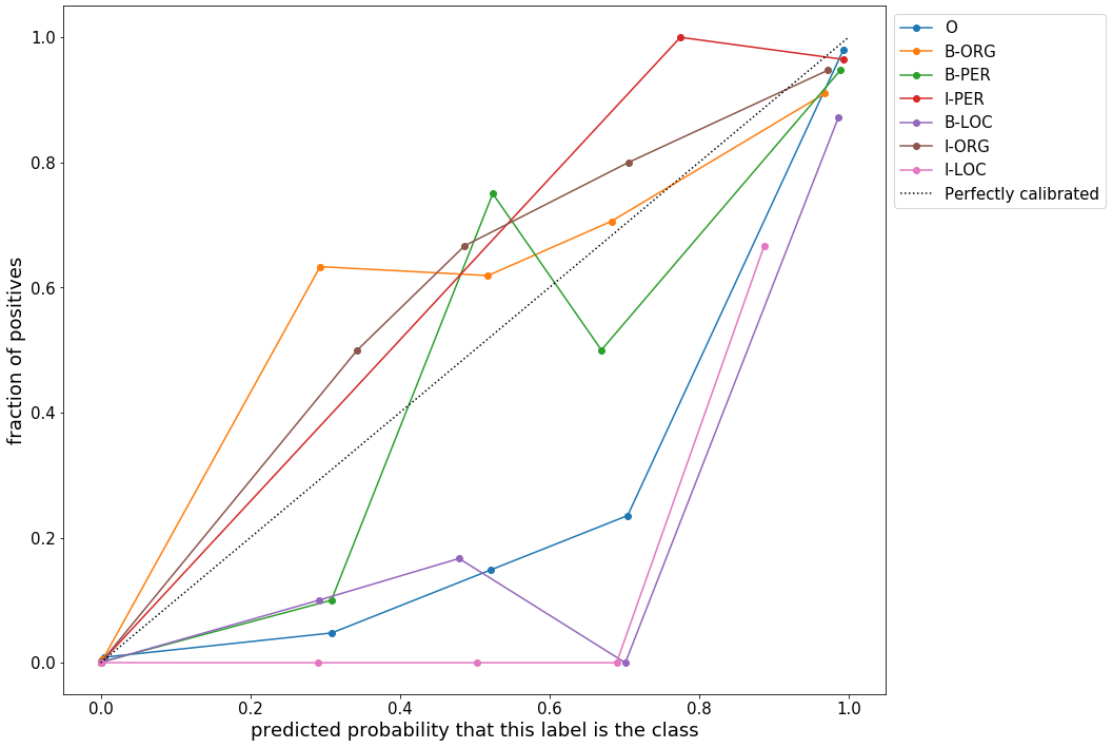


Figure 10: Calibration of Calibrated flair model predictions, split into a binary classification problem for each label.

The calibration plot for the calibrated model shown in Figure10 is unexpected, as one would think this should be more calibrated than the flair model. In some cases it is, the B-ORG, I-ORG and I-PER has become less overconfident, and is more calibrated. However, the underconfident classes has become even more underconfident.

This has an explanation, which comes from the adaption of a binary classification calibration method to a multiclass one. An underconfident predicton might have been calibrated to get more confident, making this class the most probable one among the classes. Then, if the other classes have a probability close to zero, the scaling will set the probability of the entity in the class close to one. An example can be that three classes have the probabilities 0.01, 0.01, and 0.4 after calibration of each class separately, where the class with probability 0.4 is the previously underconfident class. Since the probabilities should sum to one, the scaling of the classes would be to divide each probability with the sum of the

probabilities. In this example, the underconfident class suddenly gets a probability of $0.40/0.42 = 0.95$, changing the result of the calibration.

What is important to keep in mind, however, is that even though this method of normalising after calibration does not change any of the predictions, making the calibration an improvement of the model predictions itself.

In the Appendix, there is a full list of all the predictions that were changed when calibrating the flair model. Here are some statistics connected to the calibration.

- Number of changes: 32
- Positive changes: 17
- Negative changes: 11
- Net positive change: 6 $(= 17 - 11)$
- Change from wrong prediction to wrong prediction: 4 $(= 32 - 17 - 11)$

6 Discussion

6.1 Words tagged to be miscellaneous (MISC)

The MISC entities have caused problems for the models, and therefore, all these are printed in the Appendix. First, the BERT model doesn't seem to have been trained on data in which the label has been included, since it always predicts MISC as the least likely tag for every token. Second, it is difficult for me as a human to understand which words should be classified as MISC and which shouldn't. For example, the word "Statsministeren" (the Prime Minister), is tagged with O, while "FDBchefen" (the FDB chief) and "CNN-journalisterne" (The CNN journalists) and any persons nationality is tagged as MISC, including groups of people of the same party or nationality. One argument to show the difference could be that the journalists and the chief has a company name at the beginning of the entity. When following this rule, the Prime Minister should be a named entity only if one for example specifies the nationality, writing "den danske statsminister". However, since it is clear from context that that is indeed the Prime Minister that is referred to, it is not considered MISC. Neither is it considered a PER, even though it is a person.

There is probably some nuanced explanation coming from a linguist, but it is confusing the models more than it is helping to do the task and achieving the goal of recognizing the named entities.

6.2 Analysis of predictions in selected sentences

In order to give a better picture of how the models perform, ten example sentences will be discussed. All the plots and predictions for each model is put in the Appendix.

Each of the following subsections starts with a sentence where the target tag is put in parenthesis behind the named-entity (that means, the tag O is not shown).

The reason that the plots showing the BERT model's output scores are not showing a confidence on the interval 0-1, is that the softmax function providing such scores makes the differences between the scores that are not the highest ones become invisibly small.

Sentence 1

De mener, at Folkekongressen (B-ORG) skal give præsidenten diktatoriske (B-MISC) beføjelser.

All the models misses *Folkekongressen*, but are less sure of predicting O for this word. None of the models seem to notice that *diktatoriske* should be categorised as a named entity. Referring to the discussion on the MISC category, it is hard to understand why this word has that tag.

Sentence 2

Det mener Socialdemokratiets (B-ORG) næstformand Birte (B-PER) Weiss (I-PER) og foreslår, at de politiske ledere kommer med i det ligebehandlingsnævn, som Folketinget (B-ORG) kort før sommerferien besluttede at nedsætte.

Of all the predictions of the models, only one mistake is made, when the flair model slightly misses *Socialdemokratiets*. However, it is only a bit more than 50 % certain that this word has an O tag, while it is a bit less than 50 % certain that this word is tagged B-ORG. The calibrated flair model flips the situation, as predicts correctly by being nearly 60 % certain that it is B-ORG.

Sentence 3

Han var nemlig kandidat for og en af initiativtagerne til det grundtvigianske-socialdemokratiske (B-MISC) samarbejde, der har vundet flertallet i Skjerns (B-LOC) menighedsråd.

All the models agree on their predictions on this, and all of them are wrong. At least according to the established targets. They all miss the MISC-word, and they all predict *Skjerns menighedsråd* to be B-ORG and I-ORG, with less confidence on the first one. flair is about 60 % sure that *Skjerns* is B-ORG, and about 30 % sure that is is B-LOC. The calibrated flair is even more certain that it is B-ORG

and not B-LOC. The problem here is that one cannot deny that "Skjern" is a location, but that the two words make up an organisation is also difficult to deny. When evaluating the models, it is definitely clear that categorising these two token predictions as a complete mistake is not at all the best way to do it.

Sentence 4

Det virker "fåret", men det er interessant, hvis regeringens sikkerhedsudvalg og statsministeren ikke har været underrettet om missionen.

This sentence does not contain any named entities, and the flair models correctly predicts this as well. However, the BERT model hypothesizes that *sikkerhedsudvalg* is a B-ORG with a 92 % certainty. This is an example of how the BERT model predicts more entities among the dataset than the other models. And again, it does not seem completely correct to give the BERT model feedback that it was completely wrong, since *regeringens sikkherhedsudvalg* indeed is an organisation, although not named in this instance. The models are all really certain about its predictions overall for this sentence, an outlier other than the hypothesizing mentioned above, the flair model is around 10 % certain that *statsministeren* is a B-ORG.

Sentence 5

Det vil næppe volde rivalerne i Dansk (B-ORG) Supermarked (I-ORG) og Aldi (B-ORG) finansielle problemer at komme med et modspil i samme størrelsesorden.

BERT and Calibrated flair gets these tags right, while flair misses *Aldi*. This is another example of how the calibration of flair provides the correct prediction of the tag.

Sentence 6

600 Brugser (B-ORG) har ikke fået lov at blive Super-Brugser (B-ORG).

Again, it can be questioned whether the words are correctly annotated. Obviously, "Super-Brugsen" is an organisation, but the named entities in the sentence refers to stores, and not the organisation itself. It could maybe make sense to tag them as locations, as well as miscellaneous. All the models predicts O on these named entities, with the flair model being very uncertain of the word *Super-Brugser*. The calibration strengthes the belief that this word is indeed of the tag O, showing that calibration is not necessarily pointing the model in the correct direction in all cases.

Sentence 7

Desuden forudser FDB (B-ORG), at 200 job af sparehensyn skæres bort i de nye Super-Brugser (B-ORG).

Here, the same named entity is used as in Sentence 6. The BERT model confidently does a correct prediction on this, contrary to above. This illustrates how the model is dependent on context to do its predictions, which essentially is a good thing in NLP. flair fails, predicting with a 30 % certainty that *Super-Brugser* is tagged with O. This is fixed in the calibrated model, although it is still not certain in its prediction, and still thinks it is likely to be B-LOC as well.

Sentence 8

En ny aggressiv linje præger Super-Brugsen (B-ORG).

An example of where BERT predicts correctly, while both flair and Calibrated flair does not. Both flair models are also more certain that it is a B-LOC if it is not an O, while B-ORG is on third place.

Sentence 9

De 10 hold i Superligaen (B-ORG) skal mødes to gange i løbet af foråret, og det giver således 18 runder med afslutning den 23. juni, hvor programmet ser således ud: Lyngby-Brøndby (B-MISC) , Silkeborg-AGF (B-MISC), Vejle-Ikast (B-MISC), Frem-B (B-MISC) 1903 (I-MISC) og AaB-OB (B-MISC).

Again, BERT is the only model to recognize *Superligaen* as B-ORG. There are also a lot of football matches tagged with MISC. The models are all wrong on these, but they clearly think that they are named entities in most cases. It is difficult to evaluate which model does it best, but this shows why it is important to keep MISC as a part of the evaluation - otherwise, the models that miss the named entities will have fewer mistakes than the ones that predicted B-ORG and/or B-LOC.

6.2.1 Sentence 10

Uden dog at opnå meget mere end æren, fordi DM-guldet (B-MISC) mod sædvane ikke vil give adgang til Europa (B-MISC) Cup (I-MISC) turneringerne.

Another sentence with several MISC tags. Both flair and the Calbrated flair claims that *Europa* is B-LOC. Which is in a sense correct, but in this case, the BERT model should get more credit for understanding context, and recognizing that this is not referring to a location. However, as the model always thinks that MISC is the most unlikely prediction to make, and the other alternatives does not make sense either, it predicts O. The flair model and the Calibrated flair model

both shows uncertainty on all the MISC tags, showing that it is aware of a potential presence of a named entity.

6.3 Comparison of the models

6.3.1 Which model performs the best?

Due to factors like entities tagged with MISC, ambiguity in annotation of entities, and difference in prediction classes for the different models make it difficult to get a clear and certain way to say that one model is better than the other.

There are, without using numbers, some observations that differ the models. BERT seems to be more aware of context, as seen in Sentence 10. This is likely a result of it being made using Transformers, who provide a different amount of attention between the different words in the sentence. It does also predict the presence of named entities more often than the other models (6.1 % versus 5.6 % for flair and 5.9 % for the Calibrated flair), which gives it the highest recall, but leads to lower precision due to for example the hypothesizing in Sentence 4. BERT is also not very well calibrated, making it certain of its correct predictions, but also certain of being correct when its predictions are wrong.

The flair model is a lot more calibrated - its predictions reflects its accuracy more often. It is less likely to predict that a word is a named entity than the BERT model, but it is also often really near to predicting a named entity, and really uncertain of predicting O. Predicting less also gives flair a higher precision, but it misses more entities, providing a lower recall.

As seen in the results, the Calibrated flair model does the same predictions as the flair model, except in 32 cases. In many of these cases, as shown in some of the example sentences, the calibration gives the probability distribution over the classes for a token a slight change, making the model provide the correct prediction. Altogether, 17 changes changed the prediction to the correct tag, while 11 changes changed the prediction from a correct tag to an incorrect one. The cases where the changes improved the model, are often where the flair model did a wrong prediction and BERT did a correct one. Hence the calibrated model appears somewhere in between the two.

When looking a scores, the way the calibrated model looks like in between BERT and flair is shown in the macro average calculations of precision and recall. The macro average results are repeated below.

Precision macro average:

- BERT: 0.63
- flair: 0.73

- Calibrated flair: 0.66

Recall macro average:

- BERT: 0.65
- flair: 0.60
- Calibrated flair: 0.65

f1-score macro average:

- BERT: 0.61
- flair: 0.64
- Calibrated flair: 0.65

The weakness of using the macro average is that the five instances of I-LOC in the test set plays a very significant role in the computation of the macro average. Then again, this importance also reflects how well the models are able to avoid hypothesizing entities.

The other reason for focusing on the macro average is that the other numbers in the classification report that weren't class-specific, were the same for all models. The accuracy was 97 %, the weighted precision was 96 %, the weighted recall was 97 %, and the weighted f1-score was 96 %, all of them highly influenced by the amount of easy-to-classify outside (O) tag.

6.3.2 How MISC should play a bigger role

The MISC classes have definitely complicated the evaluation, and also worsened the performance of the models. In order to get a clearer picture of which model is the better one, the models need to be trained to recognize named entities that are neither locations, organisations or persons. This is important also if one is not interested in the MISC cases, since the models at this time are able to recognize these as named entities, but does misclassifications as they predict LOC or ORG (and sometimes PER) for words tagged with MISC. In order for this to be possible, a clear definition of what MISC should be is necessary. Although linguists might be able to provide a sensible classification of what is MISC and not, this might not be the correct way to attach the problem. The MISC category should be defined using a combination of what a person that is familiar with the models think the model is able to learn, what is as correct as possible but also makes sense for the algorithm, as well as input from linguists and other annotators. In addition, what the models think is MISC and not after training them the first times, should also

be taken into consideration. If the model thinks a word is likely to be MISC with some confidence, this word should be discussed as whether to also make MISC in the target set.

6.3.3 How one could get a more precise evaluation

Looking at NER as a merge of the binary classification problem of recognizing whether a token is a named entity or not, and the multiclass classification problem of recognizing what kind of entity a token is, gives some ideas of how to improve the evaluation. For example, if the model recognizes that a named entity is present, but misclassifies what kind of entity this is, the model should still be rewarded in some sense (not in training of the model, but when evaluating). Also, in situations where the entity is wrongly classified because of misunderstanding of the context (classifying "Brøndby" as B-LOC rather than B-ORG, even though it is clear from the context that the text is referring to the football team/organisation [3]), the model should also get some kind of recognition.

7 Conclusion

The vast diversity of language makes it difficult to put words into categories. However, the models discussed are generally good at this, recognizing entities and categorising them into locations, organisations, and persons. Calibration is definitely a tool that can improve the performance of the flair model, and might even be an even stronger tool if the model was given the ability to predict miscellaneous entities. The flair model is almost twice as fast as BERT, but all in all, the models perform similarly, and only differ in the macro average score. The model with the highest macro average f1-score was the Calibrated flair model, while flair did best on precision, and both BERT and the Calibrated flair shared the better recall score.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Aug. 2018, pp. 1638–1649, Association for Computational Linguistics.

[3] Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard, "DaNE: A named entity resource for Danish," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4597–4604, European Language Resources Association.

[4] Zhiheng Huang, Wei Xu, and Kai Yu, "Bidirectional lstm-crf models for sequence tagging," 2015.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[6] Alexandru Niculescu-Mizil and Rich Caruana, "Predicting good probabilities with supervised learning," 01 2005, pp. 625–632.

Appendix

List of words tagged to be miscellaneous

russiske (B), demokrati (B), diktatur (B), demokratiske (B), demokrati (B), diktatoriske (B), jugoslavisk (B), social-grundtvigianerne (B), indremissionske (B), grundtvigianske-socialdemokratiske (B), danskeres (B), FDBchefen (B), Ikast-Lyngby (B), B-1903-Silkeborg (B), AaB-Vejle (B), FremOB (B), Lyngby-Brøndby (B), Silkeborg-AGF (B), Vejle-Ikast (B), Frem-B (B), 1903 (I), AaB-OB (B), DM-guldet (B), Europa (B), Cup (I), Superligaen (B), litauer (B), spanske (B), Sovjet-soldater (B), litauer (B), sovjetiske (B), israelske (B), CNN-journalisterne (B), dansk (B), Brøndby'ere (B), Superliga-finalen (B), Pedal-Ove-sagen (B), The (B), Healer (I), Grammy (B), The (B), Healer (I), 40'erne (B), Dansk (B), danskere (B), Røde-Kro-løjer (B), DANSKE (B), AIDS (B), Hof (B), Tuborg (B), borgerlige (B), socialdemokratiske (B), borgerlige (B), borgerlig (B), Ariostea-mandskabet (B), Orientering (B), De (B), ringer (I), , (I), vi (I), spiller (I), KV-planen (B), borgerlige (B), EF-fiskeriministermøde (B), EF-direktiv (B), danske (B), su-pereuropæer (B), Brødrene (B), Løvehjerte (I), nazister (B), Retfærdighed (B), - (I), ikke (I), hævn (I), jøder (B), socialdemokratiske (B), UTB-praktikpladser (B), ATB-jobs (B), ATB-jobs (B), Anderledes (B), Familiebilleder (I), olymp-iske (B), irsk (B), danske (B), serbisk (B), serbisk (B), danske (B), de (B), fire (I), årstider (I), opus (B), VIII (I), Il (B), Cimento (I), dell'Armonia (I), e (I), dell'Invenzione (I), Eleva2ren (B), grønlandske (B), britiske (B), danske (B), Danmarks-turné (B), socialistiske (B), kommunismen (B), 1950'erne (B), afghanske (B), Gøngehøvdingen (B), Camel (B), svensk (B), menneskerettighed-ernes (B), Daphne-klassen (B), Søløveklassen (B), olympiske (B), Brøndby-ledelsens (B), schweizisk (B), engelske (B), danskerne (B), europæerne (B), danske (B), Porsche (B), Black (B), Celebration (I), svenskeren (B), sovjetisk (B), anden (B), verdenskrig (I), danske (B), besættelsen (B), Danfoss-lærlinge (B), Ørsted (B), Ridder (B), af (I), Dannebrog (I), Langt (B), ud (I), af (I), halsen (I), Op (B), og (I), stå (I), Verdens (B), grimmeste (I), pige (I), Stor (B), og (I), stærk (I), Dengang (B), min (I), onkel (I), Kulle (I), blev (I), skør (I), P-pillerne (B), polakkerne (B), NU-bøgerne (B), danske (B), tyskerne (B), Unix' (B), Gug'ske (B).

Confusion matrices and classification reports

BERT

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9214	0	0	1	0	4	1	2	2
B-MISC	111	0	0	0	0	10	0	0	0
I-MISC	36	0	0	1	0	0	1	0	0
B-PER	10	0	0	168	0	1	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	25	0	0	11	0	106	0	19	0
I-ORG	8	0	0	0	3	1	33	0	15
B-LOC	9	0	0	0	0	1	0	86	0
I-LOC	1	0	0	0	0	0	0	0	4

t/p	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9361	2	0	14	2	2	2
B-PER	10	168	0	1	0	1	0
I-PER	0	0	138	0	0	0	0
B-ORG	25	11	0	106	0	19	0
I-ORG	8	0	3	1	33	0	15
B-LOC	9	0	0	1	0	86	0
I-LOC	1	0	0	0	0	0	4

t/p	O	PER	ORG	LOC
O	9361	2	16	4
PER	10	306	1	1
ORG	33	14	140	34
LOC	10	0	1	90

	precision	recall	f1-score	support
O	0.98	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.93	0.93	0.93	180
I-PER	0.98	1.00	0.99	138
B-ORG	0.86	0.66	0.75	161
I-ORG	0.94	0.55	0.69	60
B-LOC	0.80	0.90	0.84	96
I-LOC	0.19	0.80	0.31	5
accuracy			0.97	10023
macro avg	0.63	0.65	0.61	10023
weighted avg	0.96	0.97	0.96	10023

	precision	recall	f1-score	support
O	0.99	1.00	1.00	9383
B-PER	0.93	0.93	0.93	180
I-PER	0.98	1.00	0.99	138
B-ORG	0.86	0.66	0.75	161
I-ORG	0.94	0.55	0.69	60
B-LOC	0.80	0.90	0.84	96
I-LOC	0.19	0.80	0.31	5
accuracy			0.99	10023
macro avg	0.81	0.83	0.79	10023
weighted avg	0.99	0.99	0.99	10023

	precision	recall	f1-score	support
O	0.99	1.00	1.00	9383
PER	0.95	0.96	0.96	318
ORG	0.89	0.63	0.74	221
LOC	0.70	0.89	0.78	101
accuracy			0.99	10023
macro avg	0.88	0.87	0.87	10023
weighted avg	0.99	0.99	0.99	10023

flair

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9222	0	0	0	0	0	2	0	0
B-MISC	112	0	0	0	0	5	0	4	0
I-MISC	37	0	0	1	0	0	0	0	0
B-PER	8	0	0	170	0	1	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	59	0	0	13	0	81	0	8	0
I-ORG	12	0	0	0	5	1	42	0	0
B-LOC	10	0	0	1	0	2	1	82	0
I-LOC	2	0	0	0	0	0	1	0	2

t/p	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9371	1	0	5	2	4	0
B-PER	8	170	0	1	0	1	0
I-PER	0	0	138	0	0	0	0
B-ORG	59	13	0	81	0	8	0
I-ORG	12	0	5	1	42	0	0
B-LOC	10	1	0	2	1	82	0
I-LOC	2	0	0	0	1	0	2

t/p	O	PER	ORG	LOC
O	9371	1	7	4
PER	8	308	1	1
ORG	71	18	124	8
LOC	12	1	4	84

	precision	recall	f1-score	support
O	0.97	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.92	0.94	0.93	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.90	0.50	0.65	161
I-ORG	0.91	0.70	0.79	60
B-LOC	0.86	0.85	0.86	96
I-LOC	1.00	0.40	0.57	5
accuracy			0.97	10023
macro avg	0.73	0.60	0.64	10023
weighted avg	0.96	0.97	0.96	10023

	precision	recall	f1-score	support
O	0.99	1.00	0.99	9383
B-PER	0.92	0.94	0.93	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.90	0.50	0.65	161
I-ORG	0.91	0.70	0.79	60
B-LOC	0.86	0.85	0.86	96
I-LOC	1.00	0.40	0.57	5
accuracy			0.99	10023
macro avg	0.94	0.77	0.83	10023
weighted avg	0.99	0.99	0.98	10023

	precision	recall	f1-score	support
O	0.99	1.00	0.99	9383
PER	0.94	0.97	0.95	318
ORG	0.91	0.56	0.69	221
LOC	0.87	0.83	0.85	101
accuracy			0.99	10023
macro avg	0.93	0.84	0.87	10023
weighted avg	0.99	0.99	0.99	10023

Calibrated flair

t/p	O	B-MISC	I-MISC	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9218	0	0	0	0	1	2	0	3
B-MISC	107	0	0	0	0	10	0	4	0
I-MISC	37	0	0	1	0	0	0	0	0
B-PER	8	0	0	169	0	2	0	1	0
I-PER	0	0	0	0	138	0	0	0	0
B-ORG	45	0	0	10	0	95	0	11	0
I-ORG	11	0	0	0	5	1	41	0	2
B-LOC	9	0	0	1	0	2	1	83	0
I-LOC	1	0	0	0	0	0	0	0	4

t/p	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	9362	1	0	11	2	4	3
B-PER	8	169	0	2	0	1	0
I-PER	0	0	138	0	0	0	0
B-ORG	45	10	0	95	0	11	0
I-ORG	11	0	5	1	41	0	2
B-LOC	9	1	0	2	1	83	0
I-LOC	1	0	0	0	0	0	4

t/p	O	PER	ORG	LOC
O	9362	1	13	7
PER	8	307	2	1
ORG	56	15	137	13
LOC	10	1	3	87

	precision	recall	f1-score	support
O	0.98	1.00	0.99	9224
B-MISC	0.00	0.00	0.00	121
I-MISC	0.00	0.00	0.00	38
B-PER	0.93	0.94	0.94	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.86	0.59	0.70	161
I-ORG	0.93	0.68	0.79	60
B-LOC	0.84	0.86	0.85	96
I-LOC	0.44	0.80	0.57	5
accuracy			0.97	10023
macro avg	0.66	0.65	0.65	10023
weighted avg	0.96	0.97	0.96	10023

	precision	recall	f1-score	support
O	0.99	1.00	0.99	9383
B-PER	0.93	0.94	0.94	180
I-PER	0.97	1.00	0.98	138
B-ORG	0.86	0.59	0.70	161
I-ORG	0.93	0.68	0.79	60
B-LOC	0.84	0.86	0.85	96
I-LOC	0.44	0.80	0.57	5
accuracy			0.99	10023
macro avg	0.85	0.84	0.83	10023
weighted avg	0.99	0.99	0.99	10023

	precision	recall	f1-score	support
O	0.99	1.00	0.99	9383
PER	0.95	0.97	0.96	318
ORG	0.88	0.62	0.73	221
LOC	0.81	0.86	0.83	101
accuracy			0.99	10023
macro avg	0.91	0.86	0.88	10023
weighted avg	0.99	0.99	0.99	10023

Changes in flair predictions when calibrating

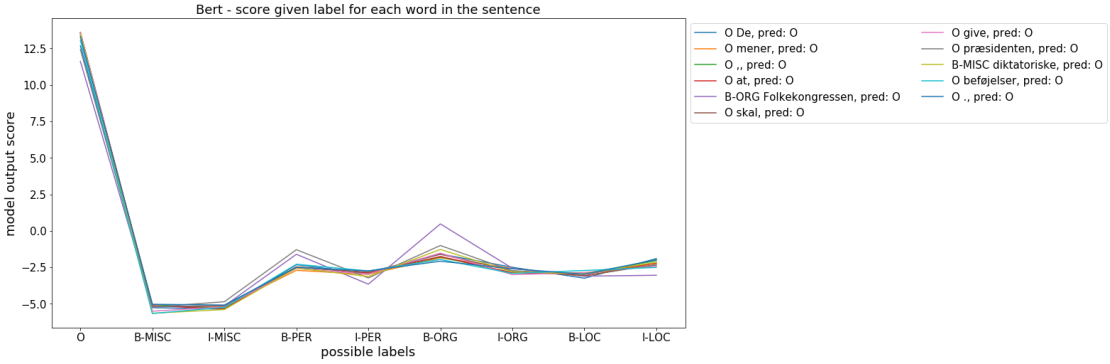
Calibrated flair	flair	Target	Target (with MISC)	Preceding words	Word	Subsequent word
B-ORG	O	B-ORG	B-ORG	Det mener	" Socialdemokratiets "	næstformand Birte
B-ORG	O	B-ORG	B-ORG	Supermarked og	" Aldi "	finansielle problemer
B-ORG	O	B-ORG	B-ORG	de nye	" Super-Brugser "	. Tidligere
B-LOC	O	B-ORG	B-ORG	linje præger	" Super-Brugsen "	. Her
B-ORG	O	O	B-MISC	Ikast-Lyngby ,	" B-1903-Silkeborg "	, AaB-Vejle
B-ORG	O	O	B-MISC	Vejle-Ikast ,	" Frem-B "	1903 og
I-LOC	O	O	O	i Brøndbys	" 500-kampsjubilar "	Bjarne Jensen
I-LOC	O	I-ORG	I-ORG	, Frederikshavn	" politi "	, som
B-ORG	O	O	O	i kulørt	" Lycra "	med Marilyn
B-ORG	O	B-ORG	B-ORG	Sællerter Nedslidte	" Levi's "	jeans til
B-ORG	O	B-ORG	B-ORG	Winter ,	" Status "	Quo ,
I-LOC	O	I-LOC	I-LOC	i Kolding	" arrest "	, efter
B-ORG	O	B-ORG	B-ORG	her triumferede	" Ariostea "	med sejr
B-ORG	O	B-ORG	B-ORG	Motorola ,	" Gatorade-Chateau-d'Ax "	, Banesto
B-ORG	O	B-ORG	B-ORG	Maskinchef på	" Ask "	var Ole
B-ORG	B-PER	B-ORG	B-ORG	at både	" Ask "	og Urd
B-ORG	O	B-ORG	B-ORG	Ask og	" Urd "	var gode
I-LOC	I-ORG	I-ORG	I-ORG	have Århus	" Havns "	bugserbåd ,
I-LOC	O	O	O	Århus Havns	" bugserbåd "	, Hermes
B-ORG	B-PER	B-ORG	B-ORG	AFSLØRINGER Siden	" Ungbo "	for tre
B-ORG	O	B-ORG	B-ORG	derfor går	" Østlandepuljen "	ind med
B-ORG	B-PER	B-ORG	B-ORG	dominerede nyhedsbureau	" Tanjug "	har tidligere
I-LOC	I-ORG	I-LOC	I-LOC	forbi Sydhavn	" S-station "	mod Sjællandsbroen
B-LOC	O	B-LOC	B-LOC	, sagde	" Serbiens "	udenrigsminister .
B-ORG	O	O	B-MISC	patruljebåde af	" Daphne-klassen "	, bygget
B-ORG	O	B-ORG	B-ORG	afleveret fra	" orlogsværftet "	i 1965-67
B-ORG	O	O	B-MISC	drøj for	" Brøndby-ledelsens "	ellers så
I-LOC	O	O	O	i Vest-Berlins	" sydvestlige "	hjørne .
B-ORG	B-PER	B-PER	B-PER	Lau ,	" H. "	Sumiyoshi og
B-ORG	O	O	B-MISC	stedet for	" P-pillerne "	, men
B-LOC	O	B-ORG	B-ORG	topklubber ,	" Besiktas "	, Fenerbache
B-LOC	O	B-ORG	B-ORG	Besiktas ,	" Fenerbache "	eller Galatasaray

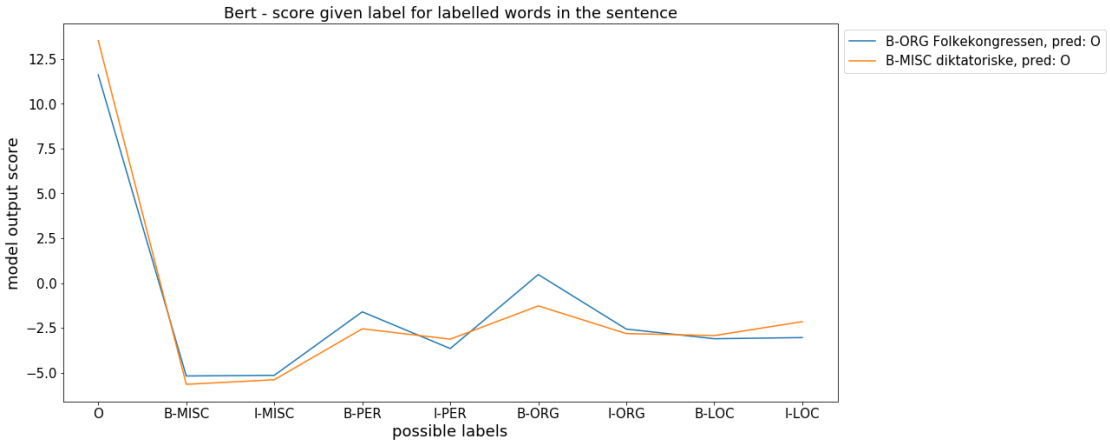
Analysis of selected example sentences

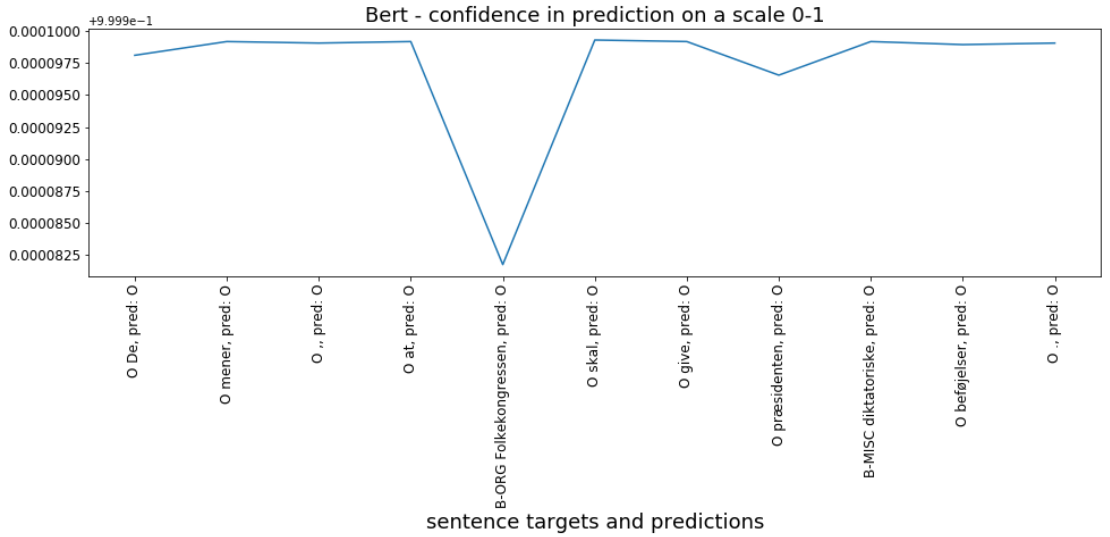
Sentence	De	mener	,	at	Folkekongressen	skal	give	præsidenten	diktatoriske	beføjelser	.
Targets	O	O	O	O	B-ORG	O	O	O	B-MISC	O	O
BERT	O	O	O	O	O	O	O	O	O	O	O
flair	O	O	O	O	O	O	O	O	O	O	O
Calibrated flair	O	O	O	O	O	O	O	O	O	O	O

Table 1: Sentence 1

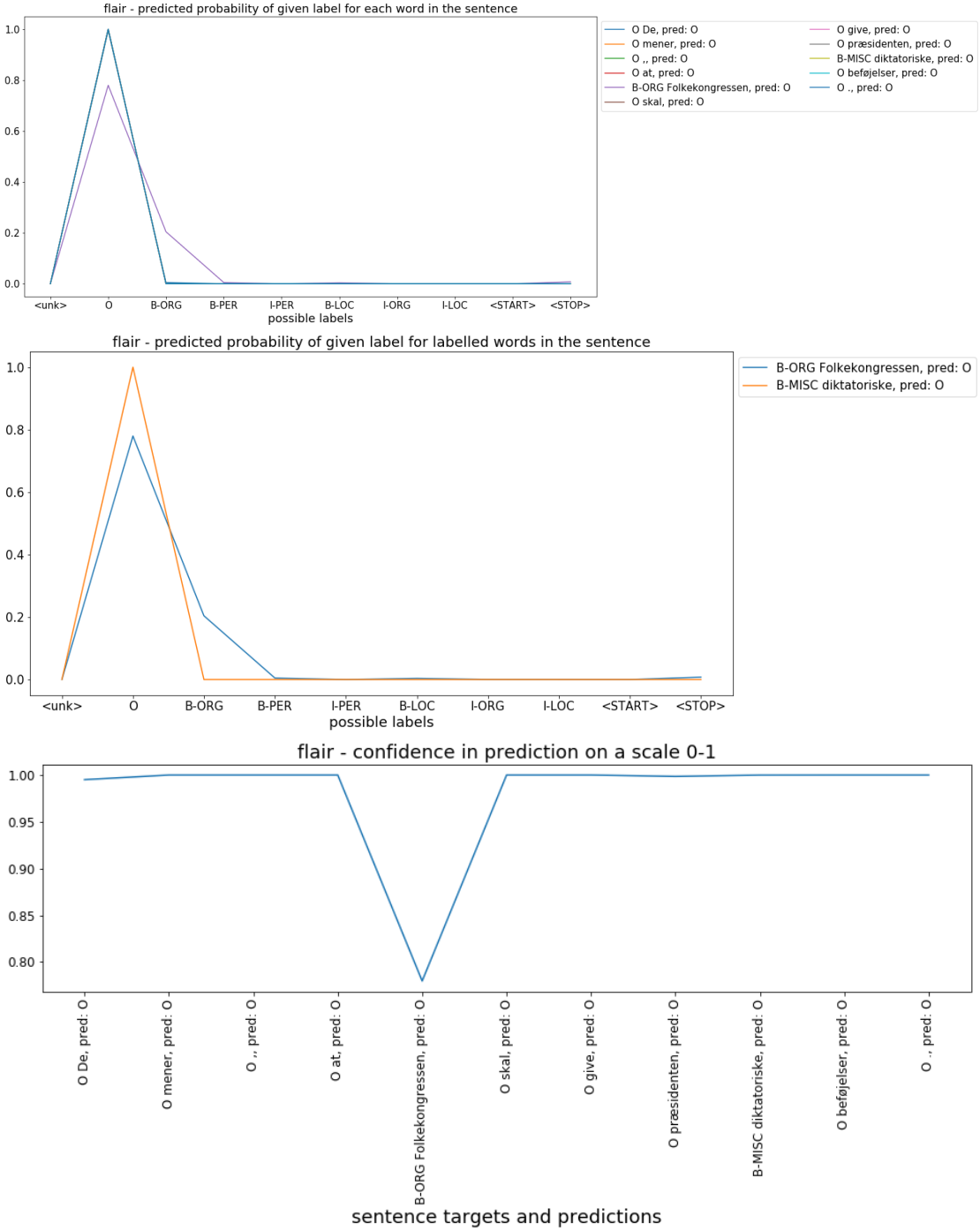
BERT - Sentence 1



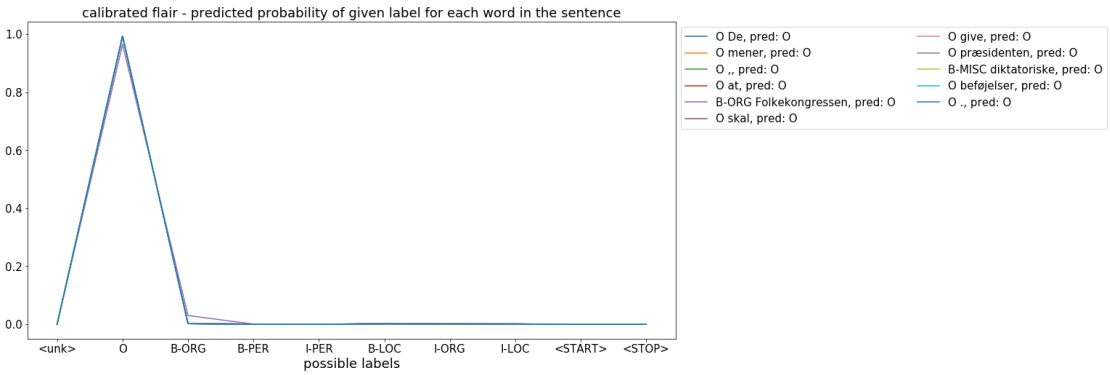


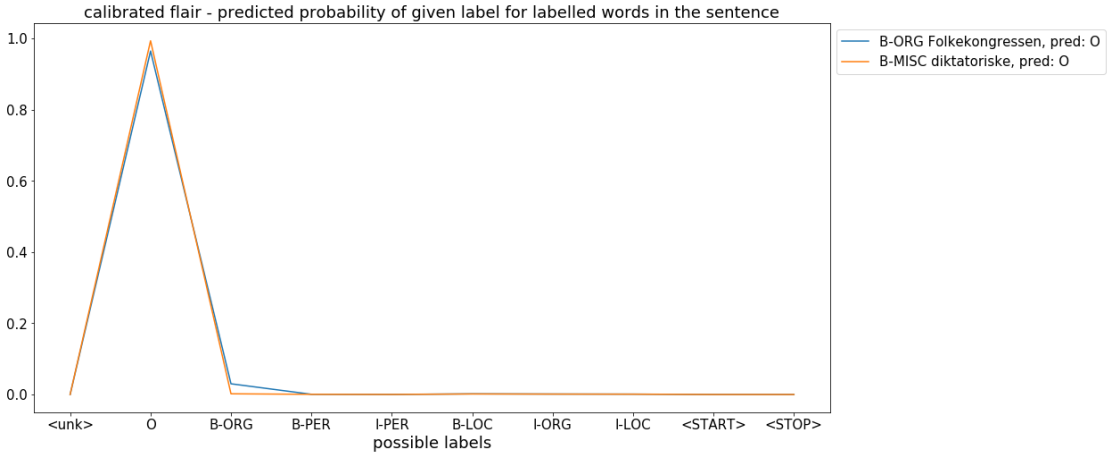


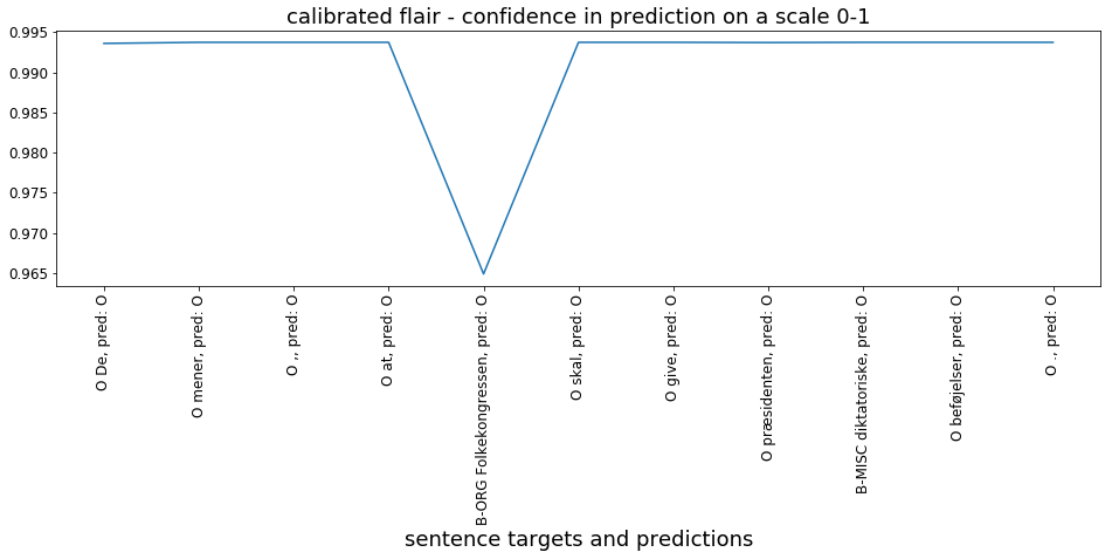
flair - Sentence 1



Calibrated flair - Sentence 1



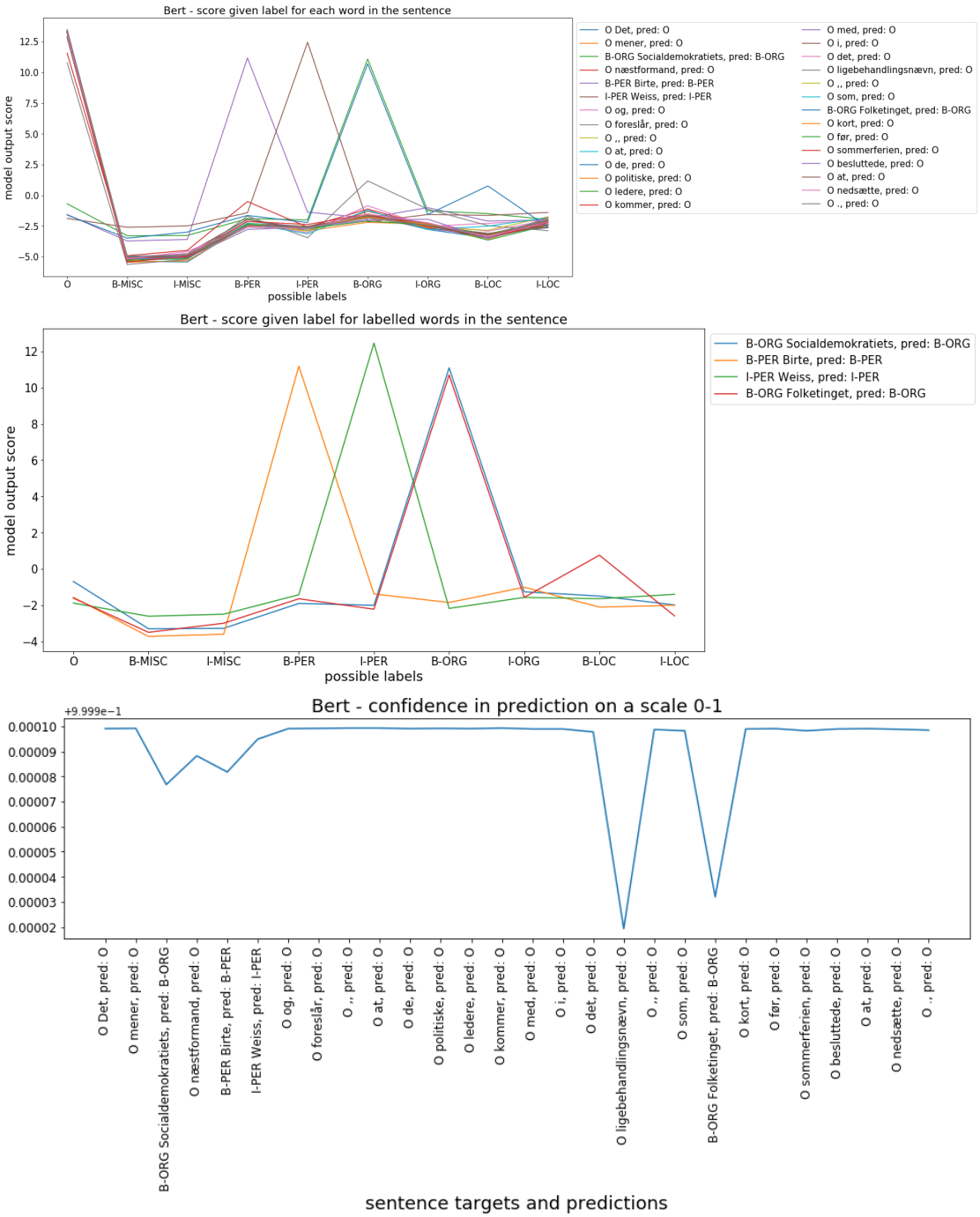




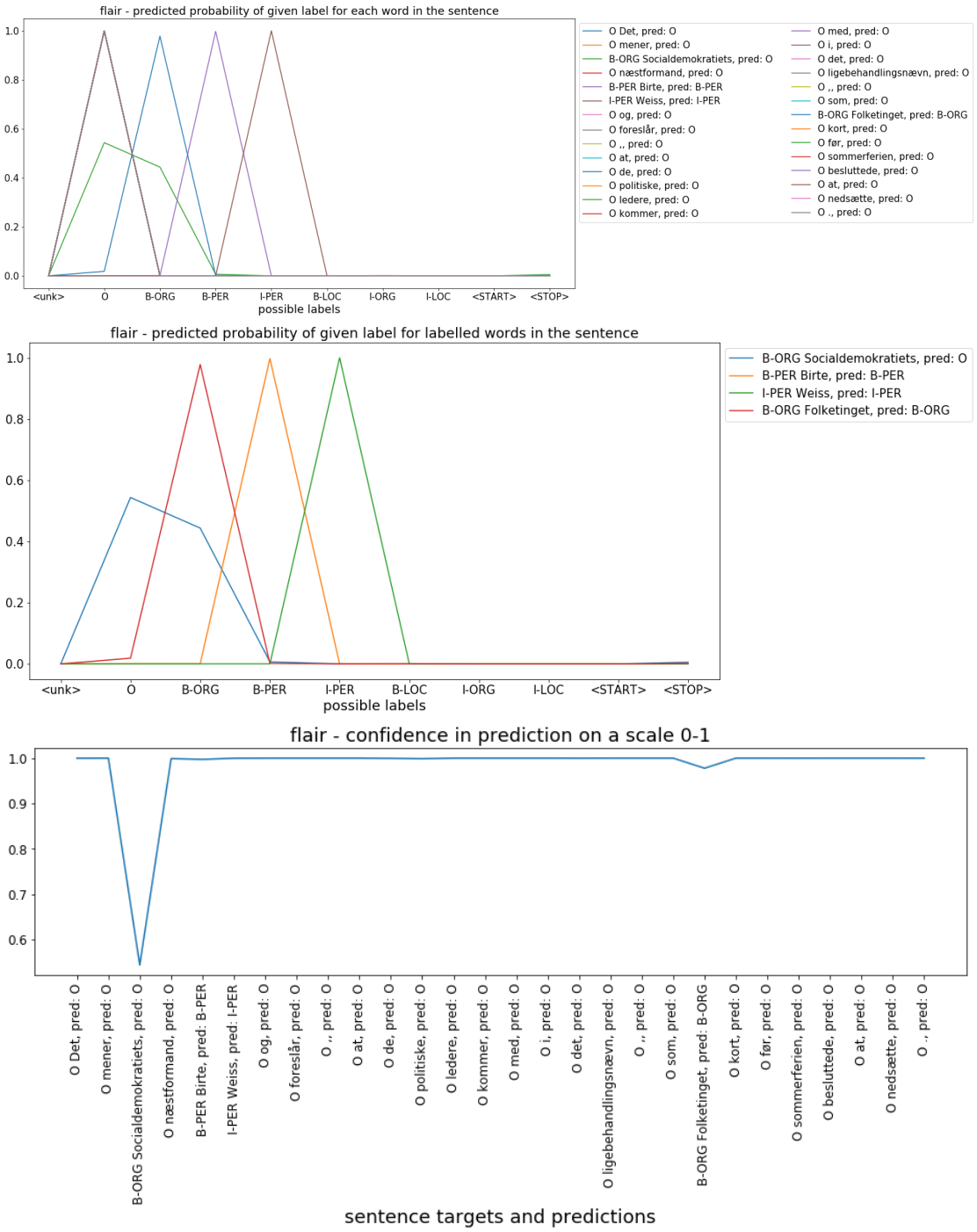
Sentence	Det	mener	Socialdemokratiets	næstformand	Birte	Weiss	og	...	det	ligebehandlingsnævn	.	som	Folketinget	kort	før	sommerferien	besluttede	at	nedsætte	.
Targets	O	O	B-ORG	O	B-PER	I-PER	O	...	O	O	O	O	B-ORG	O	O	O	O	O	O	O
BERT	O	O	B-ORG	O	B-PER	I-PER	O	...	O	O	O	O	B-ORG	O	O	O	O	O	O	O
flair	O	O	O	O	B-PER	I-PER	O	...	O	O	O	O	B-ORG	O	O	O	O	O	O	O
Calibrated flair	O	O	B-ORG	O	B-PER	I-PER	O	...	O	O	O	O	B-ORG	O	O	O	O	O	O	O

Table 2: Sentence 2

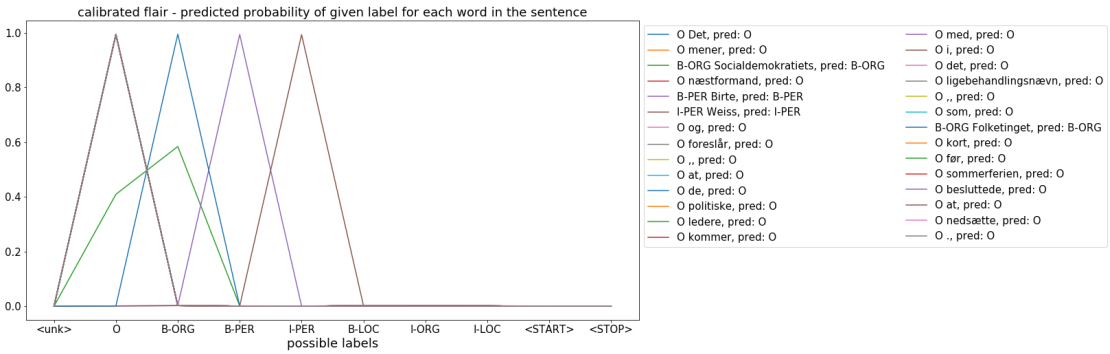
BERT – Sentence 2

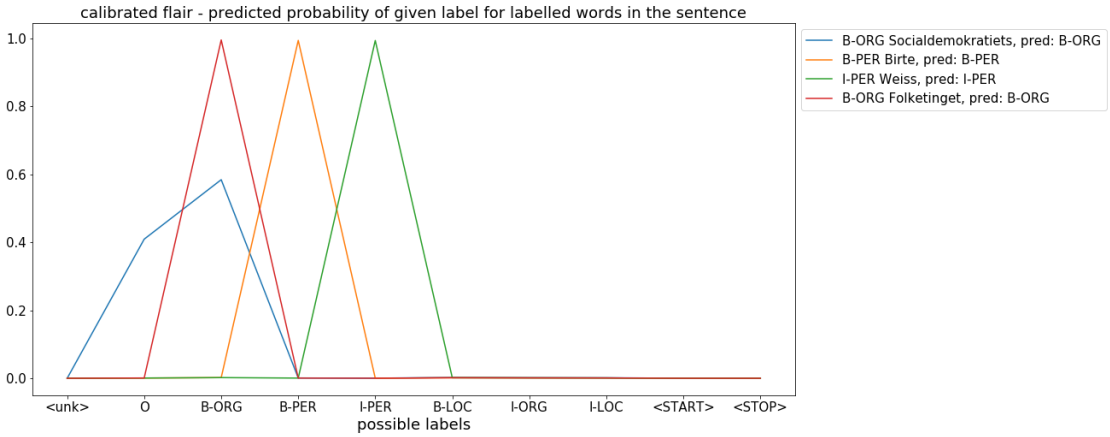


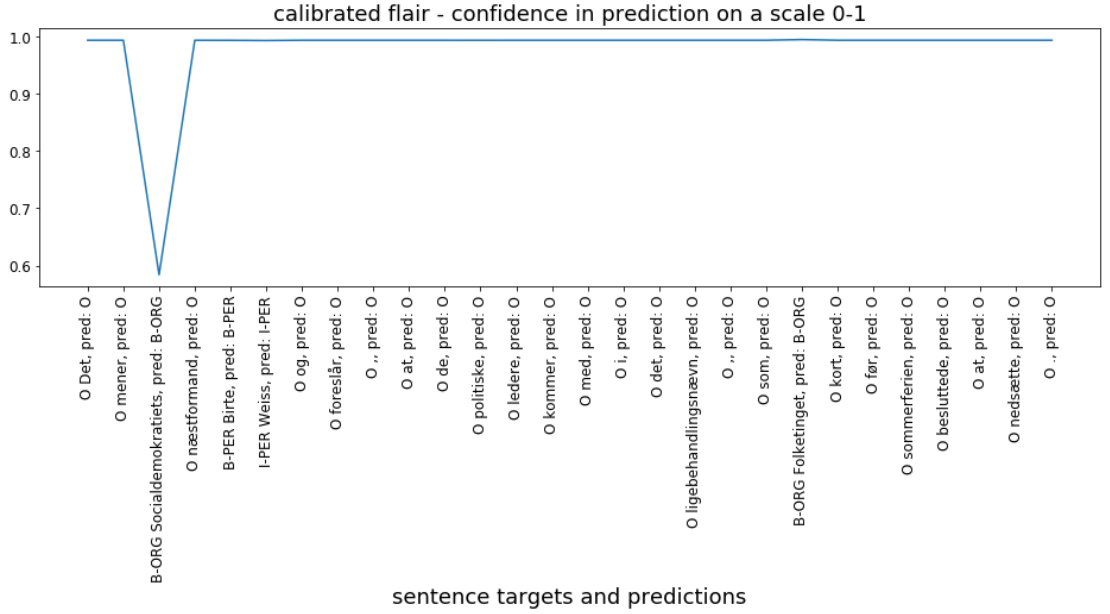
flair - Sentence 2



Calibrated flair - Sentence 2



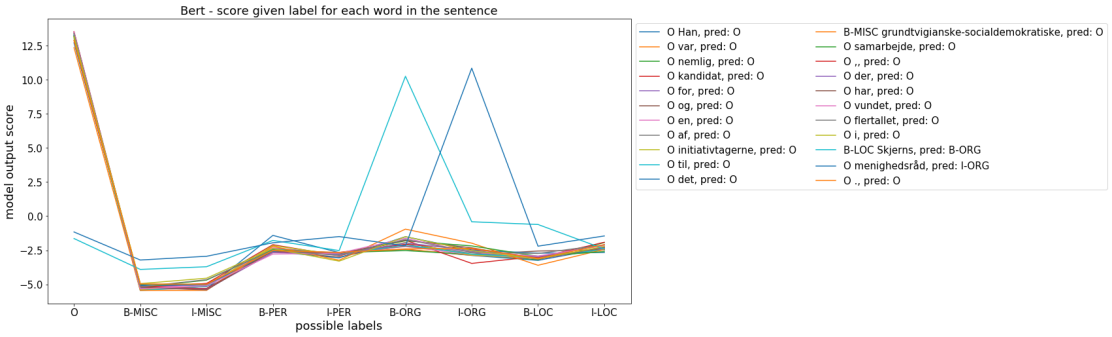


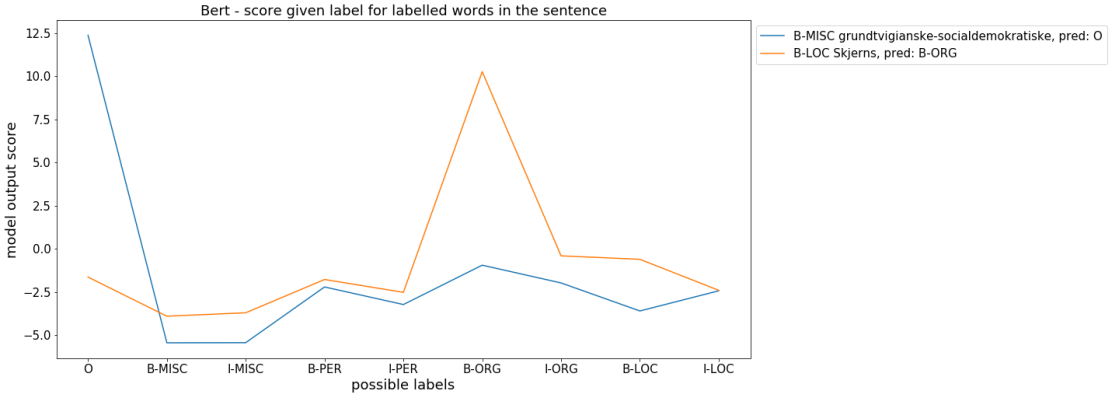


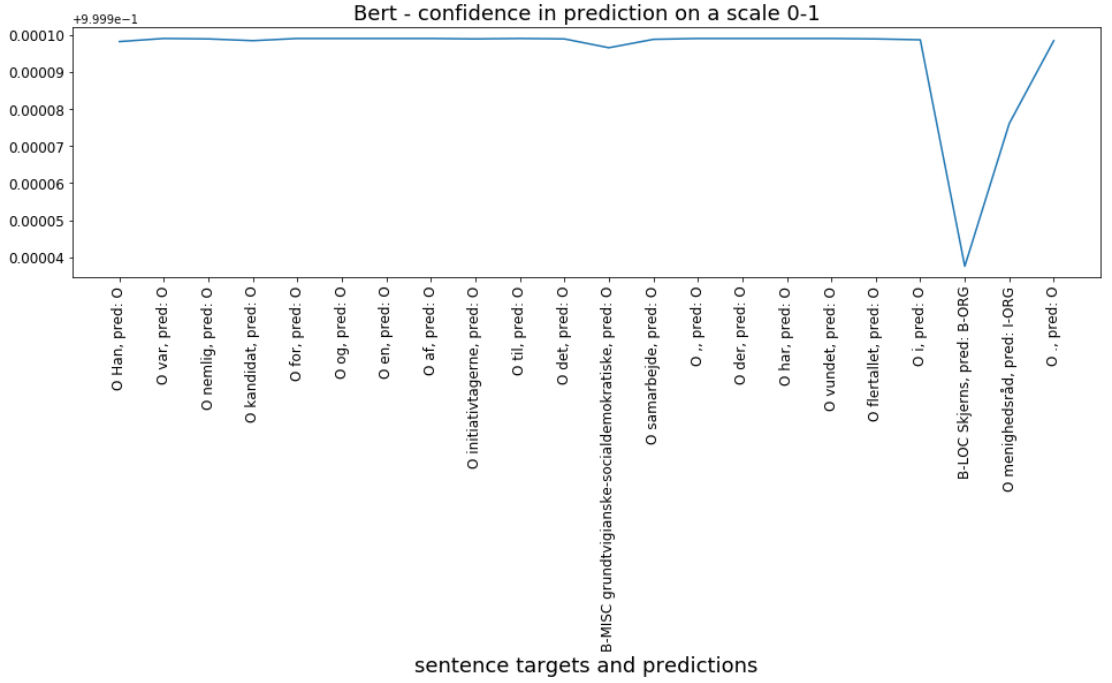
Sentence	Han	var	nemlig	kandidat	for	og	en	af	initiativtagerne	til	det	grundtvigianske-socialdemokratiske	samarbejde	,	der	har	vundet	flertallet	i	Skjems	menighedsråd	.
Targets	O	O	O	O	O	O	O	O	O	O	O	B-MISC	O	O	O	O	O	O	O	B-LOC	O	O
BERT	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-ORG	I-ORG	O
flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-ORG	I-ORG	O
Calibrated flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-ORG	I-ORG	O

Table 3: Sentence 3

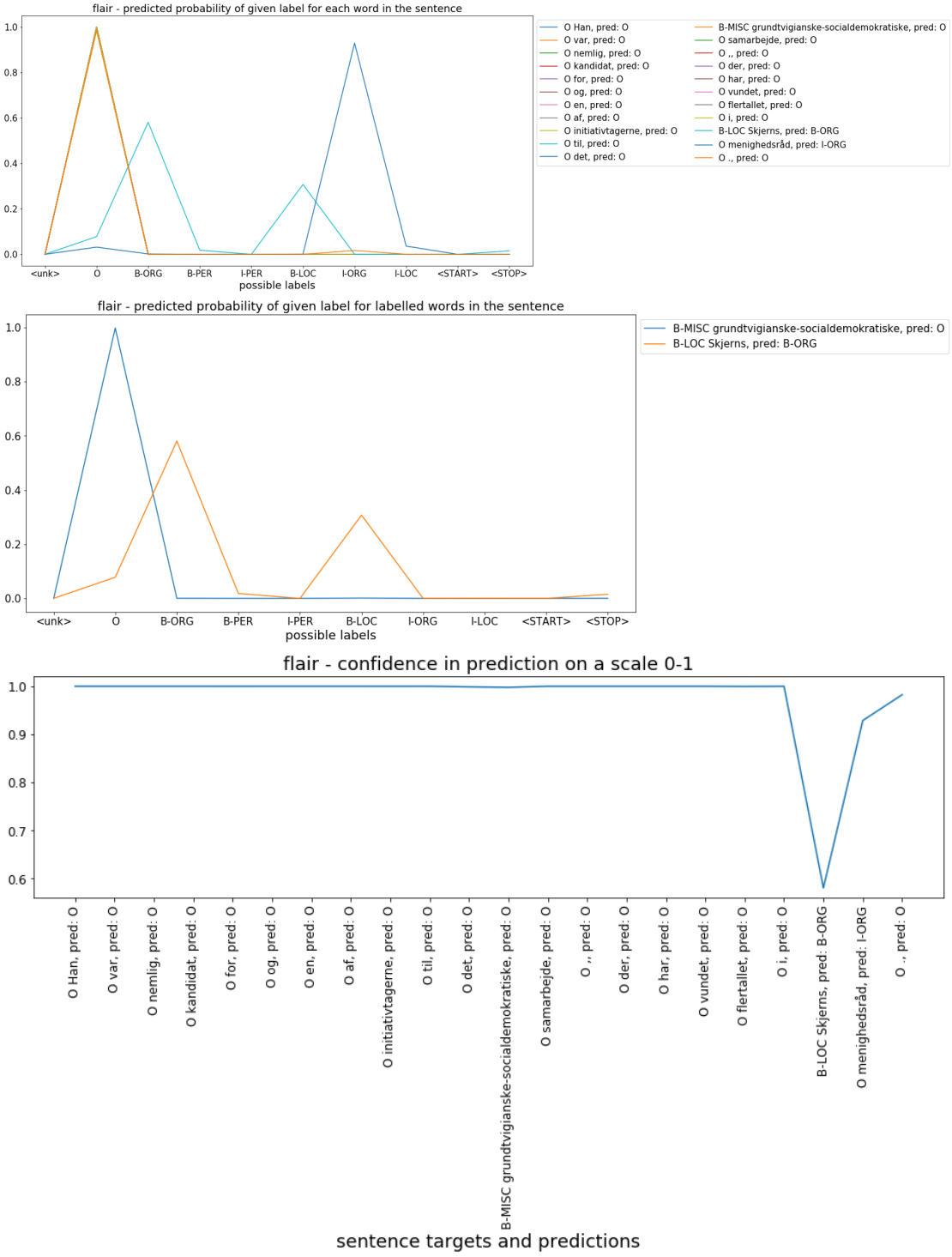
BERT - Sentence 3



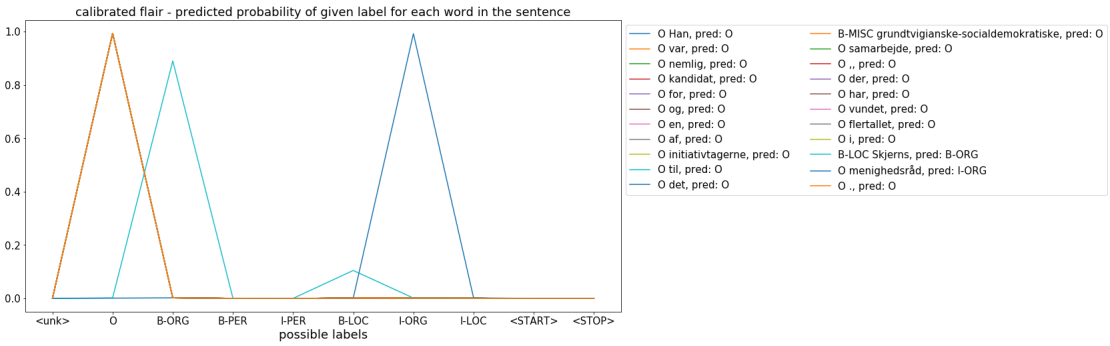


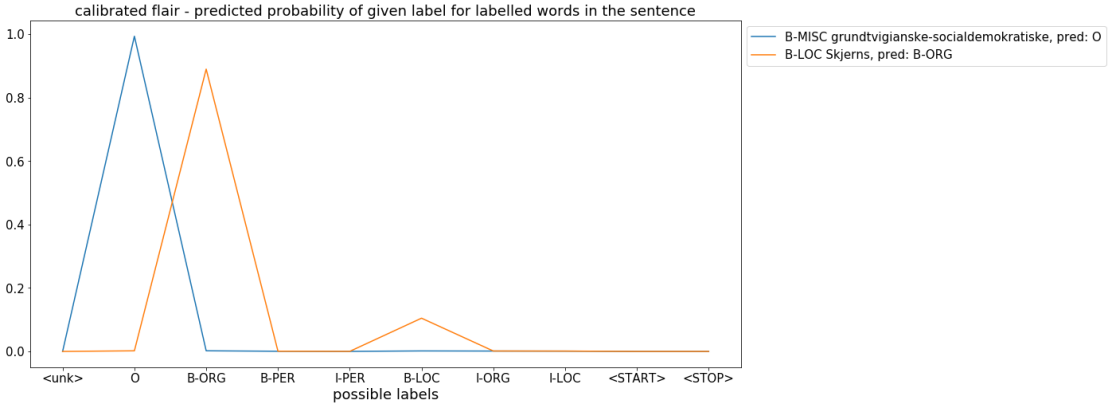


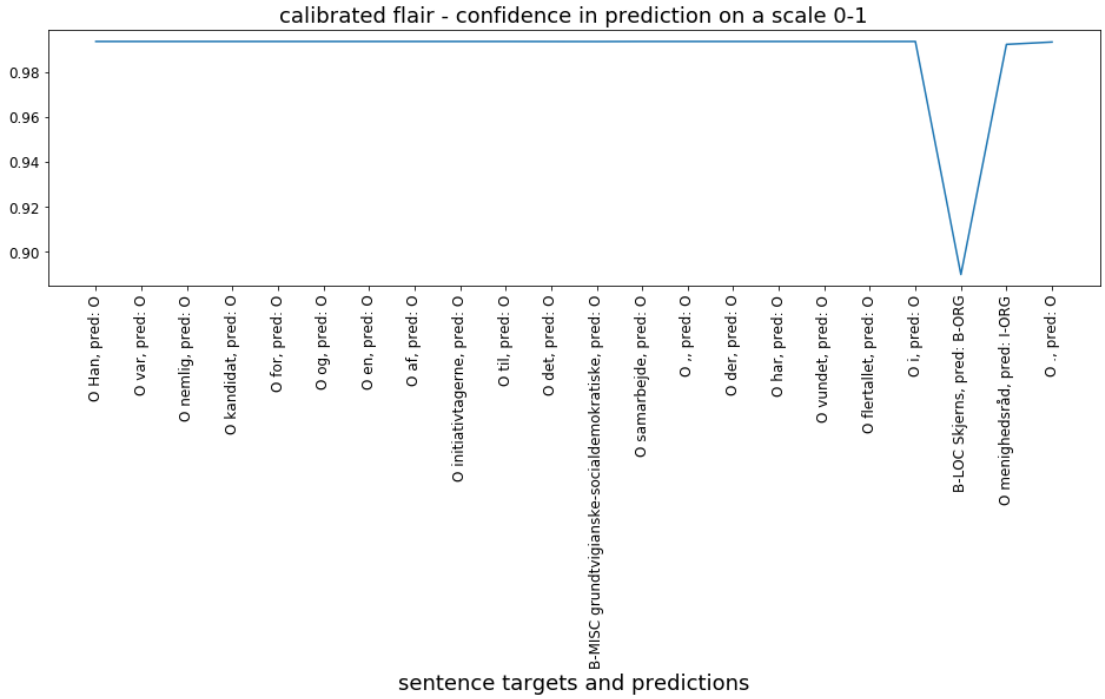
flair - Sentence 3



Calibrated flair - Sentence 3



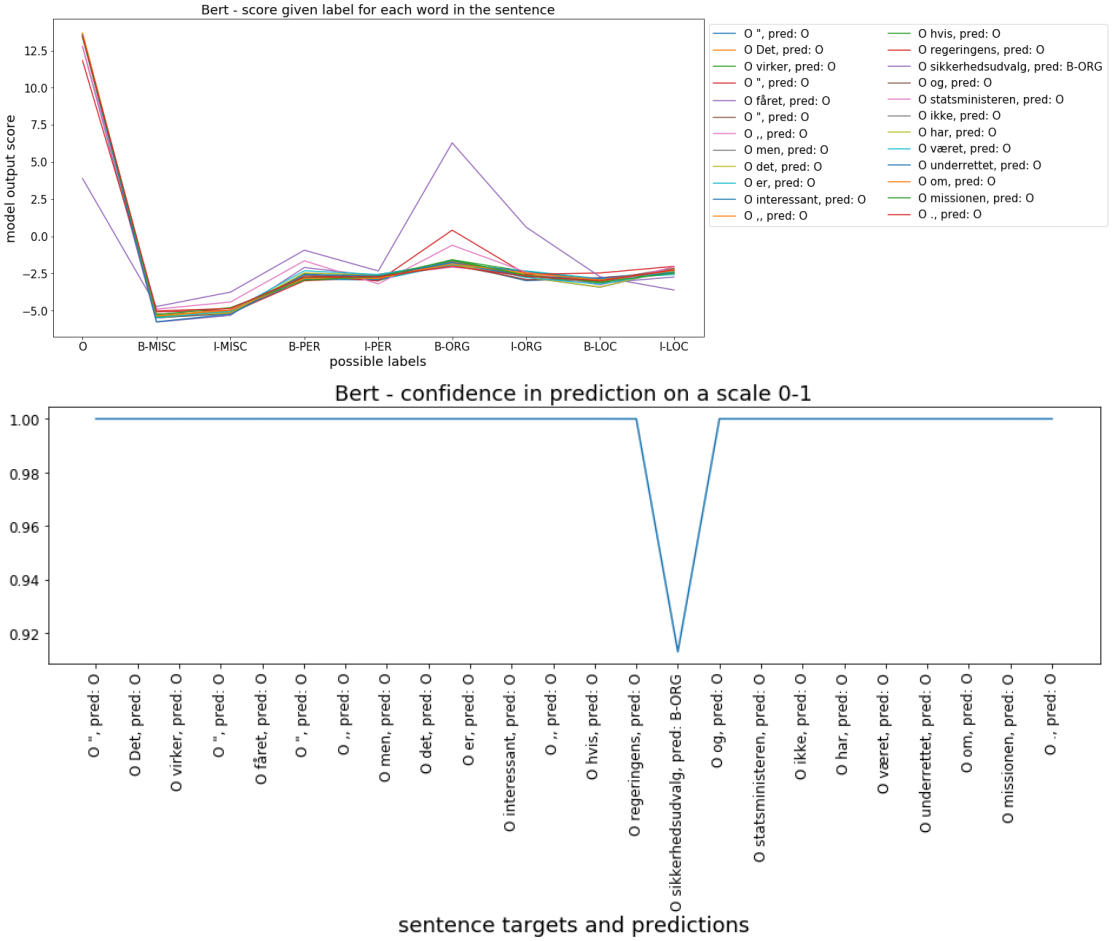




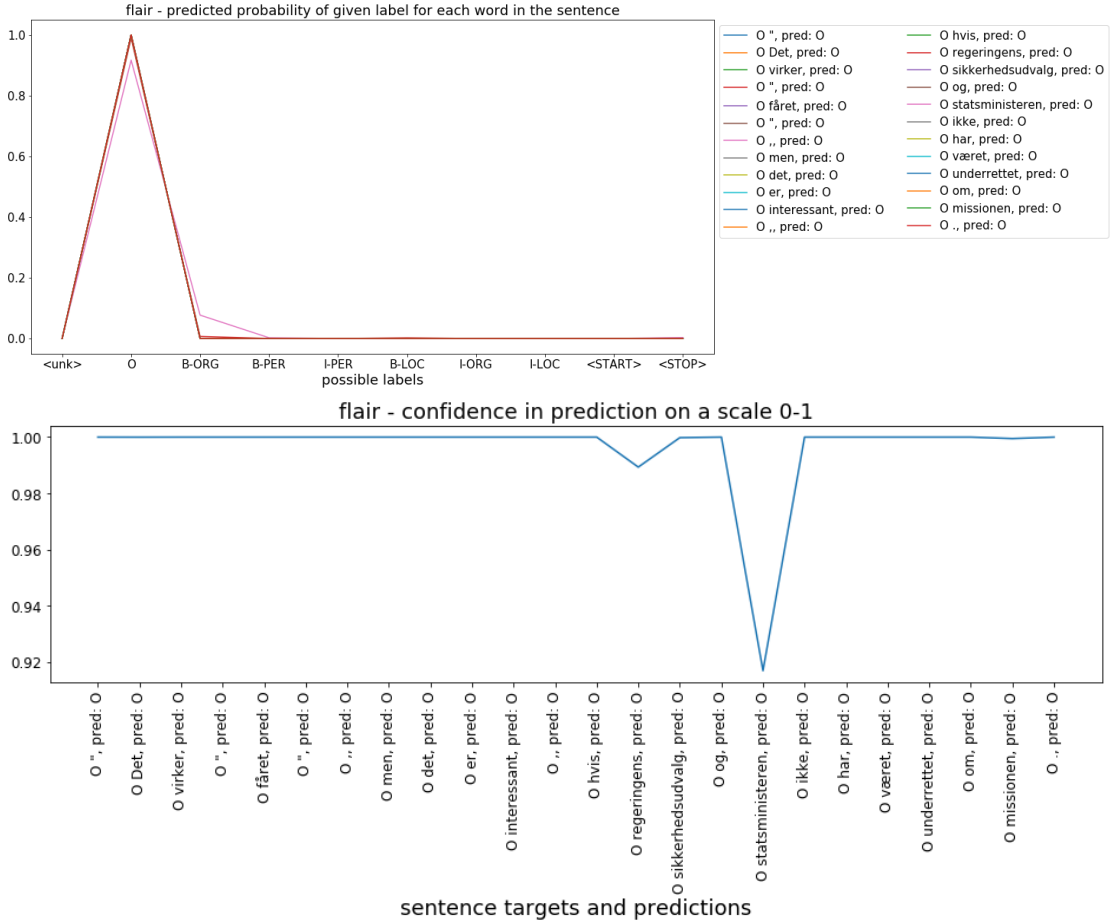
Sentence	"	Det	virker	"	føret	"	.	men	det	er	interessant	.	hvis	regeringens	sikkerhedsudvalg	og	statsministeren	ikke	har	været	underrettet	om	missionen	.
Targets	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
BERT	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-ORG	O	O	O	O	O	O	O	O	O
flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Calibrated flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Table 4: Sentence 4

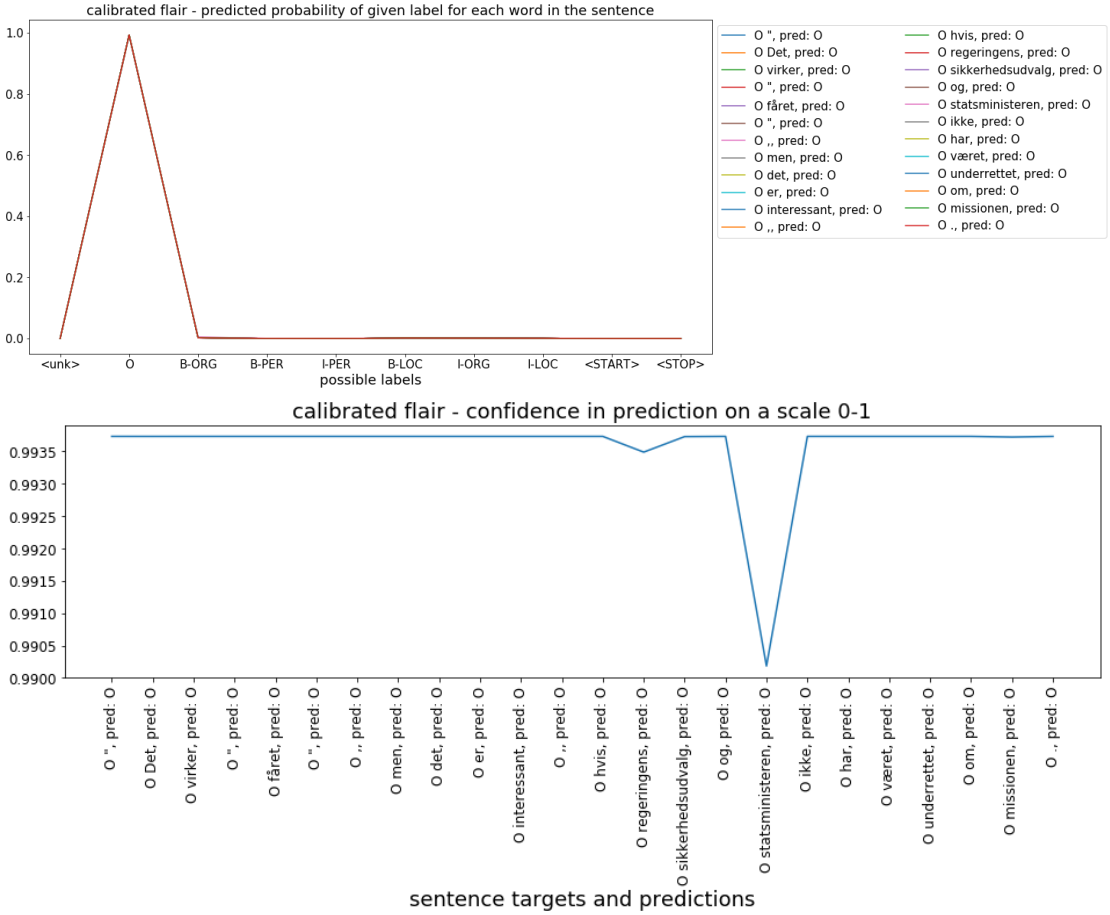
BERT - Sentence 4



flair - Sentence 4



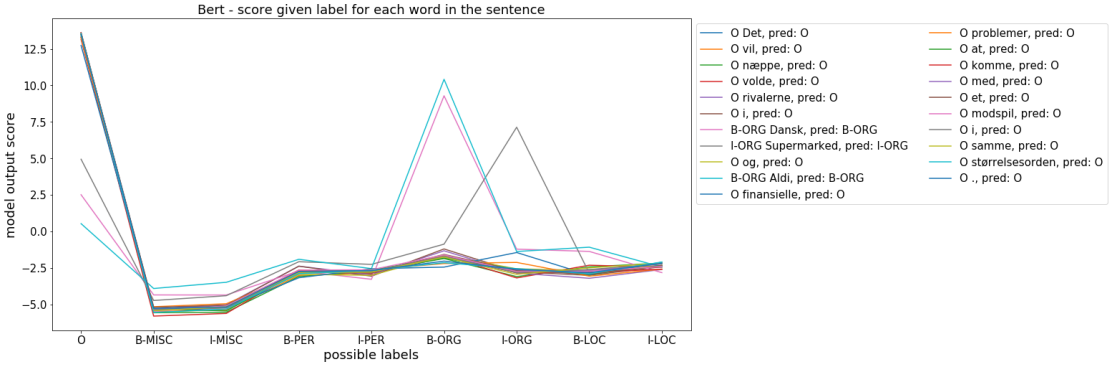
Calibrated flair - Sentence 4

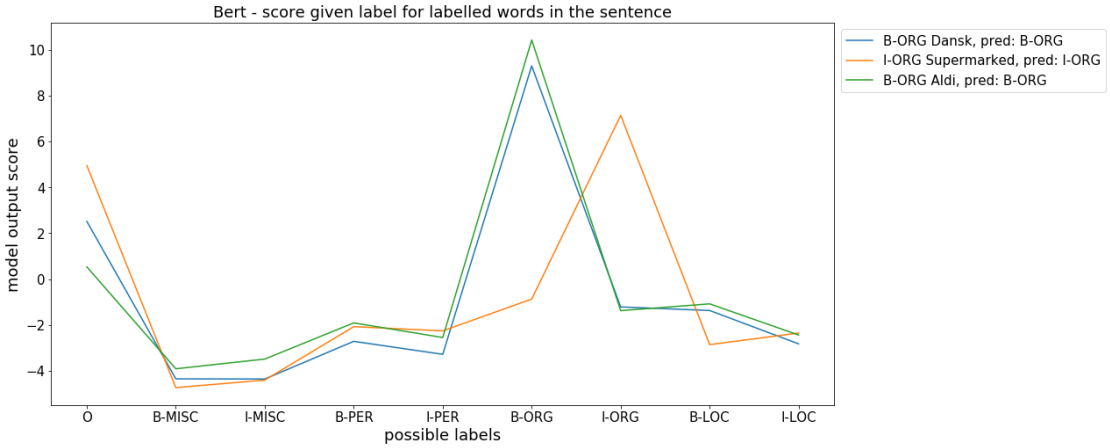


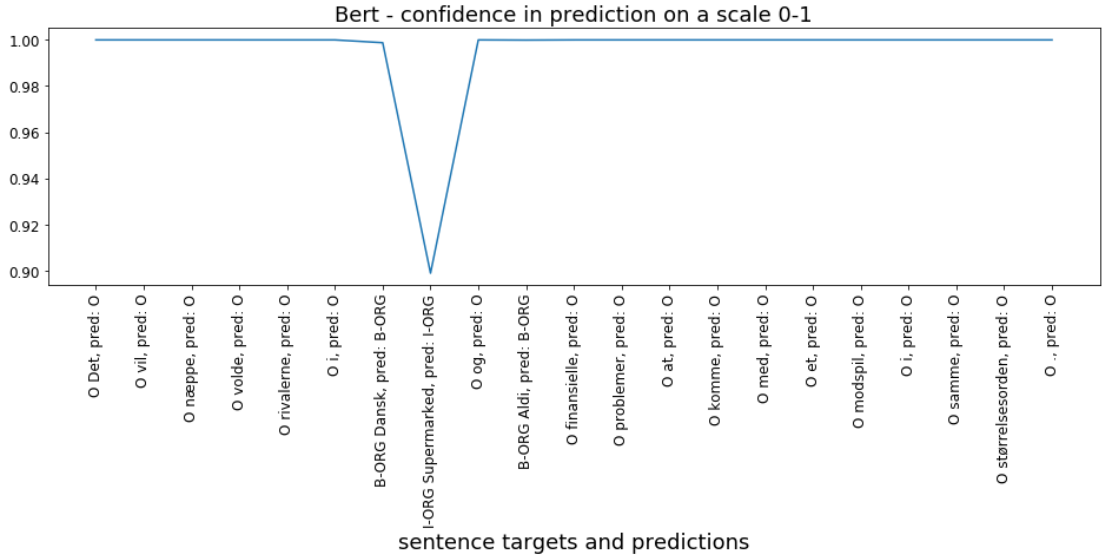
Sentence	Det	vil	næppe	volde	rivalerne	i	Dansk	Supermarked	og	Aldi	finansielle	problemer	at	komme	med	et	modspil	i	samme	størrelsesorden	.
Targets	O	O	O	O	O	O	B-ORG	I-ORG	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O
BERT	O	O	O	O	O	O	B-ORG	I-ORG	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O
flair	O	O	O	O	O	O	B-ORG	I-ORG	O	O	O	O	O	O	O	O	O	O	O	O	O
Calibrated flair	O	O	O	O	O	O	B-ORG	I-ORG	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O

Table 5: Sentence 5

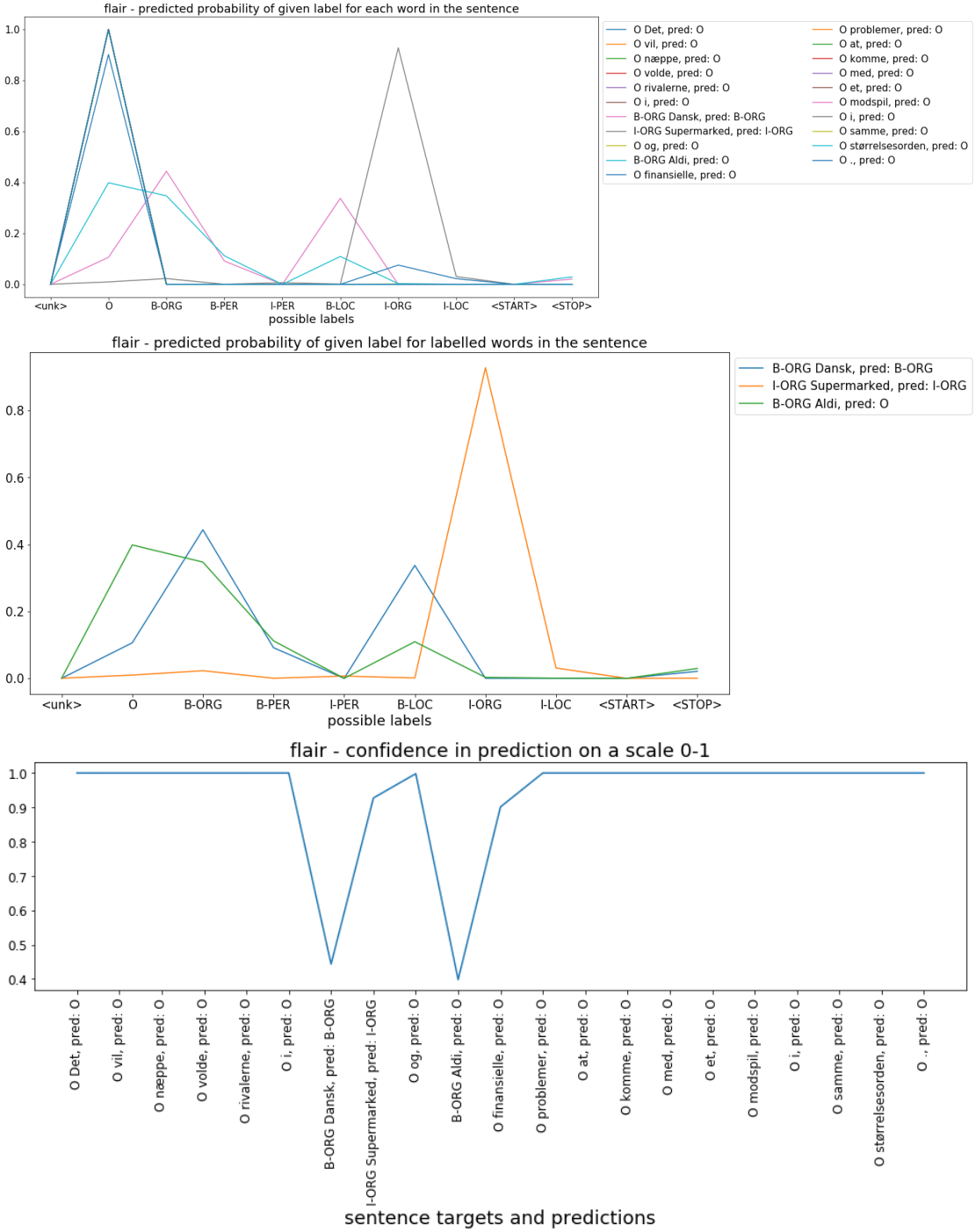
BERT - Sentence 5



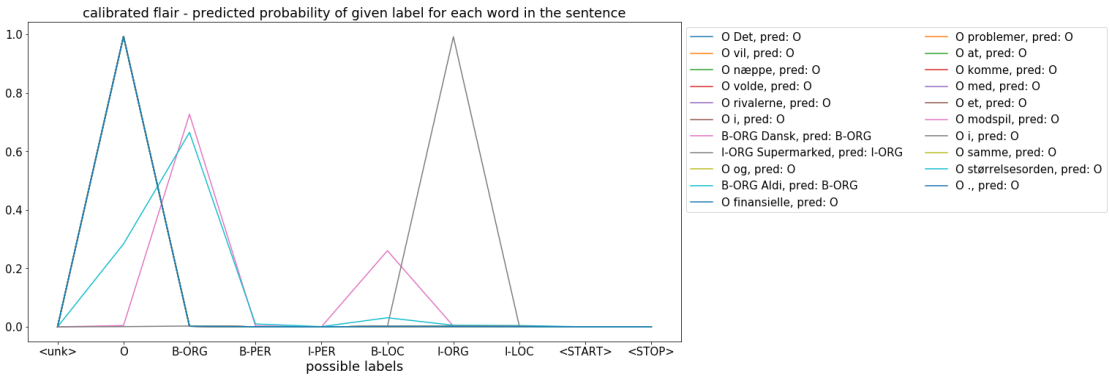


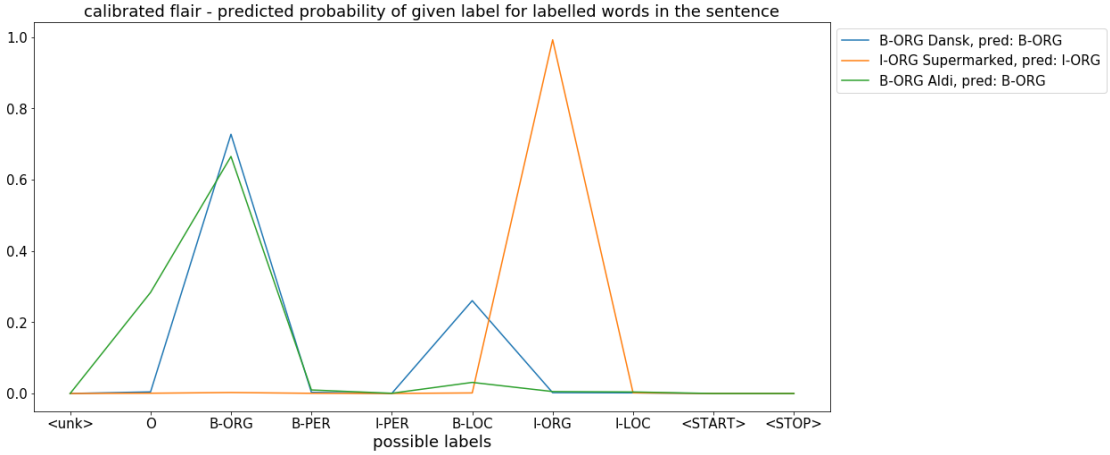


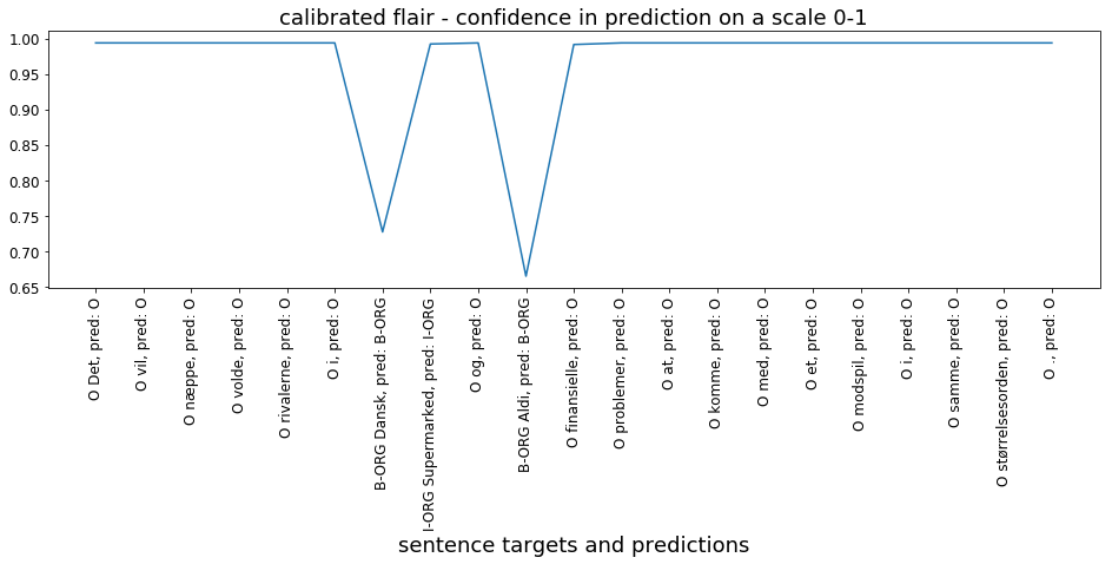
flair - Sentence 5



Calibrated flair - Sentence 5



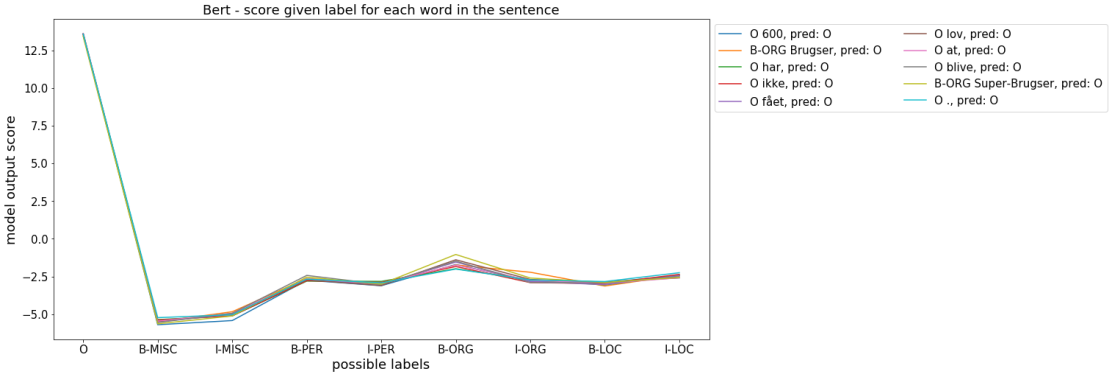


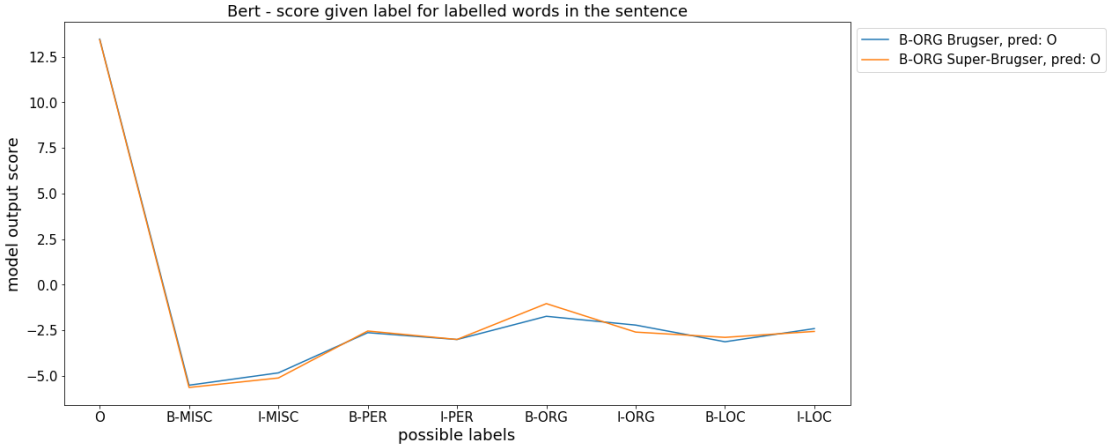


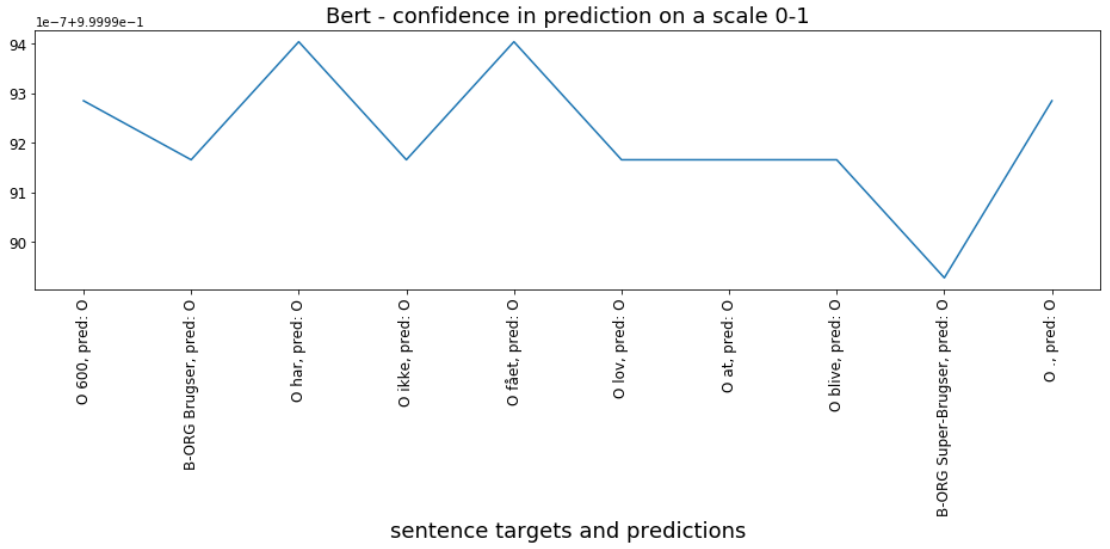
Sentence	600	Brugser	har	ikke	fået	lov	at	blive	Super-Brugser	.
Targets	O	B-ORG	O	O	O	O	O	O	B-ORG	O
BERT	O	O	O	O	O	O	O	O	O	O
flair	O	O	O	O	O	O	O	O	O	O
Calibrated flair	O	O	O	O	O	O	O	O	O	O

Table 6: Sentence 6

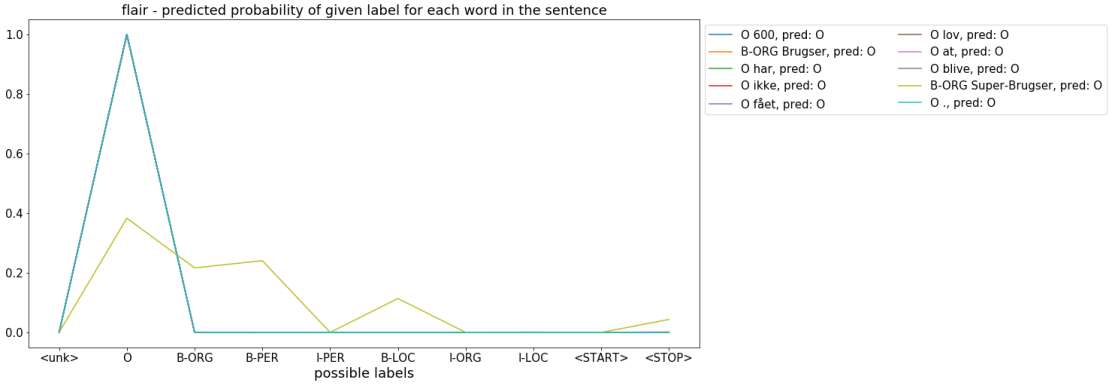
BERT - Sentence 6

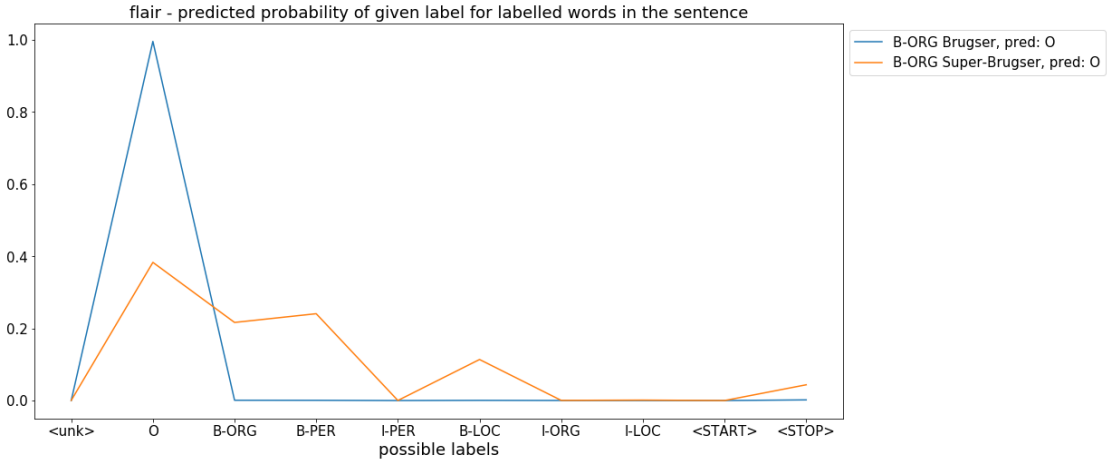


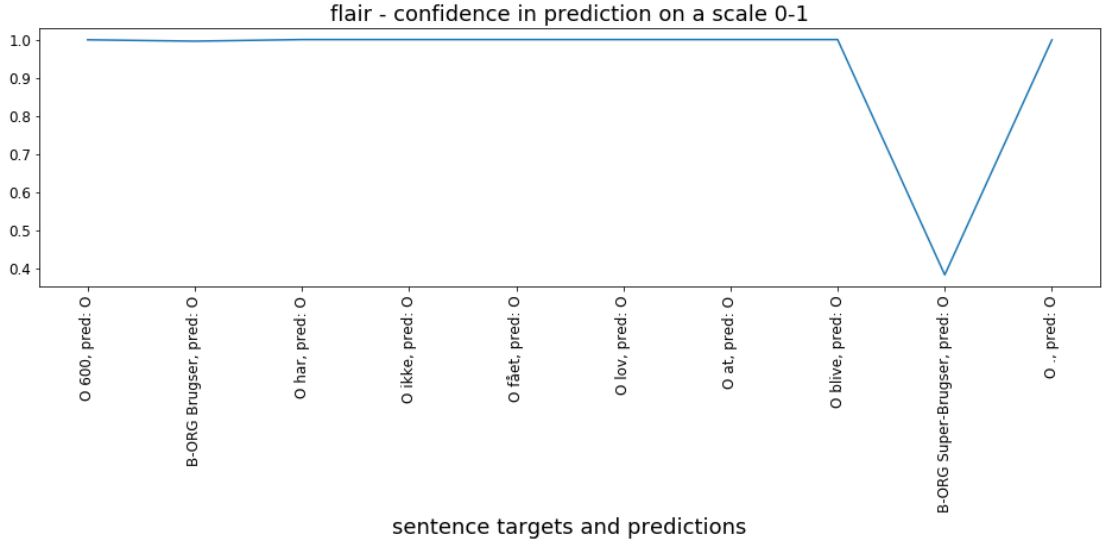




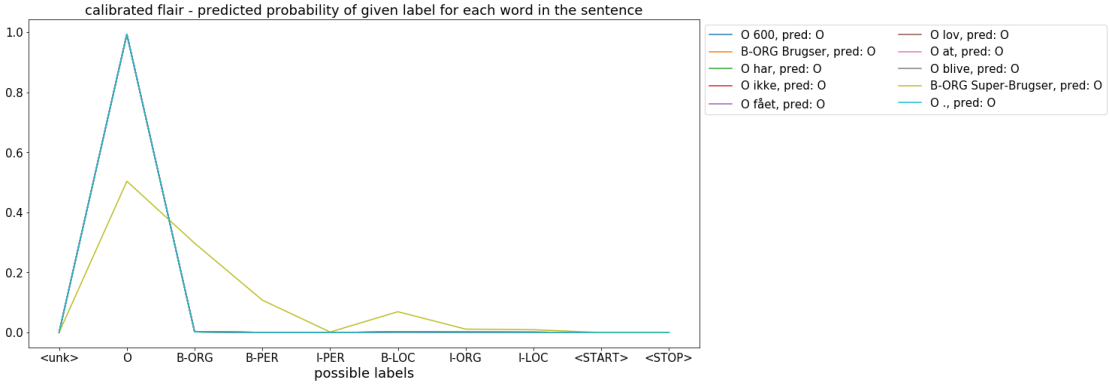
flair - Sentence 6

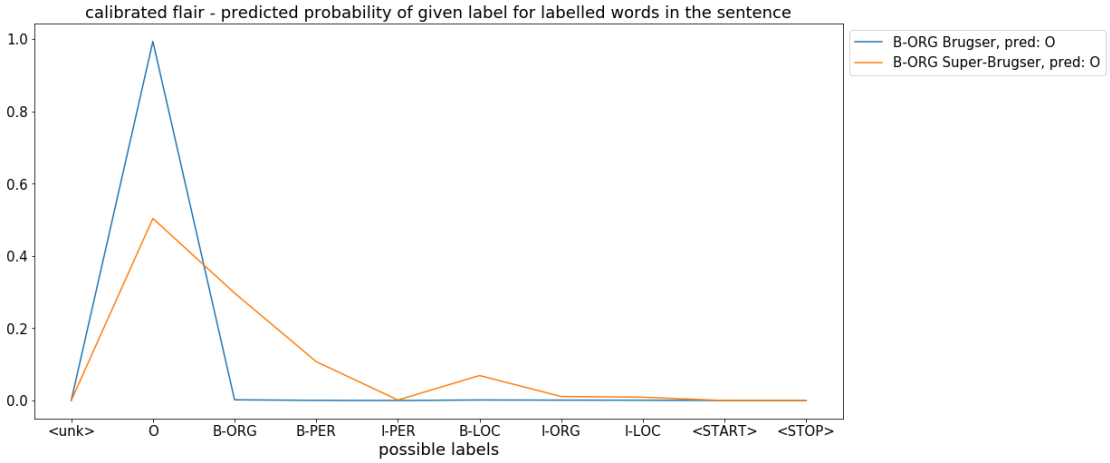


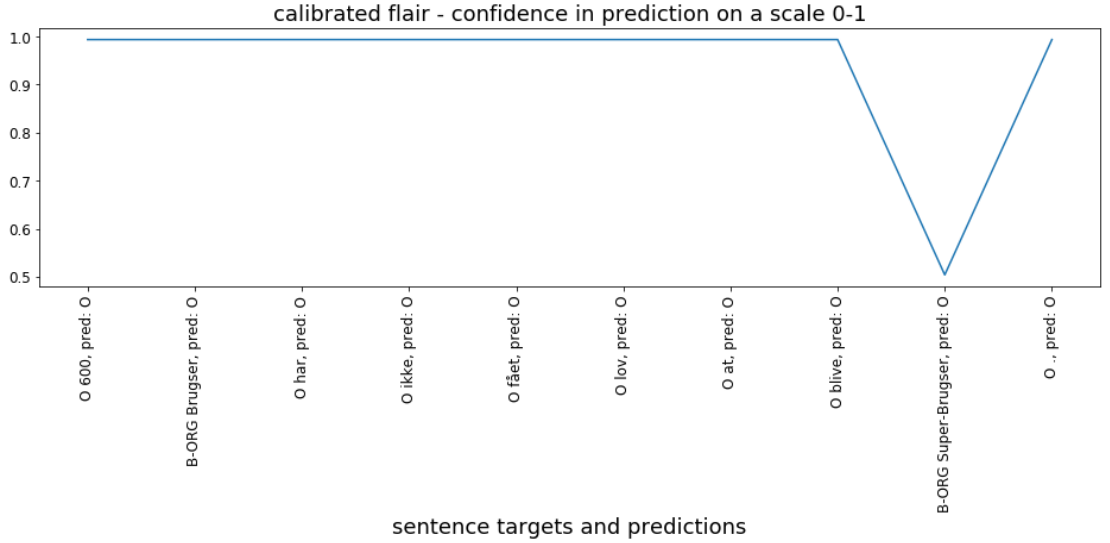




Calibrated flair - Sentence 6



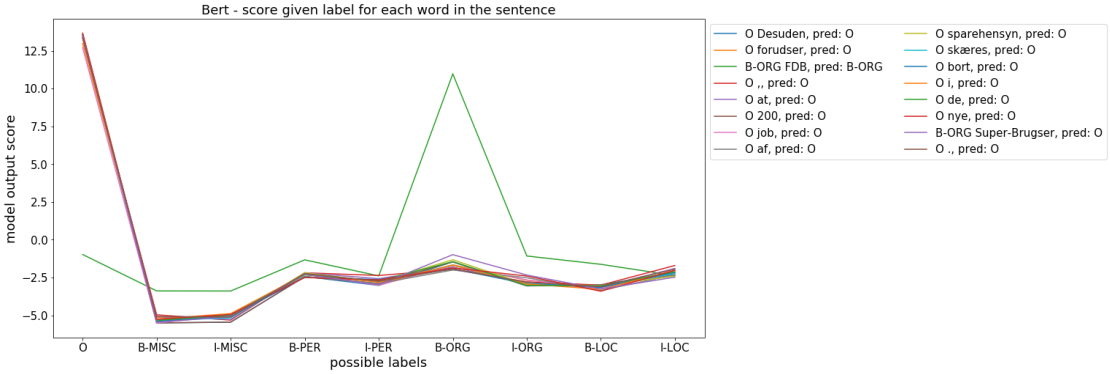


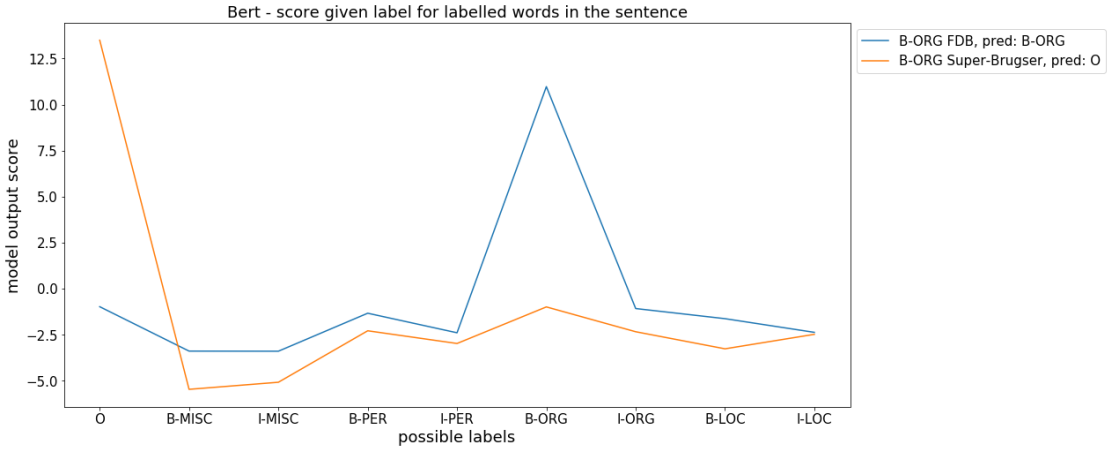


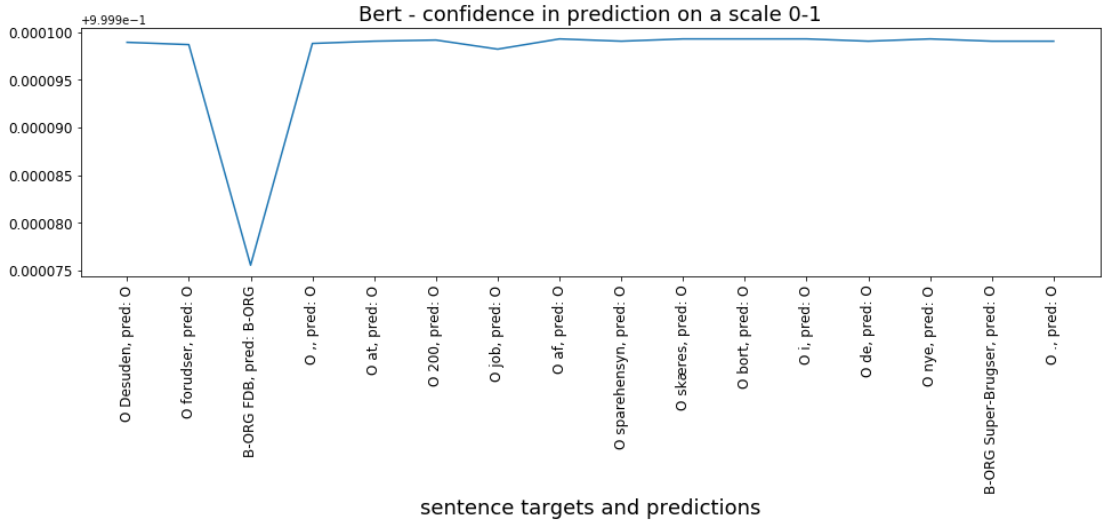
Sentence	Desuden	forudser	FDB	,	at	200	job	af	sparehensyn	skæres	bort	i	de	nye	Super-Brugser	.
Targets	O	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O	B-ORG	O
BERT	O	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O	O	O
flair	O	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O	O	O
Calibrated flair	O	O	B-ORG	O	O	O	O	O	O	O	O	O	O	O	B-ORG	O

Table 7: Sentence 7

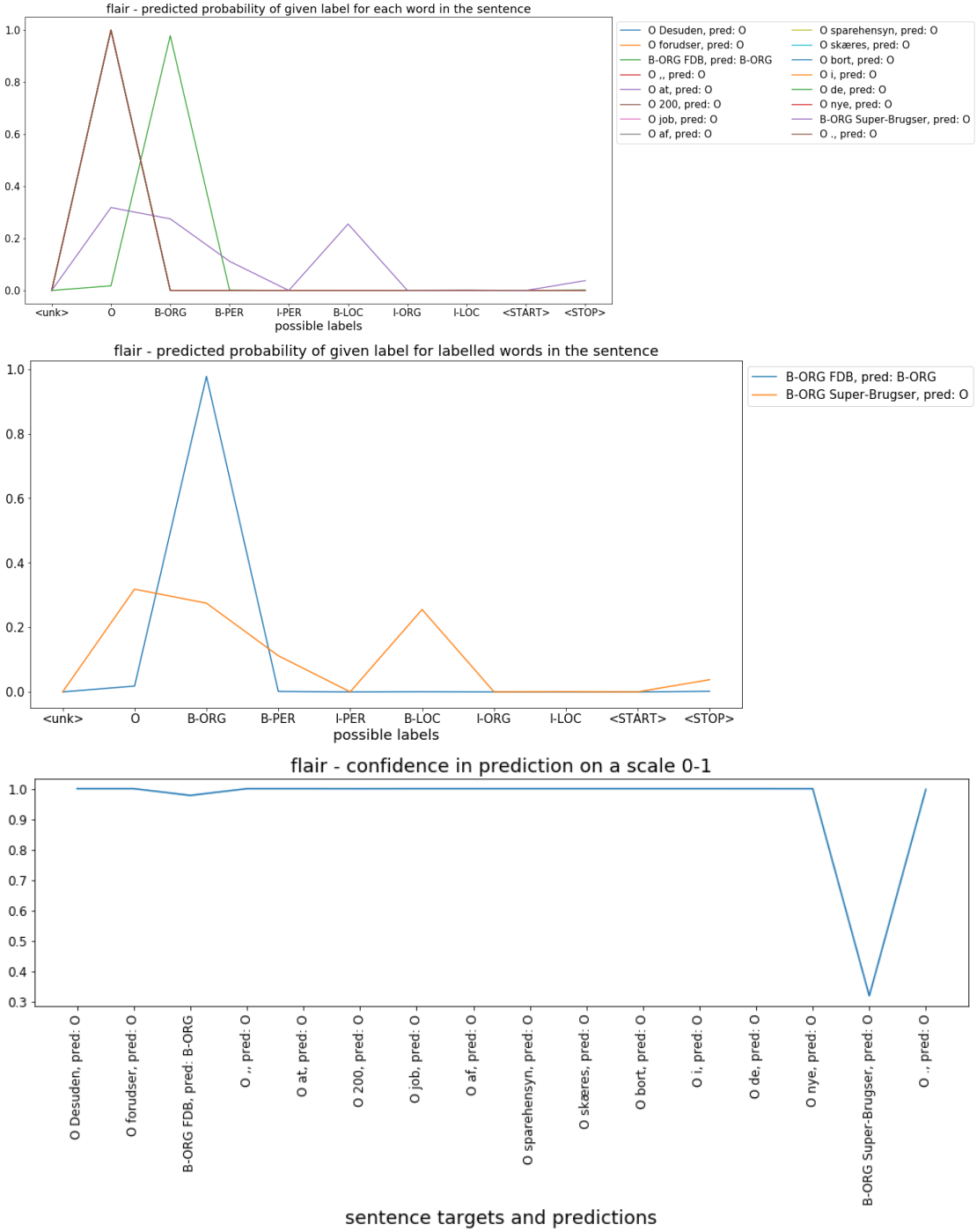
BERT - Sentence 7



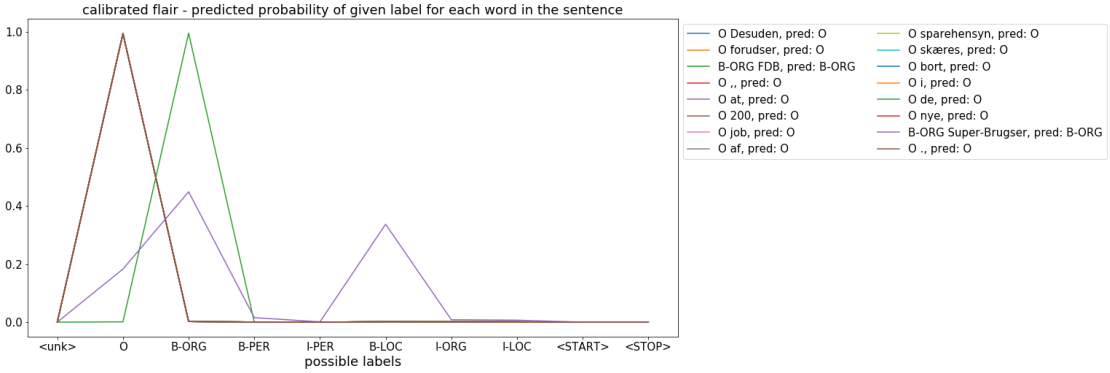


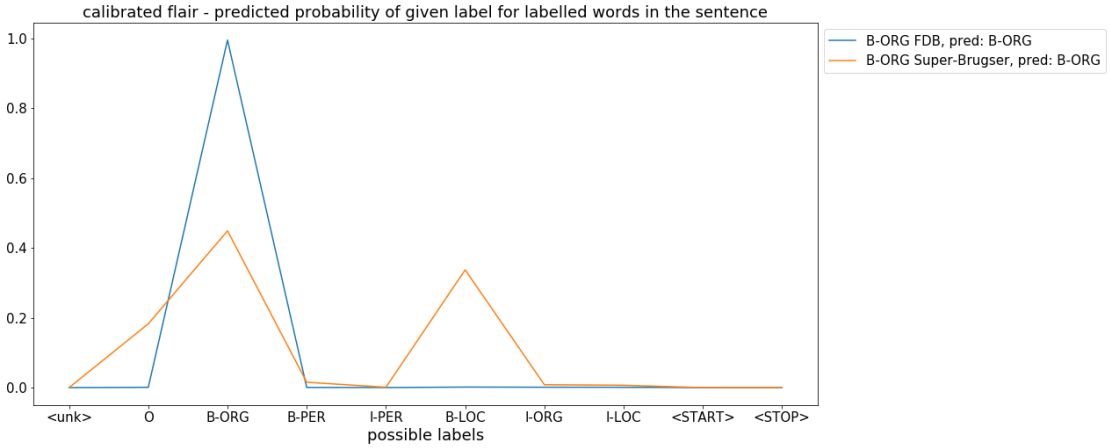


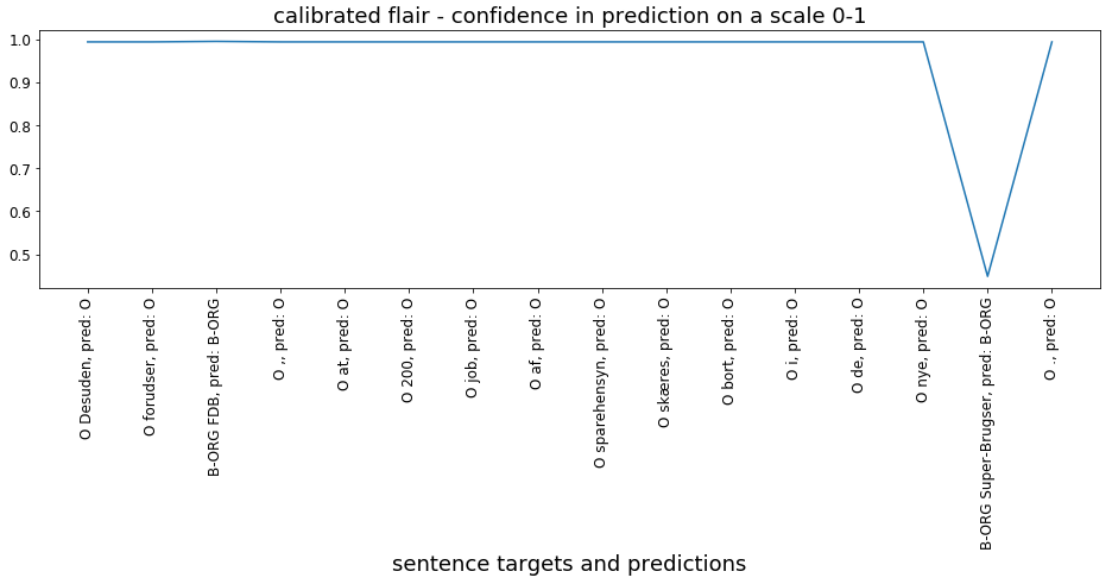
flair - Sentence 7



Calibrated flair - Sentence 7



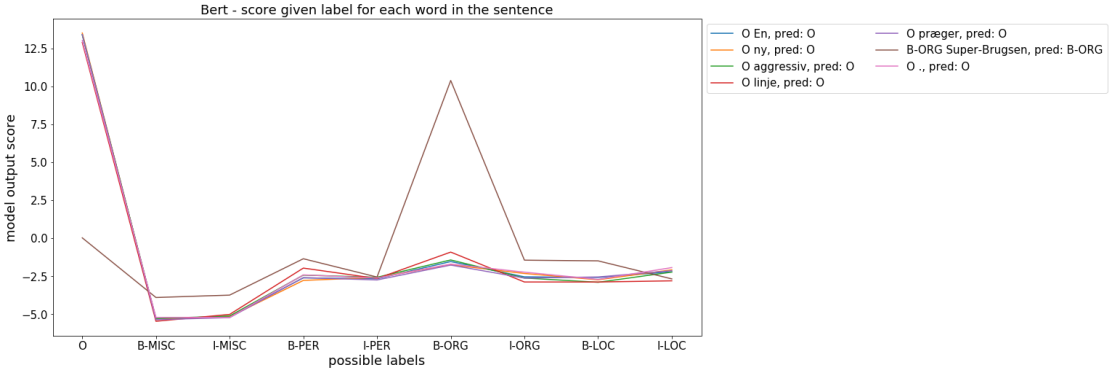


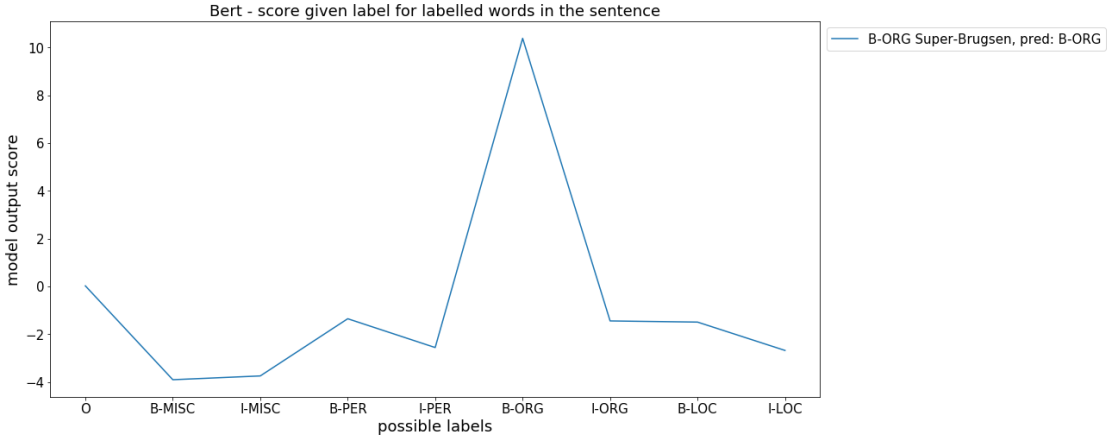


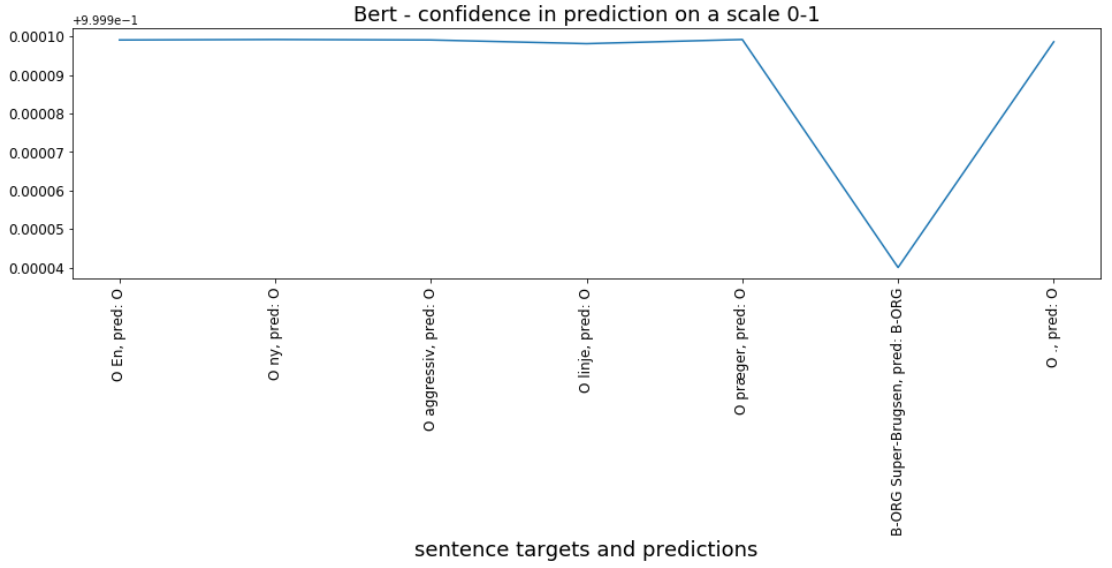
Sentence	En	ny	aggressiv	linje	præger	Super-Brugsen	.
Targets	O	O	O	O	O	B-ORG	O
BERT	O	O	O	O	O	B-ORG	O
flair	O	O	O	O	O	O	O
Calibrated flair	O	O	O	O	O	B-LOC	O

Table 8: Sentence 8

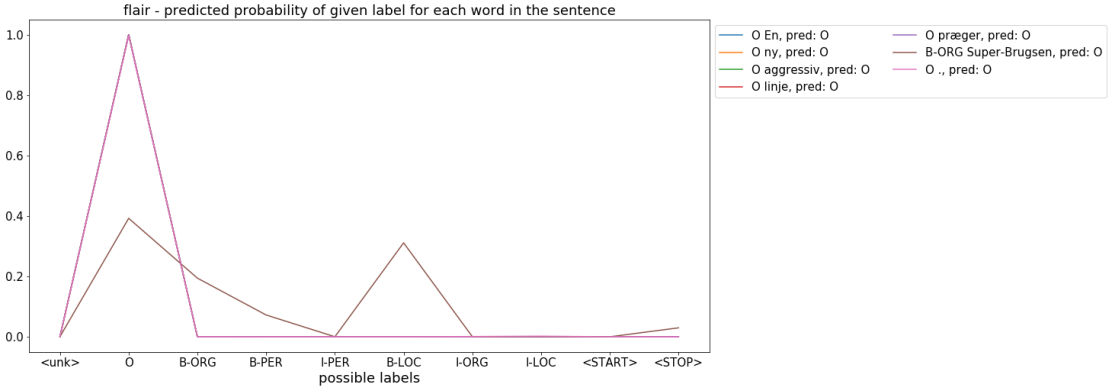
BERT - Sentence 8

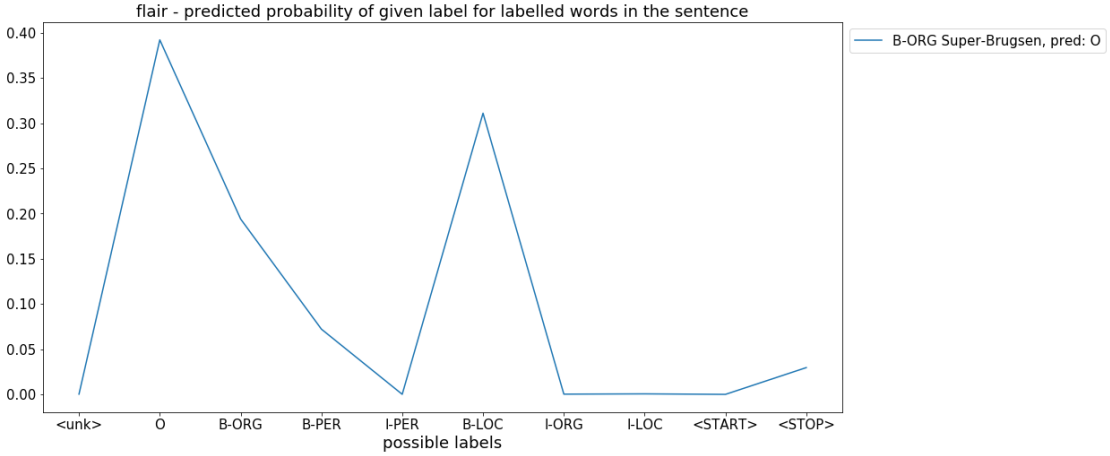


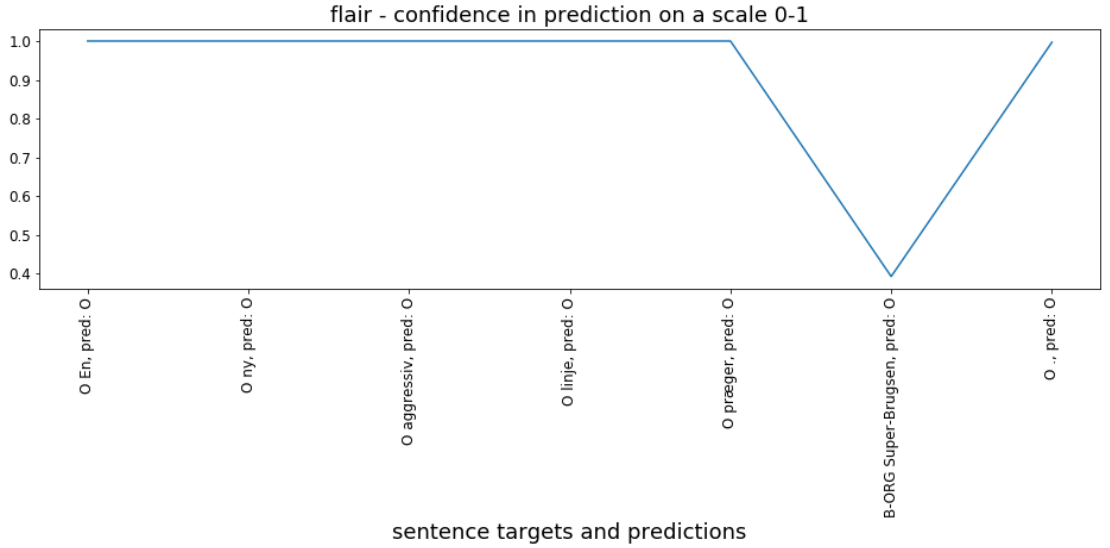




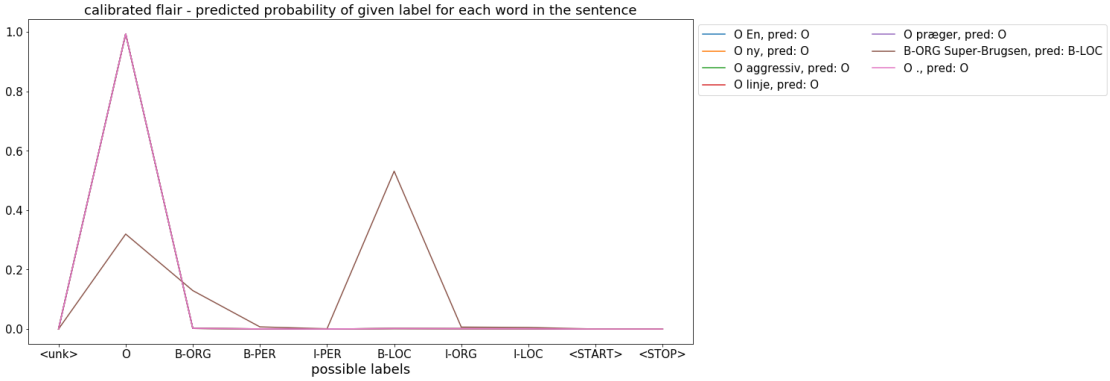
flair - Sentence 8

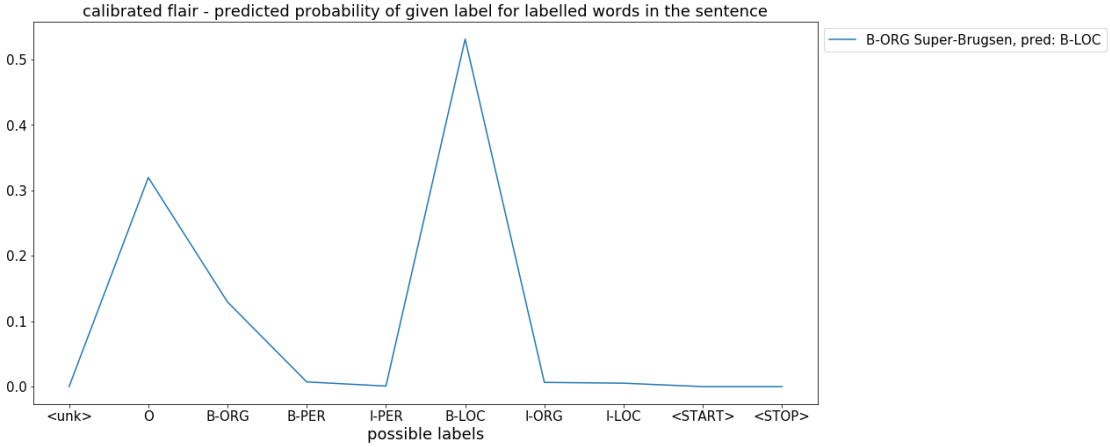


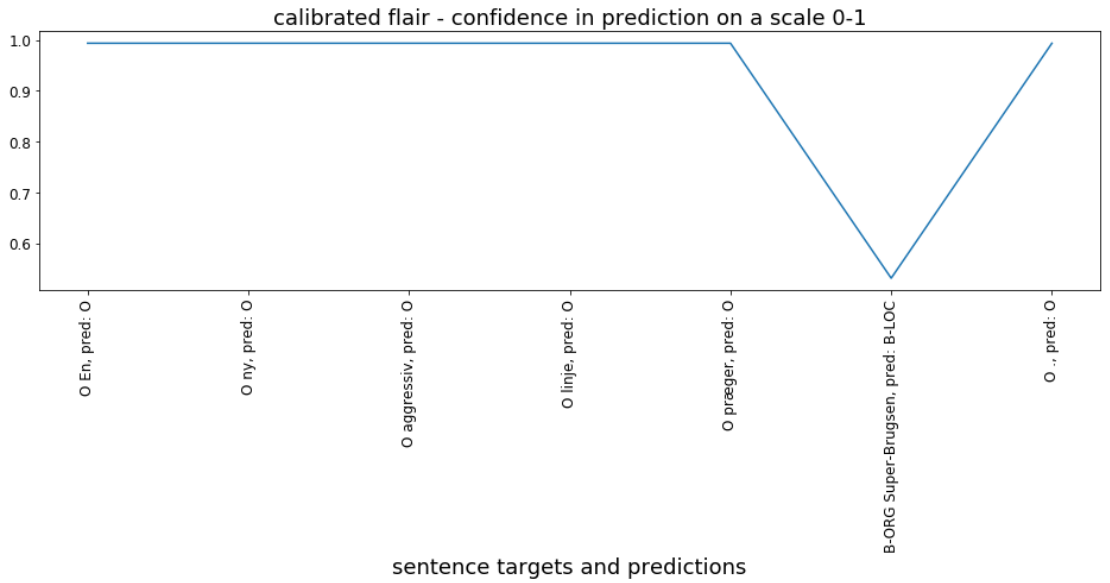




Calibrated flair - Sentence 8



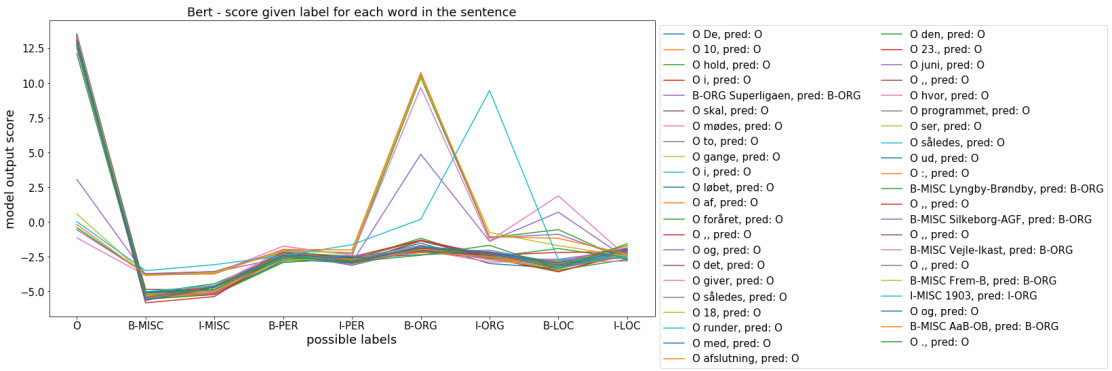


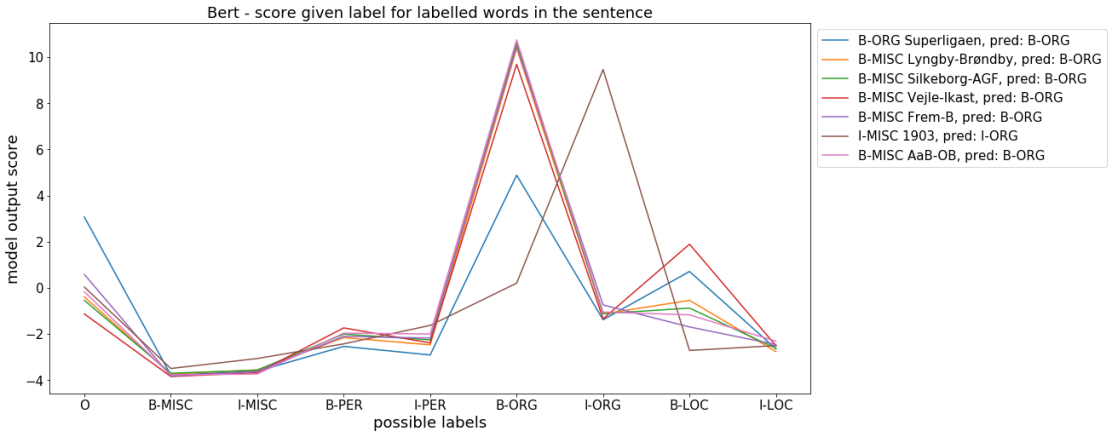


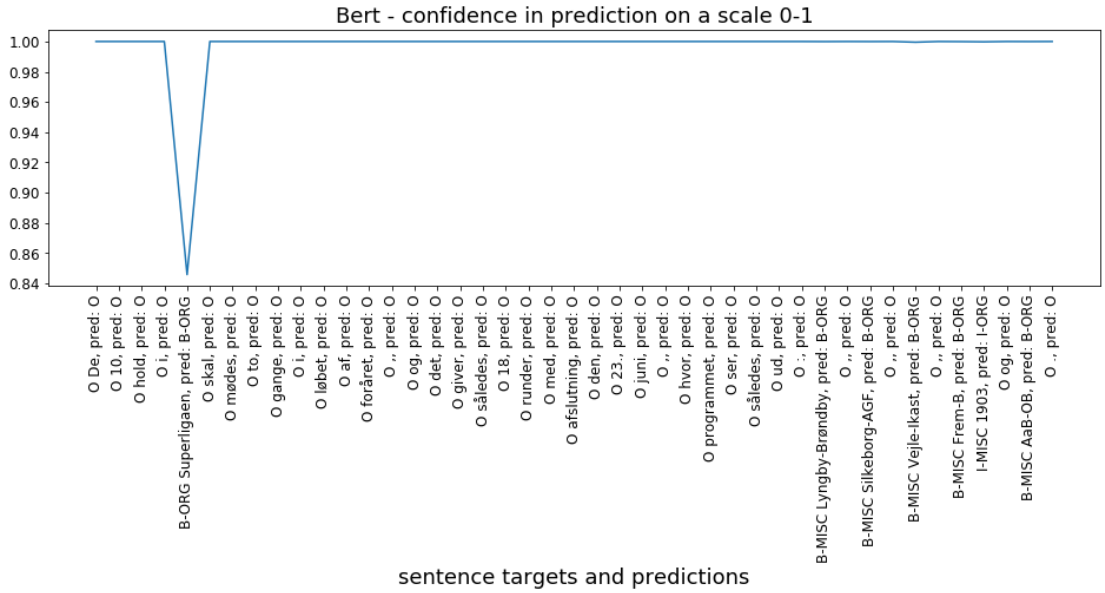
Sentence	De	10	hold	i	Superligaen	...	ud	:	Lyngby-Brøndby	,	Silkeborg-AGF	,	Vejle-Ikast	,	Frem-B	1903	og	AaB-OB	.
Targets	O	O	O	O	B-ORG	...	O	O	B-MISC	O	B-MISC	O	B-MISC	O	B-MISC	I-MISC	O	B-MISC	O
BERT	O	O	O	O	B-ORG	...	O	O	B-ORG	O	B-ORG	O	B-ORG	O	B-ORG	I-ORG	O	B-ORG	O
flair	O	O	O	O	O	...	O	O	B-LOC	O	B-ORG	O	O	O	O	O	O	B-ORG	O
Calibrated flair	O	O	O	O	O	...	O	O	B-LOC	O	B-ORG	O	O	O	B-ORG	O	O	B-ORG	O

Table 9: Sentence 9

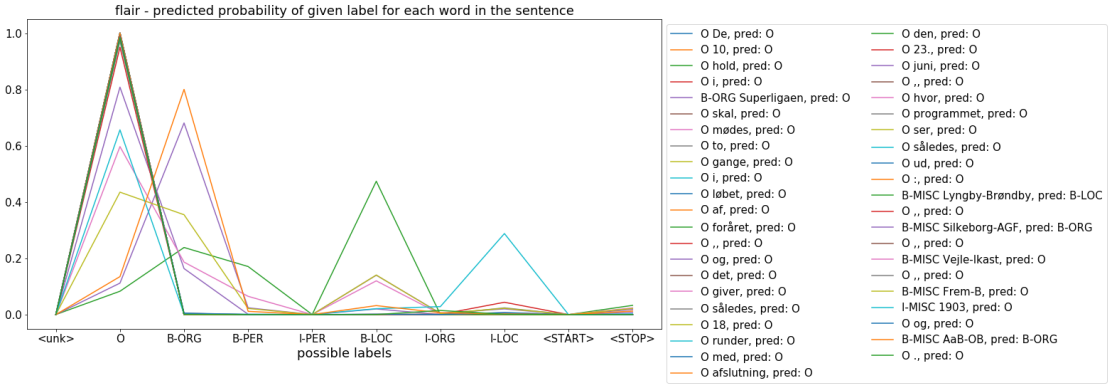
BERT - Sentence 9

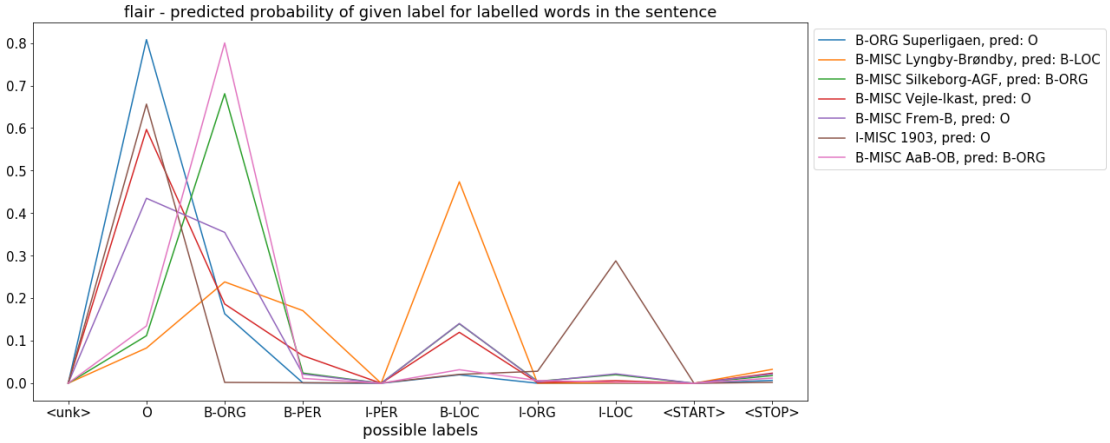


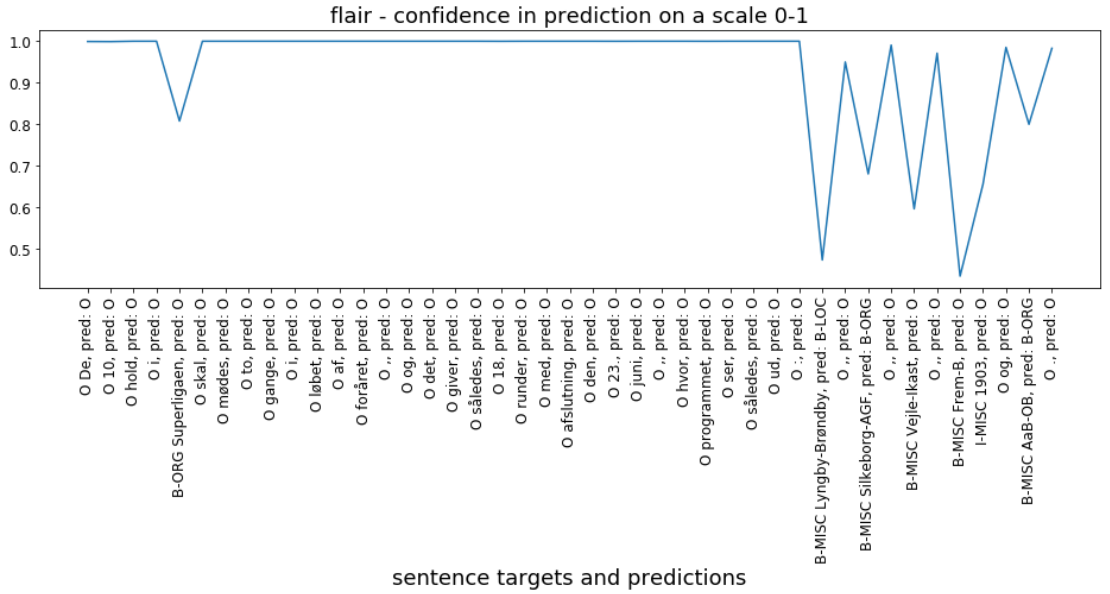




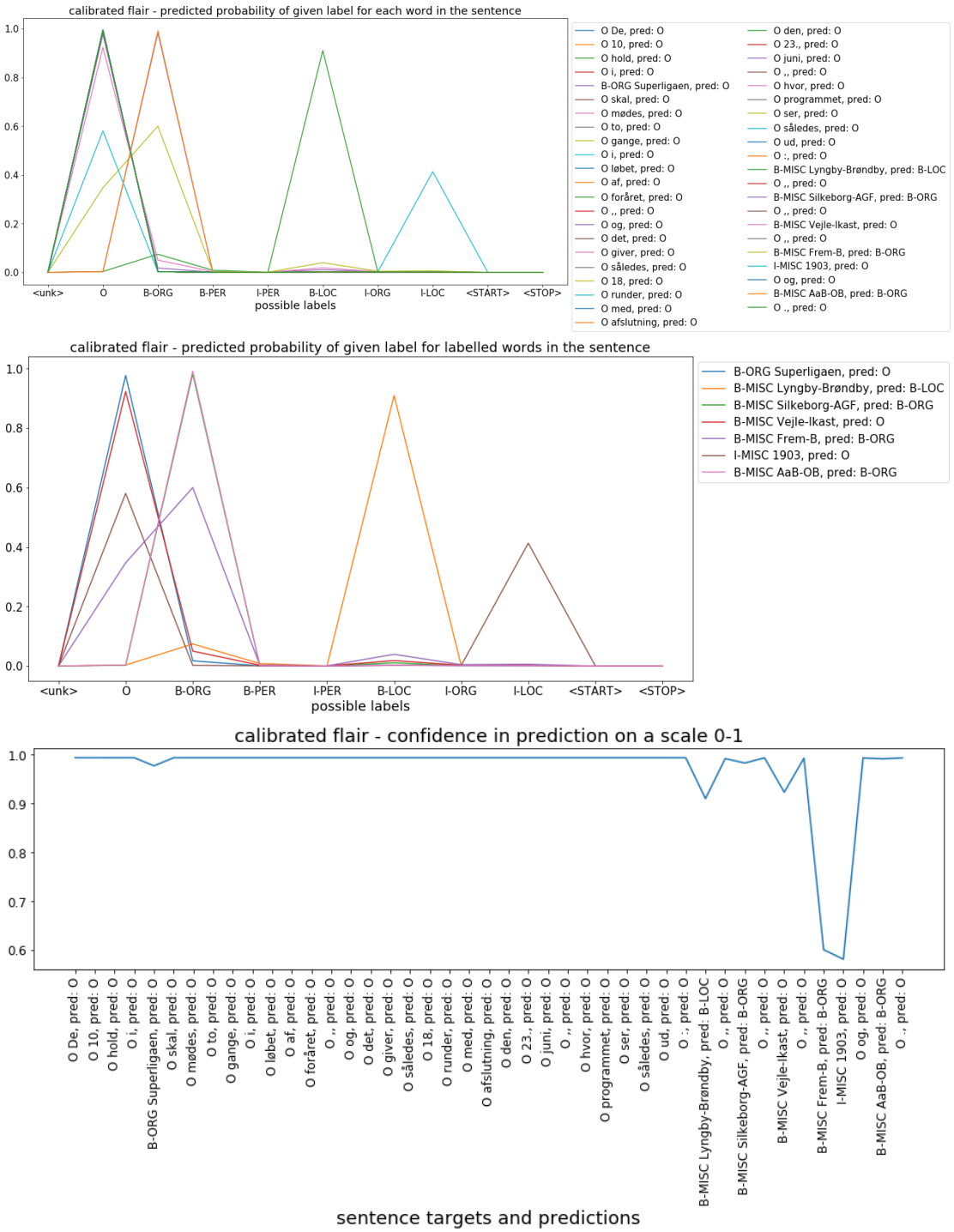
flair - Sentence 9







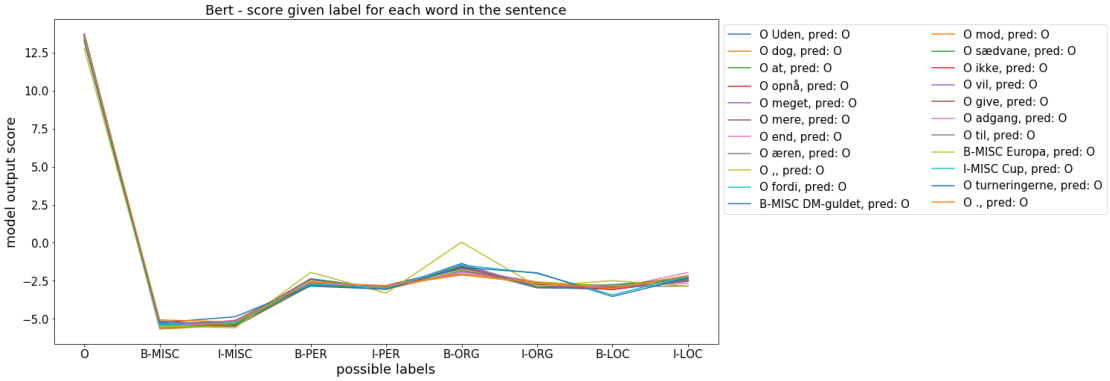
Calibrated flair - Sentence 9

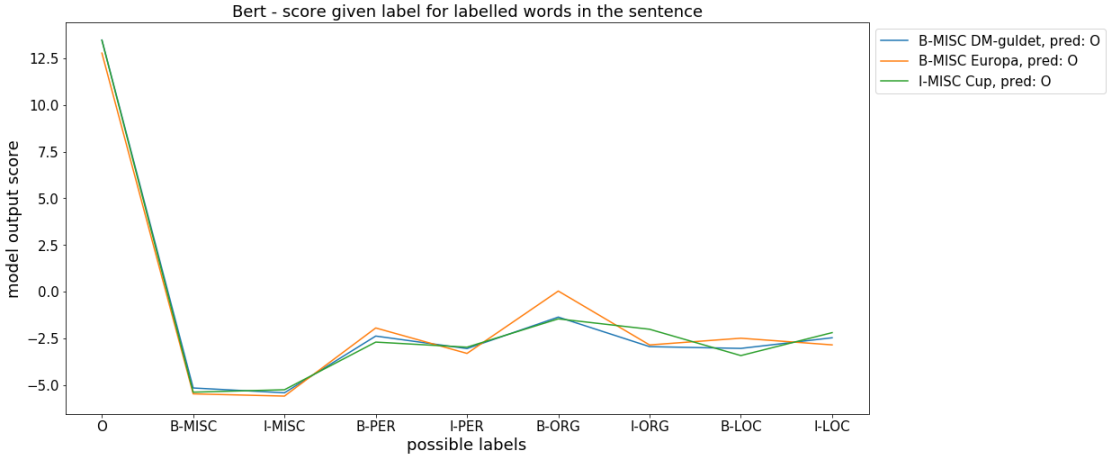


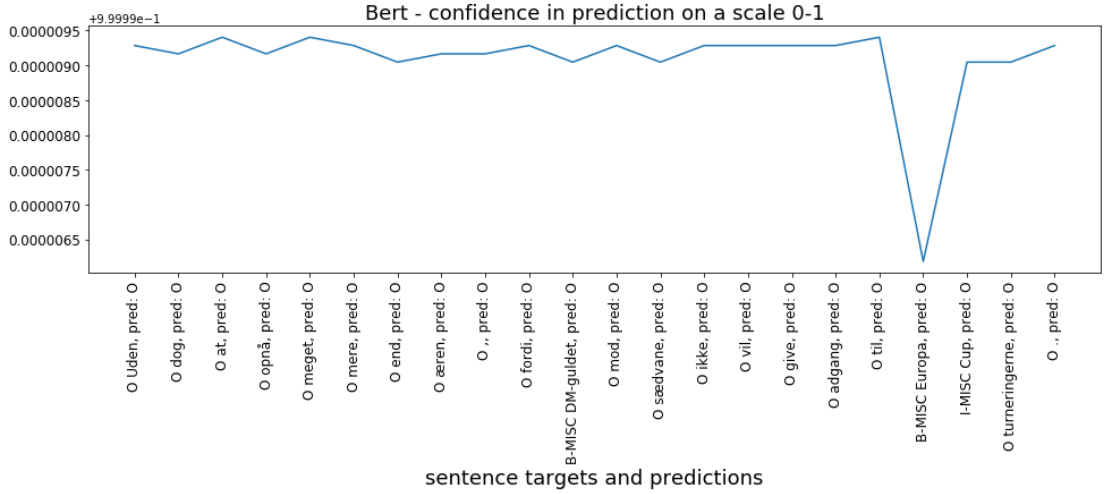
Sentence	Uden	dog	at	opnå	meget	mere	end	æren	,	fordi	DM-guldet	mod	sædvane	ikke	vil	give	adgang	til	Europa	Cup	turneringerne	.
Targets	O	O	O	O	O	O	O	O	O	O	B-MISC	O	O	O	O	O	O	O	B-MISC	I-MISC	O	O
BERT	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-LOC	O	O	O
Calibrated flair	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	B-LOC	O	O	O

Table 10: Sentence 10

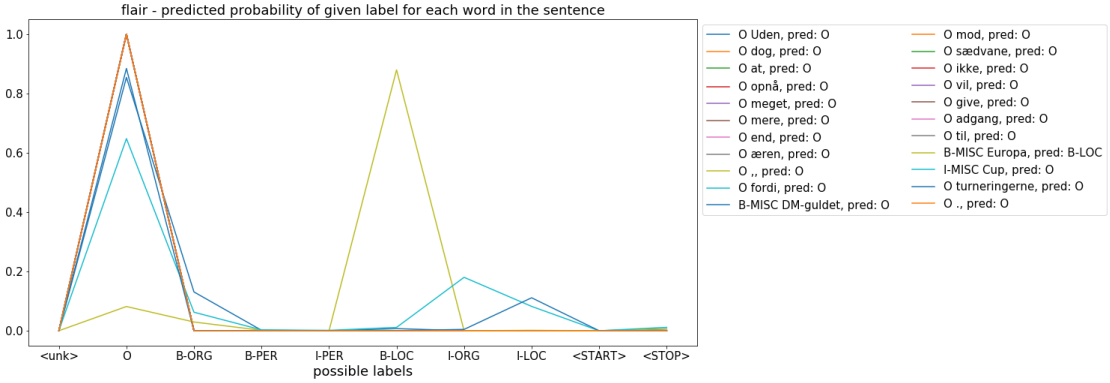
BERT - Sentence 10

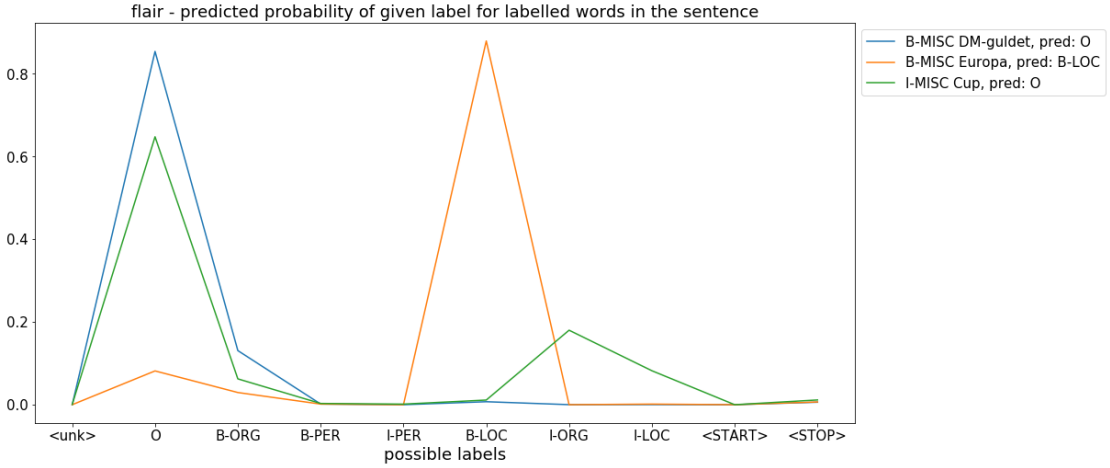


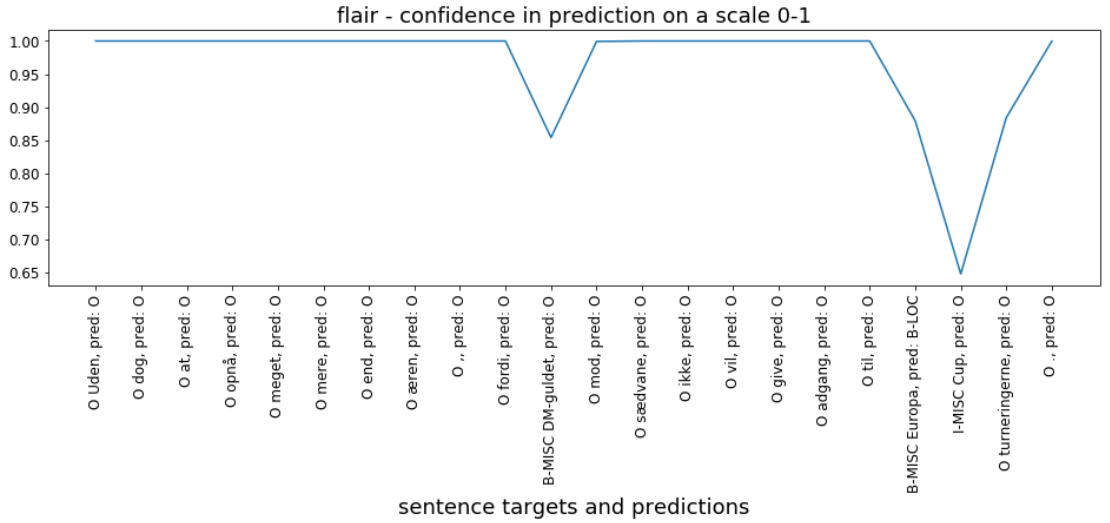




flair - Sentence 10







Calibrated flair - Sentence 10