Fully Delexicalized Contexts for Syntax-Based Word Embeddings

Jenna Kanerva[1], Sampo Pyysalo[2] and Filip Ginter[1]

[1]Dept of IT - University of Turku, Finland
[2]Lang. Tech. Lab - University of Cambridge

turkunlp.github.io

Abstract

- We propose fully delexicalized contexts derived from syntactic trees to train word embeddings

- We demonstrate and evaluate our embeddings compared to vanilla word2vec
  - Nearest neighbours
  - Correlation to human judgement
  - Dependency parsing

Outline

- Related work
- Our word embedding contexts
- Motivation
- Experiments and evaluation

Related work

- **Word2vec** Mikolov et al. (2013)
- **Word2vecf** Levy and Goldberg (2014)
  - Syntactic context

Vanilla word2vec (Mikolov et al. 2013)

Hope you enjoyed your weekend in Houston

Output layer: Hope, you, your, weekend

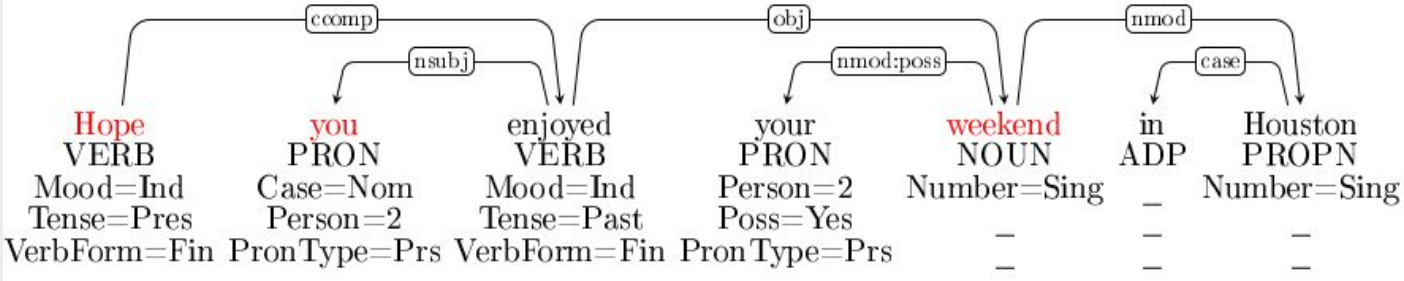
Hidden layer

Input layer: enjoyed

Hope you enjoyed your weekend in Houston

* Output layer size: vocabulary

Word2vecf (Levy & Goldberg 2014)

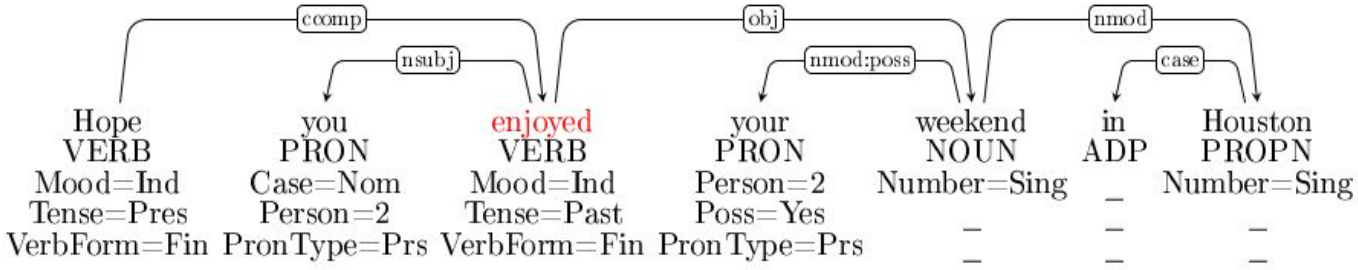


Output layer: Hope/ccomp^{gov}, you/nsubj, weekend/obj

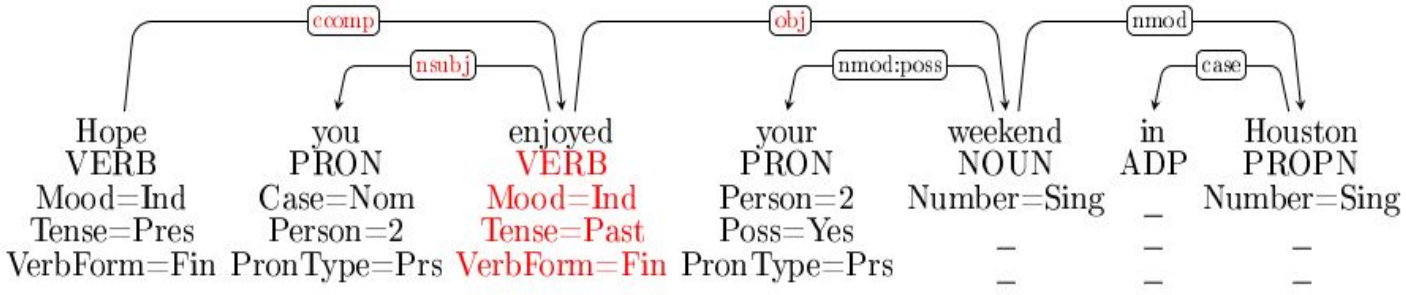
Hidden layer

Input layer: enjoyed

* Output layer size: ~vocabulary



Our method

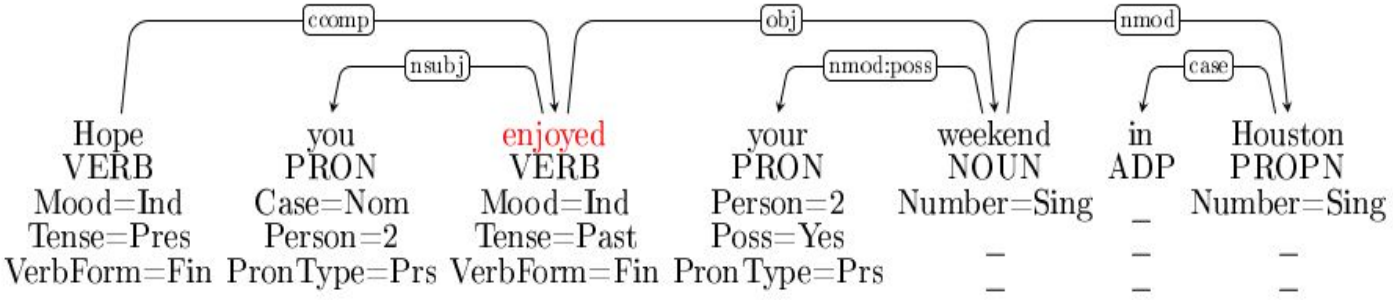


Output layer: ccomp^{gov}, nsubj, obj, VERB, Mood=Ind, Tense=Past, VerbForm=Fin

Hidden layer

Input layer: enjoyed

* Output layer size: pos tags + features + dependency types



Motivation

1) How much semantics can be learnt without the actual words?

2) Does task-specific training help?

3) Unified treebank annotations → Universal/multilingual word embeddings?

Experiments

- Training our word embeddings for 45 languages
- Inspecting nearest neighbours
- Comparing our embeddings to vanilla word2vec
  - Correlation with human judgement
  - Dependency parsing (closely related task)

Data

- Word embedding training: Automatically parsed raw text collection (Ginter et al. 2017)

- Parser training + evaluation: Universal Dependencies v2.0 treebanks (Nivre et al. 2017)

- Word similarity evaluation: evaluation service of 13 human judgement datasets (Faruqui and Dyer 2014)

Evaluation: Nearest neighbours

france	jesus	xbox	reddish	scratched	megabits
belgium	christ	playstation	brownish	knicked	megabit
luxembourg	jesus.	ps3	yellowish	bruised	kilobits
nantes	god	ps4	greenish	nicked	gigabits
marseille	ahnsahnghong	xbox360	pinkish	scuffed	mbps
bretagne	jesuschrist	wii	grayish	chewed	mbits
boulogne	y'shua	xbla	bluish	sandpapered	terabits
poitou	christ	psvita	-orange	scratches	mbit
rouen	christ.	titanfall	orangish	brusied	kbits
paris	jesus	xboxone	greyish	scraped	kilobit
toulouse	yeshua	gamecube	mid-brown	thwacked	megabytes

vanilla word2vec

france	jesus	xbox	reddish	scratched	megabits
lebanon	osama	vbox	greenish	snatched	megabytes
australia	napoleon	whitesox	grayish	touched	microseconds
england	ophelia	matchbox	bluish	punched	hectares
bolivia	gautama	firefox	greyish	deflected	tonnes
scotland	scipio	wmp	pinkish	warmed	microns
estonia	sauron	audiovox	yellowish	levelled	micrograms
switzerland	chandragupta	virtualbox	brownish	booted	litres
finland	claudius	equinox	blackish	stalked	megawatts
slovenia	jamarcus	rotax	temperate	ditched	gallons
algeria	olivia	hmp	redish	swallowed	bushels

Our delexicalized vectors

Evaluation: Nearest neighbours

france	jesus	xbox	reddish	scratched	megabits
belgium	christ	playstation	brownish	knicked	megabit
luxembourg	jesus.	ps3	yellowish	bruised	kilobits
nantes	god	ps4	greenish	nicked	gigabits
marseille	ahnsahnghong	xbox360	pinkish	scuffed	mbps
bretagne	jesuschrist	wii	grayish	chewed	mbits
boulogne	y'shua	xbla	bluish	sandpapered	terabits
poitou	christ	psvita	-orange	scratches	mbit
rouen	christ.	titanfall	orangish	brusied	kbits
paris	jesus	xboxone	greyish	scraped	kilobit
toulouse	yeshua	gamecube	mid-brown	thwacked	megabytes

vanilla word2vec

france	jesus	xbox	reddish	scratched	megabits
lebanon	osama	vbox	greenish	snatched	megabytes
australia	napoleon	whitesox	grayish	touched	microseconds
england	ophelia	matchbox	bluish	punched	hectares
bolivia	gautama	firefox	greyish	deflected	tonnes
scotland	scipio	wmp	pinkish	warmed	microns
estonia	sauron	audiovox	yellowish	levelled	micrograms
switzerland	chandragupta	virtualbox	brownish	booted	litres
finland	claudius	equinox	blackish	stalked	megawatts
slovenia	jamarcus	rotax	temperate	ditched	gallons
algeria	olivia	hmp	redish	swallowed	bushels

Our delexicalized vectors

Evaluation: human judgement

Dataset	Correlation word2vec	Correlation Our vectors	Pairs Found	Pairs Total	Reference
WordSim-353	0.7083	0.2350	353	353	Finkelstein et al. (2001)
WordSim-353-SIM	0.7677	0.4033	203	203	Agirre et al. (2009)
WordSim-353-REL	0.6691	0.1318	252	252	Agirre et al. (2009)
MC-30	0.7028	0.2929	30	30	Miller and Charles (1991)
RG-65	0.6801	0.0593	65	65	Rubenstein and Goodenough (1965)
Rare-Word	0.4250	0.1998	2006	2034	Luong et al. (2013)
MEN	0.7397	0.2027	3000	3000	Bruni et al. (2012)
MTurk-287	0.6958	0.3474	287	287	Radinsky et al. (2011)
MTurk-771	0.6406	0.1336	771	771	Halawi et al. (2012)
YP-130	0.3882	0.0464	130	130	Yang and Powers (2006)
SimLex-999	0.3376	0.1004	999	999	Hill et al. (2016)
Verb-143	0.3633	0.2425	144	143	Baker et al. (2014)
SimVerb-3500	0.2175	0.0476	3500	3500	Gerz et al. (2016)

- Our vectors not as good as word2vec but correlation is still positive

Evaluation: Dependency parsing

language	baseline	word2vec	diff to baseline	syntax-based	diff to baseline
Ancient_Greek	56.61	57.93	+1.32	58.18	+1.57
Ancient_Greek-PROIEL	72.35	72.48	+0.13	72.67	+0.32
Arabic	72.88	73.91	+1.03	74.00	+1.12
Basque	69.02	69.74	+0.72	69.93	+0.91
Bulgarian	83.90	84.29	+0.39	85.18	+1.28
Catalan	85.15	85.01	-0.14	85.31	+0.16
Chinese	68.48	68.83	+0.35	69.06	+0.58
Croatian	76.08	75.98	-0.10	77.35	+1.27
Czech-CAC	83.75	83.58	-0.17	84.54	+0.79
Czech-CLTT	69.58	68.92	-0.66	72.19	+2.61
Czech	84.47	84.24	-0.23	84.69	+0.22
Danish	75.18	74.63	-0.55	74.99	-0.19
Dutch-LassySmall	75.67	75.01	-0.66	76.68	+1.01
Dutch	74.73	75.21	+0.48	75.00	+0.27
English	79.66	80.20	+0.54	80.64	+0.98
English-LinES	74.62	74.35	-0.27	75.59	+0.97
English-ParTUT	75.72	75.21	-0.51	76.20	+0.48
Estonian	60.65	61.89	+1.24	63.22	+2.57
Finnish	75.70	75.79	+0.09	77.35	+1.65
Finnish-FTB	76.42	76.68	+0.26	77.72	+1.30
French	86.08	85.71	-0.37	86.53	+0.45
French-Sequoia	82.30	82.58	+0.28	82.65	+0.35
Galician	77.58	77.34	-0.24	78.21	+0.63
German	73.10	73.12	+0.02	72.87	-0.23
Greek	79.04	77.93	-1.11	79.93	+0.89
Hebrew	76.88	77.38	+0.50	78.52	+1.64
Hindi	87.09	86.82	-0.27	87.38	+0.29
Hungarian	65.59	66.40	+0.81	68.44	+2.85
Indonesian	74.39	72.84	-1.55	73.59	-0.80
Italian	85.44	84.98	-0.46	84.96	-0.48
Italian-ParTUT	78.21	78.74	+0.53	79.92	+1.71
Japanese	93.09	93.09	+0.00	93.23	+0.14
Korean	56.42	62.70	+6.28	63.72	+7.30
Latin-ITTB	71.15	71.72	+0.57	72.98	+1.83
Latin-PROIEL	70.08	69.76	-0.32	69.89	-0.19
Latvian	64.01	64.56	+0.55	66.16	+2.15
Norwegian-Bokmaal	83.91	83.44	-0.47	84.18	+0.27
Norwegian-Nynorsk	82.32	81.65	-0.67	81.89	-0.43
Old_Church_Slavonic	73.56	71.22	-2.34	71.40	-2.16
Persian	80.38	79.56	-0.82	80.86	+0.48
Polish	79.42	80.62	+1.20	81.21	+1.79
Portuguese-BR	85.55	86.11	+0.56	86.26	+0.71
Portuguese	83.64	84.49	+0.85	84.93	+1.29
Romanian	79.82	79.77	-0.05	80.30	+0.48
Russian	75.41	76.00	+0.59	77.48	+2.07
Russian-SynTagRus	86.76	86.58	-0.18	87.71	+0.95
Slovak	75.39	75.65	+0.26	76.55	+1.16
Slovenian	80.62	80.87	+0.25	81.38	+0.76
Spanish-AnCora	84.17	84.55	+0.38	84.31	+0.14
Spanish	84.34	83.85	-0.49	84.11	-0.23
Swedish-LinES	74.35	74.72	+0.37	75.34	+0.99
Swedish	73.39	74.25	+0.86	74.75	+1.36
Turkish	56.00	56.24	+0.24	57.75	+1.75
Urdu	76.98	76.23	-0.75	76.26	-0.72
Vietnamese	55.85	56.26	+0.41	55.22	-0.63
Average	-	-	+0.16	-	+0.88

UDPipe parser with three different pre-trained word embeddings

- Baseline: word2vec trained on treebank data
- word2vec trained on raw text collection
- Our trained on automatically analysed raw text collection

Evaluation: Dependency parsing

language	baseline	word2vec	diff to baseline	syntax-based	diff to baseline
Ancient_Greek	56.61	57.93	+1.32	58.18	+1.57
Ancient_Greek-PROIEL	72.35	72.48	+0.13	72.67	+0.32
Arabic	72.88	73.91	+1.03	74.00	+1.12
Basque	69.02	69.74	+0.72	69.93	+0.91
Bulgarian	83.90	84.29	+0.39	85.18	+1.28
Catalan	85.15	85.01	-0.14	85.31	+0.16
Chinese	68.48	68.83	+0.35	69.06	+0.58
Croatian	76.08	75.98	-0.10	77.35	+1.27
Czech-CAC	83.75	83.58	-0.17	84.54	+0.79
Czech-CLTT	69.58	68.92	-0.66	72.19	+2.61
Czech	84.47	84.24	-0.23	84.69	+0.22
Danish	75.18	74.63	-0.55	74.99	-0.19
Dutch-LassySmall	75.67	75.01	-0.66	76.68	+1.01
Dutch	74.73	75.21	+0.48	75.00	+0.27
English	79.66	80.20	+0.54	80.64	+0.98
English-LinES	74.62	74.35	-0.27	75.59	+0.97
English-ParTUT	75.72	75.21	-0.51	76.20	+0.48
Estonian	60.65	61.89	+1.24	63.22	+2.57
Finnish	75.70	75.79	+0.09	77.35	+1.65
Finnish-FTB	76.42	76.68	+0.26	77.72	+1.30
French	86.08	85.71	-0.37	86.53	+0.45
French-Sequoia	82.30	82.58	+0.28	82.65	+0.35
Galician	77.58	77.34	-0.24	78.21	+0.63
German	73.10	73.12	+0.02	72.87	-0.23
Greek	79.04	77.93	-1.11	79.93	+0.89
Hebrew	76.88	77.38	+0.50	78.52	+1.64
Hindi	87.09	86.82	-0.27	87.38	+0.29
Hungarian	65.59	66.40	+0.81	68.44	+2.85
Indonesian	74.39	72.84	-1.55	73.59	-0.80
Italian	85.44	84.98	-0.46	84.96	-0.48
Italian-ParTUT	78.21	78.74	+0.53	79.92	+1.71
Japanese	93.09	93.09	+0.00	93.23	+0.14
Korean	56.42	62.70	+6.28	63.72	+7.30
Latin-ITTB	71.15	71.72	+0.57	72.98	+1.83
Latin-PROIEL	70.08	69.76	-0.32	69.89	-0.19
Latvian	64.01	64.56	+0.55	66.16	+2.15
Norwegian-Bokmaal	83.91	83.44	-0.47	84.18	+0.27
Norwegian-Nynorsk	82.32	81.65	-0.67	81.89	-0.43
Old_Church_Slavonic	73.56	71.22	-2.34	71.40	-2.16
Persian	80.38	79.56	-0.82	80.86	+0.48
Polish	79.42	80.62	+1.20	81.21	+1.79
Portuguese-BR	85.55	86.11	+0.56	86.26	+0.71
Portuguese	83.64	84.49	+0.85	84.93	+1.29
Romanian	79.82	79.77	-0.05	80.30	+0.48
Russian	75.41	76.00	+0.59	77.48	+2.07
Russian-SynTagRus	86.76	86.58	-0.18	87.71	+0.95
Slovak	75.39	75.65	+0.26	76.55	+1.16
Slovenian	80.62	80.87	+0.25	81.38	+0.76
Spanish-AnCora	84.17	84.55	+0.38	84.31	+0.14
Spanish	84.34	83.85	-0.49	84.11	-0.23
Swedish-LinES	74.35	74.72	+0.37	75.34	+0.99
Swedish	73.39	74.25	+0.86	74.75	+1.36
Turkish	56.00	56.24	+0.24	57.75	+1.75
Urdu	76.98	76.23	-0.75	76.26	-0.72
Vietnamese	55.85	56.26	+0.41	55.22	-0.63
Average	-	-	+0.16	-	+0.88

- Green if our is better than word2vec, and difference to baseline is positive

Evaluation: Dependency parsing

- word2vec +0.16 better than baseline on average
  - diff to baseline between -1.55% and +6.28%
  - 31 treebanks positive, and 23 negative

- Our embeddings +0.88 better than baseline
  - diff to baseline between -0.80% and +7.30%
  - 45 treebanks positive, and 9 negative

Evaluation: Dependency parsing

- pre-trained embeddings does not automatically increase parsing performance across languages
- delexicalized syntactic embeddings lead to higher performance as well as generalize better across languages when evaluated in closely related task

Parser accuracy vs. quality of the embeddings

- Our word embeddings are trained on automatically parsed data
  - How does the baseline parser accuracy affect the quality of the word embeddings?

- Bootstrapping:
  - Baseline parser → parse raw text → embeddings → better parser → parse raw text → new embeddings → even better parser?

Bootstrapping on Finnish

	Baseline	Iteration 1	Iteration 2
Finnish	75.70	77.35	77.57

- Small improvement with the second iteration model

→ UDPipe not optimal parser for this study as POS tags and morphological features are not revised

Bootstrapping vol. 2

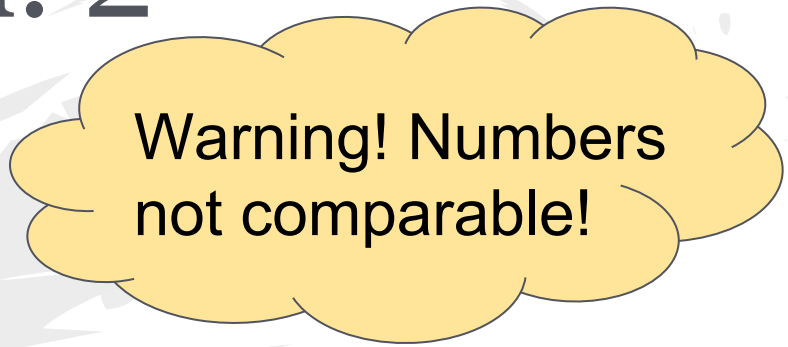
- Baseline UDPipe not competitive with state-of-the-art on Finnish
  - 75.7% compared to 83-84%

- What if we use raw data parsed with (near) state-of-the-art parser?

Bootstrapping vol. 2

- Raw data: Finnish Internet Parsebank
  - ~3.6 billion token collection of web crawled data
- Finnish-dep-parser
  - Omorfi rule-based morphological analyzer
  - Marmot tagger
  - Mate-tools graph-based dependency parser
  - UD v1.2
  - LAS estimated to be ~82%

Bootstrapping vol. 2



- Numbers are not comparable to our main result table!
  - Different version of UD (v1.2 compared to v2.0)
  - Raw text collection more than three times bigger

Bootstrapping vol. 2

- UDPipe + our embeddings trained on Finnish Internet Parsebank: 82.21%
- UDPipe + word2vec embeddings trained on Finnish Internet Parsebank: 78.35%
- UDPipe baseline: ~76.5%

Conclusions

- Fully delexicalized context for word embedding training
- Bit surprisingly, these embeddings are able to capture also semantic aspects
- Improve parsing accuracy and generalize better than standard word2vec embeddings

Thanks!

turkunlp.github.io

Academy of Finland
Kone Foundation
University of Turku Graduate School