Cross-lingual morphological analysis Literature review

Олег Сериков
Лоренцо Тони
Настя Хорошева
Влад Михайлов

Articles

- Siamese Convolutional Networks for Cognate Identification
- Using context and phonetic features in models of etymological sound change
- Morphological Analysis without Expert Annotation
- A Neural Network Based Morphological Analyser of the Natural Language

Cognate identification

- Cognates are words that come from a common ancestral language.
- Important for historical linguistics >> relationships between languages.
- Important for cross-morphology as well !
- The cognate identification task typically deals with short word lists (~ 200) and short words (~ 5)
- There is a need for developing automated cognate identification methods

SCNN

- idea of running 2 identical CNN on 2 different inputs and then comparing them = Siamese NN architecture

- idea came from DeepFace system (fb)

- siamese architectures consistently perform better than traditional linear classifier approach

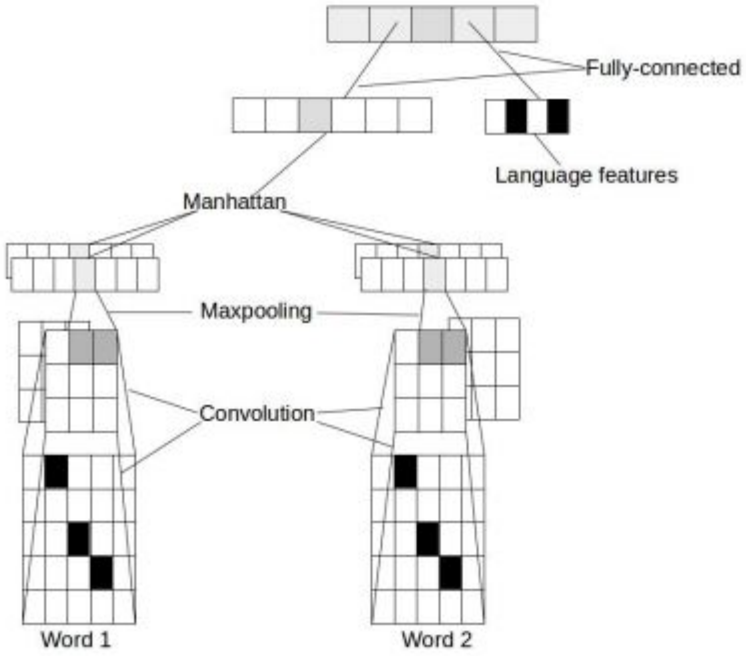


Figure 1: Illustration of Manhattan Siamese Convolutional network. We show the language features as a separate vector. Hot cells are shown in black whereas, real-valued cells are shown in grayscale.

SCNN

WHY SO IMPORTANT

- CNNs can be an alternative way to avoid explicit feature engineering through similarity computation
- SCNN is good for cognates (designed to detect similarity)

PROBLEMS

- many of the languages do not have enough corpora to train character embeddings >> hand-crafted ways of phoneme encodings to train our convolutional networks

Phoneme encoding

Features	p	b	f	v	m	8	4	t	d	s	z	c	n	S	Z	C	j	T	5	k	g	x	N	q	G	X	7	h	l	L	w	y	r	!	V
Voiced	0	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
Labial	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Dental	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alveolar	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Palatal/Post-alveolar	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Velar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Uvular	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Glottal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Stop	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0
Fricative	1	1	1	1	0	1	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0
Affricate	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nasal	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Click	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Approximant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
Lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Rhotic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Table 2: The ASJP alphabet is given in columns 2 – 35 and the phonetic value of each symbol in the ASJP alphabet. Each phoneme is a multi-hot vector of fixed dimension 16.

Implementation

Try embeddings: find cognates >> char-to-char correspondences

Using context and phonetic features in models of etymological sound change

Hannes Wettig, Kirill Reshetnikov, Roman Yangarber

aclanthology.info/pdf/W/W12/W12-0215.pdf

Что делают

Data-driven «выравнивают» этимологические данные

Датасет
- Этимологические бд, организованы как множества множеств когнатов
- i.e. StarLing (starling.rinet.ru/), часть о языках Уральской семьи

Признаки
- Контексты
- Фонологические

Модель
- Дерево решений

Как и зачем делают

Представленный алгоритм ставит в соответствие этимологическому корпусу множество извлеченных правил

Сравниваем этимологические корпуса по тому, насколько множество правил получилось емким – чем меньше правил, тем более плотный корпус

- MDL – Minimum description length

Полезная польза

Датасет когнатов, этимологические базы данных,..

Прозрение: на викисловарях есть IPA-транскрипции

Параллелизм между поиском родственных слов и машинным переводом

Morphological Analysis without Expert Annotation

GOAL

Create a morphological analyzer, which is designed to be trained on plain inflection tables.

No need for expert rule engineering or morphologically annotated corpora.

Morphological Analysis without Expert Annotation

alignment -> transduction -> re-ranking -> thresholding

Morphological Analysis without Expert Annotation

	singular	plural
nominative	ле́мма lémma	ле́ммы lémmy
genitive	ле́ммы lémmy	ле́мм lémm
dative	ле́мме lémme	ле́ммам lémmam
accusative	ле́мму lémmu	ле́ммы lémmy
instrumental	ле́ммой, ле́ммою lémmoj, lémmoju	ле́ммами lémmami
prepositional	ле́мме lémme	ле́ммах lémmax

(a) Raw Wiktionary

	singular	plural
nominative	ле́мма	ле́ммы
genitive	ле́ммы	ле́мм
dative	ле́мме	ле́ммам
accusative	ле́мму	ле́ммы
instrumental	ле́ммой	ле́ммами
prepositional	ле́мме	ле́ммах

(b) Unannotated Table

	singular	plural
nominative	N;NOM;SG	N;NOM;PL
genitive	N;GEN;SG	N;GEN;PL
dative	N;DAT;SG	N;DAT;PL
accusative	N;ACC;SG	N;ACC;PL
instrumental	N;INS;SG	N;INS;PL
prepositional	N;ESS;SG	N;ESS;PL

(c) Annotated Table

Inflected tables for M2M

->

л е м м а   L E M M A+NOMSG

л е м м ы   L E M M A+GENSG

л е м м о й   L E M M A+INSSG

Morphological Analysis without Expert Annotation

E17-2034

s	c	h	r	e	i	b	et	
s	c	h	r	e	i	b	en+2PKA	✓
s	c	h	r	e	i	b	**en+2PKE**	✓
s	c	h	r	e	i	b	en+3SIA	×
s	c	h	r	e	i	b	en+3PIE	×
s	c	h	r	e	i	b	en+2PIA	✓

M2M to DirecTL+

->

Source	Target	
schreiben + 2PKA	schriebet	×
schreiben + 2PKE	**schreibet**	✓
schreiben + 3SIA	schrieb	×
schrieben + 2PKE	schriebet	×
schreiben + 2PIA	schriebt	×

Morphological Analysis without Expert Annotation

Reranking and Thresholding criteria ->

	Description	Type
1	lemma in Corpus	binary
2	LM score	real
3	DIRECTL+ score	real
4	affix match	binary
5	no affix match	binary
6	no affix match, top-1	binary
7	mirrored	binary
8	not mirrored	binary
9	not mirrored, top-1	binary

Comparison score

	English			German			Dutch			Spanish		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DIRECTL+	**93.5**	88.9	**91.2**	**87.3**	**88.7**	**88.0**	**87.3**	**90.3**	**88.8**	**99.3**	**99.5**	**99.4**
Marmot	87.5	**94.3**	90.8	85.3	88.5	86.9	81.3	84.7	82.9	99.2	98.9	99.1

A Neural Network Based Morphological Analyser of the Natural Language

- A morphological analyser supported by a neural network to inflect words written in Polish
- The main task is to create base forms from the analysed words' forms
- The common words are inflected with a very high quality of 99.9%
- Proper nouns inflect with a quality of 93.3%

Morphological analyser

Approaches:

- A dictionary approach (dictionary)
- An algorithmic approach (set of inflection rules)

Approaches

- Dictionary approach

\+ : high quality (dictionary words)

\- : dictionary development

- Algorithmic approach

\+ : ability to analyse OOV

\- : lower quality, good set of rules development

A Neural Network Based Morphological Analyser of the Natural Language

- Dictionary approach + algorithmic approach;
- A full dictionary of the Polish language (training set);
- Inflection patterns (similar to the inflection rules);
- A decision tree – to assign appropriate inflection pattern to the given word's form;
- The main focus: to show that NN can increase quality of the analyser.

The Morphological Analyser

Inflection pattern consists of the set of affixes:

dziad**ek**, dziad**ka**, dziad**kowi**, dziad**kiem**, dziad**ku**, dziad**ki**, dziad**ków**, dziad**kom**, dziad**kami**, dziad**kach**

As we can see this word contains root *dziad-*, and can have following suffixes:

-ek, -ka, -kowi, -kiem, -ku, -ków, -kom, -kami, -kach.

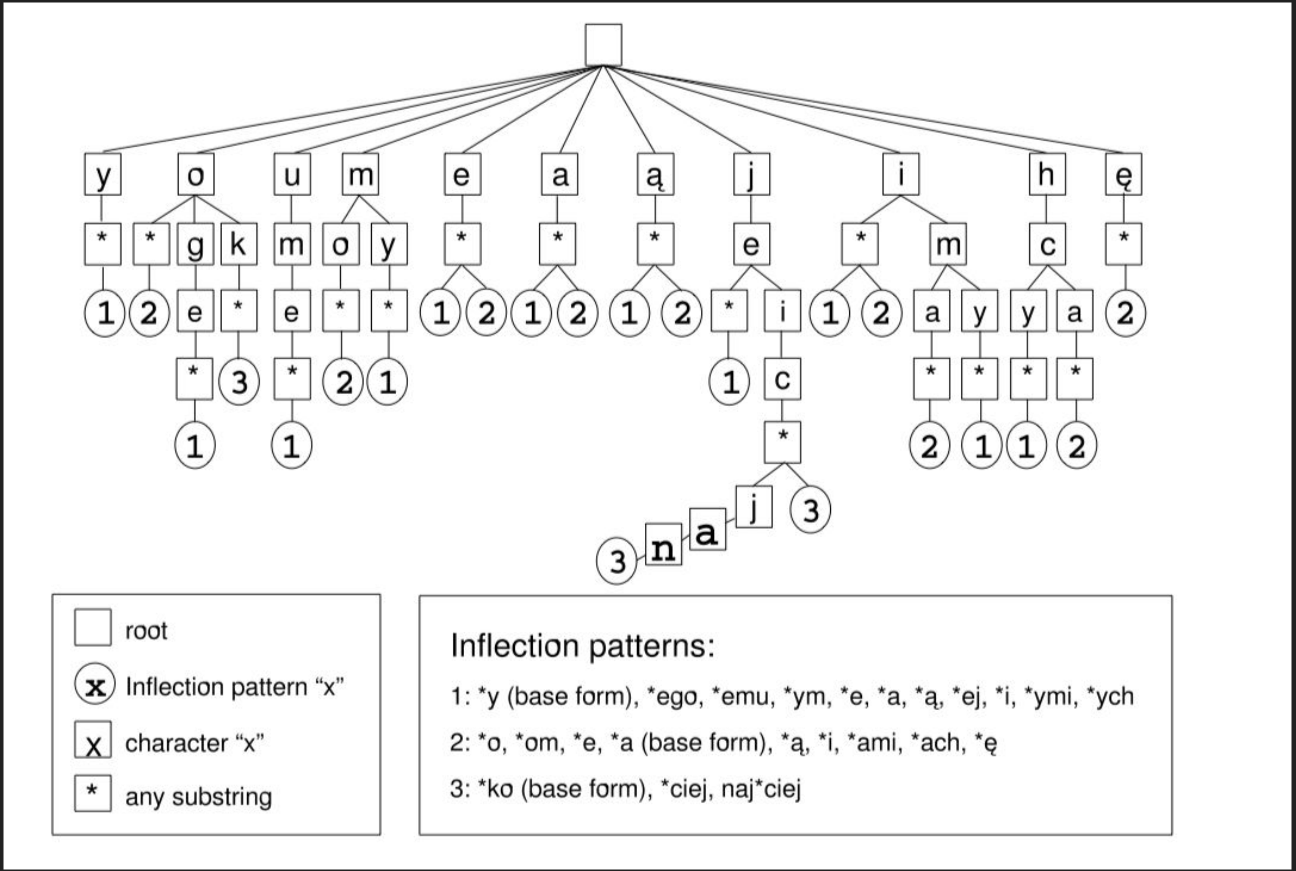
The Morphological Analyser

szyb**ko**, szyb**ciej**, **naj**szyb**ciej**

-ko, -ciej, naj-ciej

-ko [baseform + adverb], **-ciej** [adverb + comparative], **naj-ciej** [adverb + superlative]

Decision Tree of the Morphological Analyser



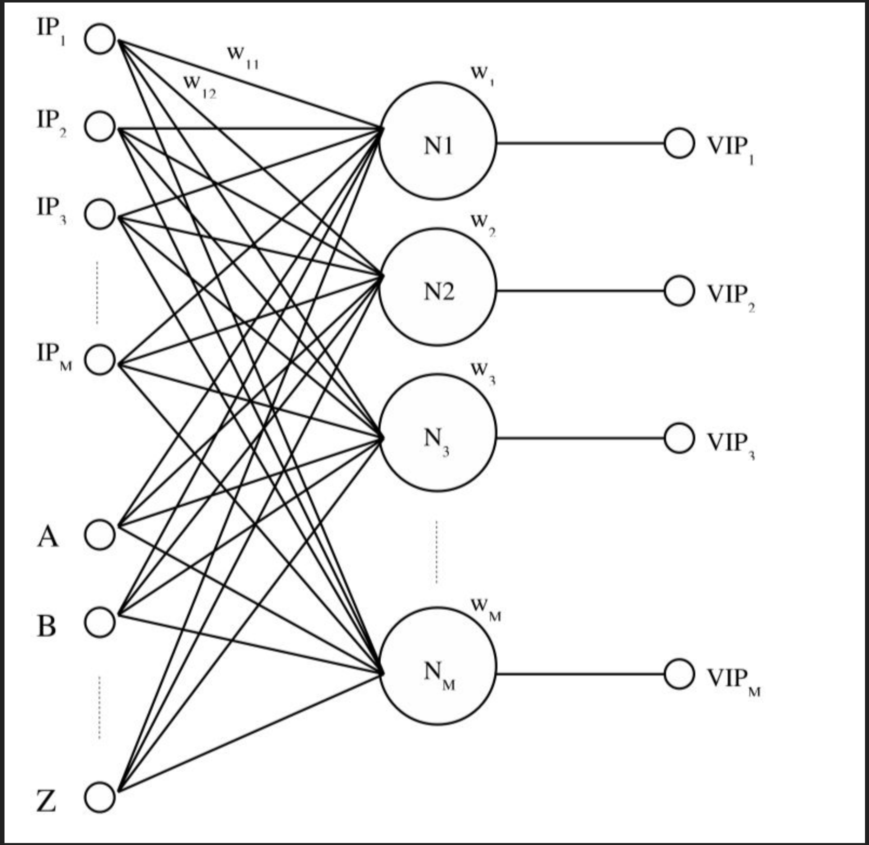
arktyczny
arktyczne
bakteria
bakterie

Nodes: decisions
Leafs: hypotheses

The problem:
candidate choice

Roots have similarities

Neural Network in the Morphological Analyser



Valid inflection pattern selection from all the candidates returned by a decision tree

The tree generates a list of candidates and stimulates the NN, and the NN -> output

The inputs of the NN layer points to the inflection patterns, stimulated by the tree

Each output of the layer points to the target inflection pattern

Neural Network in the Morphological Analyser

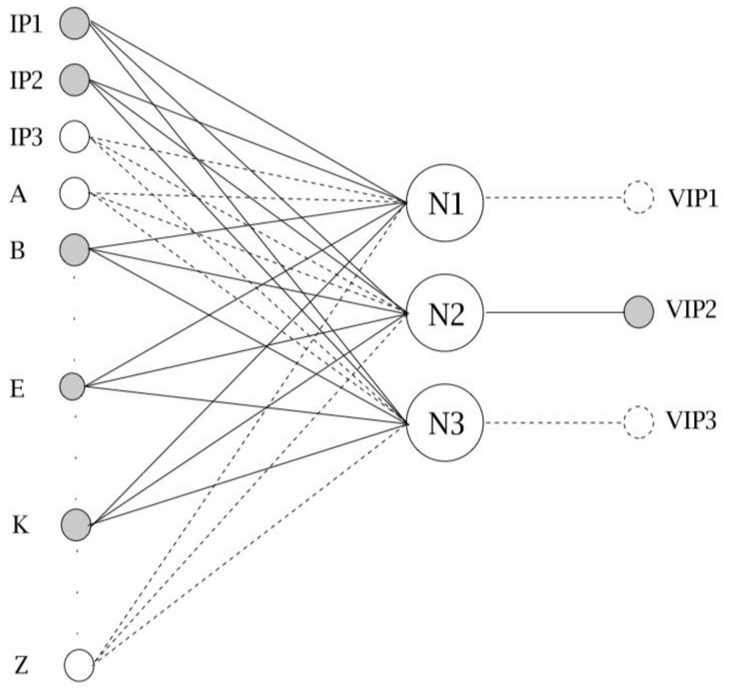


Fig. 3. Neural network analysing the Polish word "bakterie"

The training set includes all succeeding words' forms from the full dictionary

Quality = analysis of all word's forms in the dictionary

Analysis is correct if this word is converted to the valid base form

Quality measure is a ratio between a number of correctly inflected words and a number of all analysed forms

Results

- For the Polish language only 77% of all available forms (2 500 000) are inflected correctly, with the decision tree used
- Usage of a dictionary of the base forms (100 000 words) – quality of 99%
- OOV – 93.3%
- Widely used

Thank you