Probabilistic Modeling of Joint-context in Distributional Similarity

Oren Melamud, Ido Dagan, Jacob Goldberger — Bar-Ilan University

Idan Szpektor — Yahoo! Research

Deniz Yuret — Koç University

Context modeling in distributional similarity

Most distributional similarity models are based on simplified context representations, ignoring:

- Multiple context words co-occur with the target word at the same time
- The order of the context words

Mary's | son | **likes** | the | school | campus

Joint-context

Simplified representation ↓ Information is lost

Independent context features

likes	likes	likes	likes	likes
campus	Mary's	school	son	the

Traditional vector-space models

'Likes' and 'surrounds' are quite different semantically, but may share many context features.

Is this context representation discriminative enough?

Mary's son **likes** the school campus

features
Mary
son
school
campus

A forest **surrounds** the school campus

features
forest
school
campus

Initial attempts to leverage joint-contexts

The entire context around a target word is considered as a single feature.

Very high-dimensional sparse feature space.

Mary's son **likes** the school campus

features
Mary's son ___ the school campus

John's son **loves** the school campus

features
John's son ___ the school campus

More discriminative

Very sparse (requires huge corpora)

Probabilistic Distributional Similarity (PDS)

PDS is designed to model joint-contexts

Not a vector-space model

A probabilistic approach

Leverages the power of Language Models

Language Model "Plug-in"

Probabilistic Distributional Similarity (PDS) Scheme

"words are similar in meaning if they occur in similar contexts"

↓

"words are similar in meaning if they are **likely** to occur in the same contexts"

↓

"words *v* and *w* are similar in meaning if *w* is **likely** in the contexts of *v* and vice versa"

Model

How likely is target word *w* in the contexts of target word *v* ?

$$p(w|v) = \sum_{c} p(c|v) \cdot p(w|c) = E_{p(c|v)}[p(w|c)]$$

$$p(loves|likes) = \sum_{c} p(c|likes) \cdot p(loves|c) = E_{p(c|likes)}[p(loves|c)]$$

$p(c|likes)$ — $p(\text{"Mary's son \_\_ the school campus"} \mid likes)$

$p(loves|c)$ — $p(loves \mid \text{"Mary's son \_\_ the school campus"})$

Efficient non-sparse estimation for *p(w|v)*

$$p_\theta(w|v) = \frac{1}{|C(v)|} \sum_{c \in C(v)} p_\theta(w|c)$$

* $C(v)$ is a collection of contexts observed for v

Final proposed measure:

$$PDS_\theta(v,w) = \sqrt{p_\theta(w|v) \cdot p_\theta(v|w)}$$

Leveraging the power of language models:

$$p_{LM}(w|c) = \frac{p_{LM}(w,c)}{p_{LM}(*,c)}$$

Evaluation

Main evaluation setting:

Learning corpus: 100M words from Reuters RCV1

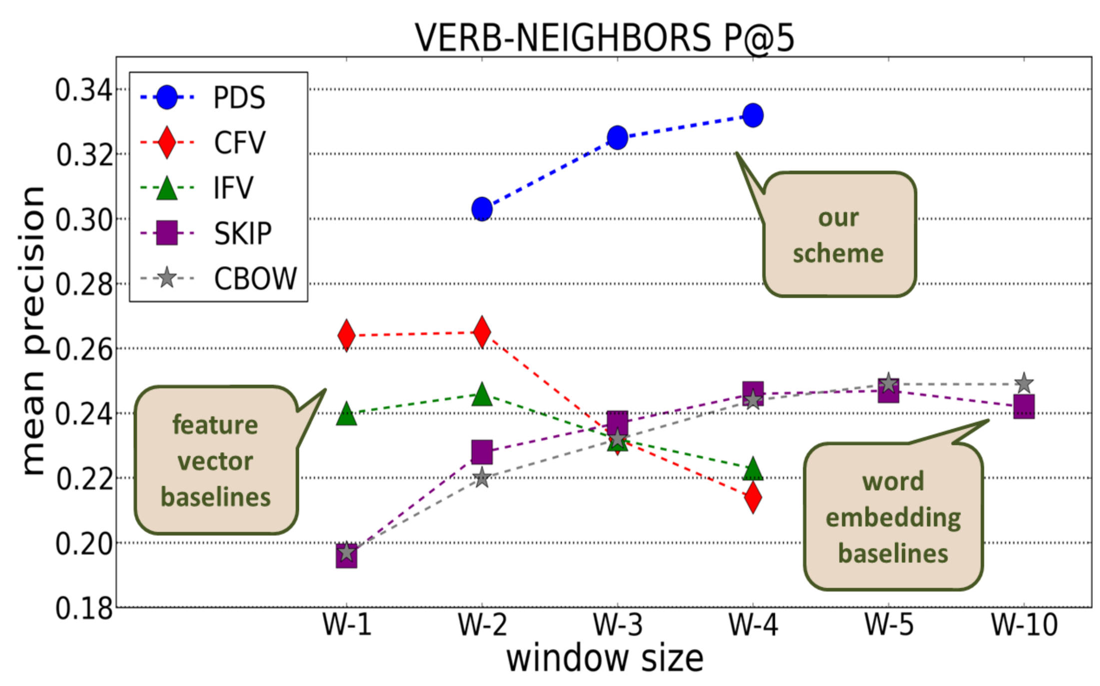
Gold standard: WordNet synonyms and semantic neighbors

Compared models:

	Model	Params
IFV $^{w-k}$	Independent Feature Vector	PPMI Vector cosine
CFV $^{w-k}$	Composite Feature Vector	
SKIP $^{w-k}$	Skip-gram word embeddings	Neg sampling 15 Dimensions 600 Vector cosine
CBOW $^{w-k}$	CBOW word embeddings	
PDS $^{w-k}$	PDS + Kneser-Ney *n*-gram LM	*n = k + 1*

3-gram

W-2 Mary's son **loves** the school campus



VerbSim ranking spearman correlations	
PDS W-4	0.616
CBOW W-5	0.528
CFV W-2	0.477
SKIP W-4	0.469
IFV W-2	0.467