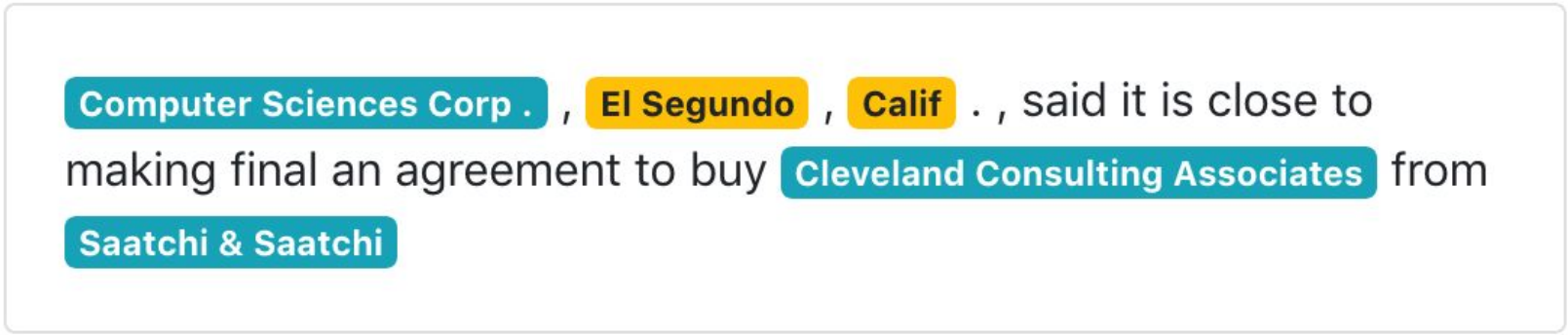
Named Entity Recognition in Noisy Domains

Valentin Malykh, Vladislav Lyalin

Neural Systems and Deep Learning Laboratory, MIPT

Motivation

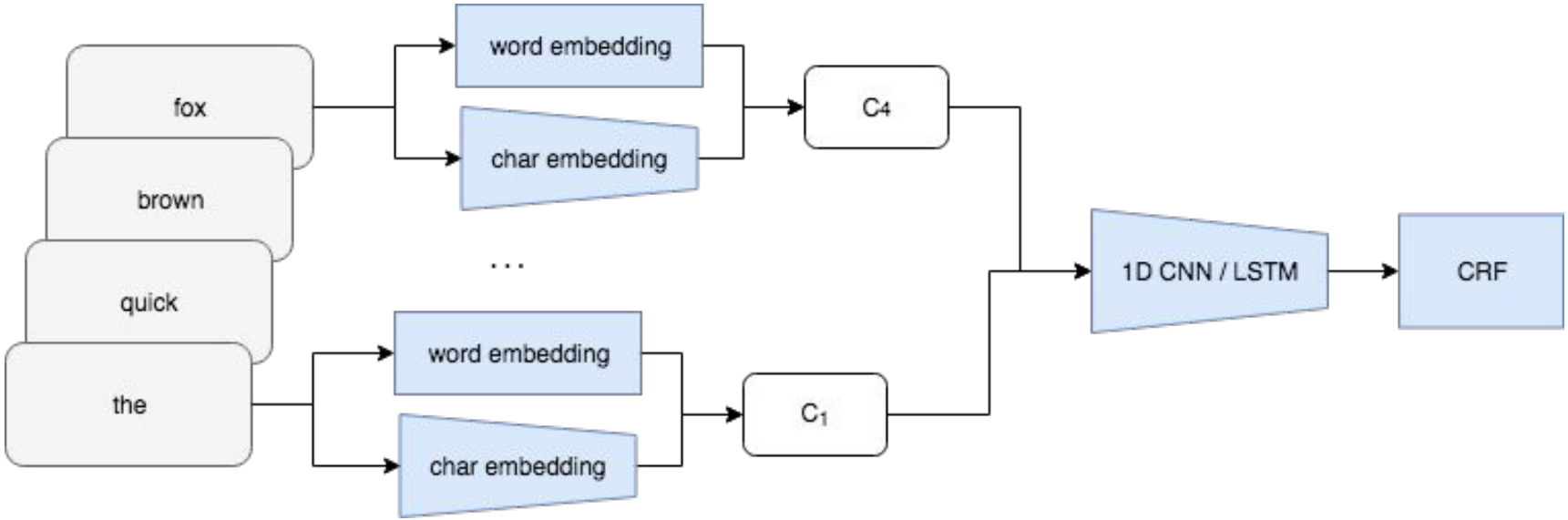
Named Entity Recognition (NER) task is an important part for conversational AI. A typical user of a conversation system has no time to check the spelling or grammar in his or her utterances. Due to that user utterances contain typos and spelling errors, so the noise robustness should be considered as a significant aspect of NER task.



Agenda

- Base model and model variations
- Experiment
  - Datasets
  - Experiment setup
  - Noise model
- Results
- Conclusion

Base model


fox
brown
quick
the
word embedding
char embedding
word embedding
char embedding
C4
C1
1D CNN / LSTM
CRF

Model variations

Word-level representations

- Embedding matrix learned jointly with model
- Fixed pre-trained matrix
- Fixed random matrix
- fastText embeddings

Char-level representations

- CNN-embeddings
- No char-level representation

Sequence processing unit

- BiLSTM
- Time dimensional CNN

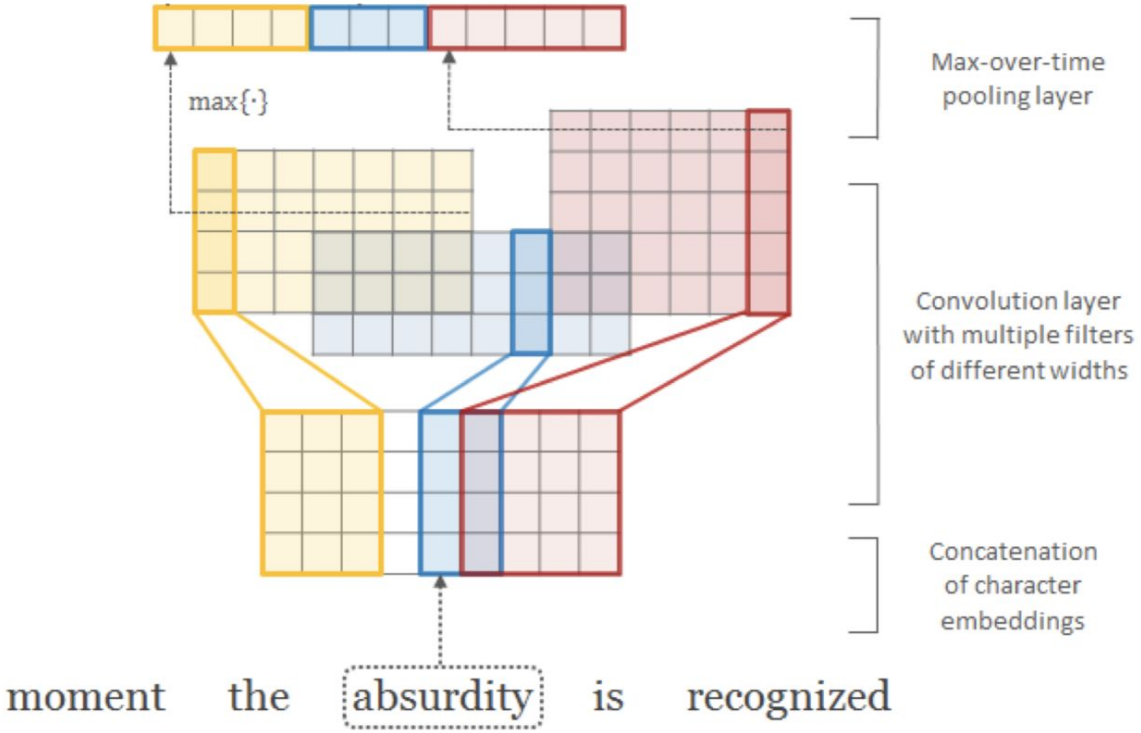
fastText

fastText is a common embedding method, which almost solves out-of-vocabulary problem. Each word is represented as sum of overlapping word n-gram vectors.

For example, n-grams of word <where>: <wh, whe, her, ere, re>, $n \in \{2, 3\}$

CharCNN

Character-level embeddings can be computed using CNN. The word is represented as sequence of character vectors, which are fed to convolutional layer(s). Next, global max-over-time pooling operation is performed.



Model variations

- **EmbedMatrix+CNN** - trainable embedding matrix and character level CNN embeddings
- **EmbedMatrix-nochar** - trainable embedding matrix, no char-level embeddings
- **FastText+CNN** - fastText vectors and character level CNN embeddings
- **FastText-nochar** - fastText vectors, no char-level embeddings
- **RandomEmbed+CNN** - non-trainable random embedding matrix and character level CNN embeddings
- **RandomEmbed-nochar** - non-trainable random embedding matrix, no char-level embeddings

Each model can use CNN or BiLSTM as sequence processing unit.
Total: 12 models

Datasets

CoNLL'03 Dataset
ORG, **LOC**, **PER** and **MISC** tags. English. News domain. 1393 documents total.

Persons-1000 Dataset
PER, **ORG**, **MEDIA**, **LOC** and **GEOPOLIT** tags. Russian. News domain. 1000 documents total. Tagging scheme and split procedure are intentionally set to be close to CoNLL.

CAp'2017 Dataset
13 types of entities. French. Social media domain. 6685 tweets total.

All datasets' annotation follow BIO-tag scheme

Preprocessing

In order to demonstrate robustness against noise we have performed automatic spell-checking of described datasets and artificially introduce noise to our datasets.

Noise model

We model the probability of inserting a letter after the current one for every letter of the input alphabet for each of the task.

Noise is modeled with probability $p \in [0.0, 0.2]$

Examples:
Noise 0.05: *spanish is the seiond most spoken languave in the uniked statys of america. forty-five million hispan\'phsnes speak spanish*

Noise 0.10: *spanish iq the secono most spoken languago in the united states of anerica. f$rty-five villion hispanopho<es speak spanish*

Experiment Setup

Three types of experiments:

1. train- and testsets are spell-checked and artificial noise in inserted
2. train- and testsets are not changed and no artificial noise is added
3. trainset is spell-checked and noised, testset is unchanged

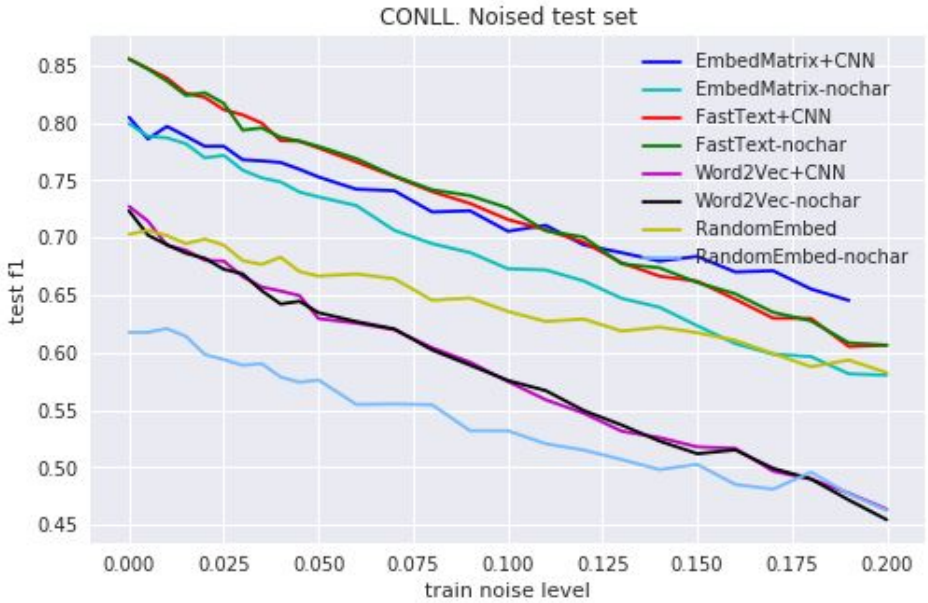
Metric:

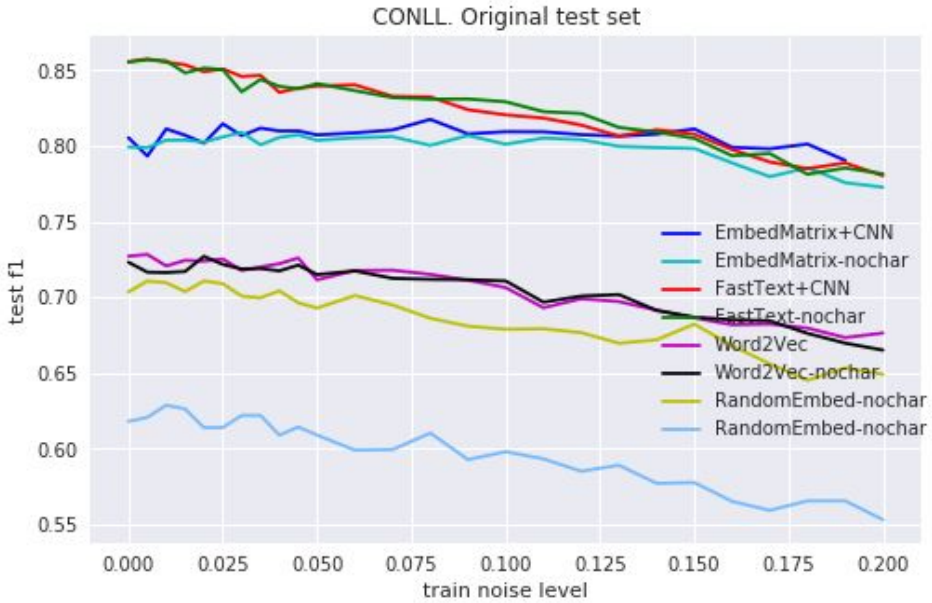
Chunkwise F1 score, macro averaging. True positive only if all words from the entity are recognized as true class.

Results. Original corpora

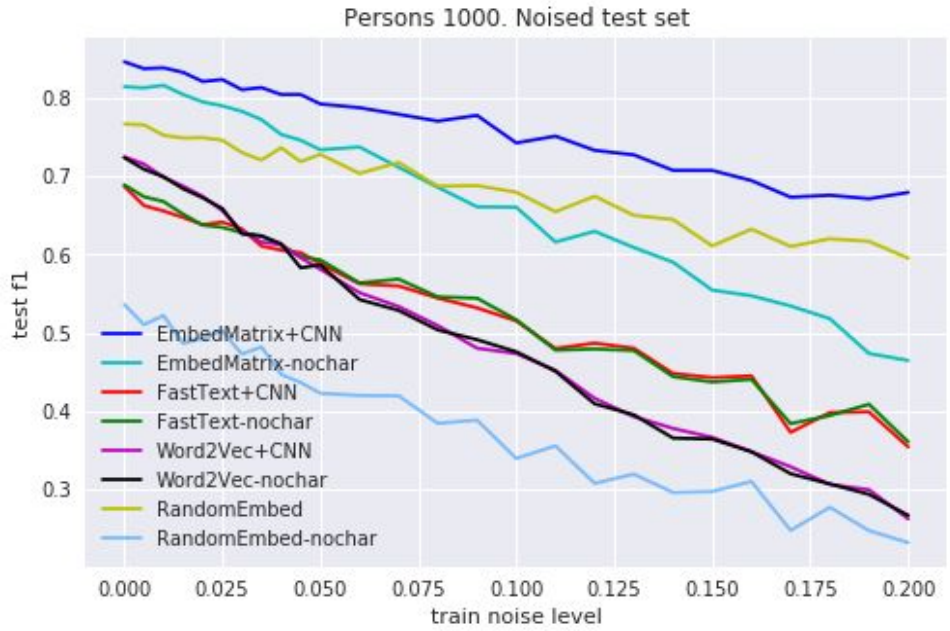
Model	CoNLL'03	Persons-1000	CAp'2017
EmbedMatrix+CNN	0.81	0.85	0.43
EmdebMatrix-nochar	0.80	0.81	0.44
RandomEmbed+CNN	0.69	0.77	0.31
RandomEmbed-nochar	0.61	0.48	0.22
FastText+CNN	0.86	0.69	0.41
FastText-nochar	0.86	0.69	0.41
Word2Vec+CNN	0.73	0.72	N/A
Word2Vec-nochar	0.72	0.72	N/A

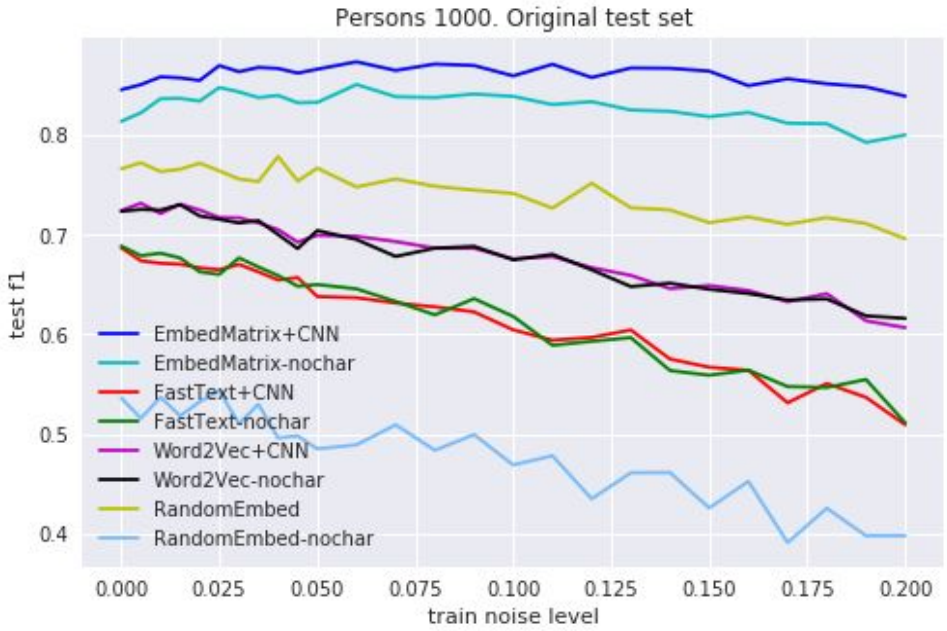
Results. CoNLL'03 (EN)



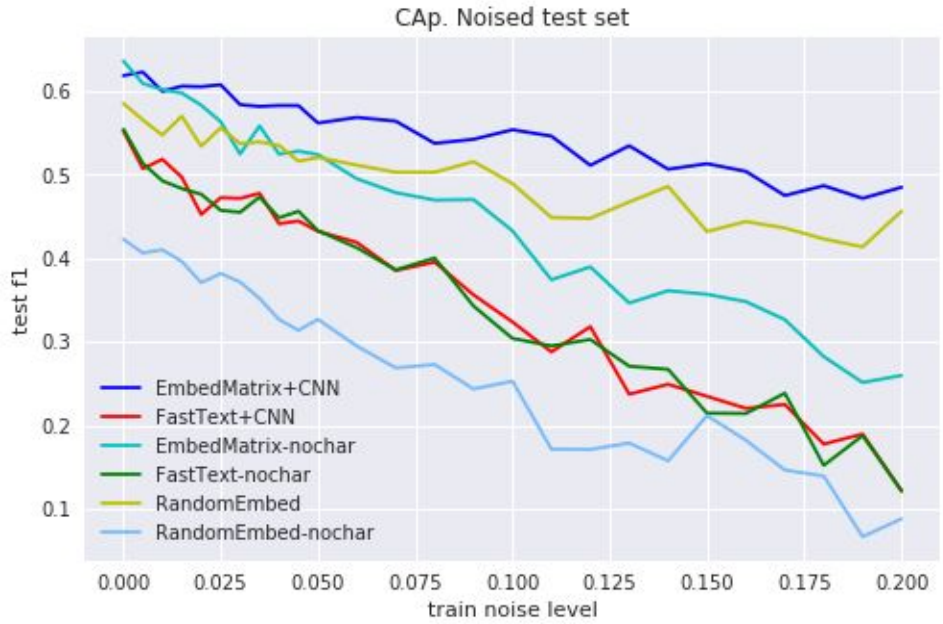


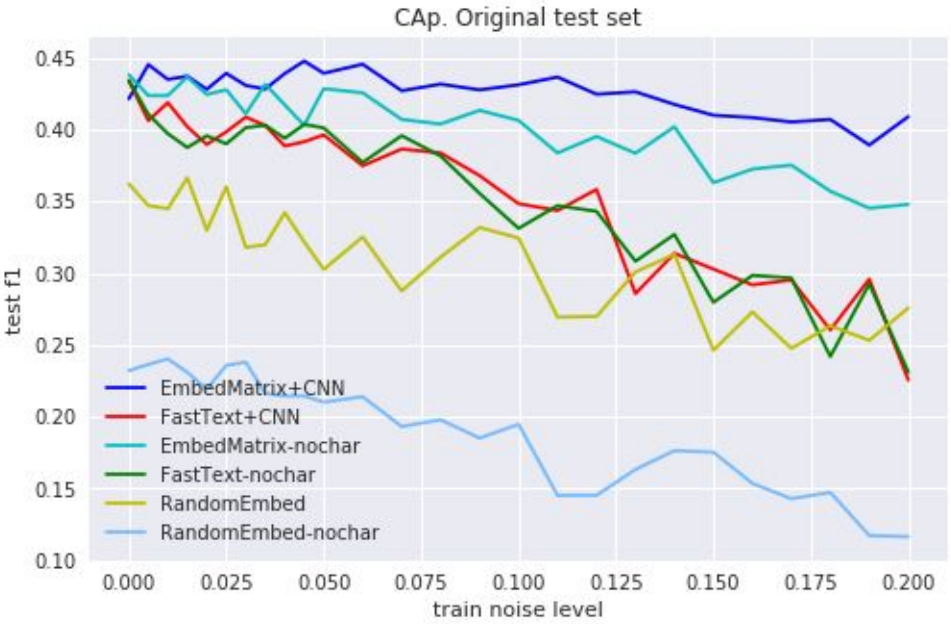
Results. Persons-1000 (RU)





Results. CAp'2017 (FR)





Results. CAp'17 state-of-the-art

Model	orig.	sp.-ch.
EmbedMatrix+CNN, CNN	0.42	0.63
EmbedMatrix-nochar, CNN	0.44	0.64
EmbedMatrix+CNN, LSTM	0.39	0.59
EmbedMatrix-nochar, LSTM	0.38	0.59
FastText+CNN, LSTM	0.52	0.67
FastText-nochar, LSTM	0.53	0.69

Previous SOTA according to [Lopez et al.] is 0.58 F1
19% relative improvement

Conclusion

We have demonstrated the robustness of several related named entity recognition architectures on three dataset of different languages. Moreover, a proposed artificial noise is demonstrated to be adequate surrogate of natural noise in the data.

Unexpectedly we have reached the new state of the art on French language dataset (without noise induction), thus it is interesting finding that spell-checking is so crucial for the Named Entity Recognition quality on this dataset.

Thank you for your attention

Conclusion

We have demonstrated the robustness of several related named entity recognition architectures on three dataset of different languages. Moreover, a proposed artificial noise is demonstrated to be adequate surrogate of natural noise in the data.

Unexpectedly we have reached the new state of the art on French language dataset (without noise induction), thus it is interesting finding that spell-checking is so crucial for the Named Entity Recognition quality on this dataset.