Named entity recognition

Domen Požrl, Jan Jenko

I. Introduction

We aim to build a system for named entity recognition in Slovenian texts. Named entities are 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.

II. Overview

A. 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 [1] to train our models. A quick review of the corpus contents can bee seen in Table I

Number of sentances	7958
Number of words	122114
Number of nouns	29748
Number of proper nouns	4631
Number of verbs	14334
Number of adjectives	14952
Number of adverbs	6521

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

B. Related work

Best performing models thus far include Hidden Markov Models and Conditional Random Fields. When applying these models, they are often combined with a dictionary of know named entities [2].

There were already several attempts at named entity recognition for Slovenian texts [3]. It focuses on peculiarities of the slovenian language, such as the numerous cases for nouns. The authors also treated the different named entity categories seperately. Specifically, they treated organizational names seperately to all other names.

III. Initial ideas

We wish to implement the model as described in [3] and see if we can improve it, by taking into account not only the context of the word in a given text, but also its context in other texts. A simple approach would be to train a classifier that works on a single text and then apply it to all the texts where the word in question appears. The resulting prediction would be the average of the predictions on individual texts. Alternatively, we could modify the word attributes, so as to take into account all texts where the word appears. This approach seems more complicated, as we would need to address the fact that different words appear in a different number of texts.

References

[1] S. Krek, K. Dobrovoljc, T. Erjavec, S. Može, N. Ledinek, N. Holz, K. Zupan, P. Gantar, T. Kuzman, J. Čibej, Š. Arhar Holdt, T. Kavčič, I. Škrjanec, D. Marko, L. Jezeršek, and A. Zajc, "Training corpus ssj500k 2.2," 2019, slovenian language resource repository CLARIN.SI. [Online]. Available: http://hdl.handle.net/11356/1210

[2] W. W. Cohen and S. Sarawagi, "Exploiting dictionaries in named entity extraction: Combining semi-markov extraction processes and data integration methods," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 89–98. [Online]. Available: https://doi.org/10.1145/1014052.1014065

[3] T. Štajner, T. Erjavec, and S. Krek, "Razpoznavanje imenskih entitet v slovenskem besedilu," *Jezikovne tehnologije*, vol. 2, no. 1, p. 58–81, 2013. [Online]. Available: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_04.pdf