An Exploration of Embeddings for Generalized Phrases

Wenpeng Yin & Hinrich Schutze

...

Prachi, 12485
Hrishikesh, 14111265

CSE IIT Kanpur

Contents

Generalized Phrases

Generalized Phrases include

- **Skip-bigrams** (SkipBs)
  For example, skip-bigrams at a distance 2 in the sentence "This tea helped me to relax" are:
  "This*helped", "tea*me", "helped*to"...
- **Continuous and non-continuous linguistic phrases**
  For example, "cold_cuts" and "White_House" are continuous phrases and "take_over" and "turn_off" are non-continuous.

Motivation

- A particular task involving a word can be solved based only on *context of word*.
- Generalized phrases can be used to infer the attributes of the context they enclose.
  For example: He helped `Xiulan` to find a flat.
- They can capture non-compositional semantics.
  For example: "keep up", "keep on", "keep from" etc.
- Embeddings of generalized phrases are better suited than word embeddings for a *coreference resolution* and *paraphrase identification*.

Embedding Learning for SkipBs

- Used `word2vec` on English Gigaword corpus.
- The corpus is represented as a sequence of sentences, each consisting of two tokens: a SkipB and a word that occurs between the two enclosing words of the SkipB.
- The distance between the two enclosing words can be $k = 2$ or $2 \leq k \leq 3$.
  - when $k = 2$, the trigram $w_{i-1}w_iw_{i+1}$ generates the single sentence "$w_{i-1} * w_{i+1} * w_i$";
  - when $2 \leq k \leq 3$, the fourgram $w_{i-2}w_{i-1}w_iw_{i+1}$ generates four sentences "$w_{i-2} * w_i * w_{i-1}$", "$w_{i-1} * w_{i+1} * w_i$", "$w_{i-2} * w_{i+1} * w_{i-1}$" and "$w_{i-2} * w_{i+1} * w_i$".

Phrase Collection

- Extracted two-word phrases defined in Wiktionary and two-word phrases defined in Wordnet.
- A collection of continuous and noncontinuous phrases of size 95218 is formed.

Identification of Phrase Continuity

- For each phrase "A_B", compute $[c_1, c_2, c_3, c_4, c_5]$ where c_i , $1 \leq i \leq 5$, indicates there are c_i occurrences of A and B in that order with a distance of i.

- If c_1 is 10 times higher than $(c_2 + c_3 + c_4 + c_5)/4$, classify "A_B" as continuous, otherwise as discontinuous.
  For example,
  "pick_off": $[1121, 632, 337, 348, 4052]$
  "Cornell_University": $[14831, 16, 177, 331, 3471]$

Sentence Reformatting

- Sentence "...A...B..." is reformated into "...A_B...A_B..." if "A_B" is a discontinuous phrase and is separated by maximal 4 words.
- Sentence "...AB..." into "...A_B..." if "A_B" is a continuous phrase.
- `word2vec` is run on the reformatted corpus.

Examples of Phrase Neighbors

turn_off	**caught_up**	**take_over**	**macular_degeneration**	**telephone_interview**
switch_off	mixed_up	take_charge	eye_disease	statement
unplug	entangled	replace	diabetic_retinopathy	interview
turning_off	involved	take_control	cataracts	conference_call
shut_off	enmeshed	stay_on	periodontal_disease	teleconference
block_out	tangled	retire	epilepsy	telephone_call
turned_off	mired	succeed	glaucoma	told
fiddle_with	engaged	step_down	skin_cancer	said

Table 1: Phrases and their nearest neighbors

Animacy classification for markables

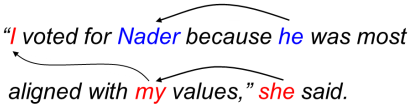


Figure : Example of markables

- A *markable* in coreference resolution refers to an entity in the real world or another linguistic expression.
- Classifying markables as animate/inanimate is useful for coreference resolution systems.
- **animate chains**: an animate pronoun markable and no inanimate pronoun markable
- **inanimate chains**: an inanimate pronoun markable and no animate pronoun markable

representation		accuracy
phrase embedding	$k = 2$	0.703
	$2 \leq k \leq 3$	0.700
word embedding	word2vec	$0.668^{*\dagger}$
	Collobert et al.	$0.662^{*\dagger}$
one-hot vectors		$0.638^{*\dagger}$

Table 2: Classification accuracy. Mark "*" means significantly lower than "phrase embedding", $k = 2$; "†" means significantly lower than "phrase embedding", $2 \leq k \leq 3$. As significance test, we use the test of equal proportion, $p < .05$, throughout.

Frequent Errors

- *Unspecific SkipBs*

  For example, "take*in" and "then*goes"

- *Untypical use of specific SkipBs*

  For example, "...the southeastern area <u>of</u> Fujian <u>whose</u> economy is the most active"

Examples of SkipB Neighbors

who*afghanistan,	**some*told**	**women*have**	**with*responsibility**	**he*worried**
had*afghanistan	other*told	men*have	of*responsibility	she*worried
he*afghanistan	two*told	children*have	and*responsibility	was*worried
who*iraq	–*told	girls*have	“*responsibility	is*worried
have*afghanistan	but*told	parents*have	that*responsibility	said*worried
fighters*afghanistan	one*told	students*have	’s*responsibility	that*worried
who*kosovo	because*told	young*have	the* responsibility	they*worried
was*afghanistan	and*told	people*have	for*responsibility	’s*worried

Table 3: SkipBs and their nearest neighbors

Paraphrase Identification Task

- Standard approaches are unlikely to assign a high similarity score to the two sentences "he started the machine" and "he turned the machine on".
- A sentence like "...A_B...A_B..." is considered as "A_B".

Methods	Accuracy	F1
baseline	0.684	0.803
word embedding	0.695	0.805
phrase embedding	**0.713**	**0.812**

Table 4: Paraphrase task results.

Comparison of Word and Phrase Embeddings

GWP	sentence 1	sentence 2
1 0 1	Common **side_effects** include **nasal_congestion**, **runny_nose**, **sore_throat** and cough, the FDA said .	The most common **side_effects** after getting the nasal spray were **nasal_congestion**, **runny_nose**, **sore_throat** and cough .
1 0 1	Douglas Robinson, a senior **vice_president** of finance, will **take_over** as chief financial officer on an interim basis .	Douglas Robinson, CA senior **vice_president**, finance, will fill the position in the interim .
1 1 0	They were being held Sunday in the Camden County Jail on $ 100,000 bail each .	The Jacksons remained **in_on** Camden County jail $ 100,000 bail .
0 0 1	The **interest_rate** sensitive two year Schatz yield was down 5.8 **basis_points** at 1.99 percent .	The Swedish **central_bank** cut **interest_rates** by 50 **basis_points** to 3.0 percent .

Table 5: Four typical sentence pairs in which the predictions of word embedding system and phrase embedding system differ. G = gold annotation, W = prediction of word embedding system, P = prediction of phrase embedding system. The formatting used by the system is shown. The original word order of sentence 2 of the third pair is "··· in Camden County jail on $ 100,000 bail".

Summary

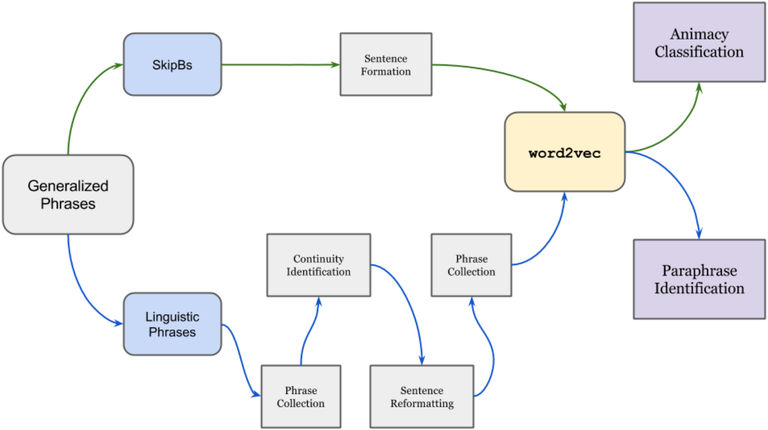


Figure : Generalized Phrases for Linguistic Tasks

Future work

- *continuous* phrases determined purely statistically, and *discontinous phrases* by dictionaries.
  - combination of two methods desirable
- to distinguish between phrases that only occur in continuous form and phrases that must or can occur discontinuously
- given a sentence containing the parts of a discontinuous phrase in correct order, how to determine that co-occurrence of the two parts constitutes an instance of the discontinuous phrase?
- which tasks benefit most significantly from the introduction of generalized phrases?

References I

W. Yin and H. Schutze, "An exploration of embeddings for generalized phrases," *ACL 2014*, 2014.

A. Neelakantan, B. Roth, and A. McCallum, "Compositional vector space models for knowledge base completion," 2015.

R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," 2013.

W. Yin and H. Schutze, "Multichannel variable-size convolution for sentence classification," *CoNLL 2015*, 2015.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NIPS 2013*, 2013.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," 2008.

End

Thank You!

Appendix

- **Appendix I**:
  - about LIBLINEAR
  - more about `word2vec`
  - `wang2vec`: improvements to `word2vec`
  - concept of *compositional vectors*.
- **Appendix II**:
  - overview of recent work by same authors
  - related work: (Socher et al., 2013)

LIBLINEAR

Fan et al. (2008)

- library large-scale linear classification.
- Homepage: http://www.csie.ntu.edu.tw/ cjlin/liblinear/
- supports logistic regression and linear support vector machines
- available in MATLAB, Octave, Java, Python, Ruby, Perl, Weka, R, Common LISP, Scilab

word2vec: Word Representations in Vector Space

(Mikolov et al., 2013b)

- **Code**: https://code.google.com/p/word2vec/
- an efficient implementation of the *continuous bag-of-words* and *skip-gram* architectures for computing vector representations
- word vectors can be successfully applied to automatic extension of facts in Knowledge Bases, and also for verification of correctness of existing facts.

word2vec Models

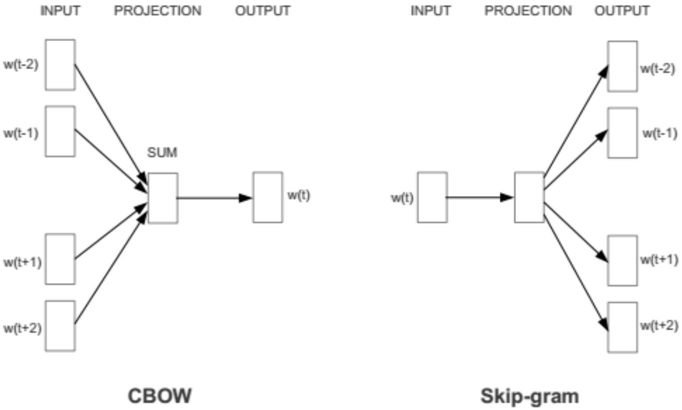


Figure : **CBOW** architecture predicts current word based on context, and **Skip-gram** predicts surrounding words given current word

wang2vec: Adaptations to `word2vec`

- **Code**: https://github.com/wlin12/wang2vec
- **structured skip-gram**: improved version of skip-gram
- **continuous window**: improved version of CBOW
- lead to improvements when used in state-of-the-art neural network systems for **part-of-speech tagging** and **dependency parsing**, relative to the original models

wang2vec Models

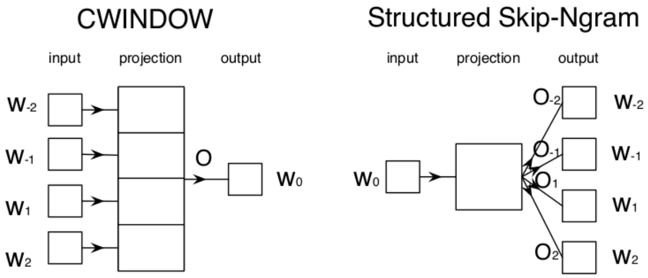
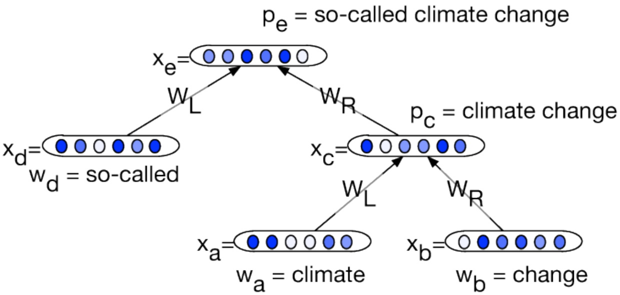


Figure : Illustration of the **Structured Skip-gram** and **Continuous Window (CWindow)** models

Compositional Vectors



Used in:

- Knowledge Base completion (Neelakantan et al., 2015)
- Parsing, in conjunction with RNN (Socher et al., 2013)

Multichannel Variable-Size Convolution for Sentence Classification

Yin and Schutze (2015)

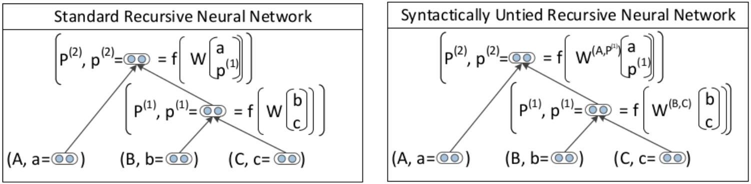
- **MVCNN**, a convolution neural network (CNN) architecture for sentence classification
    i. combines diverse versions of pretrained word embeddings
    ii. extracts features of multi-granular phrases with variable-size convolution filters.
- pretraining MVCNN is critical for good performance
- **MVCNN** achieves state-of-the-art performance on four tasks: on small-scale binary, small-scale multi-class and large-scale Twitter sentiment prediction and on subjectivity classification.

Parsing with Compositional Vector Grammars

Socher, Bauer, Manning, and Ng (2013)

- parsing model that combines the *speed* of **small-state PCFGs** with *semantic richness* of **neural word representations** and **compositional phrase vectors** item compositional vectors are learned with a new syntactically untied *recursive neural network (RNN)*
- linguistically more plausible since it chooses different composition functions for a parent node based the syntactic categories of its children

RNNs vs. SU-RNNs



(a) Tree with a **simple RNN**: same weight matrix is replicated and used to compute all non-terminal node representations. Leaf nodes are n-dimensional vector representations of words

(b) A **syntactically untied RNN** in which the function to compute a parent vector depends on syntactic categories of its children which are assumed to be given.

Figure : Comparison of RNNs and SU-RNNs