A Historical Finnish Corpus for Named Entity Recognition

November 16, 2017

Abstract

We present a corpus of Finnish news articles with a manually prepared named entity annotation. The corpus consists of 1,194 articles (260,637 word tokens) with seven named entity classes (organization, location, person, product, date, event, and person title). The articles are extracted from the archives of Digitoday, a Finnish online technology news source. The corpus is freely available for research purposes. In experiments employing the Stanford Named Entity Recognizer, the system yields an overall F1-score of 82.42. The result provides a baseline for future NER systems developed using the corpus.

1 Introduction

In this work, we present a corpus consisting of Finnish technology related news articles with a manually prepared named entity annotation. The corpus is freely available for research purposes and can be readily used for development of NER systems for Finnish.[1] To our knowledge, this is the first freely available NER corpus published for Finnish. Moreover, in addition to providing a description of the corpus, we examine it empirically using the well-known Stanford Named Entity Recognizer (**?**). The results obtained using the system provide a baseline for future NER systems developed using the corpus.

The rest of the paper is organized as follows. We first describe the corpus in Section **??**. We then present experimental results on the corpus using the Stanford Named Entity Recognizer in Section **??**. Finally, conclusions on the work are presented in Section **??**.

[1] The corpus is available at `https://github.com/mpsilfve/finer-data`

2 Corpus

2.1 Text

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 (**?**).

Person (PER) Markable person names include:

1. First names: e.g. Sauli, Barack

2. Family names: e.g. Niinistö, Obama

3. Aliases: e.g. DoctorClu, Kim Dotcom

Location (LOC) Markable locations include:

1. Buildings: e.g. Valkoinen talo (the White House)

2. Cities: e.g. Helsinki, New York

3. Continents: e.g. Eurooppa (Europe)

4. Countries: e.g. Suomi (Finland)

5. Planets: e.g. Mars

Organization (ORG) Markable organizations include:

1. Commercial companies: e.g. Nokia, Apple

2. Communities of people: e.g. Google Orkut

3. Education and research institutes: e.g. Turun Yliopisto

4. News agencies/News services/Newspapers/Newsrooms/News sites/News blogs: e.g. Reuters, Helsingin Sanomat

5. Political parties: e.g. Kokoomus (the National Coalition Party)

6. Public administration: e.g. Suomen hallitus (the Finnish Government), Euroopan Unioni (the European Union)

7. Stock exchange: e.g. New Yorkin pörssi (New York Stock Exchange)

8. Television network/station/channel: e.g. MTV3, FOX

2.3 General Annotation Guidelines

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

Multi-Word Entities and Nesting

Derivation and Compounding To be markable, each class must appear as a whole and not as a derivation or as a part of a token. For example, ”Suomi” (Finland) is considered as a markable location, whereas ”suomalainen” (Finnish) and ”Suomi-fani” (a fan of Finland) are not. An important exception to this rule is the set of cases, in which a markable named entity forms a part of a compound word while the whole compound word refers to that same entity: for example, consider the expression ”Google-ohjelmistoyhtiö” (the software company Google) instead of ”Google” or ”iPhone-puhelin” (iPhone phone) instead of ”iPhone”. In these cases, the compound word is considered markable.

Coordination *[]*

Another coordination issue addresses such cases as ”Windows 8 ja 10” (Windows 8 and 10) or ”iPhone 6 sekä 6s Plus” (iPhone 6 as well as 6s Plus). We consider these cases to be markable as single expressions. This rule also generalizes to multiple tokens: for example, consider ’**Windows XP, Vista ja 10**”.

Expression *-niminen* Written Finnish regularly employs a type of expression, in which a name is associated with a noun using a suffix ”-niminen”: for example, consider ”Sauli-niminen henkilö” (a person named Sauli) or ”iPhone-niminen puhelin” (a phone named iPhone). In these cases, we consider the complete expression markable, that is, ’Sauli-niminen henkilö” and ”iPhone-niminen puhelin” is marked as single *person* and *product* entities, respectively. We generalize this rule to also cover cases ”-merkkinen” (of brand) and ”-mallinen” (of model): for example, consider ”Tesla-merkkinen auto” (a car of the brand Tesla) and ”Samsung-mallinen puhelin” (a phone of the model Samsung).

2.4 Annotation Process

The annotation and the annotation guideline were created iteratively by a primary annotator (first author) and two auxiliary experts (second and third authors) as follows. First, an initial annotation guideline was created based on literature (**??????**) and samples from the corpus. The material was then annotated using this guideline by the primary annotator. Subsequently, the guideline and annotation were revised for two more iterative passes over the data set. The work was performed over a span of one year.

2.5 Annotation Statistics

The complete data set consists of 1,194 articles (260,637 word tokens). The counts of each named entity class in these sections are presented in Table **??**.

class	count	%
ORG	11,001	46.2
PROD	5,582	23.4
PER	2,636	11.1
LOC	2,544	10.7
DATE	1,194	5.0
TITLE	759	3.2
EVENT	110	0.5
TOTAL	23,826	100.0

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

2.6 Gazetteers

In addition to manually prepared named entity annotation, our corpus is accompanied by three different gazetteers which map words into semantically motivated categories. We construct gazetteers both in a supervised manner, using knowledge bases, and in an unsupervised manner, based on word clusters.

Our first gazetteer associates words with category labels extracted from Finnish Wikipedia articles in an automatic manner.[2] Category labels are extracted from so called *infoboxes*[3] which are fixed-format tables that present information regarding a group of articles belonging to the same category. Examples of categories include ”yritys” (company) and ”kaupunki” (town). Due to homonymy, a word can belong to several different categories. For example, the word ”Nokia” belongs both

[2] `https://github.com/mpsilfve/wikipedia_gazetteer`

[3] `https://en.wikipedia.org/wiki/Help:Infobox`

to the category ”yritys” (company) and to the category ”Suomen kunta” (Finnish municipality).

Our second gazetteer employs proper noun labels from the OMorFi morphological analyzer (**?**). OMorfi gives 8 different subcategories for proper nouns: CULT-GRP (cultural group), EVENT (event), FIRST (first name), GEO (geographical location), LAST (last name), ORG (organization), PROD (product), and MISC (miscellaneous). As in the case of Wikipedia category labels, a word can belong to several different categories. For example, ”Nokia” belongs both to the categories GEO and ORG.

Our final gazetteer is based on word clusters which are derived from the Suomi24 corpus[4] which contains text material from discussion forums of the Finnish Suomi24 online social networking website. The corpus contains material from the year 2001 up to 2015. In order to construct word clusters, we first train word embeddings for the 170K most common word forms in the Suomi24 corpus using the skip-gram model (**?**) from the Gensim toolkit (**?**). We then cluster embedding vectors into 200 clusters using the standard k-means algorithm and use cluster identity numbers as gazetteer categories. Although identity numbers do not carry any intrinsic semantic meaning, they are useful because semantically similar words frequently occur in the same cluster. For example, ”Microsoft”, ”Samsung”, ”Apple”, and ”Nokia” all belong to the same cluster (cluster number 192).

2.7 File Format

The data is presented in a two-column file format using the standard BIO notation with an empty line separating sentences. For example, consider the following sentence ”Applen Tim Cook myy iPhoneja suomalaisille.” (”Apple's Tim Cook sells iPhones to the Finns.”):

Applen	B-ORG
Tim	B-PER
Cook	I-PER
myy	O
iPhoneja	B-PRO
suomalaisille	O
.	O

[4] `http://urn.fi/urn:nbn:fi:lb-201412171`

3 Experiments

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

3.1 Data

As presented in Section **??**, the complete corpus consists of 1,194 articles (260,637 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 all articles published during 2014, while the test set consists of articles published in 2015. The resulting training and test sections contain 953 and 241 articles (210,132 and 50,505 word tokens), respectively. In addition, we form a separate development set from the training data by extracting every 10th article, starting from the 10th article. (The articles are separated using the headline information referred to in Section **??**.)

3.2 Stanford Named Entity Recognizer

The Stanford Named Entity Recognizer (**?**) is a freely available NER toolkit based on machine learning methodology.[5] 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 (**??**). We use a wide range of substring, word context and orthography features. In addition, we use the gazetteers described in Section **??** as well as a gazetteer containing the lemmas and morphological tags yielded by the FinnPos tagger (**?**).[6] The morphological tagger FinnPos contains two separate tagging and lemmatization models learned from Finnish Turku Dependency Treebank (TDT) (**?**) and FinnTreeBank (FTB) (**?**), of which we employ the latter.

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 num-

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

[6] See the configuration file.

class	precision	recall	F1
ORG	85.93	86.43	86.18
PRO	73.82	69.41	71.55
PER	70.09	75.62	72.75
LOC	86.02	89.13	87.55
DATE	88.06	86.69	87.37
TITLE	91.13	87.60	89.33
EVENT	100.00	41.18	58.33
ALL	82.77	82.08	82.42

Table 2: Precision, recall, and F-scores for each named entity class.

ber 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 are presented in Table **??**. The system yielded an overall F1-score of 82.42 with precision and recall scores of 82.77 and 82.08, respectively. In what follows, we will discuss the individual class performances.

First, as for the conventional enamex classes (organization, location, and person names), we note that, perhaps surprisingly, the person names were most challenging to learn. While organization and location names yielded F1-scores of 86.18 and 87.55, respectively, the person names yielded a much lower F1-score of 72.75. According to manual inspection, this appears to be largely due to the system's difficulty in differentiating between English person names from English company and product names which are abundant in the data. Similarly, the low F1-score of 71.55 obtained for product names stems from the system's difficulty of differentiating English product names from English organization and person names. In addition, compared to organizations and locations, product names are more versatile given that they can contain version numbers and version names according to the annotation guideline discussed in Section **??** which makes learning more tedious. Meanwhile, these difficulties are not reflected heavily in the F1-score obtained for organization class since its F1-score is elevated by frequently appearing and easily recognizable company names. such as ”Apple” and ”Google”.[7]

Second, dates and titles appear to be the easiest classes to learn. This is some-

[7] Four most frequent company names (Apple, Google, Microsoft, Samsung) comprise 17% of all marked organizations in the corpus.

what expected since they form the most restricted classes according to the annotation guideline discussed in Section **??**.

Finally, we note that the results concerning the event class should be ignored since the small amount of annotated entities (see the class statistics in Table **??**) does not enable reliable evaluation.

4 Conclusions

In this paper, we described a new corpus consisting of Finnish news articles with a manually prepared named entity annotation. The corpus contains 1,194 articles (260,637 word tokens) annotated using a set of seven named entity classes, namely, *persons*, *organizations*, *locations*, *products*, *events*, *dates*, and *person titles*. In addition, the corpus is accompanied by three gazetteers which map words into semantically motivated categories. The corpus is freely available for research purposes. In addition to description of the data, we presented experimental results on the corpus employing the Stanford Named Entity recognizer. The resulting system yielded a total F1-score of 82.42. The result provides a baseline for future NER systems developed using the corpus.