

End-to-End neural dependency parsing

(Parsing zależnościowy za pomocą sieci neuronowej)

Michał Zapotoczny

Praca magisterska

Promotor: dr Jan Chorowski

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

25 kwietnia 2017

Michał Zapotoczny

.....

.....

(adres zameldowania)

.....

.....

(adres korespondencyjny)

PESEL:

e-mail:

Wydział Matematyki i Informatyki

stacjonarne studia II stopnia

kierunek: informatyka

nr albumu: 248100

Oświadczenie o autorskim wykonaniu pracy dyplomowej

Niniejszym oświadczam, że złożoną do oceny pracę zatytułowaną *End-to-End neural dependency parsing* wykonałem samodzielnie pod kierunkiem promotora, dr. Jana Chorowskiego. Oświadczam, że powyższe dane są zgodne ze stanem faktycznym i znane mi są przepisy ustawy z dn. 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (tekst jednolity: Dz. U. z 2006 r. nr 90, poz. 637, z późniejszymi zmianami) oraz że treść pracy dyplomowej przedstawionej do obrony, zawarta na przekazanym nośniku elektronicznym, jest identyczna z jej wersją drukowaną.

Wrocław, 25 kwietnia 2017

(czytelny podpis)

Abstract

...

...

Contents

1	Introduction	7
1.1	Dependency parsing	8
1.1.1	Parsing algorithms	8
1.1.2	Evaluation of parsing algorithms	10
1.1.3	Universal Dependencies	10
1.2	Neural Networks	11
2	Neural dependency parser	13
2.1	Overview of the network architecture	13
2.1.1	Reader	13
2.1.2	Tagger	15
2.1.3	Parser	15
2.2	Training	16
2.2.1	Training criterion	16
2.2.2	Regularization, optimizer and hyperparameters	17
2.3	Multilingual training	17
3	Results	19
3.1	Single language	19
3.1.1	Impact of <i>reader</i> and <i>predictor</i>	20
3.1.2	Impact of decoding algorithm	20
3.1.3	Word pieces	20
3.1.4	Pointer softening	20

3.1.5	Recurrent state size	20
3.2	Multilanguage	20
3.3	Error analysis	20
4	Summary	23
	Bibliography	25

Chapter 1

Introduction

The ability to communicate with the user in a natural language is a major driving force in development of natural language processing algorithms. One of the basic tasks in a NLP pipeline is parsing by which we can describe sentence structure. There exist two basic parsing techniques: constituency and dependency parsing. With constituency parser we break the sentence into phrases, which can be further broken into smaller sub-phrases. Example of such parsing is shown in figure 1.1.

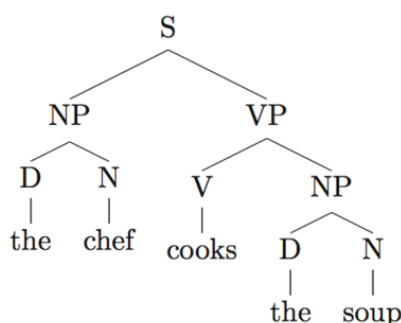


Figure 1.1: A sample constituency parse tree [ściągnięte z internetu, przegenerować wektorowo](#)

In dependency parsing each word (called *dependent*) is connected via labelled arc to another word of the sentence (called *head*) or to the special *ROOT* vertice, forming a directed tree. Example of such parsing is depicted on figure 1.2.

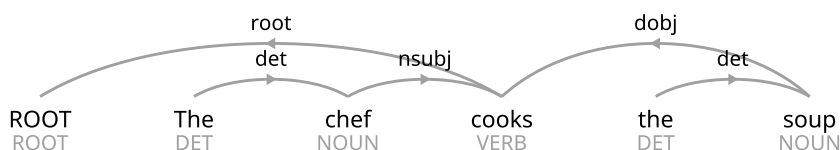


Figure 1.2: A sample dependency parse tree

1.1 Dependency parsing

A major advantage of dependency parsers over constituency ones are their independence from the word ordering. It is important in morphologically rich languages like Polish or Czech where the ordering is very flexible, and thus related words can be far apart from each other. Additionally the head-dependent relationship is a good approximation to semantic relationship between words [Covington, 2001] which is important for tasks like question answering or information extraction.

The labels of head-dependent arcs tells us about grammatical function that dependent word have in respect to the head. In table 1.1 we show some of the most popular labels for the english language.

Label	Description	Example
CASE	Case marking	<i>From</i> Friday 's Daily Star
NSUBJ	Nominal subject	<i>Musharraf</i> calls the bluff
NMOD	Nominal modifier	India defensive over Sri <i>Lanka</i>
DET	Determiner	That he missed <i>a</i> physical ?
DOBJ	Direct object	Did you know <i>that</i> ?
ADVMOD	Adverbial modifier	<i>So</i> Bush stopped flying .
AMOD	Adjectival modifier	Six weeks of <i>basic</i> training .
COMPOUND	Compound	Bush 's <i>National</i> Guard years

Table 1.1: Most popular labels from English treebank. Dependents are *italic*, while heads are **bold**

Projectivity

We can distinguish two types of dependency trees: projective and nonprojective. We say that head-dependent arc is projective if there exists a path from head to every word between head and dependent in the sentence. A dependency tree is projective if all its arcs are projective. Intuitively we can say that nonprojective trees are trees that cannot be drawn without crossing the edges (see figure 1.3). Although English dataset has less than 5% of nonprojective sentences other languages can have significant amount - like Czech (12%) or Ancient Greek (63%) [Straka et al., 2015]. It is important to acknowledge this fact because some of the presented parsing methods can only produce projective trees.

1.1.1 Parsing algorithms

In following section we will present two basic approaches to dependency parsing. Although they differ in complexity and output flexibility, both of them depend on supervised machine learning techniques.

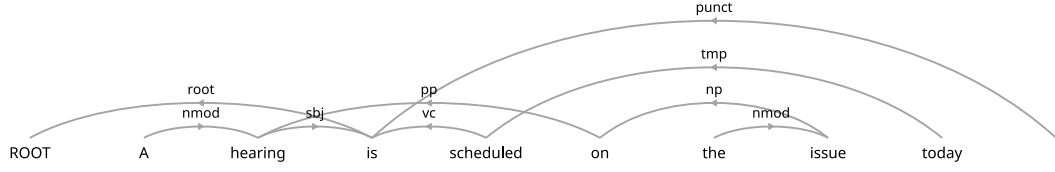


Figure 1.3: A sample nonprojective dependency parse tree. The $\text{on} \rightarrow \text{hearing}$ arc is nonprojective and thus the whole tree becomes nonprojective.

Transition based

In transition based parsing we find a transition sequence from initial configuration to terminal one. A configuration c is a triple (s, b, A) containing stack s , buffer b and set of dependency arcs A . For given sentence w_1, \dots, w_n the corresponding initial configuration would be $([\text{ROOT}], [w_1, \dots, w_n], \emptyset)$. The terminal configuration is $(\text{ROOT}, [], A)$ in which A is resulting parse tree. There are three possible transitions:

- **LEFTARC(1)** - Pop from the stack elements s_1 and s_2 , add arc $s_1 \rightarrow s_2$ with label l to the set A and push s_1 back to the stack
- **RIGHTARC(1)** - Pop from the stack elements s_1 and s_2 , add arc $s_2 \rightarrow s_1$ with label l to the set A and push s_2 back to the stack
- **SHIFT** - Move first element from the buffer to the stack

An example of parsing sequence is shown in table 1.2.

Stack	Buffer	Transition	New arcs
[ROOT]	[He has good control]	SHIFT	-
[ROOT He]	[has good control]	SHIFT	-
[ROOT He has]	[good control]	LEFTARC(nsubj)	nsubj(has, He)
[ROOT has]	[good control]	SHIFT	-
[ROOT has good]	[control]	SHIFT	-
[ROOT has good control]	[]	LEFTARC(amod)	amod(control, good)
[ROOT has control]	[]	RIGHTARC(dobj)	dobj(has, control)
[ROOT has]	[]	RIGHTARC(root)	root(ROOT, has)
[ROOT]	[]	-	-

Table 1.2: Transitions needed to parse sentence "He has good control"

Because every word of the sentence has been shifted to the stack exactly once and **ARC** operations reduce stack size by one we will have exactly $2n$ transitions from initial to terminal configuration, so the time complexity of this algorithm is $O(n)$ where n is number of words in a sentence. Using this algorithm we can represent any projective tree [Nivre, 2008] (nonprojectives cannot be represented).

The decision which transition to use between consecutive configurations is obtained through machine learning techniques. This can vary from simple linear

SVM [Nivre et al., 2005] to neural networks [Chen and Manning, 2014, Andor et al., 2016].

Graph based

Alternative approach is to use graph-based algorithms. The idea is that we define a space of candidate arcs, find a model to score each arc and then use some parsing algorithm to find dependency tree with highest score.

The candidate space and scoring depends on the used learning model, but we can distinguish two main parsing algorithms: Chu-Liu-Edmonds and Eisner.

Chu-Liu-Edmonds *Tutaj opis cle*

Eisner *Tutaj opis eisnera*

1.1.2 Evaluation of parsing algorithms

Having obtained dependency tree for a particular sentence we have to be able to compare it with gold-standard parse. There are two main evaluation metrics:

- **Unlabelled Attachment Score** (UAS), is a percentage of words with correct predicted head
- **Labelled Attachment Score** (LAS), is a percentage of words with correct predicted head **and** correct predicted label

In the experimental section we will show both scores, whenever available.

1.1.3 Universal Dependencies

Up till recently most development of dependency parsers was done in single language setting, where we have different parser for every language we want to use, each using language-specific morphosyntactical features. The Universal Dependencies project [Nivre et al., 2015] aims to provide a cross-linguistically consistent treebank annotation. It is based on previous work on universal Stanford Dependencies [De Marneffe et al., 2006], Google universal pos tags [Petrov et al., 2011] and the Intersect interlingua [Zeman, 2008]. The UD project unifies part of speech tags and dependency relations labels for 30+ languages under common CONNL-U format. The main advantage is that we can use the same parsing system for every language (data has consistent format) additionally because the morphosyntactical data has common format we can try to combine data coming from different languages (see section 2.3).

1.2 Neural Networks

A neural network is a mathematical model loosely based on how brain works. The models proven to be very flexible and obtained state of the art results in many tasks (cytaty). A neural network is defined by its topology, activation functions and parameters. The first two are chosen by the author, while the parameters are obtained by learning procedure.

The topology of the network is telling us how the layers of neurons are connected to each other. Each layer have a number of neurons which have some incoming indices i , associated activation function ϕ and some parameters w . The neuron combines the incoming signals, usually by taking weighted average and then applying the activation function: $\phi(\sum_{k=0}^{|i|} w_k i_k)$. This value is then feeded to the connected neurons from the next layer.

The activation function can have any form, but the most common are hyperbolic tangent tanh and ReLu [Nair and Hinton, 2010].

By the learning procedure we understand a optimization process that given a loss function L , data x and some parameters W to tune, tries to minimize $L(x)$. Because usually the data x size is big we cannot use simple gradient descent algorithm on all data at once because it would not fit into the memory. Instead we are using stochastic gradient descent to train only on small subset of data at once (this subset is called a minibatch).

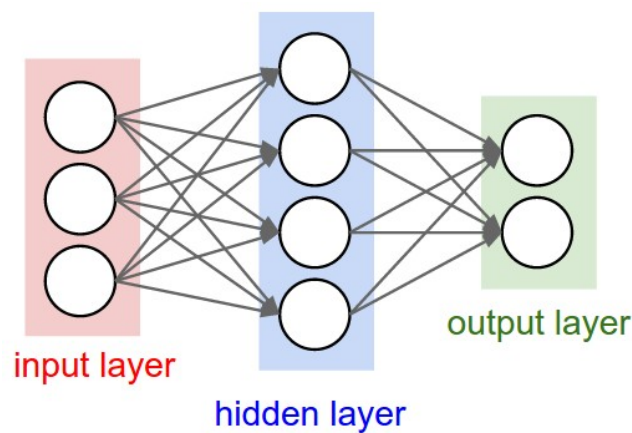


Figure 1.4: An sample neural network architectureinternet...

Chapter 2

Neural dependency parser

2.1 Overview of the network architecture

The network architecture consists of three main parts: reader, tagger and parser (see Figure 3.3). The reader subnetwork is evaluated on each individual word in a sentence, and using convolutions on their orthographic representation produces words embeddings. Next, the tagger subnetwork implemented as bidirectional RNN equips each word with a context of whole sentence. Finally parser part computes dependency tree parent for each word using attention mechanism [Vinyals et al., 2015] after which network computes appropriate dependency label. In the following paragraphs we will describe all parts in detail.

2.1.1 Reader

As stated before the reader subnetwork is run on each word producing its embedding. This architecture is based on [Kim et al., 2015]. Each word w is represented by sequence of its characters plus a special beginning-of-word and end-of-word tokens. Firstly we find low-dimensional characters embeddings which are concatenated to form a matrix C^w . Then we run 1D convolutional filters on C^w which then is reduced to vector of filter responses computed as:

$$R_i^w = \max(C^w * F^i) \quad (2.1)$$

Where F^i is i -th filter. The purpose of the convolutions is to react to specific part of words (because we have bow and eow tokens it can also react to prefixes and suffixes) which in morphologically rich languages such as Polish can depict its grammatical role.

Finally we transform filter responses R^w with a simple multi-layer perceptron ¹ obtaining the final word embedding $E^w = \text{MLP}(R^w)$.

¹Which are just linear transformations followed by non-linearity

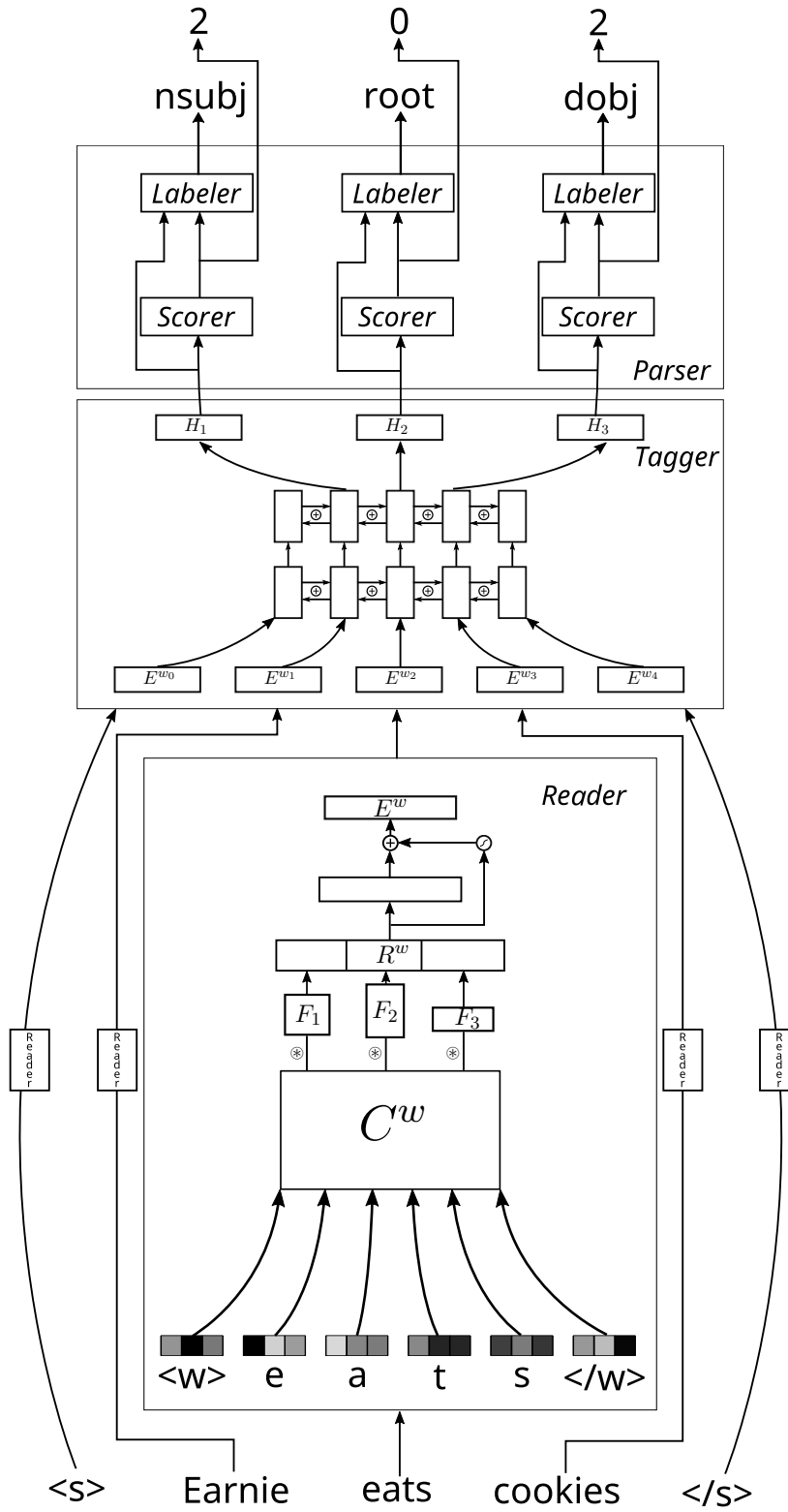


Figure 2.1: The model architecture.

2.1.2 Tagger

Having obtained the word embeddings E^w we can proceed with actually "reading" whole sentence. To do this we use multi-layer bidirectional RNN ([Schuster and Paliwal, 1997]) (we have evaluated LSTM [Hochreiter and Schmidhuber, 1997] and GRU [Cho et al., 2014] types). We combine the backward and forward passes by adding them.

POS tag predictor

To prevent overfitting and encourage network to compute morphological features we can add additional training objective. It works by taking output from one of the tagger BiRNN layers and use it to predict available part-of-speech tags for each word. The result is not feeded back to the rest of the network because we think it would introduce too much noise.

2.1.3 Parser

The last part of the network serves two purposes: to find head for each word and to compute label for that edge.

Finding head word

We use method similar to [Vinyals et al., 2015]. The input to this part are word annotations H_0, H_1, \dots, H_n (where H_0 serves as a root word) produced by the tagger. For each of the words $1, 2, \dots, n$ we compute probability distribution which tell us where the head of current word should be $(0, 1, 2, \dots, n)$. This computation (called *scorer*) is implemented as small feedforward network $s(w, l) = f(H_w, H_l)$, where $w \in 1, 2, \dots, n$, $l \in 0, 1, 2, \dots, n$.

Finding edge label

Output of the scorer can already be interpreted as pointer network, but in order to use it as attention for computing dependency label we have to normalize it:

$$p(w, l) = \sum_{i=0}^n \frac{f(H_w, H_l)}{f(H_w, H_i)} \quad (2.2)$$

The dependency label is computed by small Maxout network [Goodfellow et al., 2013], which takes the current word annotation H_w and heads annotation A_w . This part is called *labeler*. We investigated two variations of head annotation A_w

Soft attention

Which is a weighted average of normalized attention 2.2 and words annotations H .

$$A_w = \sum_{i=0}^n p(w, i) H_i$$

Hard attention

Here we use a only head annotation.

$$A_w = H_h$$

During training we use ground-truth head location, whereas during evaluation we use head word computed in previous step.

Decoding algorithm

The *scorer* give us a $n \times n + 1$ matrix of head dependency preference for each word. From this matrix we have find a set of dependencies that satisfy some constraints (exactly one root dependant, no cycles). We investigated two possibilities for such decoding: a greedy algorithm, and Chu-Liu-Edmonds [Edmonds, 1966].

2.2 Training**2.2.1 Training criterion**

Every neural network needs a training criterion which will be optimized. Here we have 3 individual training criterion combined together as linear combination. Those are:

- The negative log-likelihood loss L_h on finding proper head for each word. The training signal is propagated from scorer down to the reader.
- The negative log-likelihood loss L_l on finding dependency label. With soft attention it is propagated through the whole network (excluding pos tagging part), while with the hard attention we do not propagate through the scorer.
- The (optional) negative log-likelihood loss L_t on predicting pos tags. This error is backpropagated only through pos-predictor and part of tagger down to the reader.

So the final loss is:

$$L = \alpha_h L_h + \alpha_l L_l + \alpha_t L_t \quad (2.3)$$

2.2.2 Regularization, optimizer and hyperparameters

Regularization is an essential part of neural network training. It improves generalization and prevents overfitting. One of the most popular regularization technique is called Dropout [Srivastava et al., 2014]. Using it, we randomly drop part of the connections from a certain network layer during training. In our case dropout is applied to the *reader* output, between the BiRNN layers of the *tagger* and to the *labeller*.

The models are trained using Adadelta [Zeiler, 2012] learning rule, with weight decay and adaptive gradient clipping [Chorowski et al., 2014]. All experiments are early stopped on validation set UAS.

Hyperparameter selection is crucial for neural networks to obtain good results. For example, comparing our first experiments with architecture depicted above to the best single-language results (using the same basic architecture) gave us about 5% boost in UAS score on Polish language.

To find the best hyperparameters we have used the Spearmint system [Snoek et al., 2012] invoked on polish dataset. The chosen parameters are as follows. The *reader* embeds each character into 15 dimensions, and contains 1050 filters (50·k filters of length k for $k = 1, 2, \dots, 6$) whose outputs are projected into 512 dimensions transformed by a 3 equally sized layers of feedforward neural network with ReLU activation. The *tagger* contains 2 BiRNN layers of GRU units with 548 hidden states for both forward and backward passes which are later aggregated using addition. Therefore the hidden states of the tagger are also 548-dimensional. The *POS tag predictor* consists of a single affine transformation followed by a SoftMax predictor for each POS category. The *scorer* uses a single layer of 384 tanh for head word scoring while the *labeller* uses 256 Maxout units (each using 2 pieces) to classify the relation label [Goodfellow et al., 2013]. The training cost used the constants $\alpha_h = 0.6, \alpha_l = 0.4, \alpha_t = 1.0$. We apply 20% dropout to the *reader* output, 70% between the BiRNN layers of the *tagger* and 50% to the *labeller*. The weight decay is 0.95.

2.3 Multilingual training

The big problem of learning to parse is small number of gold standard dependency trees for many languages (including Polish). With small number of examples neural networks do not generalize well and can more easily overfit. There also exist languages with good, standardized treebank like Czech Prague Treebank [Bejček et al., 2013].

Because our model is purely neural network we can incorporate a multitask learning [Caruana, 1997]. It allows the network to learn multiple tasks at the same time, sharing common patterns and distinguishing differences. Additionally, because we are using Universal Dependencies treebanks [Nivre et al., 2015] we have common

standardized format across many languages, which allowed for easy implementation of multitask learning and experiments across different languages.

The multilingual model for n languages can be viewed as n copies of our basic model, but sharing part of the parameters. To unify input/output for each of the models we sum possible outputs for each data category (characters, POS categories, dependency labels). If some category doesn't exist within a particular language then we use a special UNK token. To actually make use of multitask learning we must share at least part of the parameters of all models. We experimented with different sharing strategies, from share-everything to only sharing the *parser* part. Additionally to prevent over-representation of some languages during training we sample (on each epoch) only portion of the available data so that each language have equal number of examples (equal to number of samples of the smallest language).

Chapter 3

Results

3.1 Single language

All experiments depicted in this section are based on configuration shown in section 2.2.2. In table 3.1 we show our baseline results for subset of the UD languages (due to limited computational resources we were not able to train models for all of them). The results are compared to SyntaxNet[Andor et al., 2016] and ParseySaurus[Alberti et al., 2017], both from Google. The ParseySaurus is based on SyntaxNet, and uses characters as input, while the original SyntaxNet uses word embeddings. Both are transition-based, and needs no external POS tagger.

language	#sentences	Ours		SyntaxNet		ParseySaurus	
		UAS	LAS	UAS	LAS	UAS	LAS
Czech	87 913	91.41	88.18	89.47	85.93	89.09	84.99
Polish	8 227	90.26	85.32	88.30	82.71	91.86	87.49
Russian	5 030	83.29	79.22	81.75	77.71	84.27	80.65
German	15 892	82.67	76.51	79.73	74.07	84.12	79.05
English	16 622	87.44	83.94	84.79	80.38	87.86	84.45
French	16 448	87.25	83.50	84.68	81.05	86.61	83.1
Ancient Greek	25 251	78.96	72.36	68.98	62.07	73.85	68.1

Table 3.1: Baseline results of models trained on single languages from UD v1.3. Our models use only the orthographic representation of tokenized words during inference and works without a separate POS tagger.

It is worth noting that SyntaxNet has different hyperparameters configuration for each language, while our parser uses the same configuration across all languages. In the following sections we will show impact of different network parameters on the result.

3.1.1 Impact of *reader* and *predictor*

Firstly we will inspect the impact of different *reader* and *predictor* settings.

training inputs	Czech		English		Polish	
	UAS	LAS	UAS	LAS	UAS	LAS
Gold POS tags						
base word	91.7	88	88.6	85.1	93.4	89.3
Predicted POS tags or no POS tags						
words	82.4	72.1	81.9	74.7	74.6	61.6
chars, soft att.	90.1	85.7	86.5	82.1	89.1	82.5
chars, tags, soft att.	89.6	82.8	86.2	81.3	90.4	83.9
chars, tags, hard att.	90.1	86.7	87.6	83.6	91.3	86

Table 3.2: Model performance on selected languages and different training inputs. Experiments were run on UD v 1.2

In the upper part "Gold POS tags" we show a favorable situation when we have access to gold pos tag (approved by human). The inputs are base words embedded in a vector - basically instead of whole *reader* part we have just simple table lookup. This setting allowed us to obtain best results (especially for Polish) but gold pos tag are not available in real-word setting so this result is noted just as a trivia.

The words and chars, soft att. settings don't use *predictor*. Using just words as inputs (embedded in the same manner as base word) gives words results. This is easy to predict because most words in the dataset occurs only once and so the network doesn't have clue about role of most of the words. Instead using character-level input allows us to obtain much better results, especially because in morphosyntactically rich languages the word spelling catches many word features that can be used in deducing the pos tag information.

3.1.2 Impact of decoding algorithm

3.1.3 Word pieces

3.1.4 Pointer softening

3.1.5 Recurrent state size

3.2 Multilanguage

3.3 Error analysis

#pieces	UAS	LAS	LAB
Polish			
25	90.29	84.91	90.17
50	90.40	85.46	90.77
75	90.23	84.87	90.40
100	89.97	84.44	90.23
Czech			
25	90.29	86.50	92.66
50	90.03	86.08	92.40
75	90.17	86.40	92.70
100	90.84	87.31	93.20

Table 3.3: Results on word pieces model. #pieces denotes how many new multi-character tokens were used. To convert word to pieces we use a greedy algorithm in which we choose the longest piece that is equal to prefix of word.

Language	Algorithm	UAS	LAS	LAB
Polish	Greedy	90.91%	86.18%	91.16%
	Edmonds	90.90%	86.15%	91.16%
Czech	Greedy	90.38%	86.79%	92.79%
	Edmonds	90.37%	86.77%	92.79%

Table 3.4: Results of label softening. Instead of using one-hot vector as groundtruth for the *scorer* we set 0.51 for the correct position and distribute the rest uniformly

Language	% of recurrent state size	UAS	LAS	LAB
Polish	100%	90.26	85.32	90.79
	50%	90.54	85.64	90.84
	25%	89.07	82.80	89.03
Czech	100%	91.41	88.18	90.79
	50%	89.98	85.92	92.31
	25%	85.56	79.16	87.88

Table 3.5: Impact of the *tager* recurrent state size on the performance. We report this value as percent of baseline rnn size.

% of recurrent state size	UAS	LAS	LAB
200%	91.45	86.36	91.05
100%	91.65	86.88	91.57
50%	89.53	83.66	89.55
25%	87.28	78.93	85.41

Table 3.6: Impact of the *tager* recurrent state size on the performance of multilanguage pl-cs model. We report this value as percent of baseline rnn size.

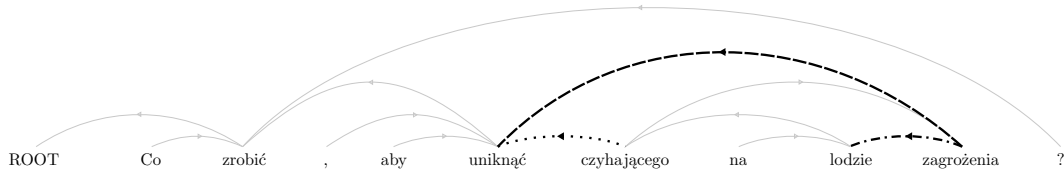


Figure 3.1: CLE Example

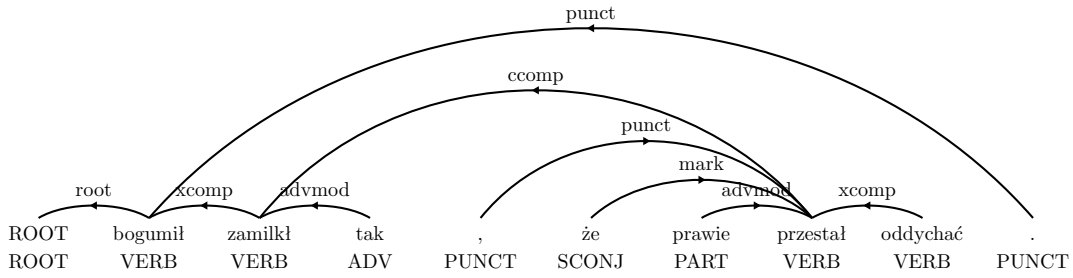


Figure 3.2: The model architecture.

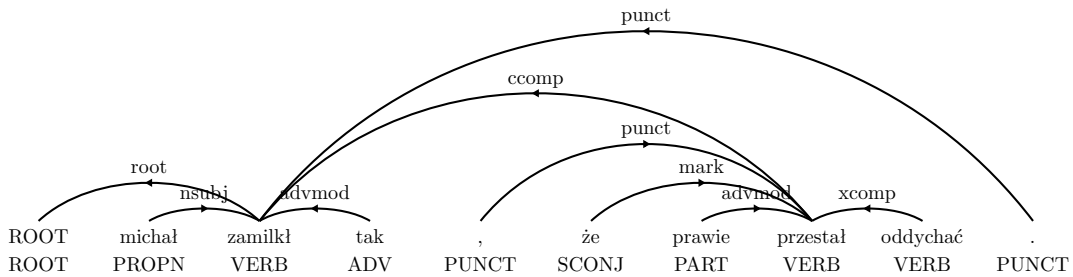


Figure 3.3: The model architecture.

Chapter 4

Summary

Bibliography

- [Alberti et al., 2017] Alberti, C. et al. (2017). SyntaxNet Models for the CoNLL 2017 Shared Task. *arXiv:1703.04929*.
- [Andor et al., 2016] Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. *arXiv:1603.06042 [cs]*. 00001 arXiv: 1603.06042.
- [Bejček et al., 2013] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague dependency treebank 3.0.
- [Caruana, 1997] Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.
- [Chen and Manning, 2014] Chen, D. and Manning, C. D. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *EMNLP*, pages 740–750. 00188.
- [Cho et al., 2014] Cho, K., van Merriënboer, B., Gülgehr, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.
- [Chorowski et al., 2014] Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y. (2014). End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. *arXiv:1412.1602 [cs, stat]*. 00000 arXiv: 1412.1602.
- [Covington, 2001] Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, pages 95–102. Citeseer. 00206.
- [De Marneffe et al., 2006] De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa.
- [Edmonds, 1966] Edmonds, J. (1966). Optimim Branchings. *JOURNAL OF RESEARCH of the National Bureau of Standards - B.*, 71B(4):233–240. 00000.

- [Goodfellow et al., 2013] Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout Networks. In *ICML*, pages 1319–1327. 00106.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- [Kim et al., 2015] Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). Character-aware neural language models. *arXiv preprint arXiv:1508.06615*. 00011
bibtex: kim_character_2015a.
- [Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- [Nivre, 2008] Nivre, J. (2008). Algorithms for Deterministic Incremental Dependency Parsing. *Comput. Linguist.*, 34(4):513–553. 00220.
- [Nivre et al., 2015] Nivre, J. et al. (2015). Universal Dependencies 1.2. <http://universaldependencies.github.io/docs/>.
- [Nivre et al., 2005] Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2005). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, page 1. 00666.
- [Petrov et al., 2011] Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. In *LREC 2012*.
- [Schuster and Paliwal, 1997] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681. 00157.
- [Snoek et al., 2012] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959. 00414.
- [Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958. 00282.
- [Straka et al., 2015] Straka, M., Hajič, J., Straková, J., and Jan Hajič, j. (2015). Parsing universal dependency treebanks using neural networks and search-based oracle. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 208–220, Warszawa, Poland. IPIPAN, IPIPAN.
- [Vinyals et al., 2015] Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2674–2682. 00010.

- [Zeiler, 2012] Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701 [cs]*. 00017 arXiv: 1212.5701.
- [Zeman, 2008] Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*.