Neural Network Based Named Entity Recognition

Jana Straková

Named Entity Recognition (NER)

Nine things about Serena Williams
Syria conflict: Opposition unveils transition plan
Working lives: The rise of Botswana’s rough diamonds

Named Entity Recognition (NER)

Nine things about Serena Williams
Syria conflict: Opposition unveils transition plan
Working lives: The rise of Botswana's rough diamonds

Nine things about **Serena Williams**
Syria conflict: Opposition unveils transition plan
Working lives: The rise of **Botswana**'s rough diamonds

Calvin and Hobbes won an award.

Vltava na všechny posluchače zapůsobila.

Viktorka je stále ve hře o titul.

Machine translation

John loves Mary.

May lost majority in June.

Grammar correction

Řepka odehrál 14 ligových zápasů.

Information retrieval

Paris - city / person

Washington - person / city / state / mountain

Thesis Content

Design, implementation and evaluation of a named entity recognition system with state-of-the-art results for Czech, focusing on neural network based techniques.

Data and Methodology

Training Data

Nine things about **Serena Williams**

Syria conflict: Opposition unveils transition plan

Working lives: The rise of **Botswana**'s rough diamonds

Classification Features

Form,

Lemma,

POS,

Is the first character capitalized?

etc.

Training Data

Nine things about **Serena Williams**
Syria conflict: Opposition unveils transition plan
Working lives: The rise of **Botswana**'s rough diamonds

Classification Features

Form,
Lemma,
POS,
Is the first character capitalized?
etc.

Training Data

Nine things about **Serena Williams**

Syria conflict: Opposition unveils transition plan

Working lives: The rise of **Botswana**'s rough diamonds

Classification Features

Form,

Lemma,

POS,

Is the first character capitalized?

etc.

Prediction on Test Data

John loves **Mary**

Data

Czech Named Entity Corpus (CNEC 2.0):

- Manually annotated sentences (mostly newspapers),
- 8993 sentences, 35220 named entities,
- 46 classes, two-level hierarchy.

English CoNLL-2003 shared task:

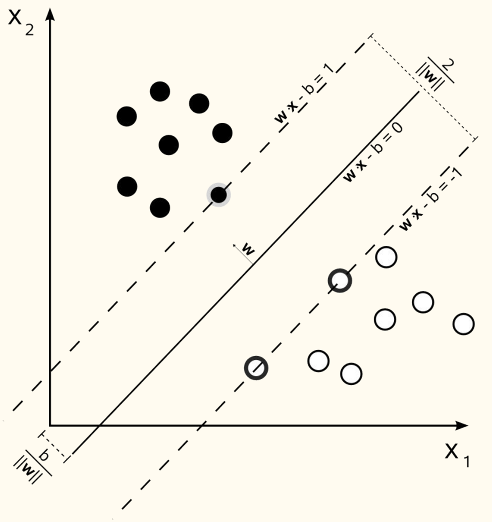
- Manually annotated sentences (mostly newspapers),
- 22137 sentences, 35062 named entities,
- 4 classes.

Thesis Content

Thesis Content

Thesis Content

Before Thesis

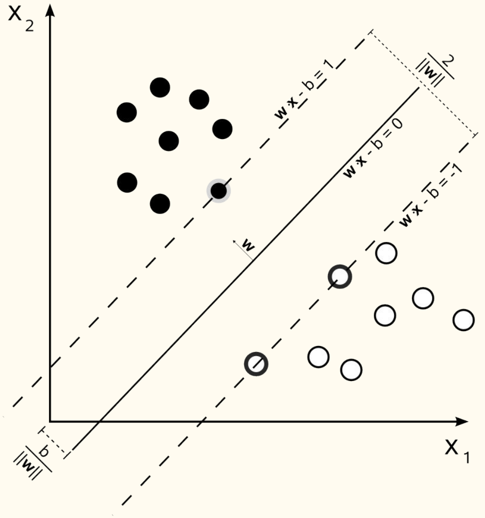


Kravalová and Žabokrtský (2009)

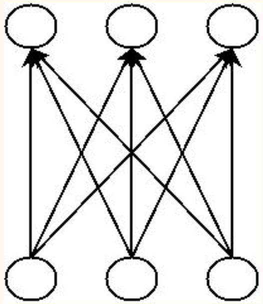
Thesis Content

Before Thesis

Research



Kravalová and Žabokrtský (2009)



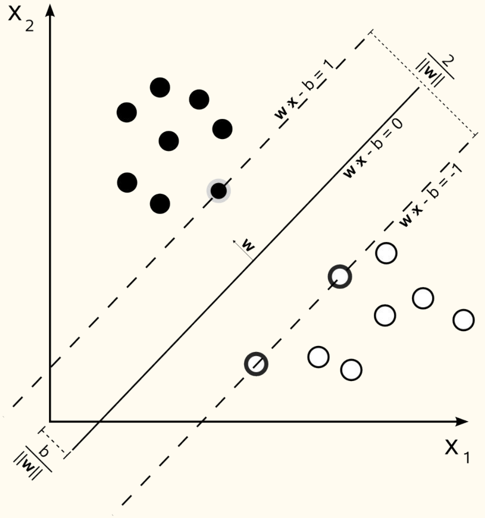
Straková et al. (2013)

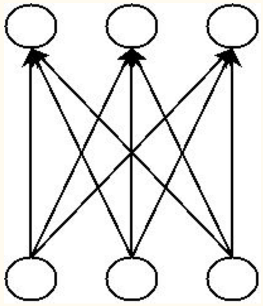
Thesis Content

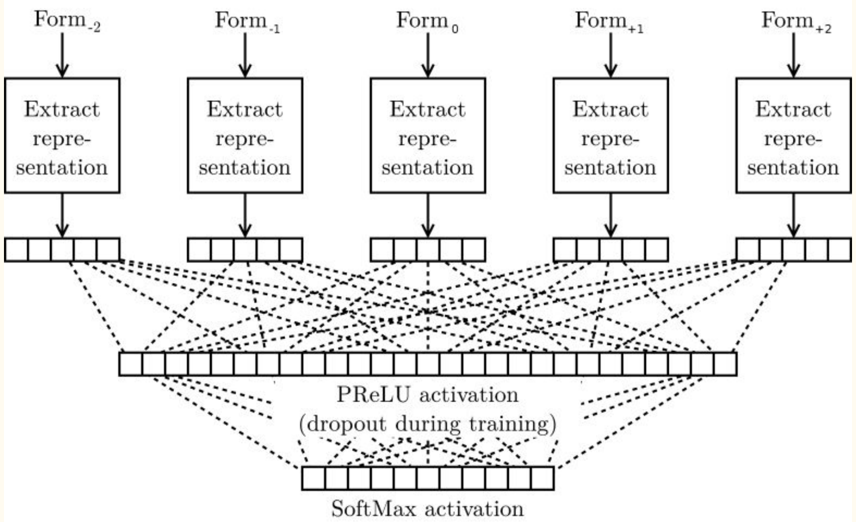
Before Thesis

Research

Thesis Achievement

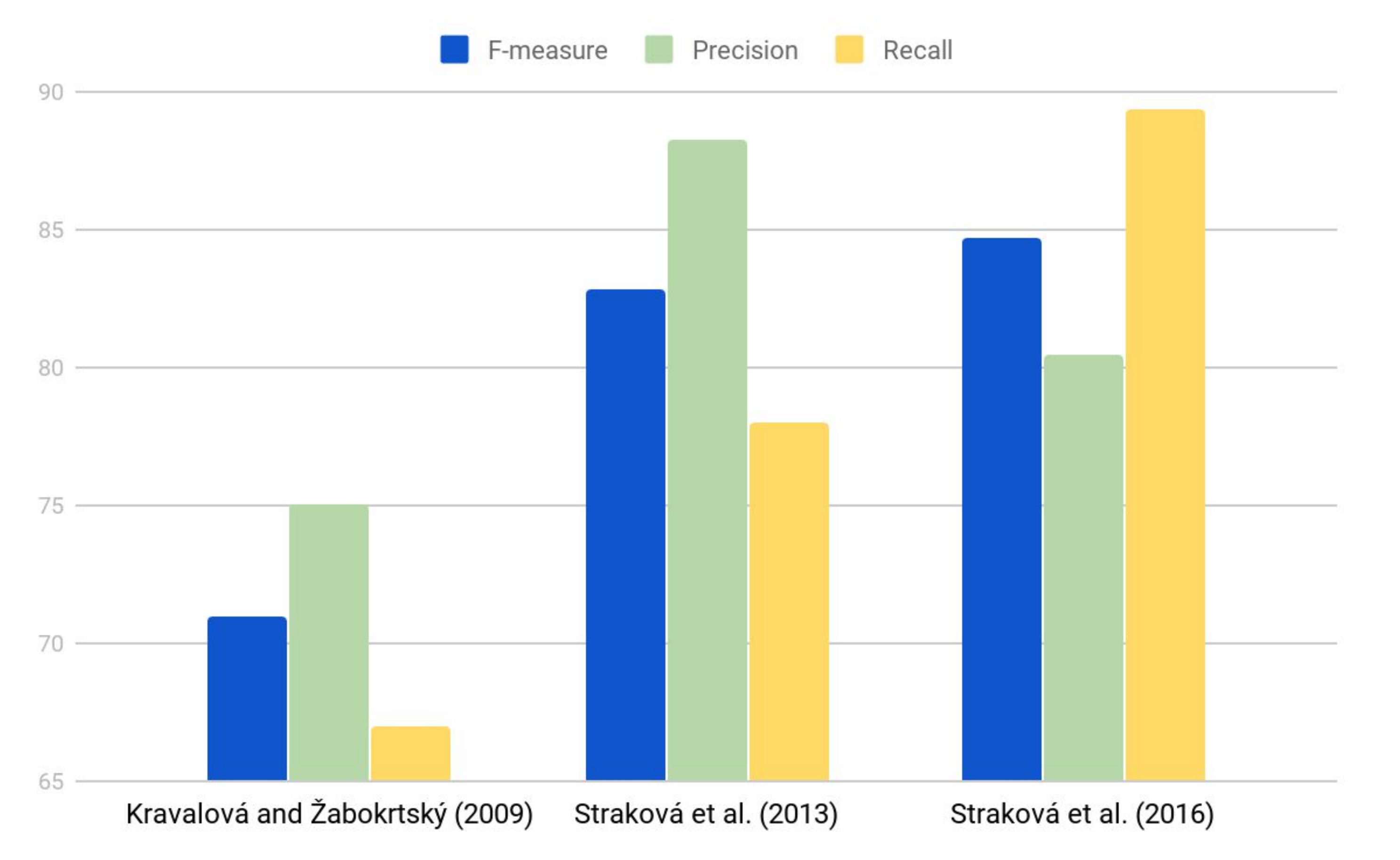






Kravalová and Žabokrtský (2009) Straková et al. (2013) Straková et al. (2016)

Incremental Results on CNEC 1.0



Featureless Neural Network for NER

$\text{Form}_{i=\{-2,-1,0,1,2\}}$	WikiPersonCategoryGazetteer
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$	WikiLocationCategoryGazetteer
$\text{POS}_{i=\{-2,-1,0,1,2\}}$	WikiOrganizationCategoryGazetteer
$\text{Window}_{i=\{-2,-1,0,1,2\}}$	WikiNamedObjectCategoryGazeteer
$\text{FirstRunLabel}_{i=\{-2,-1,0,1,2\}}$	WikiArtWorkCategoryGazeteer
PredictionHistoryWindow $(_{i=\{-500..0\}})$	WikiFilmCategoryGazeteer
$\text{FirstCapForm}_{i=\{-2,-1,0,1,2\}}$	WikiSongsCategoryGazeteer
$\text{AllCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsGazetteerGazeteer
$\text{MixedCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsMySQLGazeteer
$\text{EndsWithPeriodForm}_{i=\{-2,-1,0,1,2\}}$	FirstPersonPronounGazetteer
$\text{InternalPeriodForm}_{i=\{-2,-1,0,1,2\}}$	PersonPronounGazetteer
$\text{InternalApostropheForm}_{i=\{-2,-1,0,1,2\}}$	DayGazetteer
$\text{InternalHyphenForm}_{i=\{-2,-1,0,1,2\}}$	MonthGazetteer
$\text{InternalAmpForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003LOCGazetteer
$\text{InternalPunctuationForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003ORGGazetteer
$\text{PossessiveMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003MISCGazetteer
$\text{NegativeMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003PERGazetteer
$\text{LowercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Suffix}_{i=\{1,2,3,4\}}$
$\text{UppercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Prefix}_{i=\{1,2,3,4\}}$
$\text{TokenLengthForm}_{i=\{-2,-1,0,1,2\}}$	BrownCluster
$\text{SimplifiedPOS}_{i=\{-2,-1,0,1,2\}}$	$\text{BrownClusterPrefix}_{i=\{4,6,10,20\}}$
ManuallyCollectedCountryGazetteer	
ManuallyCollectedCityGazetteer	
ManuallyCollectedFirstNameGazetteer	
ManuallyCollectedLastNameGazetteer	

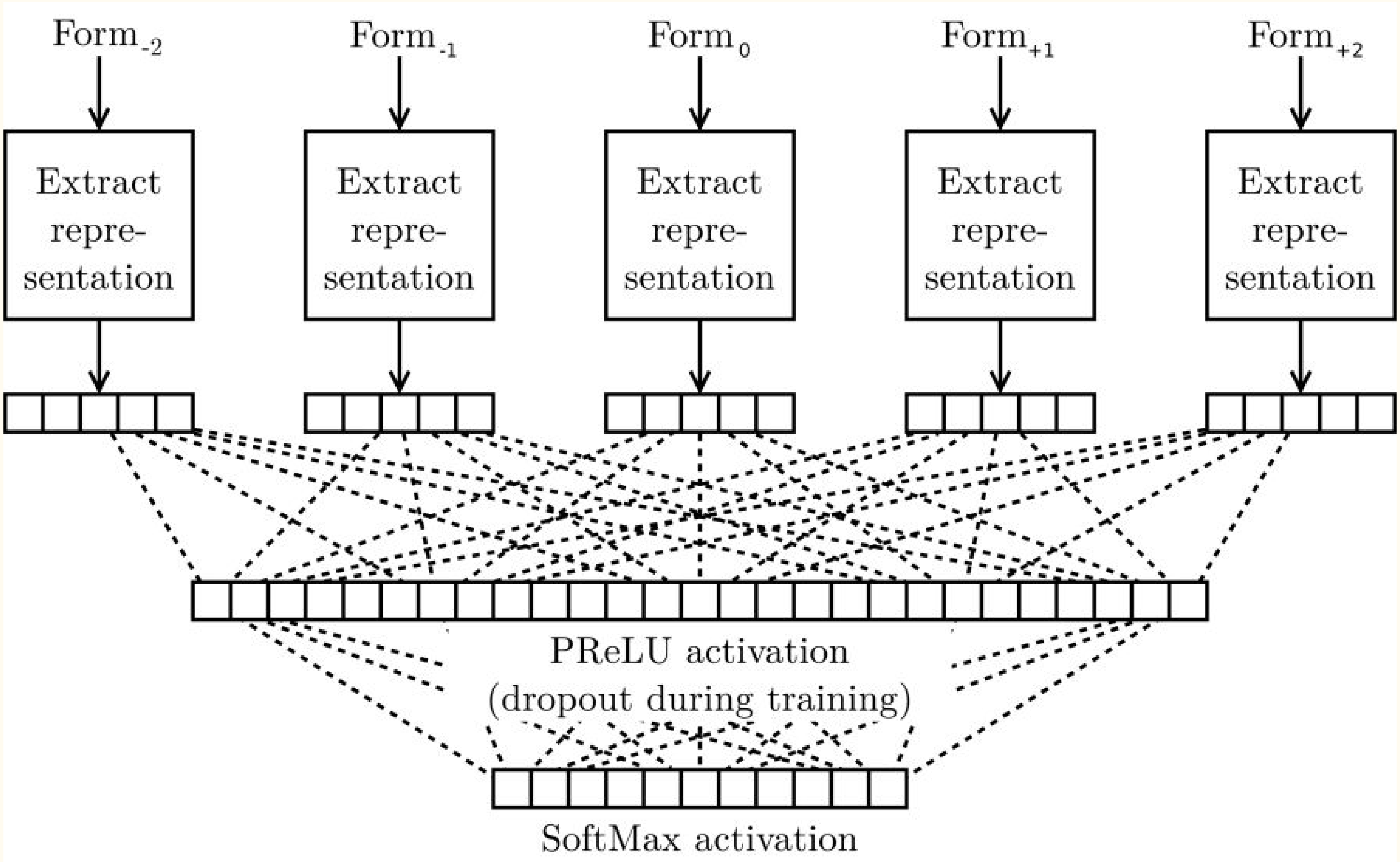
$\text{Form}_{i=\{-2,-1,0,1,2\}}$	WikiPersonCategoryGazetteer
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$	WikiLocationCategoryGazetteer
$\text{POS}_{i=\{-2,-1,0,1,2\}}$	WikiOrganizationCategoryGazetteer
$\text{Window}_{i=\{-2,-1,0,1,2\}}$	WikiNamedObjectCategoryGazeteer
$\text{FirstRunLabel}_{i=\{-2,-1,0,1,2\}}$	WikiArtWorkCategoryGazeteer
PredictionHistoryWindow $(_{i=\{-500..0\}})$	WikiFilmCategoryGazeteer
$\text{FirstCapForm}_{i=\{-2,-1,0,1,2\}}$	WikiSongsCategoryGazeteer
$\text{AllCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsGazetteerGazeteer
$\text{MixedCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsMySQLGazeteer
$\text{EndsWithPeriodForm}_{i=\{-2,-1,0,1,2\}}$	FirstPersonPronounGazetteer
$\text{InternalPeriodForm}_{i=\{-2,-1,0,1,2\}}$	PersonPronounGazetteer
$\text{InternalApostropheForm}_{i=\{-2,-1,0,1,2\}}$	DayGazetteer
$\text{InternalHyphenForm}_{i=\{-2,-1,0,1,2\}}$	MonthGazetteer
$\text{InternalAmpForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003LOCGazetteer
$\text{InternalPunctuationForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003ORGGazetteer
$\text{PossessiveMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003MISCGazetteer
$\text{NegativeMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003PERGazetteer
$\text{LowercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Suffix}_{i=\{1,2,3,4\}}$
$\text{UppercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Prefix}_{i=\{1,2,3,4\}}$
$\text{TokenLengthForm}_{i=\{-2,-1,0,1,2\}}$	BrownCluster
$\text{SimplifiedPOS}_{i=\{-2,-1,0,1,2\}}$	$\text{BrownClusterPrefix}_{i=\{4,6,10,20\}}$
ManuallyCollectedCountryGazetteer	
ManuallyCollectedCityGazetteer	
ManuallyCollectedFirstNameGazetteer	
ManuallyCollectedLastNameGazetteer	

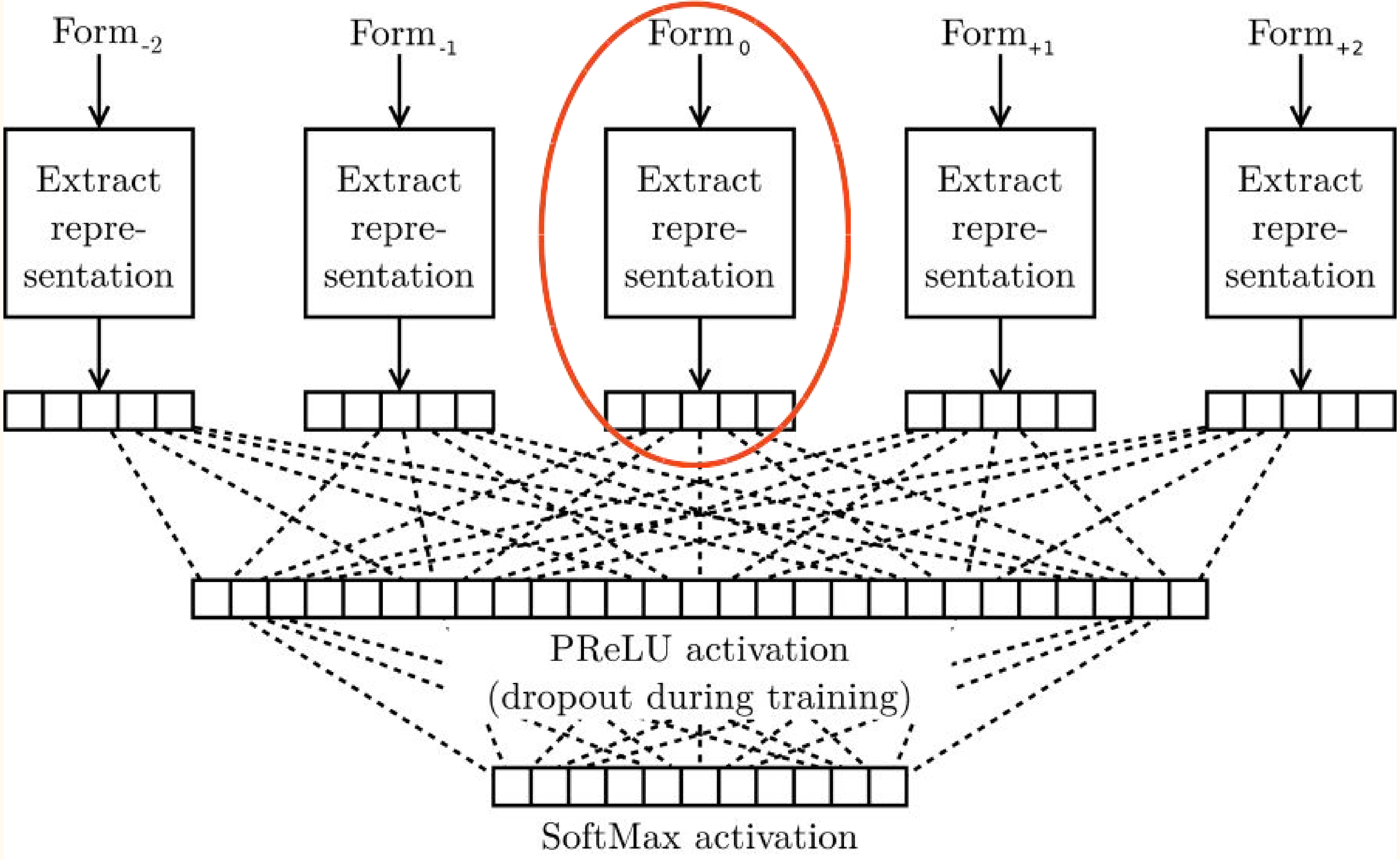
$\text{Form}_{i=\{-2,-1,0,1,2\}}$
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$
$\text{POS}_{i=\{-2,-1,0,1,2\}}$

$\text{Form}_{i=\{-2,-1,0,1,2\}}$	WikiPersonCategoryGazetteer
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$	WikiLocationCategoryGazetteer
$\text{POS}_{i=\{-2,-1,0,1,2\}}$	WikiOrganizationCategoryGazetteer
$\text{Window}_{i=\{-2,-1,0,1,2\}}$	WikiNamedObjectCategoryGazeteer
$\text{FirstRunLabel}_{i=\{-2,-1,0,1,2\}}$	WikiArtWorkCategoryGazeteer
PredictionHistoryWindow ($_{i=\{-500..0\}}$)	WikiFilmCategoryGazeteer
$\text{FirstCapForm}_{i=\{-2,-1,0,1,2\}}$	WikiSongsCategoryGazeteer
$\text{AllCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsGazetteerGazeteer
$\text{MixedCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsMySQLGazeteer
$\text{EndsWithPeriodForm}_{i=\{-2,-1,0,1,2\}}$	FirstPersonPronounGazetteer
$\text{InternalPeriodForm}_{i=\{-2,-1,0,1,2\}}$	PersonPronounGazetteer
$\text{InternalApostropheForm}_{i=\{-2,-1,0,1,2\}}$	DayGazetteer
$\text{InternalHyphenForm}_{i=\{-2,-1,0,1,2\}}$	MonthGazetteer
$\text{InternalAmpForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003LOCGazetteer
$\text{InternalPunctuationForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003ORGGazetteer
$\text{PossessiveMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003MISCGazetteer
$\text{NegativeMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003PERGazetteer
$\text{LowercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Suffix}_{i=\{1,2,3,4\}}$
$\text{UppercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Prefix}_{i=\{1,2,3,4\}}$
$\text{TokenLengthForm}_{i=\{-2,-1,0,1,2\}}$	BrownCluster
$\text{SimplifiedPOS}_{i=\{-2,-1,0,1,2\}}$	$\text{BrownClusterPrefix}_{i=\{4,6,10,20\}}$
ManuallyCollectedCountryGazetteer	
ManuallyCollectedCityGazetteer	
ManuallyCollectedFirstNameGazetteer	
ManuallyCollectedLastNameGazetteer	

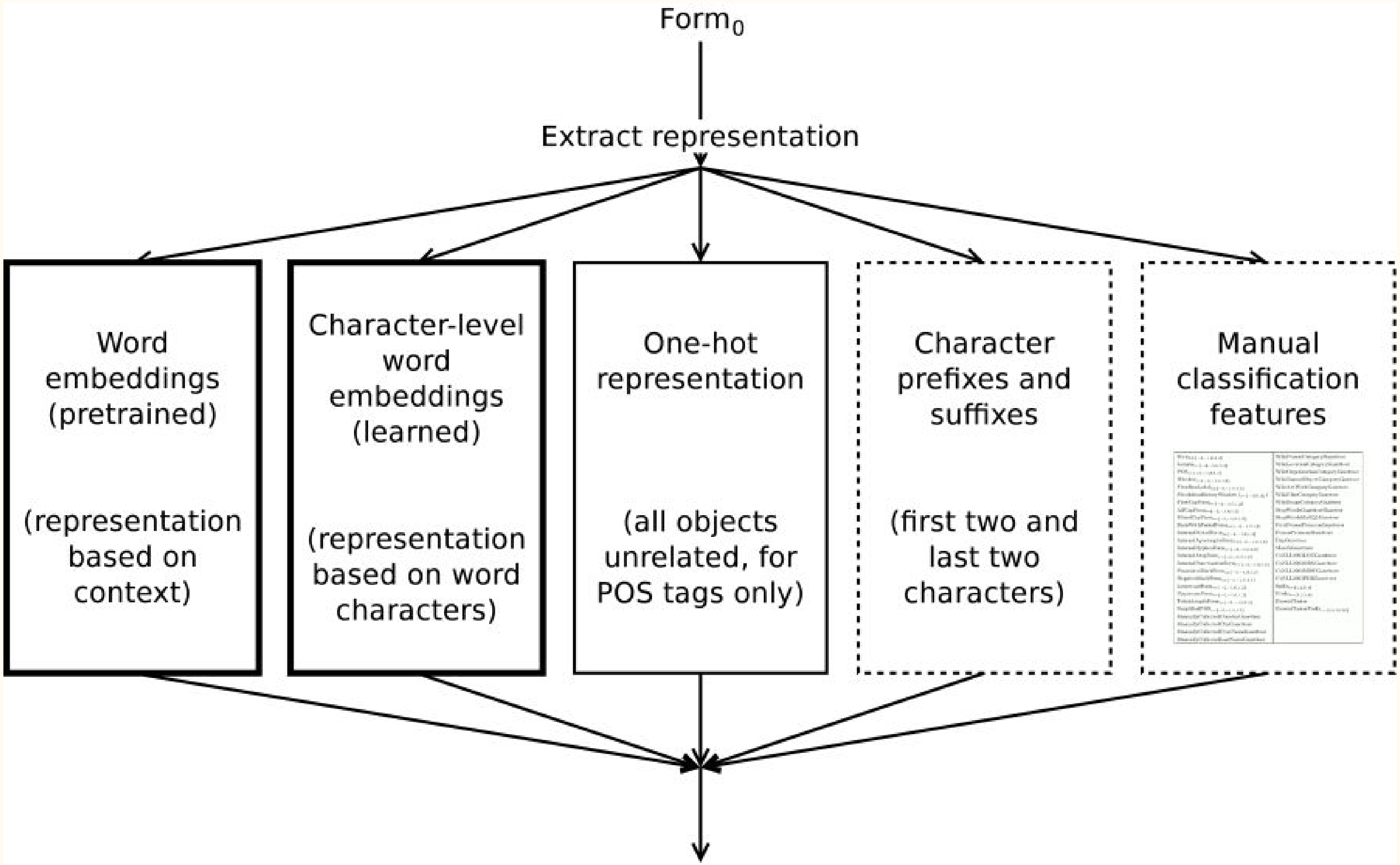
$\text{Form}_{i=\{-2,-1,0,1,2\}}$
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$
$\text{POS}_{i=\{-2,-1,0,1,2\}}$

$\text{Form}_{i=\{-2,-1,0,1,2\}}$

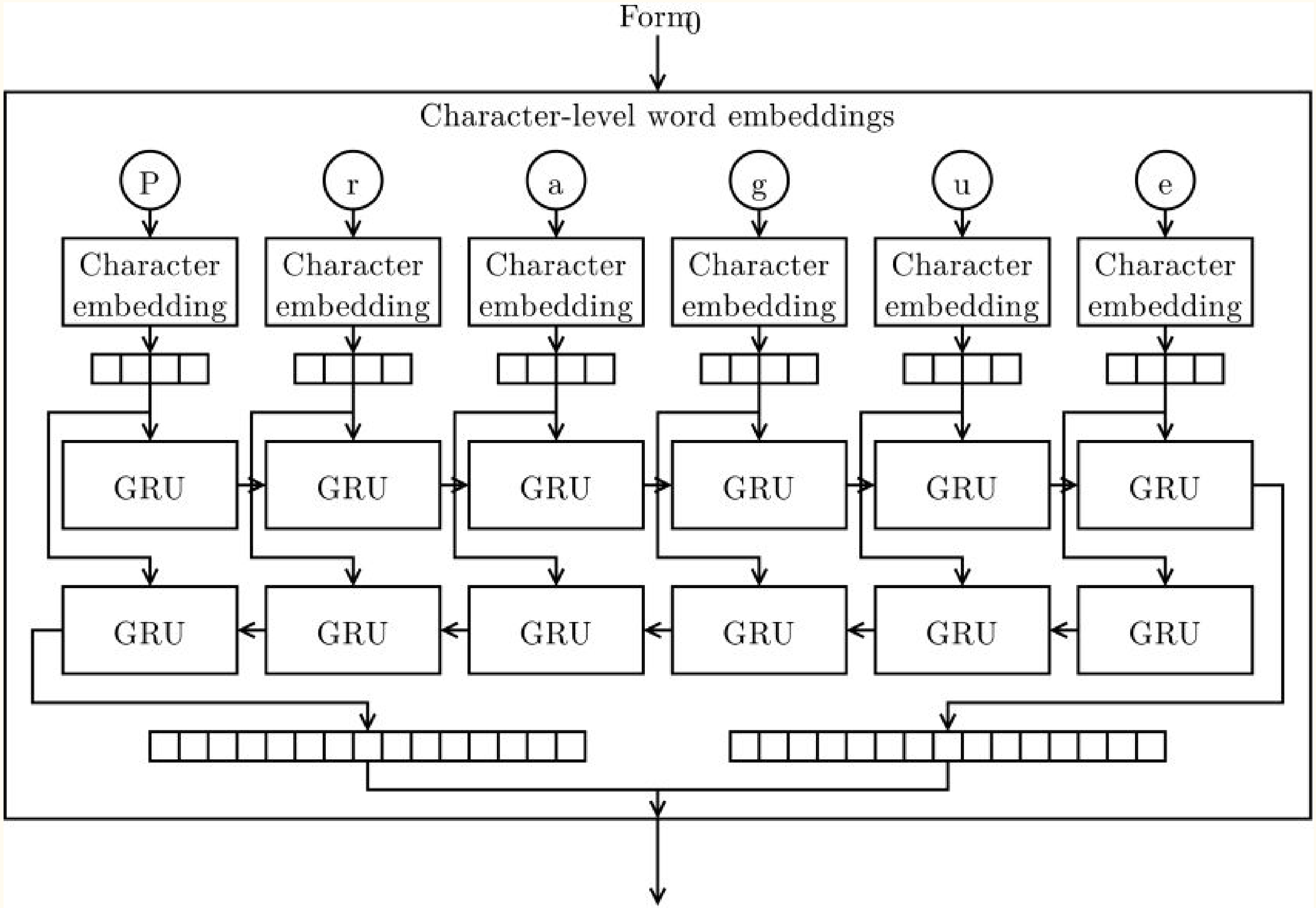

Form$_{-2}$
Form$_{-1}$
Form$_{0}$
Form$_{+1}$
Form$_{+2}$
Extract representation
Extract representation
Extract representation
Extract representation
Extract representation
PReLU activation
(dropout during training)
SoftMax activation



Form$_{-2}$
Form$_{-1}$
Form$_{0}$
Form$_{+1}$
Form$_{+2}$
Extract repre- sentation
Extract repre- sentation
Extract repre- sentation
Extract repre- sentation
Extract repre- sentation
PReLU activation
(dropout during training)
SoftMax activation



Form$_0$
Extract representation
Word embeddings (pretrained)
(representation based on context)
Character-level word embeddings (learned)
(representation based on word characters)
One-hot representation
(all objects unrelated, for POS tags only)
Character prefixes and suffixes
(first two and last two characters)
Manual classification features


Form0
Character-level word embeddings
P
r
a
g
u
e
Character embedding
Character embedding
Character embedding
Character embedding
Character embedding
Character embedding
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU

Results

Incremental Results: CNEC 1.0 F-measure (supertypes)

			CNEC 1.0 F1
(A)	forms	Word embeddings (WE)	69.61
(B)	forms	Character-level word embeddings (CLE)	76.13
(C)	forms	WE + prefixes, suffixes	74.49
(D)	forms	WE + CLE	78.11
(E)	forms	WE + CLE + prefixes, suffixes	78.32
(F)	forms	WE + CLE + prefixes, suffixes + CF	78.50
(G)	forms, lemmas, tags	WE	83.21
(H)	forms, lemmas, tags	CLE	80.88
(I)	**forms, lemmas, tags**	**WE + CLE**	**84.06**
(J)	**forms, lemmas, tags**	**WE + CLE + prefixes, suffixes + CF**	**84.68**

Results: Czech Corpora F-measure (supertypes)

	Original CNEC 1.0	Original CNEC 2.0	Extended CNEC 1.1	Extended CNEC 2.0
Kravalová et al. (2009)	71.00	-	-	-
Konkol et al. (2013)	79.00	-	74.08	-
Straková et al. (2013)	82.82	-	-	-
Konkol et al. (2014)	-	-	74.23	74.37
Demir et al. (2014)	-	-	75.61	-
Konkol et al. (2015)	-	-	74.08	-
Straková et al. (2016)	**84.68**	**82.78**	**80.88**	**80.79**

Results: English CoNLL 2003 F-measure

	English CoNLL 2003 F-measure
Ratinov and Roth (2009)	90.80
Liu et al. (2009)	90.90
Chiu et al. (2015)	90.77
Luo et al. (2015) - Joint NER + EL	91.20
Yang et al. (2016) - RNN + CRF	91.20
Lample et al. (2016) - RNN + CRF	90.94
Straková et al. (2016)	89.92

Thesis Contribution

Advance in state of the art in Czech NER

Published in peer-reviewed proceedings, book chapter and encyclopedic entry

NameTag, the named entity recognizer:

http://ufal.mff.cuni.cz/nametag

GitHub with source code for Straková et al. (2016):

https://github.com/strakova/ner_tsd2016

Questions and Answers

Neural Network Based Named Entity Recognition
Jana Straková

I miss some details about the annotation process. How many annotators participated in creation of the corpora? What was the inter-annotators agreement? How exactly were the sentences for annotation selected?

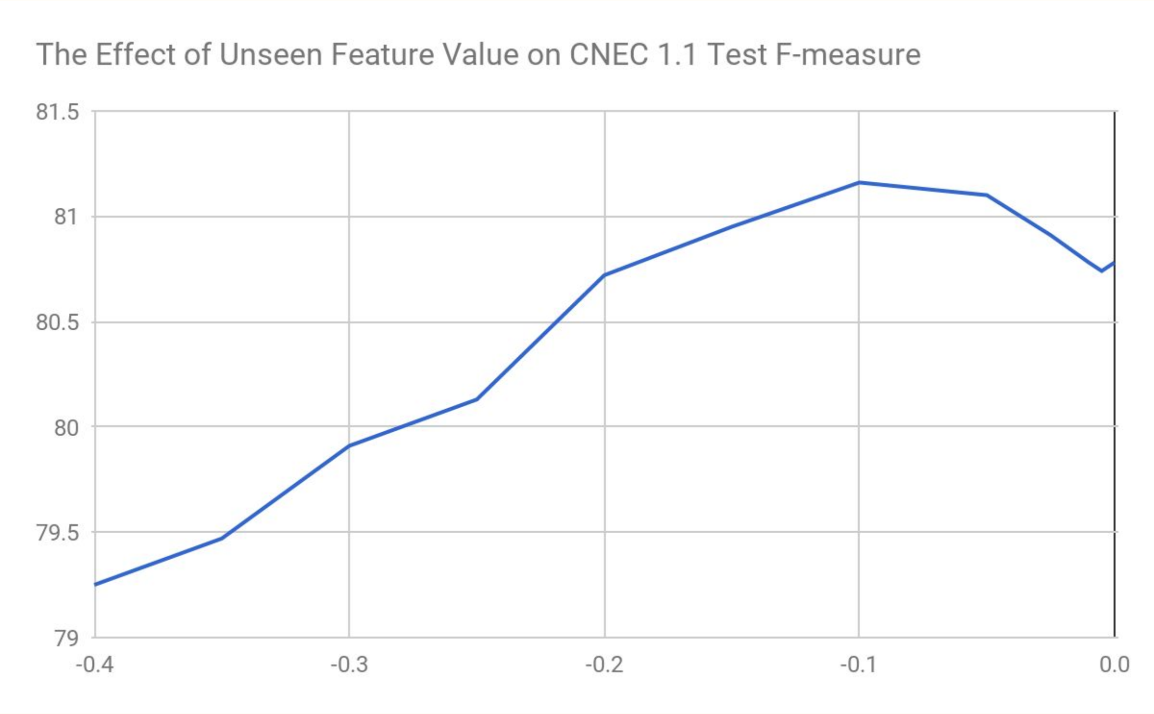
CNEC 1.0 was published by 3 authors: Magda Ševčíková, Zdeněk Žabokrtský and Oldřich Krůza (Ševčíková 2007a, Ševčíková 2007b).

It was annotated by 2 annotators in 1st round and 1 annotator in 2nd round (numbers).

I did not find inter-annotators agreement in the respective literature (listed above).

For annotation, sentences containing named entities were selected to lower the annotation costs and make the annotation process more effective. The sentences were selected so that they conform to a regular expression described in both citations and in the thesis on p. 21, section 3.1.1.

Figure 4.3 shows the effect of setting certain values to unseen weights on POS tagging and NER task performance. It seems that certain range of values can improve the classification accuracy. However, I see no experiment showing that an optimal weight established on development data can improve also the performance on test data.



Why not evaluate the system on other available NER corpora – such as Spanish, Dutch, and others?

The main goal was to develop an open-source tool for named entity recognition in Czech.

To compare the system with state of the art, I chose English.

It is true, however, that implementation and evaluation in other languages is appropriate and I suggest it as future work.

On page 72, on footnote 4, the author shows that the word2vec tool was trained with one epoch only, however, it is generally recommended to use 3-5 epochs.

This is room for improvement. When I started with word2vec, training with multiple epochs was not a general recommendation yet, but it has become now.

In section 7.2, the author states a hypothesis that character embeddings should be especially helpful for morphologically rich languages. However, I do not see any confirmation of such statement in Table 7.1. Character embeddings contribute both to Czech and English results.

Along with rapid development of knowledge about word and character-level word embeddings, my understanding of them also changes.

Nowadays, I see the word embedding in NER as “context hints for NE recognition” and character-level word embeddings as “word structure hints for NE recognition”. English words also have structural properties (morphology), therefore character-level word embeddings improve NER performance in English too.

The thesis conclusion mentions some potentially negative properties of Czech Named Entity Corpus. Would you recommend a different data selection and annotation of such corpus, with respect to manual annotation effectiveness?

Selecting sentences with higher NE density effectively decreased the annotation costs for CNEC because empty documents were not annotated.

I suggest extending this approach to a basic unit of one document instead of a sentence.

It a NE recognizer is available for the language, pre-annotation could also speed up the annotation process.

The thesis leads (especially in Chapter 7) to recently studied "end-to-end" systems based on artificial neural networks. What is your opinion on joining named entity recognition and named entity linking from this perspective - do you expect two sequential systems with different methods or do you eventually expect an end-to-end system, where the recognition step will be performed jointly in one network? What consequences would these solutions have for multilingual systems or even language independent systems?

This is a (yet) open problem, because there is no known NN-based system for entity linking (EL).

The problem is that NN would have to choose between millions of links, which is not computationally feasible.

Therefore the existing systems use "pipeline approach" with two different systems for NER and EL.

Obviously, the efforts are to solve NER and EL jointly.

Should the problem with NN for EL be solved, "end-to-end" system would be possible.

At several places in the thesis, you mention that going back to CNK would allow for reconstructing the missing long contexts. Has this been tried? If so, by whom and what were the results?

To my knowledge, this has not been tried.

In my opinion, the effort would not be worthwhile.

In development of NameTag, we abandoned the non-local features such as context aggregation and prediction history, as they increased the memory requirements and the model size.

Therefore these classification features did not receive so much attention in the thesis.

Were any feature dimensionality reduction techniques (PCA, LDA, HLDA) tried with the pre-word-embedding NN NER systems?

In Czech, this thesis reviewer, Ing. Konopík, published on this topic:

Michal Konkol, Tomáš Brychcín, and Miloslav Konopík. Latent semantics in named entity recognition. Expert Systems with Applications, 42(7):3470–3479, 2015.

Generally, dimensionality reduction techniques have been used for semantic modeling. NER uses these results as classification features.

Word embeddings are any mapping between the discrete space of vocabulary and a (smaller) continuous space of real-valued vectors. Dimensionality reduction techniques are means to create word embeddings.

Word embeddings as models of distributional, context semantics (Mikolov, Collobert) vs. word embeddings which keep as much information from the original space (H-PCA).

Context-based semantics word embeddings seem better suited for NLP tasks.

Explain, how the size of window of 500 predictions was determined and how it goes together with the „broken context“ property of CNEC.

That is a good question. The window of 500 predictions was only used for English. The respective description in the thesis should have clearly stated this.

NameTag seems to be a bit less precise than your experimental systems. Why?

NameTag is designed as a light-weight system with low resource usage and therefore some classification features are abandoned at the sake of accuracy.

It is what we believe a reasonable trade-off between software performance and recall/precision.

We keep two branches of our research: an experimental one, which is presented in papers: larger models, more classification features, larger computer memory, but usually 1-2 F-measure points gain; we then keep only those techniques which we find worthwhile.

Appendices

Neural Network Based Named Entity Recognition
Jana Straková

Classification Features

Typical set of classification features for NER

Current surface word form,

word lemma,

word stem,

POS tag,

the previous items applied to a predefined number of preceding and following words,

rule-based orthographic rules: word capitalization, first character capitalization, special characters, ...

regular expressions.

$\text{Form}_{i=\{-2,-1,0,1,2\}}$	WikiPersonCategoryGazetteer
$\text{Lemma}_{i=\{-2,-1,0,1,2\}}$	WikiLocationCategoryGazetteer
$\text{POS}_{i=\{-2,-1,0,1,2\}}$	WikiOrganizationCategoryGazetteer
$\text{Window}_{i=\{-2,-1,0,1,2\}}$	WikiNamedObjectCategoryGazeteer
$\text{FirstRunLabel}_{i=\{-2,-1,0,1,2\}}$	WikiArtWorkCategoryGazeteer
PredictionHistoryWindow $({}_{i=\{-500..0\}})$	WikiFilmCategoryGazeteer
$\text{FirstCapForm}_{i=\{-2,-1,0,1,2\}}$	WikiSongsCategoryGazeteer
$\text{AllCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsGazetteerGazeteer
$\text{MixedCapForm}_{i=\{-2,-1,0,1,2\}}$	StopWordsMySQLGazeteer
$\text{EndsWithPeriodForm}_{i=\{-2,-1,0,1,2\}}$	FirstPersonPronounGazetteer
$\text{InternalPeriodForm}_{i=\{-2,-1,0,1,2\}}$	PersonPronounGazetteer
$\text{InternalApostropheForm}_{i=\{-2,-1,0,1,2\}}$	DayGazetteer
$\text{InternalHyphenForm}_{i=\{-2,-1,0,1,2\}}$	MonthGazetteer
$\text{InternalAmpForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003LOCGazetteer
$\text{InternalPunctuationForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003ORGGazetteer
$\text{PossessiveMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003MISCGazetteer
$\text{NegativeMarkForm}_{i=\{-2,-1,0,1,2\}}$	CoNLL2003PERGazetteer
$\text{LowercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Suffix}_{i=\{1,2,3,4\}}$
$\text{UppercaseForm}_{i=\{-2,-1,0,1,2\}}$	$\text{Prefix}_{i=\{1,2,3,4\}}$
$\text{TokenLengthForm}_{i=\{-2,-1,0,1,2\}}$	BrownCluster
$\text{SimplifiedPOS}_{i=\{-2,-1,0,1,2\}}$	$\text{BrownClusterPrefix}_{i=\{4,6,10,20\}}$
ManuallyCollectedCountryGazetteer	
ManuallyCollectedCityGazetteer	
ManuallyCollectedFirstNameGazetteer	
ManuallyCollectedLastNameGazetteer	

Advanced features

Non-local features:

- Multiple appearances should be labeled with the same label (“one sense per discourse”).
- Context aggregation, extended prediction history (Ratinov and Roth, 2009).

Brown clusters

Gazetteers

Two-stage prediction

NER with Softmax NN Details

NER with Softmax Neural Networks

Straková et al. (2013)

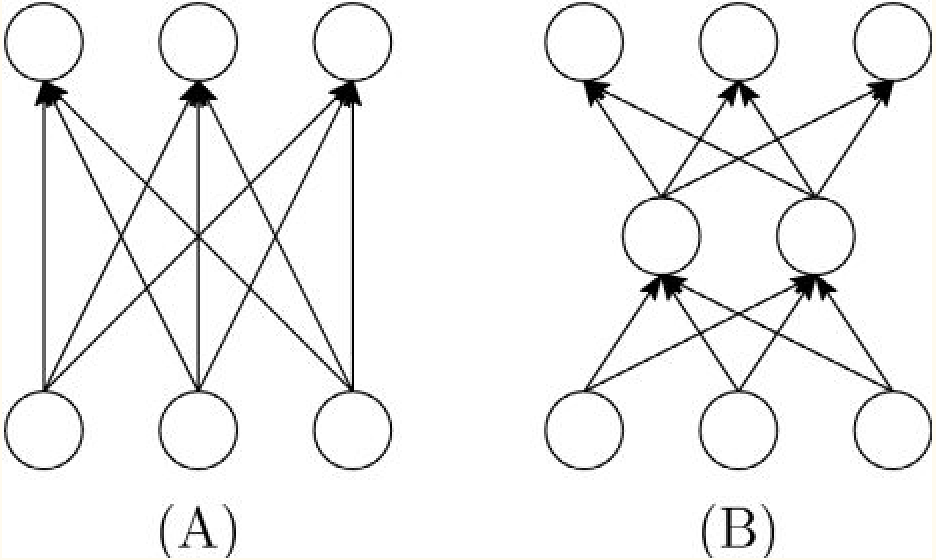
Feedforward neural network

Softmax output layer

Stochastic gradient descent (online)

Exponentially decreasing learning rate

Gaussian prior (L2 regularization term) to reduce overfitting



Log-linear model (maximum entropy)

$$P(Y = y|X = x; \theta) = \frac{1}{Z} \exp\Big(\sum_i \theta_i f_i(y, x)\Big) \tag{4.1}$$

where Z is the normalization term:

$$Z = \sum_y \exp\Big(\sum_i \theta_i f_i(y, x)\Big) \tag{4.2}$$

Softmax activation function

In this comparison, we propose replacing a standard log-linear model with a feed-forward neural network with *softmax output function*. Because our goal is to estimate probability distributions (Equation 4.1), the output layer of the network uses the softmax activation function (Bridle, 1989, 1990), a commonly known activation function producing probability distributions. If we denote q_j as the sum of inputs of an output neuron j, the output of this neuron is:

$$output_j = \frac{\exp(q_j)}{\sum_{i=1}^{\#outcomes} \exp(q_i)}.$$

Parameter Estimation - Objective Function

Our goal is to obtain a network that maximizes the likelihood of the training data. Namely, we use negative log likelihood as the loss function which we try to minimize:

$$L(\theta) = -\mathbb{E}_{x,y\sim training\ data} \log P(y|x;\theta) \tag{4.3}$$

$$\theta = \underset{\theta^*}{\operatorname{argmin}}\, L(\theta^*) \tag{4.4}$$

Parameter Estimation - Update Step

one sample (or several samples, called a *minibatch*) from the training data. Then, parameters of the model are updated to lower the loss function, by adjusting their value in a direction opposite to the partial derivative of the loss function:

$$\theta_i \leftarrow \theta_i - \lambda \frac{\partial L(\theta)}{\partial \theta_i} \tag{4.5}$$

Parameter Estimation - L2 Regularization

To reduce overfitting, we use Gaussian prior (Chen and Rosenfeld, 1999), which corresponds to L2-regularization term. The SGD update rule then becomes:

$$\theta_i \leftarrow \theta_i - \lambda \frac{\partial L(\theta)}{\partial \theta_i} - \sigma^2 \theta_i \quad (4.6)$$

Featureless NN for NER Details

Featureless NN for NER - Details

The network is trained with AdaGrad (Duchi et al., 2011).

We use dropout (Sristava et al., 2014) on the hidden layer.

Implemented in Torch7 (Collobert et al., 2011a).

Window size = 2.

Hidden layer size = 200, dropout = 0.5, minibatches size = 100.

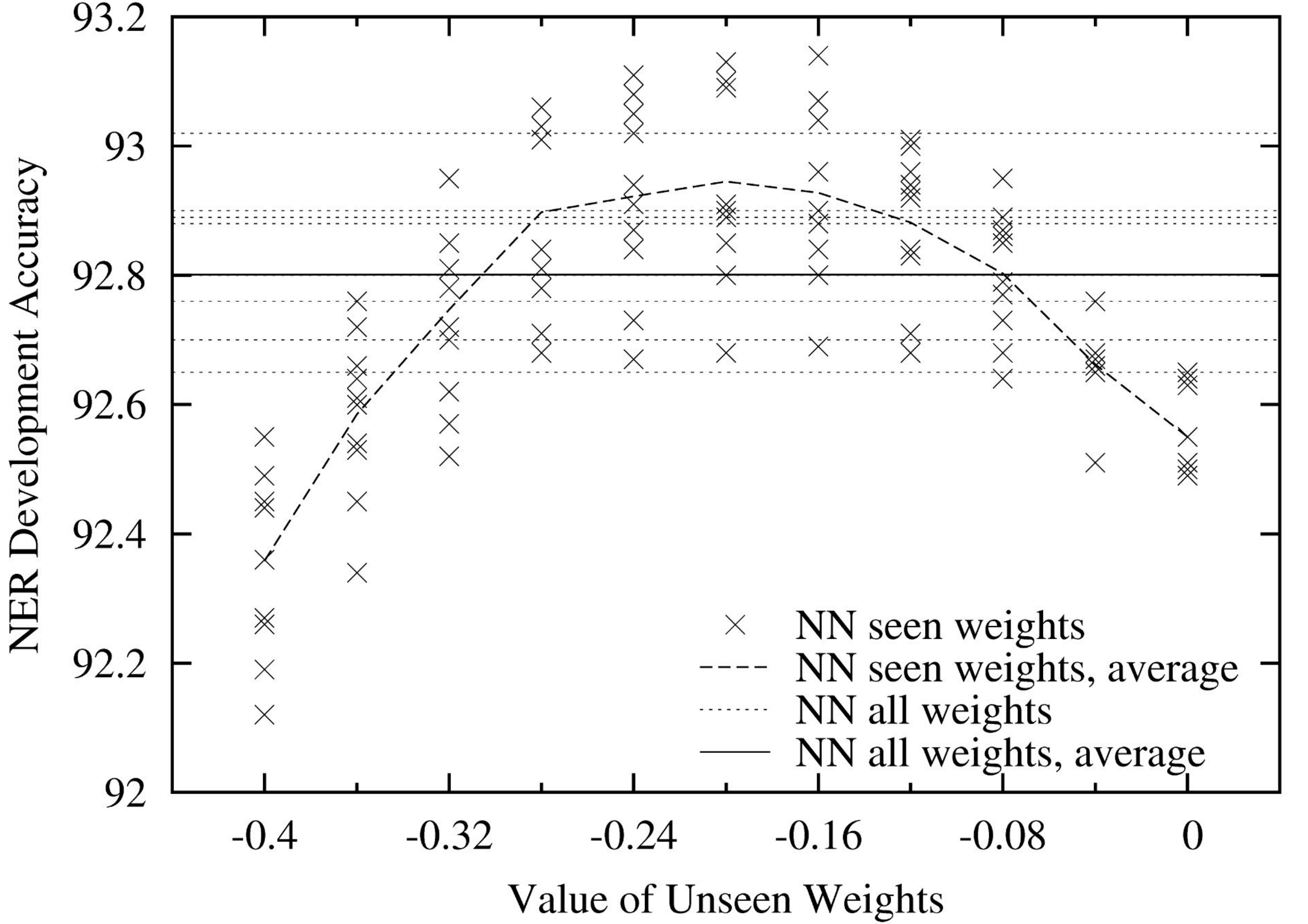
Learning rate = 0.02 with decay.

Character-level word embeddings dimension is 32 or 64.

Word embeddings dimension is 200 (English - GigaWord, Czech - Czech SYN).

Ensemble of 5 networks.

Figures


NER Development Accuracy
Value of Unseen Weights
93.2
93
92.8
92.6
92.4
92.2
92
-0.4
-0.32
-0.24
-0.16
-0.08
0
NN seen weights
NN seen weights, average
NN all weights
NN all weights, average

Tables

NER with Softmax NN - F-measure on CNEC

	Types F1	Supertypes F1
Kravalová and Žabokrtský (2009)	68.00	71.00
Konkol and Konopík (2011)	NA	72.94
Straková et al. (2013)	79.23	82.82

NER with Softmax NN - F-measure on CoNLL-2003

	F-measure
Finkel et al. (2005)	86.86
Chieu and Ng (2003)	88.31
Florian et al. (2003)	88.76
Ando and Zhang (2005)	89.16
Straková et al. (2013)	89.16
Suzuki and Isozaki (2008)	89.92
Ratinov and Roth (2009)	90.80