Historical Named Entity Recognition

November 22, 2017

Abstract

1 Introduction

The rest of the paper is organized as follows.

2 Corpus

2.1 Text

[Text description: Newspaper text, manually corrected?]

271 pages. Simple tokenization (special characters removed apart from .,!?, special characters assigned to their own lines)

2.2 Named Entities

The set of named entity classes contains the three fundamental entity types, *person*, *organization*, *location*, collectively referred to as the *enamex* since MUC-6 competition (Grishman and Sundheim, 1996).

Location (LOC) Markable 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: e.g.

10. Buildings: Puutarhakadun rukoushuone, Turun tuomiokirkko, Nikolain kirkko, Rauman kirkko, Ilomantsin pappila, Lehtisten kartano, Kakkaraisten puustelli

11. Railroads, railway stations: e.g. Porin rata, Pietarin-Riihimäen rautatie, Karjaan asema

Person (PER) Markable person names include:

1. First names: e.g. Elias, Liisa

2. Family names: e.g. Lönnrot, Ylitalo

Organization (ORG) Markable 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

1. Senates, parliaments, governments: e.g. Suomen keisarillinen senaatti, Englannin parlamentti, Venäjän hallitus

1. Bureaus: e.g. Konginkankaan pastorinvirasto, Tuuloksen kirkkoherranvirasto, Heinolan kaupungin maistraatti, Hämeen läänin maakanslia, Raaseporin kihkakunnanoikeus, Turun hovioikeus, Kuopion raastuvanoikekus

1. Armies, regiments, battalions: e.g. Ruotsin armeija, Porin rykmentti, Turun pataljoona

1. Congregations, dioceses: e.g. Kuopion seurakunta, Kuopion hiippakunta

1. Chapters: e.g. Turun tuomiokapituli, Porvoon hiippakunnan tuomiokapituli

1. Judicial districts : e.g. Rannan tuomiokunta

1. Storest, factories, inns, hotells, restaurants: e.g. O. Jalanderin kirjakauppa, Ruikan kestikievari, Bahnen puoti, Daalintehdas, Phoenix-hotelli, Phoenix-ravintola

1. Companies, enterprises: Kirkollinen kirja- ja paperikauppa O Y, Werner Söderström Osakeyhtiö, Turun Rautakalutehdas-yhtiö, Georg Segerstrålan Lakiasiain toimisto, Hämeen Sanomain kirjapaino

1. Banks: e.g. Suomen pankki, Englannin pankki, Pohjoismaiden säästöpankki

1. Newspapers and journals: e.g. Suomen Wiikkolehti, Pyhäkoululehti, Uusi Suometar

2.3 Comments on Annotation Guidelines

In the following, we address some general aspects of the annotation.

Abbreviations Abbreviated word tokens are considered markable if they appear in a multi-token entity with one or more non-abbreviated word tokens. 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 markable: ”E. L.”, ”M. Laurinp. Y.”.

Multi-Word Entities and Nesting The marked entities can span multiple word tokens, for example, consider "Elias Lönnrot" or ”Suomen Pankki”. 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 Pankki” (Bank of Finland) is marked as an organization while its sub-part ”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 Pankki" or ”O. Jalanderin kirjakauppa” (bookstore of O. Jalander) where the organization entity contains three tokens in addition to the nested person entity ”O. Jalander”.

In previous work, the non-nested annotation has been the prevalent annotation strategy, apart from a few expections (Byrne, 2007; Benikova et al., 2014). This is most likely due to the added annotation effort required by the nested approach.

2.4 Annotation Process

We first annotated manually 170 pages (248,544 word tokens). Subsequently, we learned a Stanford NER system using these pages, tagged the remaining 101 pages (211,034 word tokens) using the resulting system and manually corrected the automatically annotated pages.

2.5 Annotation Statistics

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

class	count	%
PER	5102	48.88
LOC	6285	39.68
ORG	1471	11.44
Total	12858	100.0
PER	5355	50.85
LOC	6981	39.00
ORG	1394	10.15
Total	13730	100.0

Table 1: Counts and relative portions of named entity classes. The upper and lower blocks correspond to manual and semi-manual annotation, respectively.

2.6 Gazetteers

In addition to manually prepared named entity annotation, our corpus is accompanied by a gazetteers which map words into semantically motivated categories. *[how were the gazetteers made?]*

3 Experiments

In this section, we present experimental results on the corpus employing the Stanford Named Entity Recognizer toolkit (Finkel et al., 2005). In what follows, we will first describe the utilized training and test set splits in Section 3.1. We then discuss the Stanford Named Entity Recognizer in Section 3.2 and the employed evaluation measures in Section 3.3. Finally, we present and discuss the obtained results in Section 3.4.

3.1 Data

As presented in Section 2.5, the complete corpus consists of 271 pages (459,578 word tokens). In order to carry out the experiments, we separate the data into two non-overlapping sections, training and test sets. In the training set, we include 136 manually annotated pages and all 101 semi-manually annotated pages. Meanwhile, the test set consists of the remaining 34 manually annotated pages. The resulting training and test sections, therefore, contain 237 and 34 pages (381,356 and 67,223 word tokens), respectively. Furthermore, we create a second version of the test pages, where the text has been produced by an automatic OCR system instead of manually recognized (the NE annotation are manually corrected).

3.2 Stanford Named Entity Recognizer

The Stanford Named Entity Recognizer (Finkel et al., 2005) is a freely available NER toolkit based on machine learning methodology.[1] 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 (Lafferty et al., 2001; Finkel et al., 2005). We use a wide range of substring, word context and orthography features. In addition, we use the gazetteers described in Section 2.6.

3.3 Evaluation Measures

We evaluate the system using the standard **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 the number of all annotated entities in data), and **F1-score** (the harmonic mean of precision and recall).

3.4 Results

Obtained results on the groundtruth test set are presented in Table 2. The system yielded an overall F1-score of 82.01 with precision and recall scores of 87.90 and 76.87, respectively.

Obtained results on the OCR test set are presented in Table 3. The table contains two sets of values. The upper block presents the loss in recall caused by

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

class	precision	recall	F1	# found	# gold standard
LOC	0.9006	0.8577	0.8786	1740	1826
PER	0.8478	0.7627	0.8030	1084	1205
ORG	0.8636	0.4409	0.5838	242	473
Totals	0.8790	0.7687	0.8201	3066	3504

Table 2: Precision, recall, and F-scores for each named entity class on the Groundtruth test set.

class	precision	recall	F1	# found	# gold standard
LOC	1.0000	0.8719	0.9315	1591	1826
PER	1.0000	0.9046	0.9499	1088	1205
ORG	1.0000	0.6871	0.8145	324	473
Totals	1.0000	0.8582	0.9237	3003	3504
LOC	0.8896	0.7234	0.7979	1485	1826
PER	0.8068	0.6515	0.7208	973	1205
ORG	0.7669	0.2643	0.3931	163	473
Totals	0.8512	0.6367	0.7285	2621	3504

Table 3: Precision, recall, and F-scores for each named entity class on the OCR test set. The upper block shows the performance loss due to the OCR. The lower block shows the overall recognition performance.

the OCR process and hence the upper limit to the NER performance. The lower block, meanwhile, shows the final NER performance taking into account both errors yielded by the OCR process and the Stanford NER system.

4 Conclusions

References

Darina Benikova, Chris Biemann, and Marc Reznicek. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2524–2531, 2014.

Kate Byrne. Nested named entity recognition in historical archive text. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 589–596, 2007.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 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)*, pages 363–370, 2005.

Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: A brief history. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING 1996)*, volume 96, pages 466–471, 1996.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, 2001.