Dependency-based modeling of joint-context in Distributional Similarity

Emmanuele Chersoni

Aix-Marseille Université

University of Pisa

4th VARIAMU Workshop, Hong Kong PolyU

6-7th January 2016

.

.

Distributional Semantic Models

- Distributional Semantic Models (DSM) follow the Distributional Hypothesis: *words occurring in similar contexts tend to have similar meanings*
- Target words represented as high-dimensional vectors, whose dimensions are (a function of) target frequencies in context
- Choose a word window around the target, break into a set of individual words and register the co-occurrence counts. Example[1]:

  Mary's son likes the school campus

  The forest surrounds the school campus

Target	son-n	Mary-n	school-n	campus-n	forest-n
like-v	1	1	1	1	0
surround-v	0	0	1	1	1

[1] The examples are taken from (Melamud *et al.*, 2014)

Distributional Semantic Models

.

- Dimensions capture the association between the target and the contexts, indipendently of one another
- Missing information about inter-relations of words
- Models dealing with this limitation: (Ruiz-Casado *et al*: 2005), (Agirre *et al.*: 2009) and (Melamud *et al.*: 2014)
- Joint contexts = entire windows as contextual features: *e.g to work* represented by features like *I* ____ *for a major tech company*, *Obama* ____ *as a civil rights attorney*.

Distributional Semantic Models

Target	son-n	Mary-n	school-n	campus-n	forest-n
like-v	1	1	1	1	0
surround-v	0	0	1	1	1

.

- In traditional DSM, targets may share individual features (recall the *like* and *surround* examples)...
- … but individual features may not be sufficient to reflect semantic differences between words
- On the other hand, joint contexts alone may not provide sufficient evidence for measuring similarities → data sparseness

The Probabilistic Distributional Similarity model (Melamud *et al.*, 2014)

.

- Based on a corollary of the DH: *words are similar in meaning if they are likely to occur in the same contexts*
- Given two words *a* and *b*, assign them a high score if *a* is likely to occur in the contexts of *b* and vice versa
- *Sim* (*a*, *b*) is a function of p(*a* | contexts of *b*) and of p (*b* | contexts of *a*)

The Probabilistic Distributional Similarity model (Melamud *et al.*, 2014)

.

- PDS makes use of joint contexts: p (*score* | The Arsenal forward _____ a goal) is the probability of *to score* to fill the placeholder
- Given *a* = *like*, *b* = *love* and the collections of contexts C_{like} and C_{love}, compute the probability of *like* in C_{love} = {Mary's son ____ the school campus, Micheal ____ to play football...} and vice versa

The Probabilistic Distributional Similarity model (Melamud *et al.*, 2014)

.

Formula for computing *p* (*b* | *a*):

$$p(b|a) = \frac{1}{Ca} * \sum_{c \epsilon Ca} p(b|c)$$

Similarity measure for the target word types:

$$sim(a,b) = \sqrt{p(b|a) * p(a|b)}$$

The Probabilistic Distributional Similarity model (Melamud *et al.*, 2014)

.

- Word windows of order *k* around a target word, not crossing sentence boundaries
- probabilities computed with the Kneser-Ney language model (Kneser, Ney: 1995)
- p (*b* | *c*) = p (*b*, *c*) / p (*c*)

  p (*b*, *c*) = probability of *c* with *b* filling the placeholder

  p(*c*) = probability of p(*, *c*)
- PDS works very well for verbs and is close to the state of the art for nouns

A dependency-based model of Distributional Similarity

- PDS includes in the joint contexts all words within the word window = even words with no relation with the target are included in the semantic representation
- A syntactic notion of joint context: given a target in a context, build joint contexts by extracting its dependencies from a parsed corpus

Often	often	RB	1	4	ADV
infected	infected	JJ	2	3	NMOD
people	people	NNS	3	4	SBJ
are	be	VBP	4	0	ROOT
rejected	reject	VVN	5	4	VC
by	by	IN	6	5	LGS
family	family	NN	7	6	PMOD
.	.	SENT	8	4	P

A dependency-based model of Distributional Similarity: motivation

- Verbs and arguments arranged into a web of mutual expectations (McRae *et al.:* 1998; Hare *et al.,* 2009)
- Knowledge of typical participants of events reflected in knowledge of typical arguments combinations (=joint contexts)
- Common practice in DSMs to extract typical fillers of verb argument positions as static lists... but the occurrence of one or more arguments of a joint context influences the expectations for the others

$C_{drink+subj}$ (student-n, baby-n, worker-n...) $C_{drink+obj}$ (water-n, wine-n, milk-n...)

E.g. if we start hearing a sentence like *the baby drinks, milk* is a very predictable obj; if subj = *the student*, then *wine* or other alcoholic drinks would be more predictable

A dependency-based model of Distributional Similarity: experimental settings

- Corpus: BNC (Burnage, Dunlop: 1992), parsing with Malt Parser (Nivre *et al.*: 2005)
- Focus on verbs
- Extraction of a list of verb-argument dependencies

  pred = “present-v” sbj = “she-p” obj = “report-n”

  pred = “oversee-v” sbj = “Kate-n” obj = “course-n”
- Personal pronouns and proper names mapped onto single tags:

  Kate-n → ProperName she-p → PersPro

A dependency-based model of Distributional Similarity: experimental settings

For each verb occurrence, generate the following contexts:

pred = “acknowledge-v” subj = “ProperName-n” obj=”failure-n”

- a single context for each of the dependencies

  acknowledge-v ProperName-n+subj 1

  acknowledge-v failure-n+obj 1

- a joint context, obtained by joining all the dependencies

  acknowledge-v ProperName-n+subj_failure-n+obj 1

- combined single dependencies and joint contexts:

  hope to add richer contextual information and to limit sparseness issues

A dependency-based model of Distributional Similarity: experimental settings

- Each verb associated to a dictionary of contexts (with their frequencies):

  $C_{accelerate}$ = {growth-n+obj 4, activity-n+obj 10, car-n+subj 23, agreement-n+subj_development-n+obj 3 …}

- Use the dictionaries to compute verb similarities:

$$p(b|a) = \frac{1}{Ca} * \sum_{c \epsilon Ca} p(b|c) * w(a,c)$$

w(*a, c*) is a measure of the association between the context *c* and the word *a*

A dependency-based model of Distributional Similarity: experimental settings

$$p(b|c) = \frac{p(b,c)}{p(c)}$$

- p(b, c) = frequency of context c with b satisfying the dependency constraints in c, divided by the total number of verb-argument extractions
- Example: p (*accelerate*, agreement-n+subj_development-n+obj) is the frequency of *accelerate* with agreement as subj and development as obj, divided by the total number of extractions
- p(c) = frequency of context c divided by the total number of extractions
- Word similarity computed as

$$sim(a,b) = \sqrt{p(b|a) * p(a|b)}$$

A dependency-based model of Distributional Similarity: experimental settings

- In (Melamud *et al.*: 2014), two types of evaluations: synonym retrieval in a Wordnet-derived dataset, and score comparison with the VerbSim gold standards (Yang, Powers: 2006)
- VerbSim dataset: 130 verb pairs annotated with an average of 6 human judgements of semantic similarity

brag-v	boast-v	4.0
merit-v	deserve-v	3.667
refer-v	explain-v	1.833

- Extracted a subset of 112 verbs with more than 100 occurrences in the training corpus and measured Spearman correlation

A dependency-based model of Distributional Similarity: experimental settings

- We compare two different models: a traditional vector space model (DS: cosine as a measure of similarity) and PDS
- Experiments with and without joint contexts to test their impact on performance

Melamud et al,, 2014	**Our model**
Joint context as a word window of order *k*	Joint context as a list of target dependencies (sbj, obj, iobj)
No context weighting	Context weighting (PPMI, LMI)
Kneser-Ney smoothing to compute Jcs probabilities	No smoothing
Random sampling of 10000 contexts for each word	No context sampling

Results

Model	Corpus	Joint contexts	Verb pairs	Spearman
DS + PPMI	BNC	No	112	0.484
PDS + PPMI	BNC	No	112	0.540
DS + PPMI	BNC	Yes	112	**0.606**
PDS + PPMI	BNC	Yes	112	0.585

. Melamud *et al.*: 2014

Model	Corpus	Window order	Verb pairs	Spearman
PDS	RCV vol. 1	4	107	**0.616**
CFV	RCV vol. 1	2	107	0.477
IFV	RCV vol. 1	2	107	0.467
Skip Gram	RCV vol. 1	4	107	0.469
CBOW	RCV vol. 1	5	107	0.528

A dependency-based model of Distributional Similarity: conclusions

- A syntax-aware notion of joint context improves the performance
- Cosine similarity-based models seem to benefit more from joint contexts
- Next steps: VerbSim evaluation with the same setting of (Melamud *et al.*: 2014), WordNet evaluation
- More experiments with smoothing and contexts sampling

Bibliography

.

- E. Agirre *et al.*, *A study on similarity and relatedness using distributional and wordnet-based approaches*, in *Proceedings of the NAACL Conference*, 2009;
- P. Blache, *Chunks et activation: un modèle de facilitation du traitement linguistique,* in *Proceedings of TALN*, 2013;
- G. Burnage, D. Dunlop, *Encoding the British National Corpus,* in *English Language Corpora: Design, Analysis and Exploitation,* 1992;
- M. Hare *et al.*, *Activating event knowledge*, in *Cognition,* vol. 111, no. 2, 2009, pp. 151-167;
- Z. Harris, *Distributional structure*, in *Word*, vol. 10, n. 23, 1954;
- R. Kneser, H. Ney, *Improved backing-off for m-gram language modeling,* in *Conference on Acoustics, Speech and Signal Processing,* 1995;

Bibliography

- K. McRae *et al., Modeling the influence of thematic fit in online sentence comprehension,* in *Journal of Memory and Language,* vol. 38, 1998, pp. 283-312.
- O. Melamud *et al.*, *Probabilistic modeling of joint-context in Distributional Similarity,* in *Proceedings of the CONLL Conference*, 2014;
- J. Nivre *et al.*, *Malt Parser: a language-indipendent parser for data-driven dependency parsing,* in *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories,* 2005;
- M. Ruiz-Casado *et al.*, *Using context-window overlapping in synonymy discovery and ontology extension*, in *Proceedings of the RANLP Conference,* 2005;
- D. Yang, D. Powers, *Verb similarity on the taxonomy of WordNet,* in *Proceedings of the 3rd International Wordnet Conference,* 2006.