Named Entity Recognition for Spoken Finnish

Dejan Porjazovski
dejan.porjazovski@aalto.fi
Aalto University

Juho Leinonen
juho.leinonen@aalto.fi
Aalto University

Mikko Kurimo
mikko.kurimo@aalto.fi
Aalto University

MeMAD
Methods for Managing Audiovisual Data

Presented by: Dejan Porjazovski

What is named entity recognition?

- Natural Language Processing (NLP) task
- The goal is to find entities in a text and classify them to predefined categories
- Some of the categories include: person, location, organization, product, date

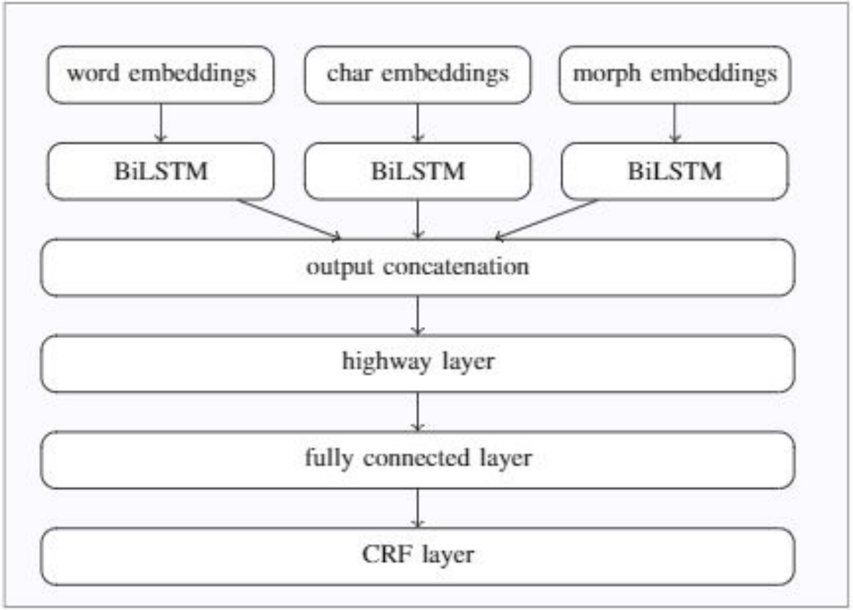
Challenges in NER

- Named entity ambiguity
- Data sparsity
- Unstructured data

Data

- Digitoday dataset consisting of online technological articles
- Parliament dataset, containing Finnish parliament sessions, annotated using a rule-based system
- Yle Pressiklubi dataset, containing popular TV shows, annotated using a rule-based system
- Estonian dataset, used for transfering tags

Methods



word embeddings
char embeddings
morph embeddings
BiLSTM
BiLSTM
BiLSTM
output concatenation
highway layer
fully connected layer
CRF layer

Knowledge transfer

- Transferred tags from Estonian to Finnish
- Multilingual embeddings aligned in a single vector space, provided by MUSE
- Nearest neighbor search from Estonian to Finnish language
- Use thresholding to keep only the good translations
- Translate the word and take the tag from Estonian
- Person and locations names are almost the same in Finnish and Estonian - directly copy them

Results

- Micro average F1 score for the Digitoday dataset

architecture	precision	recall	F1
FiNER	90.41	83.51	86.82
GÜNGÖR-NN	83.59	85.62	84.59
word+char+morph-LSTM	85.52	83.74	84.62
word+char+morph-LSTM+transfer	85.27	84.19	84.73

Aalto University

Results

- Micro average F1 score for the Wikipedia test set

architecture	precision	recall	F1
FiNER	85.17	72.47	78.31
GÜNGÖR-NN	62.98	55.89	59.22
word+char+morph-LSTM	71.34	56.38	62.98
word+char+morph-LSTM+transfer	74.55	61.93	67.66

Aalto University

Results

- Micro average F1 score for the ASR datasets

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	46.11	89.25	60.81	80.00	85.71	82.76
LOC	14.53	69.39	24.03	76.92	86.96	81.63
ORG	/	/	/	55.56	26.79	36.14
avg	28.26	82.39	42.09	76.25	72.91	74.54

Results

- Micro average F1 score for the Parliament dataset, where the model is trained without removing capitalization and punctuation

TAG	precision	recall	F1
PER	72.73	8.60	15.38
LOC	19.57	18.37	18.95
avg	29.82	11.97	17.09

Aalto University

Results

- Micro average F1 score for the ASR datasets, comparing only entities found by the rule-based system

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	98.81	89.25	93.79	85.04	85.71	85.38
LOC	100.00	69.39	81.93	89.55	86.96	88.24
ORG	/	/	/	78.95	26.79	40.00
avg	99.15	82.39	90.00	85.92	72.91	78.88

Results

- Micro average F1 score for the manually annotated subsample of the ASR datasets

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	91.43	84.21	87.67	91.11	85.42	88.17
LOC	77.27	80.95	79.07	84.62	84.62	84.62
ORG	/	/	/	100.00	32.14	48.65
avg	85.96	83.05	84.48	90.00	70.59	79.12

Aalto University

Conclusion

- Subwords usually help with modeling agglutinative languages like Finnish
- Knowledge transfer technique improved the results on the out-of-domain dataset
- Training the system with lowercase data improved the results on the dataset that did not have capitalization and punctuation