Reverse Engineering a Rule-Based Finnish Named Entity Recognizer

Miikka Silfverberg

Department of Modern Languages
University of Helsinki

June 9, 2015

Introduction

How well can the behavior of a rule-based NE recognizer be reproduced in a supervised machine learning setting?

- I used the Finnish Named Entity Recognizer FiNER to label text from the Finnish wikipedia.
- I used a CRF-based morphological tagging toolkit FinnPos to train a statistical NE recognizer and evaluated it using held-out data.
- Additionally, I conducted an experiment on out-of-domain data (Europarl).

FiNER: A Finnish Named Entity Recognizer

```
Define PersonName
       (PropFirst [WSep CapWord]* WSep) PropLast
       EndTag(EnamexPrsHum) ;
```

- A rule-based Named Entity Recognizer based on recursive transition networks.
- Created using `hfst-pmatch`.
- Uses the morphological analyzer OMorFi as a gazetteer.

FiNER: A Finnish Named Entity Recognizer

O Näin	O ollen	O esimerkiksi	B-Loc Jordanin	I-Loc alue	O säästyi	O .
Thus e.g. the Jordan area was spared.						

- FiNER annotates 15 different entities. E.g. person names, titles, companies, date and time expressions.
- The entities belong to five broader categories: location, measure, time, person and organization.

FinnPos: Morphological Tagging for Finnish

Noun+Sg+Ess	Num+Card	Prop+Sg+Nom	V+Act+Past+Sg3
Vuosi	1275	Amsterdam	saada
Vuonna	**1275**	**Amsterdam**	**sai**
In the year 1275 Amsterdam got...			

- A statistical tagging toolkit intended for morphological tagging of morphologically rich languages such as Finnish.
- Fast to train with a customizable feature set.
- A discriminative model based on Conditional Random Fields.
- `github.com/mpsilfve/FinnPos/`

Experiments

- Eight different models: 100KW to 800KW training data from the Finnish Wikipedia.
- 100KW for development and 100KW for testing.
- Morphological labels provided by FiNER and semantic tags provided by OMorFi.
- Additionally, an out-of-domain experiment using 20KW from the Finnish Europarl corpus.
- Evaluation: Precision, recall and F1-score on **complete** NEs.

Feature Set + OMorFi

B-Time	I-Time	B-Loc	0
Noun+Sg+Ess	Num+Card	Prop+Sg+Nom	V+Act+Past+Sg3
-	-	GEO	-
Vuosi	1275	Amsterdam	saada
UC	-	UC	LC
Vuonna	1275	**Amsterdam**	sai
In the year 1275 Amsterdam got...			

- Word forms and morphological tags in a five word window.
- Semantic tags in a three word window.
- Lemma, suffixes and prefixes.
- Capitalization in three word window.
- Digits and dashes.

Results: Total

Training Data	Unlabeled F1-Score	Labeled F1-Score
100 KW	89.4%	85.0%
200 KW	90.6%	86.9%
300 KW	92.1%	89.1%
400 KW	92.2%	89.4%
500 KW	93.1%	90.9%
600 KW	93.1%	91.3%
700 KW	93.5%	92.0%
800 KW	**93.8%**	**92.4%**

Results: Element-wise

	Precision	Recall	F1-Score	NE Frequency
Location	95.6%	97.3%	96.5%	39.7%
Measure	97.2%	87.5%	92.1%	0.5%
Time	97.5%	95.1%	96.3%	9.1%
Person	94.4%	92.9%	93.6%	34.6%
Organization	91.9%	**77.7%**	84.2%	16.2%

Error Analysis

- I analyzed some 40 differences in labeling of organization entities.
- Differences are grouped into missing elements, extra elements and partial matches.

Missing Elements
Fischlamin kylän kansakoulu *The primary school of Fischlam village.*
Extra Elements
Panepistimiou-kadulla sijaitsevat yliopisto ... *The University and ... on Panepistimiou street.*
Partial Matches
Fugakukai International Association Fugakukai International Association

Error Analysis

	FiNER Correct	FinnPos correct	Count
Partial Matches	70%	30%	7
Extra	30%	70%	3
Missing	60%	30%	32

- FiNERs decision is correct in about 60% of all differences.

Out-of-Domain Experiment

- The 800 KW Wikipedia model was applied on approx. 20KW from the Finnish Europarl corpus.

Wikipedia

	R	P	F1
Unlabeled	92.4%	95.3%	93.8%
Labeled	91.9%	94.8%	93.3%

Europarl

	R	P	F1
Unlabeled	78.4%	94.5%	85.7%
Labeled	77.5%	93.4%	84.7%

Conclusions

- A high F1-score of 92.4% shows that it is easier to find FiNER NEs that real NEs.
- Long and very varied NEs such as organization names are tricky.
- Changing domain seems to have a severe effect on recall but doesn't really affect precision.
- Even more training data would probably improve results slightly.