Tagging Named Entities in 19th century Finnish Newspaper Material with a Variety of Tools

Kimmo Kettunen and Teemu Ruokolainen

National Library of Finland, Centre for Preservation and Digitization, Mikkeli, Finland
kimmo.kettunen@helsinki.fi
teemu.ruokolainen@helsinki.fi

Abstract Named Entity Recognition (NER), search, classification and tagging of names and name like frequent informational elements in texts, has become a standard information extraction procedure for textual data. NER has been applied to many types of texts and different types of entities: newspapers, fiction, historical records, persons, locations, chemical compounds, protein families, animals etc. In general a NER system's performance is genre and domain dependent and also used entity categories vary (Nadeau and Sekine, 2007). The most general set of named entities is usually some version of three partite categorization of locations, persons and organizations. In this paper we report large scale trials and evaluation of NER with data out of a digitized Finnish historical newspaper collection Digi. Experiments, results and discussion of this research serve development of the Web collection of historical Finnish newspapers.

Digi collection contains 1,960,921 pages of newspaper material from years 1771–1910 both in Finnish and Swedish. We use only material of Finnish documents in our evaluation. The OCRed newspaper collection has lots of OCR errors; its estimated word level correctness is about 70–75 % (Kettunen and Pääkkönen, 2016). Our baseline NER tagger is a rule-based tagger of Finnish, FiNER, provided by the FIN-CLARIN consortium. We show also results of limited category semantic tagging with tools of the Semantic Computing Research Group (SeCo) of the Aalto University. Two other tools, Finnish Semantic Tagger, and Connexor's NER software are also evaluated. We report also development work of statistical tagger of Finnish and a new evaluation and learning corpus for NER of historical Finnish.

Keywords: named entity recognition, historical newspaper collections, Finnish

Introduction

Digital newspapers and journals, either OCRed or born digital, form a growing global network of data that is available 24/7, and as such they are an important source of information. As the amount of digitized journalistic data grows, also tools for harvesting the data are needed to gather information. Named Entity Recognition (NER) has become one of the basic techniques for information extraction of texts since the mid-1990's (Nadeau and Sekine, 2007). In its initial form NER was used to find and mark semantic entities like person, location and organization in texts to enable information extraction related to this kind of material. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software (Nadeau and Sekine, 2007). In this paper we report evaluation results of NER for historical 19th century Finnish. Our historical data consists of an evaluation collection out of an OCRed Finnish historical newspaper collection 1771-1910 (Kettunen and Pääkkönen, 2016).

Kettunen et al. (2016) have earlier reported NER evaluation results of the historical Finnish data with two tools, FiNER and ARPA. Both tools achieved maximal F-scores of about 60 at best, but with many categories the results were much weaker. Word level accuracy of the evaluation collection was about 73 %, and thus the data can be considered very noisy. Results for modern Finnish NER have not been reported extensively so far. Silfverberg (2015) mentions a few results in his description of transferring an older version of FiNER to a new version. With modern Finnish data F-scores round 90 are achieved. In this paper we add two more analysis tools to our earlier NER repertoire. Finnish Semantic tagger (FST) is not a NER tool as such; it has first and foremost been developed for semantic analysis of full text. FST assigns a semantic category to each word in text employing a comprehensive semantic category scheme (USAS Semantic Tagset, available in English[1] and also in Finnish[2]). Connexor's NER tool[3] is commercial tool for modern Finnish.

[1] http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

[2] https://github.com/UCREL/Multilingual-USAS/raw/master/Finnish/USASSemanticTagset-Finnish.pdf

Results for the historical data

Our historical Finnish evaluation data consists of 75 931 lines of manually annotated newspaper text. Most of the data is from the last decades of 19th century. Earlier NER evaluations with this data have achieved at best F-scores of 50–60 in some name categories (Kettunen et al., 2016). Our baseline tagger, FiNER, is described more in Kettunen et al. (2016). Shortly described it is a rule-based NER tagger that uses morphological recognition, morphological disambiguation, gazetteers (name lists), pattern and context rules for name tagging.

We evaluated performance of our different NER tools using the *conlleval*[4] script used in Conference on Computational Natural Language Learning (CONLL). *Conlleval* uses standard measures of precision, recall and F-score, the last one defined as 2PR/(R+P), where P is precision and R recall (Manning and Schütze, 269). Its Evaluation is based on "exact-match evaluation" (Nadeau and Sekine, 2007). In this type of evaluation NER system is evaluated based on the micro-averaged F-measure (MAF) where precision is the percentage of correct named entities found by the NER software; recall is the percentage of correct named entities present in the tagged evaluation corpus that are found by the NER system. In the strict version of evaluation named entity is considered correct only if it is an exact match of the corresponding entity in the tagged evaluation corpus: "a result is considered correct only if the boundaries and classification are exactly as annotated" (Poibeau and Kosseim, 2001). As FST and Connexor's, tagger do not distinguish multipart names with their boundaries only a comparable loose evaluation without entity boundary detection is reported here.

Table 1 shows F-score results of the four evaluations of locations and persons in our evaluation data. We performed two evaluations with FST: one with the words as they are, and the other with *w*→*v* substitution. Variation of *w* and *v* is one of the most salient features of 19th century Finnish. Modern Finnish uses w mainly in foreign names like *Wagner*, but in 19th century Finnish w was used many times instead of v (Kettunen and Pääkkönen, 2016).

	<EnamexPrsHum>		**<EnamexLocXxx>**	
	F-score	Number of found tags	F-score	Number of found tags
ARPA	52.90	3636	52.35	2933
Connexor	56.40	5321	60.87	1802
FiNER	58.10	2681	57.47	1541
FST	34.75	897	67.11	1420
FST w/v	76.10	908	59.12	1536

Table 1. Evaluation of four tools with loose criteria and two categories in the historical newspaper collection. W/v stands for w to v substitution in words.

Substitution of *w* with *v* decreased number of unknown words to FST with about 3 % units and has a noticeable effect on detection of locations and a small effect on persons. Overall locations are recognized better; their recognition with w/v substitution is about 3.5 per cent units better than without substitution. FST's performance with locations equals that of FiNER's, but its performance with person names is clearly inferior. Performance of the taggers is expectedly not very good, as the data is very noisy.

It is evident that the main reason for low NER performance is the quality of the OCRed texts. If we analyze the tagged words with a morphological analyzer (Omorfi v. 0.3[5]), we can see that wrongly tagged words are of lower quality than those that are tagged right. Figures are shown in Table 2. Thus improvement in OCR quality will most probably bring forth a clear improvement in NER of the material.

	Locations	**Persons**
ARPA right tag, word unrec.	1.9	4.5
Connexor right tag, word unrec.	10.2	25.0
FiNER right tag, word unrec	6.3	12.8
FST right tag, word unrec. rate w/v	4.1	0.06

[3] https://www.connexor.com/nlplib/?q=technology/name-recognition

[4] http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt, author ErikTjong Kim Sang, version 2004-01-26

[5] https://github.com/flammie/omorfi

ARPA wrong tag, word unrec.	22.7	29.3
Connexor wrong tag, word unrec.	53.5	57.4
FiNER wrong tag, word unrec	38.3	34.0
FST wrong tag, word unrec. rate w/v	33.9	28.4

Development of a new statistical tagger

Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Discussion

We have shown in this paper results of NE tagging of historical OCRed Finnish with four tools, FiNER ARPA, a Finnish Semantic tagger, FST and Connexor's NE software. FiNER and Connexor's tool are dedicated NER tools for Finnish, but FST is a general semantic tagger and ARPA a semantic web linking tool. Our results show they all tag names of locations at the same level in the noisy OCRed historical newspaper collections. FiNER is clearly better with names of persons in with the historical data.

In general our results show that NE tagging in a noisy historical newspaper collection can be done to some extent with tools that have been developed for modern Finnish. Anyhow, it seems obvious, that better results could be achieved with a new tool, that is trained with the noisy historical data. We have ongoing development work with regards to this.

Acknowledgements

This work is funded by the EU Commission through its European Regional Development Fund, and the program Leverage from the EU 2014–2020.

References

Kettunen, K. and Pääkkönen, T. (2016), "Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means", in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf.

Kettunen, K., Mäkelä, E., Kuokkala, J., Ruokolainen, T. and Niemi, J. (2016). Modern Tools for Old Content - in Search of Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910. LWDA 2016. http://ceur-ws.org/Vol-1670/paper-35.pdf

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P, Nykänen, A. and Varantola, K. (2005), "A semantic tagger for the Finnish language", available at http://eprints.lancs.ac.uk/12685/1/cl2005_fst.pdf.

Manning, C. D., Schütze, H. (1999), *Foundations of Statistical Language Processing*. The MIT Press, Cambridge, Massachusetts.

Nadeau, D., and Sekine, S. (2007), "A Survey of Named Entity Recognition and Classification", *Linguisticae Investigationes*, Vol. 30 No. 1, pp. 3–26.

Poibeau, T. and Kosseim, L. (2001), "Proper Name Extraction from Non-Journalistic Texts", *Language and Computers*, Vol. 37 No. 1, pp. 144–157.

Silfverberg, M. (2015), "Reverse Engineering a Rule-Based Finnish Named Entity Recognizer", paper presented at Named Entity Recognition in Digital Humanities Workshop, June 15, Helsinki available at: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/KielipankkiEventNERWorkshop2015/Silfverberg_presentation.pdf (accessed April 5 2016).