Named entity recognition

Jan Jenko, Domen Požrl

Abstract

We present a new approach to constructing attributes for use with the Conditional Random Field model in Named entity recognition as well as evaluate it on the ssj500k dataset.

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 final group of the 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.

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 tells us whether the word has any of the following properties:

1. First letter capital
2. Only capital letters
3. Capital letters inside of the word
4. Numbers in the word

	#
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.

5. Only numbers in the word
6. Numbers and mathematical operations in the word
7. Only numbers and letters in the word (alphanumeric)
8. Roman numerals
9. Includes hyphen or dash
10. Only capital letters seperated by a dot
11. Single capital letter followed by a dot
12. Single letter
13. Single capital letter
14. Punctuation
15. Quotation marks
16. Only lowercase letters

The second group of attributes tells us whether the word is present in the following lists:

1. Slovene female names
2. Slovene male names
3. Slovene surenames
4. Slovene area names
5. Slovene cities
6. Capital cities of every country in the world
7. Every country in the world
8. Slovene names of days in the week
9. Slovene month names

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 Results

(Štajner et al., 2013) differentiated the named entities between person, location, organisation and miscellaneous. The corpus contains 5 different types of named entities. The already mentioned 4 and the 5th derived person type. The words connected with this type of a named entity are usually those expressing ownership of a certain object like "Darwinova [teorija]", "Nadina [hiša]". Because the article does not go into detail on how they handled this 5th type we devised two separate experiments. In the first experiment (2) we include all 5 types of named entities.

	Accuracy	Recall	F1	#
per	0.7390	0.8912	0.8080	340
org	0.4118	0.3889	0.4000	108
misc	0.0000	0.0000	0.0000	27
loc	0.8769	0.8358	0.8559	341
deriv-per	0.7273	0.5854	0.6486	41
combined	0.7287	0.763	**0.7425**	857

Table 2: Results of the first experiment.

We ranked the models by a wheighted average of the F1 metrics on individual classes. The weight of each class is it's frequency in the dataset. We get similar, slightly better results than those in (Štajner et al., 2013) (F1 = 0.73) however it is worth noting that we didn't use cross validation to evaluate the model.

When eveluating the model it seems reasonable to take into account the class of the words that are

	Accuracy	Recall	F1	#
per	0.7371	0.8824	0.8032	340
org	0.4175	0.3981	0.4076	108
misc	0.1000	0.0370	0.0541	27
loc	0.8785	0.8270	0.8520	341
deriv-per	0.6944	0.6098	0.6494	41
notpropn	0.9988	0.9983	0.9985	43414
combined	0.9936	0.9937	0.9936	44271

Table 3: Results of the first experiment with "not a proper noun" class.

not named entities. That is because when we use the model on an unknown word, we have no way of telling whether that word even is a named entity. If we want to maximize the successfulness of the model on 5 of the original classes and adding a single "not a proper noun" class we get slightly different results, as seen in 5

In the second experiment we combine the person and derived person type. Merging the pesonal named entity type and the derived personal type improves the models, but only marginally. Optimizing for performance on four classes only again yields a different model compared to when optimizing for performance on five classes (including “not proper name”).

When optimizing for performance on four classes only, we get an average F1 score of 0.745 (compared to 0.743 from before.

	Accuracy	Recall	F1	#
per	0.7035	0.9029	0.7908	381
org	0.5323	0.3056	0.3882	108
misc	0.0476	0.0370	0.0417	27
loc	0.8761	0.8504	0.8631	341
combined	0.7299	0.779	**0.7452**	857

Table 4: Results of the second experiment.

When adding the “not proper name”, we get an average F1 score of 0.9936, which is the same as before. Since the majority class has more than 98% support and already has a very high F1 score, we can’t expect mush improvement here.

	Accuracy	Recall	F1	#
per	0.7354	0.8609	0.7932	381
org	0.4057	0.3981	0.4019	108
misc	0.0714	0.0370	0.0488	27
loc	0.8892	0.8240	0.8554	341
notpropn	0.9988	0.9982	0.9985	43313
combined	0.9936	0.9936	**0.9936**	44271

Table 5: Results of the second experiment with "not a proper noun" class.

References

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.

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