Name the Name – Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data

Teemu Ruokolainen and Kimmo Kettunen

The National Library of Finland, DH research, Saimaankatu 6, FI-50100 Mikkeli
teemu.p.ruokolainen@gmail.com,
kimmo.kettunen@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. Performance of a NER system is usually quite heavily genre and domain dependent. Entity categories used in NER may also vary. The most used set of named entity categories is usually some version of three partite categorization of locations, persons, and organizations.

In this paper we report evaluation results with data extracted from a digitized Finnish historical newspaper collection Digi using two statistical NER systems, namely, Stanford Named Entity Recognizer and LSTM-CRF NER model. The OCRed newspaper collection has lots of OCR errors; its estimated word level correctness is about 70–75%. Our NER evaluation collection and training data are based on ca. 500 000 words which have been manually corrected from OCR output of ABBYY FineReader 11. We have also available evaluation data of new uncorrected OCR output of Tesseract 3.04.01.

Our Stanford NER results are mostly satisfactory. With our ground truth data we achieve F-score of 0.89 with locations and 0.84 with persons. With organizations the result is 0.60. With re-OCRed Tesseract output the results are 0.79, 0.72, and 0.42, respectively. Results of LSTM-CRF are similar.

Keywords: Named Entity Recognition, Evaluation, Historical Newspapers, Finnish, OCR Data.

1 Introduction

The National Library of Finland has digitized and put available online the historical newspapers and journals published in Finland between 1771 and 1929. This collection contains 7.61 million pages in Finnish and Swedish. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as

Digi. An open data package of the collection's 1771–1910 part was released during the year 2016[1].

Digi is part of the growing global network of digitized newspapers and journals, and historical newspapers are considered more and more as an important source of historical knowledge. As the amount of digitized journalistic information grows, also tools for harvesting the information are needed. Named Entity Recognition has become one of the basic techniques for information extraction of texts since the mid-1990s [1]. 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 these kinds of entities. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software [1–2].

Our goal with the usage of NER is to provide the users of Digi better means for searching and browsing the newspapers and journals, i.e. new ways to structure, access, and also enrich information. Different types of names, especially person names and names of locations are used frequently as search terms in different newspaper collections [3]. They can provide also browsing assistance to collections if the names are recognized and tagged in the newspaper data and put into the index [4]. Thus named entity annotation of newspaper text allows a more semantically-oriented exploration of the contents of a large archive.

We have earlier reported NER results for 19th and early 20th century Finnish with different tools [5]. These tools were mostly rule-based tools for analysis of modern Finnish. None of them had advanced capability of handling 19th century Finnish. The best results we were able to achieve were F-scores of around 0.60. As we performed the evaluation with our heavily erroneous OCR data (with word recognition rate of about 73%), quite low scores were to be expected. Nevertheless, we gained invaluable experience in usage of NER and setting up an evaluation corpus. We became also much more familiar with our data.

For this evaluation, however, we started from scratch. We had now available a 500 000 word token OCRed and manually checked ground truth wordlist for our re-OCR process [6–7]. This data contains our old OCR, manually corrected ground truth (GT), and Tesseract v. 3.04.01 OCR data. Out of the GT data we created a new evaluation and training corpus for NER. The training data was tagged first manually and subsequent additions were made semi-manually (cf. section 2.4). As our primary NER tool we use a standard trainable statistical tagger, Stanford NER[2] [8]. We show also results of a state-of-the-art LSTM-CRF NER engine of Lample et al. [9] for comparison.

The rest of the paper is organized as follows. Section 2 introduces our corpus, its tagging and principles of tagging. Section 3 discusses experiments and their results. Finally, Section 4 makes concluding remarks.

[1] Available from https://digi.kansalliskirjasto.fi/opendata/submit?set_language=en

[2] https://nlp.stanford.edu/software/CRF-NER.shtml

2 Corpus

2.1 Text

Our NER evaluation and training data set is based on our re-OCR ground truth data. This data consists of 479 pages of both journals and newspapers from time period 1836–1918. Most of the data is from 1870 onwards, as the majority of publications in the collection is from 1870–1910 [10]. In the final GT data selection 56% of the pages are from journals and 44% from newspapers. Journal data has about 950 K of characters, newspaper data 3.06 M. The final ground truth text was corrected manually in two phases: first correction was performed by a subcontractor from output of ABBYY FineReader v.11 and the final correction was done in house at the National Library of Finland. The resulting GT is not errorless, but it is the best reference available [11].

Out of the OCR GT data we included a subset of 271 pages in our NE annotated data set, and for this we applied the following tokenization procedure. First, we discarded all non-alphanumeric characters apart from the following: comma, period, exclamation mark, question mark, and colon. Subsequently, we assigned each remaining punctuation character to its own token. For example, the following sentence fragment *Alennusmyynti Herra J. Olssonin puodissa!* (‘Sale at the store of Mister J. Olsson’) would be tokenized as

Alennusmyynti
Herra
J
.
Olssonin
puodissa
!

As a result of tokenization, the complete set of 271 pages contains 459 578 tokens.

2.2. Named Entities

The set of named entity classes we use contains the three fundamental entity types, person (PER), location (LOC), and organization (ORG), collectively referred to as the *enamex* since the MUC-6 competition [2]. In what follows, we provide a structured list of these entities marked in the data.

Location (LOC). Marked locations include:

1. Cities, towns, villages, municipalities, provinces: e.g. *Pori, Porin kaupunki, Salon kauppala, Jämsän piiri, Peltomaan torppa, Hangon kylä, Anttolan pitäjä, Suomenniemen kappeli*
2. Farms, crofts: e.g. *Häntälän rustholli, Jussilan tila, Hagan kuninkaankartano*
3. Countries: e.g. *Suomi, Kiina*
4. Other geographical areas: e.g. *Baltistan, Kasmir*
5. Continents: e.g. *Eurooppa, Aasia*

6. Seas, lakes: e.g. *Itämeri, Päijänne*
7. Streets, roads: e.g. *Aleksanterikatu, Kuopion-Hämeenlinnan maantie*
8. Islands, peninsulas: *Alsen saari, Balkanin niemimaa*
9. Buildings: *Puutarhakadun rukoushuone, Turun tuomiokirkko, Nikolain kirkko, Rauman kirkko, Ilomantsin pappila, Lehtisten kartano, Kakkaraisten puustelli*
10. Railroads, railway stations: e.g. *Porin rata, Pietarin-Riihimäen rautatie, Karjaan asema*

Person (PER). Marked person names include:

1. First names: e.g. *Elias, Liisa*
2. Family names: e.g. *Lönnrot, Ylitalo*

Organization (ORG). Marked organizations include:

1. Societies, associations: e.g. *Suomen evankelis-luterilainen pyhäkouluyhdistys, Pelastusarmeija, Airiston purjehdusseura, Wenäjän Palovakuutusseura*
2. Schools, academies, universities: e.g. *Suomen yliopisto, Kodiksamin kansakoulu, Hämeenlinnan kutomakoulu*
3. Senates, parliaments, governments: e.g. *Suomen keisarillinen senaatti, Englannin parlamentti, Venäjän hallitus*
4. Bureaus: e.g. *Konginkankaan pastorinvirasto, Tuuloksen kirkkoherranvirasto, Heinolan kaupungin maistraatti, Hämeen läänin maakanslia, Raaseporin kihkakunnanoikeus, Turun hovioikeus, Kuopion raastuvanoikekus*
5. Armies, regiments, battalions: e.g. *Ruotsin armeija, Porin rykmentti, Turun pataljoona*
6. Congregations, dioceses: e.g. *Kuopion seurakunta, Kuopion hiippakunta*
7. Chapters: *e.g. Turun tuomiokapituli, Porvoon hiippakunnan tuomiokapituli*
8. Judicial districts *: e.g. Rannan tuomiokunta*
9. Storest, factories, inns, hotells, restaurants: e.g. *O. Jalanderin kirjakauppa, Ruikan kestikievari, Bahnen puoti, Daalintehdas, Phoenix-hotelli, Phoenix-ravintola*
10. Companies, enterprises: *Kirkollinen kirja- ja paperikauppa O Y, Werner Söderström Osakeyhtiö, Turun Rautakalutehdas-yhtiö, Georg Segerstrålan Lakiasiain toimisto, Hämeen Sanomain kirjapaino*
11. Banks: e.g. *Suomen pankki, Englannin pankki, Pohjoismaiden säästöpankki*
12. Newspapers and journals: e.g. *Suomen Wiikkolehti, Pyhäkoululehti, Uusi Suometar*

2.3 Specifications on Annotation Description

In the following, we address three special cases of annotation, namely, how we addressed abbreviations, multi-word entities, including overlapping (nested) entities, and the expression *-niminen*.

Abbreviations. Abbreviated tokens are considered markable if they appear in a multi-token entity with one or more non-abbreviated word token. For example, the following person names are considered markable as a whole: *E. Lönnrot*, *Elias L.*, *Matti*

Laurinp. Ylitalo. Meanwhile, the following are not considered marked: *E. L.*, *M. Laurinp. Y.*. Equivalently, we consider *St. Louis* markable, whereas e.g. *N. Y.* referring to the city *New York* would not be considered markable.

Multi-Word Entities and Nesting. In general, marked entities can span multiple tokens, for example, consider *Turun tuomiokirkko* ('Turku Cathedral'), *Elias Lönnrot*, and *Suomen evankelis-luterilainen pyhäkouluyhdistys* ('the evangelical Lutheran Sunday school association of Finland'). As for multi-token entities, it is in general possible to employ either a nested (overlapping) or non-nested (non-overlapping) annotation approach. In the nested case, an expression such as *Suomen evankelis-luterilainen pyhäkouluyhdistys* is marked as an organization while its subpart *Suomen* is marked as a location. In our annotation, we follow this nested annotation approach in cases of multi-token organizations, that is, in case a multi-token organization entity contains location or person entities, the latter are also considered markable. For example, consider the previous example *Suomen evankelis-luterilainen pyhäkouluyhdistys* or *O. Jalanderin kirjakauppa* ('bookstore of O. Jalander') where the organization entity contains three tokens with the nested person entity *O. Jalander*.

While the most well-known named entity corpora employ the non-nested approach [2, 12], this results in a loss of information [13–15]. Therefore, we adopt the nested approach described above because of the prominent number of location and person names included in the marked organization entities (of all marked organizations in the complete data, roughly 46% have a nested location name and 7% a nested person name).

The Expression -niminen. Written Finnish uses occasionally expression *-niminen* ('named or called') which has the form *NAME-niminen NOUN*. For example, consider *Bogskär-nimiset kalliot* ('an outcrop called Bogskär') or *Butler-niminen mies* ('a man called Butler'). These cases are marked as a whole, that is, we consider *Bogskär-nimiset kalliot* and *Butler-niminen mies* as a two-token location entity and a two-token person entity, respectively.

2.4 The Annotation Process

We first performed a preliminary annotation of 170 pages (248 544 tokens) which yielded the first version of the set of entities to be included (Section 2.2) and the rule set considering abbreviations and multi-word entities (Section 2.3). This work was performed by the first author. Subsequently, the set of entities and rules were refined during a discussion period between the first and second author, and the annotation was refined correspondingly. Finally, a second discussion and refinement iteration was performed, after which the annotation was considered converged. The complete process was performed during the course of eleven months: the preliminary annotation phase took roughly six, the first refinement iteration roughly four, and the second refinement iteration roughly one month, respectively. While during this time both the primary and secondary author were employed full-time, full working hours were not assigned to the annotation project, but were divided among other project responsibilities as well. In practice, the annotation was performed using a standard spreadsheet software by assigning the tokenized text into the first column, the location and person annotation to the second, and the organization annotation to the third column.

Subsequently, we trained a Stanford NER system using the 170 manually annotated pages, tagged the remaining 101 pages (211 034 tokens) using the resulting system, and manually corrected the automatically annotated pages. The correction was again performed by the first author. This semi-manual annotation phase was considerably faster compared to the completely manual phase, and took roughly a month.

As the annotation was in effect performed by a single annotator, we tested the quality of the annotation description presented in Sections 2.2 and 2.3 as an annotation guideline by enlisting an additional annotator and measuring inter-annotator agreement. The additional annotator annotated independently five uniformly sampled pages (5,341 tokens in total) from the GT data based on the annotation description. This gave us a second set of annotations for this portion of the data set which we then compared with the pre-existing manual annotation. The pre-existing annotation contained 229 named entities in total (75 persons, 126 locations, and 28 organizations). To compare annotations, we used Cohen's κ [16], a measure for inter-annotator agreement commonly applied in natural language processing [17]. For a more detailed description, see e.g. the inter-annotator experiment in [18]. For all entities (persons, locations, organizations), the agreement score is 0.80. This agreement can be considered strong [19] indicating that the annotation description in Sections 2.2 and 2.3 can be applied by an external annotator and that the annotations in the corpus are consistent with the description.

2.5 Annotation Statistics

The complete data set consists of 170 manually annotated pages (248 544 tokens) and 101 semi-manually annotated pages (211 034 tokens). The counts of each named entity class in these sections are presented in Table 1.

Table 1. Counts and relative portions of named entity classes.

Manual annotation	**Count**	**%**
PER	5 102	48.88
LOC	6 285	39.68
ORG	1 471	11.44
Total	**12 858**	**100.0**
Semi-manual annotation		
PER	5 355	50.85
LOC	6 981	39.00
ORG	1 394	10.15
Total	**13 730**	**100.0**

In total there are 10 457 entities of person (39.33%), 13 266 entities of location (49.89%) and 2 865 entities of organization (10.78%) in the resulting data set.

2.6 Gazetteers

In addition to the prepared named entity annotation, our corpus is accompanied with three gazetteers which map words into semantically motivated categories. We compiled two of the gazetteers, person names and locations, by combining different open source word lists. E.g. the following sources for person names were used:

- *family names* from the Institute of Languages of Finland[3], Wiktionary page[4], Genealogia.fi[5]
- *first names* of men and women from Wikipedia page[6, 7].

Also other lists mentioned in Wikipedia of name lists[8] were used; especially old first names that do not belong to current name calendar were harvested. The resulting name list is a combination of the different lists, from which multiple occurrences of the same name were sorted out.

[3] http://kaino.kotus.fi/sukunimientaivutus/index.php?s=hakemisto

[4] https://fi.wiktionary.org/wiki/Luokka:Suomen_kielen_sukunimet

[5] http://www.genealogia.fi/nimet/nimi62s.htm

[6] https://fi.wikipedia.org/wiki/Luettelo_suomalaisen_nimip%C3%A4iv%C3%A4kalenterin_naisten_nimist%C3%A4

[7] http://fi.wikipedia.org/wiki/Luettelo_suomalaisen_nimip%C3%A4iv%C3%A4kalenterin_miesten_nimist%C3%A4

[8] https://fi.wikipedia.org/wiki/Wikipedia:Luettelo_etunimiluetteloista

Our gazetteer of person names contains about 456 700 entries. About 18 600 of them are names in base form, the rest are automatically generated most important twelve inflected case forms of the names formed with a noun form generator FCG_12 described in [20][9]. Inflected forms were included in the gazetteers to take care of the variation of word forms in the texts of Finnish, which is a highly inflectional language. We used generation of inflected forms for the gazetteer rather than lemmatizing of the target texts because lemmatization or part-of-speech tagging of the OCR text itself would have been more error prone.

Names of locations were gathered mainly from the following sources:

- *names of countries* from the Institute of Languages of Finland[10]
- *names of municipalities* from Wikipedia[11], names of former municipalities[12], names of cities in Finland[13], Swedish names in Finland[14]
- *other geographical names* from data of The National Land Survey of Finland[15]

Our gazetteer of locations contains about 333 670 entries. About 20 000 of them are names in base form, the rest are automatically generated most important inflected forms of the names.

As our organization name gazetteer we used the name list of Finnish organizational names in the Finto ontology[16], as names of organizations are not available in any other source publicly. Finto contains 43 355 concepts and the entries in the list were used as such.

3 Experiments

This section presents experimental results on the corpus employing the Stanford Named Entity Recognizer toolkit [8] and the LSTM-CRF model presented by Lample et al. [9]. In what follows, we will first describe the utilized training and evaluation set splits in Section 3.1. We then describe the Stanford Named Entity Recognizer and LSTM-CRF in Sections 3.2. The employed evaluation measures are described in Section 3.3. In section 3.4, we present and discuss the obtained results. Section 3.5 discusses the errors in performance of Stanford NER.

3.1 Data

As presented in Section 2.5, the complete corpus consists of 271 pages (459 578 tokens). In order to carry out the experiments, we separated the data into three non-

[9] The twelve different forms are the most frequent inflected forms of Finnish nouns in texts based on textual statistics.

[10] http://kaino.kotus.fi/maidennimet/index.php?s=hakemisto&h=fi

[11] https://fi.wikipedia.org/wiki/Luettelo_Suomen_kunnista

[12] http://fi.wikipedia.org/wiki/Luettelo_Suomen_entisist%C3%A4_kunnista

[13] http://fi.wikipedia.org/wiki/Luettelo_Suomen_kaupungeista

[14] http://kaino.kotus.fi/svenskaortnamn/

[15] http://www.maanmittauslaitos.fi/en/e-services/open-data-file-download-service

[16] https://finto.fi/cn/en/

overlapping sections, namely, training, and development and evaluation sets. In the training set, we included 136 manually annotated pages and 84 semi-manually annotated pages. The remaining 17 semi-manually annotated pages were assigned to the development set. The evaluation set consists of the remaining 34 manually annotated pages. The resulting training, development, and evaluation sections, therefore, contain 220, 17, and 34 pages (351 859, 29 596, and, 67 223 tokens), respectively.

Finally, we created a second version of the evaluation pages, in which the text has been produced by an automatic OCR system instead of manually recognized and the NE annotation is manually verified to take into account the errors introduced by the OCR system. Specifically, we employ the 50% rule, i.e., we remove NE annotation from tokens with more than 50% character errors [21].

3.2 Stanford Named Entity Recognizer

The Stanford Named Entity Recognizer[17] is a freely available established NER toolkit based on machine learning methodology. Given an annotated training data set and a feature extraction scheme specification, the toolkit can be employed to learn new NER models. Specifically, the toolkit contains an implementation of an arbitrary order conditional random field (CRF) model [8, 22].

Stanford NER has been used so far successfully in many NER evaluations, and primarily with English (e.g. [23]). Results for other languages exist, too. These include e.g. Chinese [23], Dutch, French and German [24]. Importantly, named entity recognition with the Stanford NER tool has been reported in the Europeana historical newspaper project, and the results have been good [4, 24]. Standford NER's performance with low quality OCR data has been evaluated e.g. in Rodriquez et al. [25]. They compare four available NER tools with OCRed data of The European Holocaust Research Infrastructure's transcripts. Stanford NER achieves best results in the comparison in general.

We employed the package in an out-of-the-box manner using the sample feature set presented in the "1. How can I train my own NER model?" section of the Stanford NER CRF FAQ page[18]. The feature set includes tag context, word context, and features describing word orthography. In addition, we employ the gazetteers described in Section 2.6. Our preliminary experimentation with expanded feature settings showed no consistent improvement in accuracy. The learning method (stochastic gradient descent) terminates when the average improvement in the objective function decreases below a set threshold (as opposed to an early stopping criterion based on the model accuracy on the separate development set). Therefore, the combined training and development set of 237 pages (381 356 tokens) is employed in the model training.

The Stanford system can incorporate part-of-speech tags and lemmas provided by an external toolkit as a feature. We experimented using the morphological tagger and lemmatizer at our disposal, the FinnPos system [26]. However, the quality of the analysis was not sufficiently high. The main reason for this is that the FinnPos system

[17] Available at http://nlp.stanford.edu/software/CRF-NER.shtml

[18] https://nlp.stanford.edu/software/crf-faq.shtml

is designed for modern Finnish. In addition, the analysis is further hindered by the OCR errors.

Given the nested annotation described in Section 2.3.2, we perform the model training and predictions on evaluation set in two stages. First, we learn a model for location and person name classes simultaneously. We then learn a second pipeline model for organizations, in which the location and person tags (gold standard during training and model predictions during testing) are given as features to the model. The Stanford package's interface provides a simple mean to accomplish this. We found that the use of the pipeline resulted in an improvement of around 3-4% in F-score of organizations.

3.3 LSTM-CRF

The Stanford NER system has the benefit of being able to perform well in an out-of-the-box usage scenario and being capable of incorporating gazetteers. However, it should be noted that the current state of the art in NER and in NLP in general is yielded by utilizing deep learning methodology [27]. In particular, we refer to the LSTM-CRF model presented by Lample et al. [9] which was shown to yield state-of-the-art accuracy for multiple languages (English, Dutch, German, Spanish) on standard NER data sets. The LSTM-CRF model is a bidirectional LSTM network with a sequential conditional random field layer above it. This approach differs from the Stanford model in two important respectives. First, instead of manually designed specific orthographic features, the model learns character-based representations of words [28] from training data. Second, instead of external hand-crafted gazetteers, the model utilizes pre-trained word embeddings [29] learned from an external unannotated corpus to capture distributional properties of words. We utilize a freely available implementation of the system used in the experiments of Lample et al. [9].[19]

Ideally, we would learn the word embeddings utilized by the LSTM-CRF model from the Digi collection in order to achieve maximal vocabulary coverage. However, the quality of OCR text is currently not sufficiently high. Therefore, we instead utilize the pre-trained word embeddings provided by the Finnish Internet Parsebank project[20]. These embeddings are learned from the Finnish Internet Parsebank containing 1.5 billion tokens (116 million sentences) using the word2vec software [30]. Roughly 59% of the word forms in the training set were found in these embeddings and were used to initialize the word embeddings of the model. The embeddings of word forms which do not have pretrained embeddings receive random initializations. On the development set, we found that using the embeddings results in an increased convergence speed as well as increase in the model accuracy by roughly up to 5%.

The preprocessing steps of the corpus described in Section 2.1 included tokenization but not sentence segmentation. Sentence segmentation was not performed by default because the Stanford NER system used during the corpus development is not sensitive to the length of “sentence” segments (in this case, a single “sentence” consists of all tokens on a single page). However, during actual experimentation present-

[19] https://github.com/glample/tagger

[20] http://bionlp-www.utu.fi/fin-vector-space-models/fin-word2vec.bin

ed in this section it turned out that the LSTM-CRF implementation of Lample et al. [9] is sensitive, so that its learning is completely prohibited by too long “sentences”. Therefore, in order to enable the use of the LSTM-CRF model, we first tried to segment the corpus into sentences using the Apache OpenNLP[21] library and the sentence segmentation model employed in the development of the Turku Dependency Treebank [31]. However, this model did not perform sufficiently well due to the abundant usage of abbreviations in the historical text. Specifically, roughly 57 per cent of multi-token person names (31 per cent of all person names) in the corpus contain an abbreviated token (e.g. *E. Lönnrot*, *W.G.F. Lundgren*, *Timo Laurinp. Peltonen*, etc.). Moreover, the difficulty of deciding if a period marks an abbreviation or the end of a sentence becomes more pronounced in the presence of OCR errors. In consequence, we decided to employ the following simple segmentation heuristic instead: we segment text 1) at each exclamation point, question mark, and colon (since these characters were never contained in any named entity) and 2) at any position where the current segment would exceed 100 tokens. This approach provides us sufficiently short “sentence” segments to enable use of the LSTM-CRF model while yielding a tolerably small probability of segmenting a multi-token named entity, thus, immediately introducing a named entity recognition error.

Given the segmented data sets and the word embeddings from the Finnish Internet Parsebank project, we employ the LSTM-CRF model in an out-of-the-box manner using the default parameter settings. The LSTM-CRF implementation evaluates the model accuracy on the development set after each pass over the training data. We learn two separate models for locations and persons, and organizations, respectively. The out-of-the-box implementation runs the training for 100 passes over the training data, after which we apply the parameter settings yielding best performance on development set to the test set.

Finally, there has recently been a renewed interest in developing NER models for specifically the nested annotation case using neural network models [32–35]. However, implementations of these approaches are not generally available. Therefore, we opted for the work of Lample et al. [9] as their implementation worked well in our preliminary experiments with little to no additional work.

3.4 Evaluation Measures

We evaluate the systems using the standard measures of precision (the number of correctly recognized entities divided by the number of all recognized entities), recall (the number of correctly recognized entities divided by the number of all annotated entities in data), and F1-score (the harmonic mean of precision and recall) [36].

3.5 Results

Results of the Stanford NER system. Obtained results on the ground truth evaluation set using the Stanford NER are presented in Table 2. The system yielded an

[21] https://opennlp.apache.org/index.html

overall F1-score of 0.8200 with **precision** and recall scores of 0.8696 and 0.7758, respectively.

Table 2. Precision, recall, and F-scores for each named entity class on the ground truth evaluation set.

Class	**Precision**	**Recall**	**F1**	**Found entities**	**Entities in the gold standard**
LOC	0.8872	0.8566	0.8716	1764	1826
PER	0.8408	0.7801	0.8093	1118	1205
ORG	0.8740	0.4536	0.5972	246	473
Totals	0.8696	0.7758	0.8200	3128	3504

These values, however, are overly optimistic since in a real use case the recognition has to be performed on a lower quality OCRed text. Therefore, we will next discuss results obtained in this scenario. We begin by noting that some NEs can be lost due to the OCR process itself due to tokens contained in NEs being not recognized properly. This leads to a reduction in recall of NEs since no NER system can recover these lost entities.

Table 3. shows the final NER performance taking into account both errors yielded by the OCR process and the Stanford NER system.

Table 3. Precision, recall, and F-scores for each named entity class on the OCR evaluation set.

Class	**Precision**	**Recall**	**F1**	**Found entities**	**Entities in the gold standard**
LOC	0.8527	0.7322	0.7879	1485	1826
PER	0.7856	0.6631	0.7192	1017	1205
ORG	0.8012	0.2896	0.4255	171	473
Totals	0.8247	0.6487	0.7262	2756	3504

The system yielded an overall F1-score of 0.726 with precision and recall scores of 0.825 and 0.649, respectively. Locations and persons achieve a F1 score of, 0.788 and 0.726, respectively. Result of organizations is only 0.43, which is low.

Results of LSTM-CRF. Results of the LSTM-CRF for GT data are shown in Table 4 and results of the OCR data in Table 5.

Table 4. Precision, recall, and F-scores for each named entity class on the ground truth evaluation set with the LSTM-CRF model.

Class	Precision	Recall	F1	Found entities	Entities in the gold standard
LOC	0.8918	0.8202	0.8545	1682	1826
PER	0.8653	0.7938	0.8280	1115	1205
ORG	0.7937	0.5404	0.6430	326	473
Totals	0.8719	0.7731	0.8196	3123	3504

Table 5. Precision, recall, and F-scores for each named entity class on the OCR evaluation set with the LSTM-CRF model.

Class	Precision	Recall	F1	Found entities	Entities in the gold standard
LOC	0.8598	0.6884	0.7646	1471	1826
PER	0.8212	0.6822	0.7452	1022	1205
ORG	0.6816	0.3214	0.4368	227	473
Totals	0.8309	0.6450	0.7262	2720	3504

Results overall. Overall, the Stanford NER system and the LSTM- CRF model yield very similar F-scores between 0.70 and 0.80 in the OCR data. Stanford is slightly better in locations and Lample in persons, but the differences are small and rather meaningless from a practical point of view, 2–3% units. Results are consistent in both GT and OCR data. As for organizations, both systems perform badly, achieving only F-scores of 0.43–0.44 with the OCR data.

3.6 Error analysis of Stanford NER results

To be able to pinpoint some of the problems of our data for the NE taggers, we performed error analysis of the output of the Stanford tagger. Ehrmann et al. [37] suggest that application of NE tools on historical texts faces three challenges: i) noisy input texts, ii) lack of coverage in linguistic resources, and iii) dynamics of language. Lack of coverage in linguistic resources can be e.g. be missing old names in the lexicons of the NER tools. With dynamics of language Ehrmann et al. refer to different rules and conventions for the use of written language in different times. In this respect late 19th and early 20th century Finnish is not that different from current Finnish, but obviously also this can affect the results.

In our earlier NER evaluation [5] especially Ehrman’s first point, noisy input, was the obvious reason for low performance of evaluated NER tools. Now that we have available a good quality ground truth evaluation collection along with a lower quality re-OCRed version of the same data, we can see more clearly effects of OCR quality on the results. With the GT data performance of persons and locations can now be considered good. As was shown in Tables 2 and 3, our new improved OCR quality impairs results with 9–10% units when compared to GT data NER. With locations and persons this result is now anyhow acceptable and useful. Results of organizations are still far from even good enough.

We have used gazetteers, i.e. extra linguistic resources, with Stanford NER. Their impact on the results is, however, not straightforward. As we stated in section 2.6, especially the compiled gazetteers of location and person names are very comprehensive. We evaluated their effect on the results of tagging in different phases of cumulating training data with our earlier NER evaluation data of 75 000 tokens [5]. When we had relatively little training data for the Stanford NER model, about 50–60 000 tokens, it seemed that the gazetteers had an impact of about +5% units with locations and persons. However, in the final evaluation with our current NER evaluation data, the effect of gazetteers on the result is very small, maximally about 2% units. Stanford NER's documentation does not make it possible to analyze impact of gazetteers on the tagging in detail, but in our data their impact seems to be quite minimal.

As organizations are performing worst, we tried to improve their performance by adding a name list of 3000+ historical Finnish limited liability companies of the time period 1865–1912 obtained from a web page containing historical stock exchange information, *Pörssitieto*.[22] Addition of these names did not improve performance probably because names of organizations in our data are mostly other entities than limited liability companies.

Reason for bad performance of organizations is not completely clear to us. Bad performance with organizations was also one of the findings in our earlier NER evaluation [5] with all evaluated taggers. It is possible that class ORG as we have defined it is too broad a category to be found well. Organizations are also much scarcer in the training data compared to locations and persons: only about 11% of the training data entities are marked as an organization. The used gazetteer of organizations is not optimal, either. Even a short overview of the Finto ontology shows that the resulting list is very heterogeneous and contains very different types of organizations. It has been compiled from the names of publishers and target audiences of Finnish publications in the National bibliography of Finland.

We performed also detailed error analysis on results of locations and persons in GT and OCR evaluation data to pinpoint problems of our data and Stanford NER's performance in it. We found 857 misclassifications in the results of locations and persons in the GT evaluation data. In OCR evaluation data there were 1016 errors. Error classes and their counts are shown in Table 6.

[22] https://www.porssitieto.fi/yhtiot/

Table 6. Amounts of errors in the tagged data.

Error	**Amount in the GT data**	**Amount in the OCR data**
PER missed	241	252
LOC missed	224	204
NULL marked as PER	114	212
NULL marked as LOC	106	162
LOC marked as PER	58	76
PER marked as LOC	40	46
Confused beginnings and endings of PER	65	61
Confused beginnings and endings of LOC	9	3
	857	**1016**

As the four first content rows in the table show, about 80% of the errors in both data are either missing entity tags or marked entities in case, where there should be none. Missed persons and locations are most common, and cover about 54% of the errors in the GT data and about 44% in the OCR data. Marking of nonexistent persons and locations is more frequent with OCRed output.

Common possible causes for errors are the following:

- spelling variants of words (variant/common): *Itaalia*/*Italia, Buda-Pestiä/Budapestiä, Amsterdami/Amsterdam*, *Tukholmi/Tukholma*, *Kiöpen-hawni/Köpenhamina, Kalefornia*/*Kalifornia*
- spelling errors or erroneous OCR (*Vulgarian* pro *Bulgarian*, *Insbuckissä* pro *Innsbruckissa*, *seppo* pro *Seppo*)
- foreign names that are rare*: Cassagnac, Buchhoz, Henszelman, Bergamasco* etc.
- broken lines (e.g. *Hel- sinki* broken to two separate lines due to hyphenation)

In general, the orthography and morphology of words provide strong cues to the NER system regarding the “nameness” of tokens: for example, consider capitalization and the suffix *-nen* commonly present in Finnish surnames. Therefore, we would expect spelling and OCR errors to hamper the recognition performance.

Indeed, this effect can be seen, for example, by analyzing correctly and incorrectly marked words with a morphological analyzer. In a set of correctly tagged person names of the GT data, 89% of the names are recognized by a morphological analyzer. In a set of wrongly tagged person names, 55% of the names are recognized. In the set of correctly tagged location names, 93% of the names are recognized by a morphological analyzer. With wrongly tagged location names 67% of the names are recognized. The effect of different errors or orthographical variants on tagging is thus clear, as it was in Kettunen et al. [5], where tagging was usually based on morphological analysis of words.

One case of errors, persons marked as locations or vice versa due to ambiguity of names, occurs surprisingly infrequently. Many Finnish surnames can be also names of locations, either names of municipalities, villages or houses. Surprisingly these kinds of errors are quite rare in the data with Finnish names; it appears that the information provided by the context is sufficient to disambiguate between the location and person senses. A few of them, like *Haapala*, *Ylitalo* and *Suonio*, are confused, but most errors of this kind occur with foreign last and first names.

4 Conclusions

We have reported in this paper usage of two standard statistical NER tools, Stanford NER and LSTM-CRF model, for annotation of OCRed Finnish historical newspaper and journal data. We have created an evaluation collection of 67 223 tokens and trained the NER systems with manually and semi-manually tagged data of 381 356 tokens. The results we achieve are mostly good. Using the Stanford NER system, we obtained F-scores of 0.79 and 0.72 for locations and persons respectively on our re-OCRed output. Meanwhile, for organizations the system yielded an F-score of only 0.43. Using the LSTM-CRF, we obtained F-scores of 0.76 and 0.75, respectively, on the re-OCRed data. For organizations, the model yielded an F-score of 0.44.

Our results show now clearly, what we predicted after our earlier experiments: improved OCR quality data will also improve NER results (Kettunen et al., 2017). We have now available OCR data out of which about 90% of the words are recognized by a morphological recognizer; in the old data the percentage is 81% and it was even lower in Kettunen et al [5], about 73%. We now consider the NER performance of locations and persons to be of sufficient quality to be used in our online journalistic collection Digi. This result is in accordance with most of the results in NER of historical or OCRed data. NER experiments with OCRed data in other languages show usually improvement of NER when the quality of the OCRed data has been improved from very poor to somehow better (see e.g. [21, 38–39]). Miller et al. [39] show that rate of achieved NER performance of a statistical trainable tagger degraded linearly as a function of word error rates. On the other hand, results of Rodriquez et al. [25] show that manual correction of OCRed material that has 88–92 % word accuracy does not increase performance of four different NER tools significantly.

Finally, a note about usage of Named Entity Recognition is in order. Named Entity Recognition is a tool that needs to be used to some useful purpose. We have now acceptable recognition rate for locations and persons, and we need to decide, how we are going to use extracted names in Digi. In our case extraction of person and place names is primarily a tool for improving access to the Digi collection.

Some exemplary suggestions of NER usage are provided by archive of Italian newspaper La Stampa[23] that covers years 1867–2005 of the newspaper and Australian Trove Names [40]. La Stampa style usage of names provides informational filters after a basic search has been conducted. User can further look for persons, locations and organizations mentioned in the article results. This kind of approach enhances browsing access to the collection: users can "wander around" in the collection and perhaps find things that they were not searching for in the first place [41–42]. Trove Names' name search takes the opposite approach: user searches first for names and then gets articles where the names occur. We believe that the La Stampa style of usage of names in the GUI of a newspaper collection is more informative and useful for users, as the Trove style can be achieved with the normal search function in the GUI of a newspaper collection.

Stanford NER performs so far only basic recognition and classification of names, which is the first stage in named entity analysis [43]. To be of more general practical use names would need both intra document reference entity linking as well as multiple document reference entity linking [43–44]. We intend to explore this work in the future. One more possible use for NER is usage with tagging and classification of images published in the newspapers. Most of the images (photos, illustrations, graphs etc.) have short title texts. It seems that many of the images represent locations and persons, with names of the objects mentioned in the image title. As image recognition and classifying of low quality print images may not be very feasible, image texts may help in classifying at least a reasonable part of the images.

Acknowledgements

This work was supported by the Academy of Finland as part of the project Computational History and Transformation of Public Discourse in Finland, 1640–1910, decision number 293341.

References

1. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes, 30, 3–26 (2007).
2. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A brief history. In: Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING 1996), volume 96, 466–471 (1995).
3. Crane, G., Jones, A.: The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. In Proceedings of JCDL'06, June 11–15, (2006), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.6257&rep=rep1&type=pdf
4. Neudecker, C., Wilms, L., Faber, W. J., van Veen, T.: Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition. In: Proceedings of IFLA 2014,

[23] https://www.lastampa.it/archivio-storico/index.jpp

(2014), http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf

5. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L.: Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. Digital Humanities Quarterly 11(3), (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html
6. Koistinen, K., Kettunen, K., Pääkkönen, T.: Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In: Proceedings of NoDaLiDa 2017, (2017), http://www.ep.liu.se/ecp/131/Title_Pages.pdf
7. Koistinen, K., Kettunen, K., Kervinen, K.: How to Improve Optical Character Recognition of Finnish Historical Newspapers Using Open Source Tesseract OCR Engine. In: Vetulani, Z., Paroubek, P. (eds.) Human Language Technologies as a Challenge for Computer Science and Linguistics: 8th Language & Technology Conference, pp. 279–283 (2017).
8. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the Fourty-Third Annual Meeting on Association for Computational Linguistics (ACL 2005), 363–370 (2005).
9. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. In: Proceedings of NAACL-HLT, 260–270, (2016), https://www.aclweb.org/anthology/N16-1030/
10. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. In: Calzolari, N. et al. (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 956–961, https://www.aclweb.org/anthology/L16-1152
11. Kettunen, K., Kervinen, J., Koistinen, M.: Creating and using ground truth OCR sample data for Finnish historical newspapers and journals. In: Proceedings of Digital Humanities in the Nordic Countries, 3rd Conference (2018), http://ceur-ws.org/Vol-2084/shortplus1.pdf
12. Sang, E. F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition (2003), https://www.aclweb.org/anthology/W03-0419.pdf
13. Ohta, T., Tateisi, Y.,Kim, J. D.: The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In: Proceedings of the second international conference on Human Language Technology Research, pp. 82–86. Morgan Kaufmann Publishers Inc. (2002).
14. Byrne, K.: Nested named entity recognition in historical archive text. In: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 589–596 (2007).
15. Benikova, D., Biemann, C., Reznicek, M.: NoSta-D named entity annotation for German: Guidelines and dataset. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2524–2531 (2014).
16. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1(20), 37–46 (1960).
17. Artstein R.: Inter-annotator Agreement. In: Ide N., Pustejovsky J. (eds.) Handbook of Linguistic Annotation. Springer, Dordrecht (2017).
18. Ruokolainen, T., Kauppinen, P., Silfverberg, M., Lindén, K.: A Finnish News Corpus for Named Entity Recognition. Language Resources and Evaluation (2019), https://link.springer.com/article/10.1007/s10579-019-09471-7

19. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics, 1(33), 159–174 (1977).
20. Kettunen, K., Arvola, P.: Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish. In: Salampasis, M., Larsen, B. (eds.), Advances in Multidisciplinary Retrieval, IRFC 2012, LNCS 7356, pp. 113−126 (2012).
21. Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K., Jensen, L. S.: Extracting Person Names from Diverse and Noisy OCR Text. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM, (2010), http://dl.acm.org/citation.cfm?id=1871845
22. Lafferty, J., McCallum, A., Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), pp. 282–289 (2001).
23. Wang, M., Che, W., Manning, C.D.: Effective Bilingual Constraints for Semi-supervised Learning of Named Entity Recognizers. Association for the Advancement of Artificial Intelligence (AAAI) (2013), https://nlp.stanford.edu/pubs/aaai13-wang.pdf
24. Neudecker, C.: An Open Corpus for Named Entity Recognition in Historic Newspapers. In: LREC 2016, Tenth International Conference on Language Resources and Evaluation, (2016), http://www.lrec-conf.org/proceedings/lrec2016/pdf/110_Paper.pdf
25. Rodriguez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of Named Entity Recognition Tools for raw OCR text. In: Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna September 21, pp. 410–414 (2012).
26. Silfverberg, M., Ruokolainen, T., Lindén, K.., Kurimo, M.: FinnPos: an open-source morphological tagging and lemmatization toolkit for FinnishLang Resources & Evaluation 50: 863–878, (2016), https://doi.org/10.1007/s10579-015-9326-3
27. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers (2017).
28. Ling, W., Dyer, C., Black, A.W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., Luis, T.: Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1520–1530 (2015).
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013).
30. Kanerva, J., Luotolahti, J., Laippala, V., Ginter, F. Syntactic n-gram collection from a large-scale corpus of Internet Finnish. In: Proceedings of the Sixth International Conference Baltic HLT 2014, pp. 184–191. IOS Press (2014).
31. Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Nissilä, A., Ojala, S., Salakoski, T., Ginter, F.: Building the essential resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation 48(3), 493–531 (2014).
32. Sohrab, M.G., Miwa, M.: Deep exhaustive model for nested named entity recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 2843–2849 (2018).
33. Katiyar, A., Cardie, C.: Nested named entity recognition revisited. In: Proceedings of The Sixteenth Annual Conference of the North American Chapterof the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018), pp. 861–871 (2018).
34. Ju, M., Miwa, M., Ananiadou, S.: A neural layered model for nested named entity recognition. In: Proceedings of The Sixteenth Annual Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018), pp. 1446–1459 (2018).

35. Wang, B., Lu, W., Wang, Y., Jin, H.: A neural transition-based model for nested mention recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 1011–1017 (2018).
36. Manning, C.D., Schütze, H.: Foundations of Statistical Language Processing. The MIT Press, Cambridge, Massachusetts (1999).
37. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic Evaluation of NER Systems on Old Newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), pp. 97–107 (2016), https://www.linguistics.rub.de/konvens16/pub/13_konvensproc.pdf
38. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces, 35: 482–489 (2013).
39. Miller, D., Boisen, S., Schwartz, R. Stone, R., Weischedel, R.: Named entity extraction from noisy input: Speech and OCR. In: Proceedings of the 6th Applied Natural Language Processing Conference, 316–324, Seattle, WA, (2000), http://www.anthology.aclweb.org/A/A00/A00-1044.pdf
40. Mac Kim, S., Cassidy, S.: Finding Names in Trove: Named Entity Recognition for Australian. In: Proceedings of Australasian Language Technology Association Workshop (2015), https://aclweb.org/anthology/U/U15/U15-1007.pdf
41. Bates, M.: What is Browsing – really? A Model Drawing from Behavioural Science Research. Information Research 12, (2007), http://www.informationr.net/ir/12-4/paper330.html
42. Toms, E.G.: Understanding and Facilitating the Browsing of Electronic Text. International Journal of Human-Computer Studies, 52, 423–452 (2000).
43. McNamee, P., Mayfield, J.C., Piatko, C.D.: Processing Named Entities in Text. Johns Hopkins APL Technical Digest, 30, 31–40 (2011).
44. Ehrmann, M., Nouvel, D., Rosset, S.: Named Entity Resources – Overview and Outlook. In: Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation, (2016), http://www.lrec-conf.org/proceedings/lrec2016/pdf/987_Paper.pdf