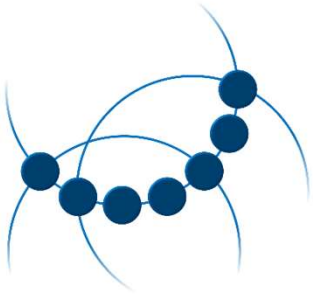
FIN-CLARIN

CLARIN
Common Language Resources and Technology Infrastructure



Nokia – Person, Location, Organisation, Product, Event, Time or Common Noun?

Automatically Recognizing and Categorizing Names in Finnish Text

FIN-CLARIN

What is the Entity Named Nokia?

CLARIN

- “Mr. Nokia: Ongelma huipulla” /*Mr. Nokia: Problem at the Top*/ (Uusi Suomi, 9/2010) [**Person**]
- “Yksi ihminen loukkaantui vakavasti ulosajossa Nokialla” /*One person was severely injured as his car derailed in Nokia*/ (Aamulehti 11/2019) [**Location**]
- “Nokia Oyj on suomalainen maailmanlaajuisesti toimiva tietoliikennealan yhtiö” /*Nokia Corporation is a Finnish multinational telecommunications company*/ (Wikipedia) [**Organisation**]
- “Uusi Nokia DNA:lta” /*New Nokia from DNA*/ (DNA web page) [**Product**]
- “Vuoden 2019 Nokia-viikkoa vietettiin 18.-26.5.2019.” /*The 2019 Nokia Week took place on May 18-26, 2019.*/ (Nokia municipality web page) [**Event**]
- “Nokia eli soopeli (Martes zibellina), näätäeläin” /The sable (Martes zibellina) is a species of marten/ (Wikipedia) [**Common Noun**]

FIN-CLARIN

Who needs Named Entity Recognition

CLARIN

Information extraction to provide

- Message routing, e.g. for news providers or customer support
- Efficient search algorithms
- Content recommendations
- Pseudonymization services
- Creation of linked open data
- …

FIN-CLARIN

What is a Named Entity

CLARIN

Main categories:

- Person
- Location
- Organization
- Product
- Event
- Date

FIN-CLARIN

Can’t we just have a list?

CLARIN

Person names include:

1. First names: e.g. Sauli, Barack
2. Family names: e.g. Niinistö, Obama
3. Aliases, nicknames: e.g. DoctorClu, Kim Dotcom
4. Combinations of 1-3: e.g. Sauli Niinistö, Marko ”Fobba” Forss
5. Fictional/mythological characters: e.g. Joulupukki (Santa Claus)

FIN-CLARIN

Can’t we just have a list? …

CLARIN

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

FIN-CLARIN

Can’t we just have a list? …

CLARIN

Organizations include:

1. Commercial companies: e.g. Nokia, Apple, Time Warner
2. Communities/groups of people: e.g. Google Orkut, The Kinks, Stop the cyborgs - kansalaisliike
3. 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)
4. Judicial organizations: 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)
5. 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), New Yorkin poliisi (New York police department)

FIN-CLARIN

Can’t we just have a list? …

CLARIN

6. News agencies/News services/Newspapers/Newsrooms/News sites/News blogs: e.g. Reuters, Helsingin Sanomat, Foss Patents -blogi
7. Political parties: e.g. Kokoomus (the National Coalition Party)
8. Public administration: e.g. Suomen hallitus (the Finnish Government), Euroopan Unioni (the European Union), ulkoministeriö (the Finnish Ministry for Foreign Affairs), Tulli (Finnish Customs), Helsingin kaupunki (City of Helsinki)
9. Sport leagues: e.g. National Hockey League (NHL)
10. Stock exchange, banks: e.g. New Yorkin pörssi (New York Stock Exchange), Suomen Pankki (the Bank of Finland)
11. Television networks/stations/channels: e.g. MTV3, FOX
12. Websites (referring to the underlying organization): e.g. Amazon.com, Verkkokauppa.com
13. …

FIN-CLARIN

Annotating the whole name

CLARIN

- Expressions: -niminen, -mallinen, -merkkinen, ... e.g.
  - “Sauli-niminen henkilö” (a person named Sauli) or
  - “iPhone-merkkinen puhelin” (a phone branded iPhone)
- Compounds and derivatives:
  - “Google-ohjelmistoyhtiö” (the software company Google) instead of “Google”
  - “iPhone-puhelin” (iPhone phone) instead of “iPhone”
- Coordination:
  - “Windows- ja Linux-käyttöjärjestelmät” (Windows and Linux operating systems)
  - “Windows 8 ja 10” (Windows 8 and 10) or
  - “iPhone 6 sekä 6s Plus” (iPhone 6 as well as 6s Plus)
  - ’Windows XP, Vista ja 10” (Windows XP, Vista and 10)
- Abbreviations and acronyms:
  - ”Yhdysvaltain Tiedusteluvirasto (NSA)”, ”Yhdysvaltain Tiedusteluvirasto NSA”

FIN-CLARIN

Nested Entities

CLARIN

- A nested entity cannot be of the same length as its enclosing entity, i.e. Nokia cannot be left ambiguous with regard to location and organization, e.g. $[[\text{Nokia}]_{\text{LOC}}\ \text{week}]_{\text{EVENT}}$
- We did not allow a nested entity to be of the same class as its enclosing entity, i.e. $[\text{Microsoft Research}]_{\text{ORG}}$
- A top-level entity can have more than one consecutive nested entity or levels of nesting

FIN-CLARIN

Training and Test Data

CLARIN

- Digitoday, a Finnish online technology news source.
- Training: 953 articles (193,742 word tokens) published in 2014
- Test: 240 articles (46,363 word tokens) published in 2015
- Statistics:

Entity	Top	Nested
ORG	9137	279
PER	2214	113
LOC	2022	410
DATE	955	2
PRO	4446	0
EVENT	93	0
TOTAL	18863	804

Entity	Top	Nested
ORG	1879	154
LOC	511	90
PER	406	27
DATE	238	2
PRO	1072	0
EVENT	18	0
TOTAL	4123	273

FIN-CLARIN

Methods

CLARIN

- **FiNER – rule-based system** developed by FIN-CLARIN Kielipankki (Ruokolainen et al., 2019) for predicting **both top-level and nested entities** with a capture mechanism that allows **online adaptation.** https://link.springer.com/article/10.1007/s10579-019-09471-7
- **GÜNGÖR-NN – state-of-the-art neural network** architecture of Güngör et al. (2018) predicting **top-level entities** while benefitting from word embeddings trained on the complete Finnish internet, incl. Wikipedia.
- **SOHRAB-NN – state-of-the-art neural network** architecture of Sohrab and Miwa (2018) **also predicts nested entities** while benefitting from word embeddings.

FIN-CLARIN

Test Results

CLARIN

- Digitoday (in-domain)

FINER			GÜNGÖR-NN			SOHRAB-NN		
pre	rec	F1	pre	rec	F1	pre	rec	F1
90.79	80.25	**85.20**	84.04	80.73	82.35	86.30	77.23	81.51

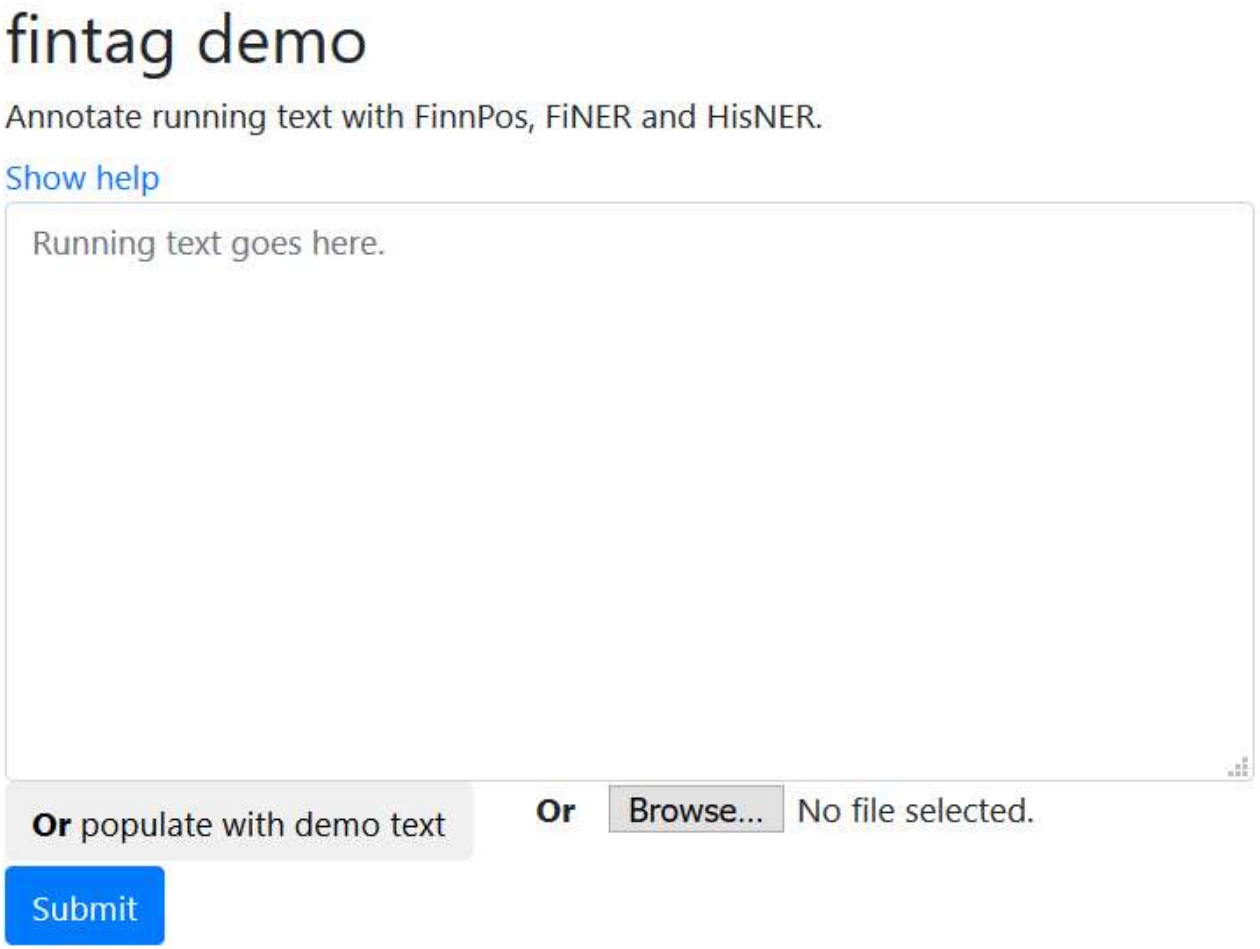
- Wikipedia (out of domain)

FINER			GÜNGÖR-NN			SOHRAB-NN		
pre	rec	F1	pre	rec	F1	pre	rec	F1
88.66	**72.74**	**79.91**	67.46	55.07	60.64	63.79	44.63	52.52

FIN-CLARIN

https://www.kielipankki.fi/tools/demo/

CLARIN



FIN-CLARIN

Availability

CLARIN

- Persistent Identifier of the tool:
  - http://urn.fi/urn:nbn:fi:lb-201908161
- Software access location:
  - http://urn.fi/urn:nbn:fi:lb-201908162
- Demo:
  - https://www.kielipankki.fi/tools/demo/
- Publication:
  - https://link.springer.com/article/10.1007/s10579-019-09471-7