Named entity recognition

Jan Jenko, Domen Požrl
University of Ljubljana
Faculty for computer and information science
Večna pot 113, SI-1000 Ljubljana
jj0549@student.uni-lj.si, *dp2576@student.uni-lj.si*

Abstract

We explore a supervised learning approach at classifying named entities, a widely studied problem in language processing. We followed (Štajner et al., 2013) and recreated their experiments using the CRF model. We went a step further in investigating the importance of different groups of attributes. Finally, we applied a neural network based BERT model to the same problem and found that it performs better, depending on the context of our problem.

1 Introduction

Named entities are proper nouns, words such as personal and geographic names, organisation and institution names, titles, etc. Named entity recognition plays an important role in text analysis, since named entities convey more information than simply the words that comprise them - they refer to fixed entities that remain constant throughout the text, and often even in between different texts. As such, they are an excellent means of linking texts. For example, the name of a public figure can be used to link together all newspaper articles in which the person in question is mentioned, helping us in constructing a narrative about that person through time. With this purpose in mind, there now exist indexes of named entities.

2 Related work

(Štajner et al., 2013) already worked on this exact problem which is why for now we implement their solution with minor changes in attribute construction. Their approach uses Conditional Random Fields for supervised learning on a set of attributes that are acquired from every word in the dataset. The first group of attributes comes from the patterns of the letters in the word. They check if there are any numbers in the word or if there are any capital letters, any punctuation and so on. The second group of attributes is a dictionary of known words related to some of the proper nouns we are trying to identify. Such words include but are not limited to world countries, capital cities of every country, all Slovenian cities, names of the days in the week, names of the months and so on. The third group are all morphosyntactic properties that are available for each word in the dataset. The fourth and final group of attributes is one attribute that describes the length of the word and another that is a logical conjunction of the attributes of the neighbouring words.

3 Data

Current systems for named entity recognition are trained using supervised learning, for which large corpora of labeled texts must exist. We will use the SSJ500k corpus (Krek et al., 2019) to train our models. It is important to note, that only a part of the corpus is tagged for named entities and so, we will only use this part of the corpus. In Table 1 we can see, alongside simple statistics, that the corpus is imbalanced in regards to number of proper nouns versus other words, the last of which represent about 98% of the entire corpus. This could potentially present a difficulty which one of the approaches looks to resolve by balancing the dataset.

The version of the SSJ500k corpus available to us had 5 classes of named entities: personal, derived personal, organizational, location and miscellaneous. However, the older version that (Štajner et al., 2013) used did not include the derived personal named entity. These are similar to personal named entities, but signify possession. To be able to compare our results to results in (Štajner et al., 2013), we decided to merge the personal and derived personal classes.

	#
Sentences	7742
at least 1 named entity	1846
exactly 1 named entity	1294
exactly 2 named entities	346
3 or more named entities	206
Words	147165
Other words	144337
Proper nouns	2828
Person proper nouns	1295
Location proper nouns	972
Organisation proper nouns	334
Person derived proper nouns	128
Miscellaneous proper nouns	99

Table 1: This table shows a few simple statistics about the SSJ500k corpus.

4 Attributes construction

As described in the Related works section, we used a very similar method of attribute construction. The first group of the attributes, as seen in Table 2, tells us whether the word has any of the stated properties.

The first group of attributes
First letter capital
Only capital letters
Capital letters inside of the word
Numbers in the word
Only numbers in the word
Numbers and mathematical operations in the word
Only numbers and letters in the word (alphanumeric)
Roman numerals
Includes hyphen or dash
Only capital letters seperated by a dot
Single capital letter followed by a dot
Single letter
Single capital letter
Punctuation
Quotation marks
Only lowercase letters

Table 2: This table shows the attributes of the first group

The second group of attributes tells us whether the word is present in lists of words from Table 3.

The second group of attributes
Slovene female names
Slovene male names
Slovene surenames
Slovene area names
Slovene cities
Capital cities of every country in the world
Every country in the world
Slovene names of days in the week
Slovene month names

Table 3: This table shows the attributes of the second group

The third group of attributes is the morphosyntactic properties available in the corpus. All of the described groups of attributes so far are very similar to those in (Štajner et al., 2013). The last group of attributes is where we introduce a different approach. We still take into account the length of the word but instead of logically combining the neighbouring words we use the entire set of attributes of the neighbouring words. That means that the final attributes vector for an ith word are all of the attributes vectors of words from $i-2$ to $i+2$. This way we give more importance to bigger parts of sentences and not just a single word.

5 Methods

In the first part, we used the CRF model to try reproduce the methods and results from (Štajner et al., 2013). We then further investigated the importance of groups of attributes by running multiple experiments. First, we used all attributes, as described in section 4. Notice that the attributes do not only contain properties of the target word (the word we are trying to classify), but also of its neighbourhood, that is of words that are at most two places away from the target word. In the second experiment we strip away the attributes with properties of words that are two places away from the target word, and in the third experiment we also strip away attributes with properties of words that are one place away from the target word. Surprisingly, we got the best results in the last experiment so we used the stripped data (attributes with properties of the target word only) in the subsequent experiments with the CRF model. Next, we eliminated a single attribute that carries the morphosyntactic properties of the word. The rationale here is that the data is unlikely to contain this property and also not contain the named entity classification we are after. This had a significant negative effect on the classification accuracy. In the last CRF experiment, we tried to somewhat balance the classes. The dataset is highly imbalanced, with vast majority of words not being a named entity. To counter this imbalance, we removed from our training data the sentences that did not contain any named entities.

For each experiment we first tuned the hyper-parameters using a grid search with data split into train (70%) and test (30%) subsets. We then used the models with optimal hyper-parameters and evaluated them using a five fold cross validation.

In the second part, we tested a BERT model (Devlin et al., 2018) pretrained for slovene, croatian and english (Ulčar and Robnik-Šikonja; Žitnik, 2020) on a modified version of the constructed attributes. We used individual words and their morphosyntactic properties to update the whole model weights for named entity recognition with 3 epochs. The data was split into train (80%) and test (20%) subsets and then the test data was split in half for validation. Later the performance was evaluated in terms of precision, recall and the F1-score.

6 Results

Table 4 shows the best results achieved in (Štajner et al., 2013). The combined figure is the weighted average score, where weights are the supports of each class.

Depending on the application, this table can be considered incomplete, as it does not show the classification accuracy for words that are not a named entity. We believe this is an important aspect of a classifier, so we will also include this figure.

	Precision	Recall	F1	#
per	0.85	0.91	0.88	1922
org	0.63	0.58	0.60	785
misc	0.39	0.30	0.33	406
loc	0.82	0.80	0.81	1284
combined	0.74	0.72	**0.73**	4397

Table 4: Best results achieved in (Štajner et al., 2013) using the CRF model.

6.1 CRF model

The results of the first experiment where we used original data is shown in table 5. The overall performance is comparable to results in table 4, but there are significant differences in individual classes. Organizational named entities and miscellaneous classes have significantly worse results in our experiment, but are also very underrepresented. This might contribute to the poor performance (as the model didn't have many significant learning samples), but it also means the overall score is not severely impacted.

	Precision	Recall	F1	#
per	0.79	0.82	0.81	1423
org	0.43	0.38	0.39	334
misc	0.20	0.09	0.12	99
loc	0.84	0.82	0.84	972
combined	0.74	0.74	**0.75**	2828
not-propn	0.998	0.999	0.999	144337
combined	0.993	0.994	**0.994**	147165

Table 5: Classification accuracy using original data.

As seen in table 6, the classification accuracy did not change much in the second experiment, where we did not consider properties of words that were two places away from the target word.

	Precision	Recall	F1	#
per	0.79	0.84	0.82	1423
org	0.47	0.35	0.39	334
misc	0.19	0.09	0.12	99
loc	0.84	0.83	0.84	972
combined	0.75	0.75	**0.75**	2828
not-propn	0.998	0.999	0.999	144337
combined	0.993	0.999	**0.994**	147165

Table 6: Classification accuracy of the CRF model working on data containing properties of the target word and its direct neighbors.

We then went further and disregarded properties of all words except the target word. The performance improved marginally, as seen in table 7

Next, we removed the attribute containing the morphosyntactic properties of the words. The performance worsened significantly, as seen in table 8.

Finally, we attempted to improve the performance by balancing the training data. As the model performs on whole sentences it was impossible to balance the classes completely (nor would we want to, given how few words are actually a named entity). We removed the sentences with no named entities and cut the original 7742 sentences

	Precision	Recall	F1	#
per	0.78	0.90	0.84	1423
org	0.76	0.27	0.40	334
misc	0	0	0	99
loc	0.87	0.84	0.85	972
combined	0.78	0.79	**0.76**	2828
not-propn	0.998	0.999	0.999	144337
combined	0.994	0.995	**0.994**	147165

Table 7: Classification accuracy of the CRF model working on data containing only properties of the target word.

	Precision	Recall	F1	#
per	0.81	0.40	0.46	1423
org	0	0.01	0	334
misc	0	0	0	99
loc	0.98	0.56	0.71	972
combined	0.75	0.39	**0.48**	2828
not-propn	0.989	0.996	0.992	144337
combined	0.984	0.984	**0.983**	147165

Table 8: Classification accuracy of the CRF model working with data without the morphosyntactic properties.

down to 1846 significant ones. In the original data, only 2% of all words are named entities. In the balanced training data, this figure is 7%.

As seen in table 9, the balancing did not have a significant impact on the classification performance.

	Precision	Recall	F1	#
per	0.79	0.84	0.82	1423
org	0.47	0.35	0.39	334
misc	0.19	0.09	0.12	99
loc	0.84	0.83	0.84	972
combined	0.75	0.75	**0.75**	2828
not-propn	0.998	0.999	0.999	144337
combined	0.993	0.994	**0.994**	147165

Table 9: Classification accuracy of the CRF model learning on a more balanced dataset.

6.2 BERT model

In Table 10 we can see that the BERT model performs quite differently compared to the CRF model. The individual named entity classes were predicted significantly better, but it also performed significantly worse in predicting words that are not named entities. All CRF models predicted the 'non-propn' class with F1 score of at least 0.99, while BERT only achieved F1 of 0.80. Because the 'non-propn' class carries such weight, the overall performance of the model is worse. Usually this kind of method uses a lot of data (Walia) so perhaps using more data could improve the performance even further, but we only had the current corpus at our disposal. In Figure 1 we can see how validation and training loss change with respect to the number of epochs. It is worth noticing that the validation loss does not decrease drastically therefore there is no reason to think that increasing the number of epochs would improve the performance of the model significantly.

	Precision	Recall	F1	#
per	0.81	0.89	0.85	142
org	0.83	0.71	0.77	28
misc	0.58	0.93	0.72	15
loc	0.98	0.93	0.95	132
combined	0.86	0.88	**0.87**	317
not-propn	0.78	0.81	0.80	317
combined	0.78	0.81	**0.80**	634

Table 10: Classification accuracy of the BERT model.

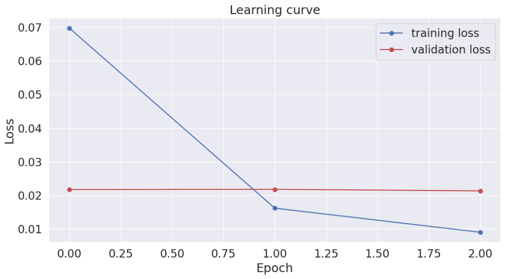

Figure 1: Training loss and validation loss with respect to number of epochs.

7 Conclusion

We managed to reproduce the results from (Štajner et al., 2013) using the CRF model and further analysed the importance of certain groups of attributes. We found that including attributes with properties of neighboring words does not significantly improve the performance, in fact, the model performed better overall if we did not include those attributes. The CRF model relies heavily on the attribute containing the morphosyntactic properties of words. If we ommit this attribute, the overall F1 score (weighted average) accross all four classes of named entities drops from the usual 0.75 to 0.48. Balancing the training dataset does not improve the result - CRF itself seems to account for data imbalance well.

We then tried to classify named entities based on the original data, but with a neural network based BERT model. The new model performed better in classifying the words that were indeed named entities, but did worse overall since it did not perform well in classifying words that are not named entities as such.

Which model to use depends on the context of the task. If we know that a word is a named entity and want to classify it, BERT performs better. If we need overall performance on all words, we would choose CRF.

References

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.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI.

Ulčar and Robnik-Šikonja. Pretrained model.

Abhinav Walia. [link].

Slavko Žitnik. 2020. Neural sequence tagging (pytorch).

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. *Jezikovne tehnologije*, 2(1):58–81.