A Finnish News Corpus for Named Entity Recognition

Teemu Ruokolainen Miikka Silfverberg Krister Lindén

February 1, 2018

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

Named entity recognition (NER) is a fundamental textual information extraction task, in which the aim is to locate and classify named entity expressions into predefined classes (Nadeau and Sekine, 2007). The set of classes typically include such entities as *persons*, *locations*, and *organizations*. While the core NER task as known today was originally proposed already in the late 1990s (Grishman and Sundheim, 1996), developing novel NER systems and techniques continues to be an active field of research in natural language processing to this day (Lample et al., 2016; Strauss et al., 2016; Leaman and Lu, 2016; Nguyen et al., 2016; Van Tran et al., 2017).

Development of automatic NER systems requires manually annotated data, that is, a corpus of running text with manually located and classified named entity tags. In addition to being employed to evaluate system performances, the data is necessary for training models with machine learning techniques.[1] Such data resources

[1]While there does exist a body of work aiming at learning named entity recognition systems from unannotated data in an unsupervised manner, supervised learning from annotated data has been the prevalent approach in literature.

are currently available for multiple Western languages. For example, consider work on English (Grishman and Sundheim, 1996; Tjong Kim Sang and De Meulder, 2003), German (Tjong Kim Sang and De Meulder, 2003; Faruqui et al., 2010; Benikova et al., 2014), Swedish (Dalianis and Åström, 2001; Kokkinakis et al., 2014), and French (Petasis et al., 2001; Poibeau, 2003).

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.[2] 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 (Finkel et al., 2005). 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 2. We then present experimental results on the corpus using the Stanford Named Entity Recognizer in Section 3. Finally, conclusions on the work are presented in Section 4.

2 Corpus

In this section, we describe the corpus in terms of chosen text material, set of named entity classes, and annotation process. When appropriate, we provide context for our specific choices using examples of NER corpora published for English, German, and Swedish (Grishman and Sundheim, 1996; Tjong Kim Sang and De Meulder, 2003; Faruqui et al., 2010; Benikova et al., 2014; Dalianis and Åström, 2001; Kokkinakis et al., 2014).

2.1 Text

The text material is extracted from the archives of Digitoday, a Finnish online technology news source. The material covers a variety of technology related topics, such as business, science, and information security. The material is extracted from articles published between years 2014 and 2015 under the Creative Commons (CC BY-ND-NC 1.0 FI) licence. The extracted text contains 1,194 articles (260,637 word tokens) in total. In addition to raw word forms, the corpus includes meta data describing the publishing date of the articles and if the word tokens belong to a headline, an ingress, or an article body. Figures, tables, and respective captions were not included in the corpus.

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

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). These three classes appear in the vast majority of published work on NER (Nadeau and Sekine, 2007) and for these three classes our annotation guideline follows the previous work. In addition, our annotation contains the classes *products*, *events*, *dates*, and *person titles* which have been employed in previous work to varying extent. In the following, we discuss these classes in more detail.

Person (PER) Description.
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) Description.
Markable locations include:

1. Buildings: e.g. Valkoinen talo (the White House), Lasipalatsi, World Trade Center
2. Cities, towns, city districts: e.g. Helsinki, New York, Pälkäne, Katajanokka
3. Continents: e.g. Eurooppa (Europe)
4. Countries, states: e.g. Suomi (Finland), Kalifornia (California)
5. Geographical areas: e.g. Latinalainen Amerikka (Latin America), Pohjoismaat (Nordic countries), Itä-Eurooppa (eastern Europe), manner-Kiina (mainland China)
6. Parks: Huis Ten Boch -teemapuisto, Yosemiten kansallispuisto (Yosemite national park)
7. Planets, celestial objects: e.g. Mars, Maapallo (Earth), Kuu (Moon)
8. Seas, lakes, rivers: Atlantti (the Atlantic), Volga

Organization (ORG) Description.
Markable organizations include:

1. Commercial companies: e.g. Nokia, Apple, Time Warner
2. Commissions: e.g. Kalifornian Public Utilities -komissio
3. Communities and groups of people: e.g. Google Orkut, The Kinks
4. Education, research, and scientific institutes: e.g. Turun Yliopisto, Carnegie Mellon University, Poliisiammattikorkeakoulu (the (Finnish) Police University College), Euroopan avaruusjärjestö (the European Space Association), Suomen Ilmatieteen laitos (The Finnish Meteorological Institute)
5. Judicial systems: e.g. Helsingin hovioikeus (Helsinki court of appeals), Yhdysvaltain korkein oikeus (The Supreme Court of the United States), Euroopan Unionin tuomioistuin (The Court of Justice of the European Union)
6. Law enforcement organizations: e.g. Keskusrikospoliisi (the (Finnish) National Bureau of Investigation), Yhdysvaltain liittovaltion poliisi (the Federal Bureau of Investigation), Australian poliisi (Australian police force)
7. News agencies/News services/Newspapers/Newsrooms/News sites/News blogs: e.g. Reuters, Helsingin Sanomat, Foss Patents -blogi
8. Political parties: e.g. Kokoomus (the National Coalition Party)
9. Public administration: e.g. Suomen hallitus (the Finnish Government), Euroopan Unioni (the European Union), Ulkoministeriö (the (Finnish) Ministry for Foreign Affairs), Tulli ((Finnish) Customs)
10. Sport leagues: e.g. National Hockey League (NHL),
11. Stock exchange, banks: e.g. New Yorkin pörssi (New York Stock Exchange), Suomen Pankki (the Bank of Finland)
12. Television networks/stations/channels: e.g. MTV3, FOX
13. Websites (referring to the underlying organization): e.g. Amazon.com, Verkkokauppa.com

Product (PROD) Markable products include:

1. Artifacts: e.g. IPhone 6, Lumia-puhelin, Hubble-avaruuskaukoputki (Hubble telescope), Compute Stick, Playstation, Snapdragon-suoritin, Watson-supertietokone
2. Laws: e.g. Patriot Act-laki (the Patriot Act law)
3. Networks (other than television network): e.g. Tor
4. Platforms: e.g. Google Play, Kickstarter
5. Product series/collections: e.g. -sarja, -mallisto
6. Programming languages: e.g. Java, C++
7. Projects/Programs: e.g. Vitja, VideoLan-projekti, Blue Shield -sairausvakuutusohjelma
8. Protocols: e.g. pop3, imap
9. Services/Platforms: e.g. Apple Store, Google Play, Tor Mail -sähköpostipalvelu
10. Software: e.g. Windows 10, Trojan-Banker.Win32.Chthonic, Dropbox
11. Systems: e.g. Alipay-järjestelmä
12. Technologies: e.g. HoloLens-teknologia
13. Vehicles/Vessels: e.g. Tesla Model S, Opportunity, Kansainvälinen avaruusasema (International Space Station), Helsingin Metro
14. Websites: e.g. Amazon.com-sivusto, Tvkaista.fi-verkko-osoite
15. Works/Art: e.g. Tuntematon Sotilas, Hurt Locker, Angry Birds, American Idol

Version numbers of products are considered markable parts of the whole expression if they appear in the immediate context of the product name: for example, consider ”Windows 10”, ”iPhone 6”, ”Android 5.0”, ”Androidin versio 5.0” (version 5.0 of Android). In contrast, version names are considered markable even if they appear individually: for example, consider ”Vista” and ”Lollipop” referring to ”Windows Vista” and ”Android Lollipop”, respectively.

Entities considered here to belong in the *product* class have been incorporated in previous work to varying extents. For example, in the English CoNLL-2003 data (Tjong Kim Sang and De Meulder, 2003), projects/programs are included in

an entity class labeled as *miscellaneous*. Similarly, in the German NoSta-D corpus (Benikova et al., 2014), products are included in an entity class labeled *other*. Meanwhile, in the Swedish HSFT-SweNER system, Kokkinakis et al. (2014) define classes *artifact* (food/wine products, prizes, means of communication (vehicles), etc.) and *work&art* (printed material, names of films, novels and newspapers, sculptures, etc.), both of which are subsets of our *product* class.

Event (EVENT) Markable events include:

1. Expos: e.g. CES-messut
2. Explicitly marked events: e.g. Mobile World-tapahtuma

Again, entities considered in this work as *events* have been incorporated in previous work to varying extents. For example, the English and German CoNLL-2003 data (Tjong Kim Sang and De Meulder, 2003) assign events to the class labeled *miscellaneous*, while the German NoSta-D corpus (Benikova et al., 2014) assigns events to the class *other*. Meanwhile, the Swedish HSFT-SweNER system (Kokkinakis et al., 2014) has a separate class for events.

Date (DATE) Markable date expressions include:

1. 1.10.2016
2. lokakuussa (in October)
3. 1. lokakuuta (1st October)
4. 1. päivä lokakuuta (1st day of October)
5. lokakuun 1. (1st October)
6. lokakuun 1. päivä (1st day of October)
7. 2016
8. vuonna 2016 (in the year 2016)
9. syyskuussa 2016 (in October 2016)
10. vuoden 2016 syyskuussa (in October 2016)
11. 1. ja 2. lokakuuta (1st and 2nd October)

12. lokakuun 1. ja 2. päivä (1st and 2nd October)
13. vuosina 2000 ja 2001 (during the years 2000 and 2001)
14. 1.-10. lokakuuta (from 1st to 10th October)
15. 1. - 10. lokakuuta (from 1st to 10th October)
16. tammikuusta lokakuuhun (from January to October)
17. tammi-lokakuussa (between January and October)
18. 2000-2001
19. 2000 - 2001
20. vuodesta 2000 vuoteen 2010 (from the year 2000 to 2001)
21. vuoden 2000 lokakuusta vuoden 2001 tammikuukuuhun (from October 2000 to January 2001)
22. vuoden 2000 tammikuusta lokakuuhun (from January to October in 2000)

According to our definition employed here, a date is a time expression which can be expressed as a triplet (day, month, year): for example, consider ”3.10.2016” and ”3. lokakuuta 2016” (3rd October 2016). To be markable, at least the month or year has to be explicitly specified: for example, consider ”3. lokakuuta” (3rd October), ”lokakuussa 2016” (in October 2016), and ”2016”. Therefore, expressions such as ”3. päivä tätä kuuta” (3rd of this month) are not markable. Numerics values can be spelled out or expressed using digits.

Expressions such as ”vuonna 2016” (in the year 2016) and ”3. päivä lokakuuta" (3rd day of October) are considered markable as a whole. This is because the expressions ”3. päivä lokakuuta” and ”vuonna 2016” have the same meaning as ”3. lokakuuta” and ”2016”, respectively. Months may sometimes have a purely nominal use and are not markable in those instances: for example, consider ’Tammikuu on kylmä kuukausi” (January is a cold month).

Expressions employing coordination such as ”lokakuun 1. ja 2. päivä” (1st and 2nd October) and ”vuosina 2000 ja 2001” (during the years 2000 and 2001) are considered markable as a whole. Similarly, from-to expressions are markable as a single expression: for example, consider ”2000-2010”, ”2000 - 2010”, ”vuodesta 2000 vuoteen 2010”, ”1.-3. lokakuuta”, ”lokakuun ensimmäisestä päivästä kolmanteen”. Adpositions within from-to expressions are considered markable: for example, consider ”vuosien 2010 ja 2011 aikana” (during the years 2010 and 2011).

Dates, or more generally temporal expressions, have been incorporated in previous work to varying extents. For example, in the English and German CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and the German NoSta-D (Benikova et al., 2014) data, temporal expressions are not considered. Meanwhile, the MUC-6 and MUC-7 data sets and the HSFT-SweNER system (Kokkinakis et al., 2014) take into account a wide range of temporal expressions, including absolute expressions such as ”3.10.2016” and relative ones such as ”next week”. In addition, it should be noted, that recognition of time expressions and temporal relations has been a research topic of its own interest: see, for example, the TempEval competition (Verhagen et al., 2007, 2010; UzZaman et al., 2013).

Title (TITLE) Titles for people are markable if they appear in the immediate precontext of proper names. For example, in the following, markable titles are bolded: **presidentti** Sauli Niinistö (**president** Sauli Niinistö), **perustaja** ja **varapuheenjohtaja** Ville Oksanen (**founder** and **vice president** Ville Oksanen). While applied to lesser extent in named entity recognition data, title annotation can be found, for example, in a related work on semantic webs (Ehrmann et al., 2017).

2.3 General Annotation Guidelines

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

Multi-Word Entities and Nesting The marked entities can span multiple word tokens, for example, consider "Sauli Niinistö" or ”Nokia Solutions and Networks”. 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 ”Manchester United” is marked as an organization while its sub-part ”Manchester” is marked as a location. In our annotation, however, we follow the non-nested annotation approach, that is, only the longer span is considered markable. In previous work, this has been the prevalent annotation strategy, apart from a few expections (Byrne, 2007; Benikova et al., 2014).

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 Written Finnish regularly employs a type of coordination between compound words which share a common part. For example, the expression ”Windows-käyttöjärjestelmä ja Linux-käyttöjärjestelmä” (Windows operating system and Linux operating system) is written in a more concise manner as ”Windows- ja Linux-käyttöjärjestelmät” (Windows and Linux operating systems). In these cases, the tokens ”Windows-" and ”Linux-käyttöjärjestelmät” are both considered markable. This rule generalizes to multiple tokens as well: for example, consider ”**Windows-**, **Linux**- ja **OSX-käyttöjärjestelmät**”.

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 (Grishman and Sundheim, 1996; Tjong Kim Sang and De Meulder, 2003; Faruqui et al., 2010; Benikova et al., 2014; Dalianis and Åström, 2001; Kokkinakis et al., 2014) 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 1.

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.[3] Category labels are extracted from so called *infoboxes*[4] 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 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 (Pirinen, 2008). OMorfi gives 8 different subcategories for proper nouns: CULTGRP (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

[3] https://github.com/mpsilfve/wikipedia_gazetteer

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

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[5] 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 (Mikolov et al., 2013) from the Gensim toolkit (Řehůřek and Sojka, 2010). 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

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.

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

3.1 Data

As presented in Section 2.5, 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 2.)

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.[6] 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 as well as a gazetteer containing the lemmas and morphological tags yielded by the FinnPos tagger (Silfverberg et al., 2016).[7] The morphological tagger FinnPos contains two separate tagging and lemmatization models learned from Finnish Turku Dependency Treebank (TDT) (Haverinen et al., 2014) and FinnTreeBank (FTB) (Voutilainen, 2011), 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 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 are presented in Table 2. 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.

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

[7] 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.

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 2.2 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".[8]

Second, dates and titles appear to be the easiest classes to learn. This is somewhat expected since they form the most restricted classes according to the annotation guideline discussed in Section 2.2.

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 1) does not enable reliable evaluation.

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

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.

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.

Hercules Dalianis and Erik Åström. SweNam - Swedish named entity recognizer. *Technical Report, Kungliga Tekniska Högskolan*, 2001.

Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. JRC-Names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295, 2017.

Manaal Faruqui, Sebastian Padó, and Maschinelle Sprachverarbeitung. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of Die Zehnte Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2010)*, pages 129–133, 2010.

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.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531, 2014.

Dimitrios Kokkinakis, Jyrki Niemi, Sam Hardwick, Krister Lindén, and Lars Borin. HFST-SweNER–A new NER resource for Swedish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2537–2543, 2014.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Robert Leaman and Zhiyong Lu. TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2013)*, pages 3111–3119, 2013.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. J-NERD: Joint named entity recognition and disambiguation with rich linguistic features. *Transactions of the Association for Computational Linguistics*, 4:215–229, 2016.

Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D Spyropoulos. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of the Thirty-Ninth Annual Meeting on Association for Computational Linguistics (ACL 2001)*, pages 426–433, 2001.

Tommi Pirinen. Automatic finite state morphological analysis of Finnish language using open source resources (in Finnish). Master's thesis, University of Helsinki, 2008.

Thierry Poibeau. The multilingual named entity recognition framework. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics (EACL 2003)*, volume 2, pages 155–158, 2003.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. FinnPos: An open-source morphological tagging and lemmatization toolkit for Finnish. *Language Resources and Evaluation*, 50(4):863–878, 2016.

Benjamin Strauss, Bethany E Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the Second Workshop on Noisy User-generated Text (WNUT 2016)*, pages 138–144, 2016.

Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning (HLT-NAACL 2003)*, volume 4, pages 142–147, 2003.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, volume 2, pages 1–9, 2013.

Cuong Van Tran, Tuong Tri Nguyen, Dinh Tuyen Hoang, Dosam Hwang, and Ngoc Thanh Nguyen. Active learning-based approach for named entity recognition on short text streams. In *Multimedia and Network Information Systems. Advances in Intelligent Systems and Computing*, volume 506, pages 321–330. 2017.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 Task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluation (SEMEVAL 2007)*, pages 75–80, 2007.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SEMEVAL 2010)*, pages 57–62, 2010.

Atro Voutilainen. FinnTreeBank: Creating a research resource and service for language researchers with constraint grammar. In *Proceedings of the NODALIDA 2011 Workshop Constraint Grammar Applications*, pages 41–49, 2011.