Distributional Semantics of Multi-Word Expressions

Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab

COST Action IC1207 PARSEME Meeting
Warsaw, September 16, 2013

Thanks to Marco Baroni for the permission to use EACL 2012 tutorial slides.

MWE and semantics

Rayson *et al.* (2010)

(...) in order to develop more efficient [MWE extraction] algorithms, we need deeper understanding of the structural and semantic properties of MWEs, such as morpho-syntactic patterns, semantic compositionality, semantic behavior in different contexts, cross-lingual transformation of MWE properties etc. Compositionality determines the strategy needed to interpret and translate MWEs. In particular, the semantics of a highly compositional MWE can be interpreted by aggregating that of its constituent words, whereas for a highly idiomatic MWE, we would need to resort to contextual information and specific knowledge resources.

- Either way we need semantics: to detect non-compositional MWEs or to model the meaning of compositional MWEs

Semantic compositionality of MWEs

- Compositionality: degree to which the features of the parts of an MWE combine to predict the features of the whole
- Decomposability: degree to which the semantics of an MWE can be ascribed to those of its parts (Baldwin *et al.*, 2003)
  1. non-decomposable MWEs
     kick the bucket, *hot dog*, *shoot the breeze*, *take a haircut*
  2. idiosyncratically decomposable
     spill the beans, *let the cat out of the bag*
  3. simple decomposable = "institutionalised"
     traffic light, *motor car*, *house boat*
- MWEs populate a continuum between compositional and non-compositional expressions (Bannard *et al.*, 2003)

Distributional semantics (DS)

- Distributional semantics (DS) models lexical meaning with high coverage and low development costs
- DS does not readily scale up to represent meaning of phrases (and sentences)
- However, there is much recent work on distributional semantics composition (DSC) and on unifying DS and formal semantics

DS + MWE ?

- DS can be viewed as a data-driven framework for bottom-up modeling and analysis of MWE meaning
- DS models can cover both extremes in the semantic transparency continuum:
  - detect non-compositional MWEs via DSC or similarity-based measures
  - model the meaning of semantically transparent MWEs via DSC

Next: A brief overview of both aspects

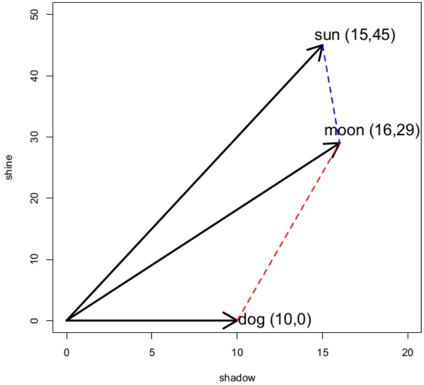
Outline

Distributional semantics

- Representation of word meaning based on distributional hypothesis (Harris, 1954):
  - correlation between similarity of words' contexts and words' semantic similarity
- Words represented as vectors of context features obtained from corpus
- Semantic similarity predicted via vector similarity
- Distributional semantic models used in many applications (Turney and Pantel, 2010)

Distributional semantic models

	planet	night	full	shadow	shine	crescent
moon	10	22	43	16	29	12
sun	14	10	4	15	45	0
dog	0	4	2	10	0	0



Distributional semantic models

- Parameters:
  - **context elements:**
    documents, words in a window, words linked by a dependency path, . . .
  - **weighting:**
    raw frequency counts, mutual information, log-likelihood, tf-idf, . . .
  - **dimensionality reduction:**
    none, SVD, topic modeling, column filtering, random indexing
- Typical models:
  - **VSM**: documents, raw counts, no reduction
  - **LSA**: VSM + SVD
  - **HAL**: words, column filtering
  - **COALS**: words + weighting (+ SVD)
- Parameter exploration by Bullinaria and Levy (2012) and Lapesa *et al.* (2013)

Outline

Why non-compositionality detection?

(1) MWE extraction (Lin, 1999; Schone and Jurafsky, 2001)
- extraction of non-compositional/institutionalised MWEs (Lin, 1999)
- non-compositionality as one of the features (Schone and Jurafsky, 2001)

(2) Non-compositional MWEs could/should be treated differently
- single units in IR (Acosta et al., 2011) or MT (Carpuat and Diab, 2010)
- special treatment in semantic tasks such as SRL
  Sporleder and Li (2009): MWEs violate selectional restrictions, subcategorization constraints, change assignment of roles, etc.
- ...

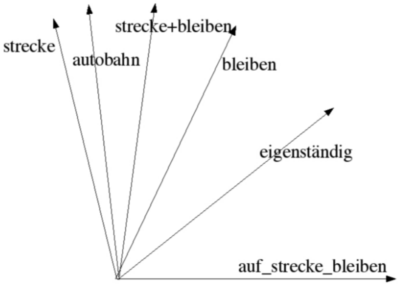
(3) DS models could also profit from treating non-compositional MWEs as single units (Krčmář *et al.*, 2013)

Baldwin *et al.* (2003)

- Compositional MWEs are generally endocentric (dependents narrow the meaning of the head)
  - *house boat* is a hyponym of *house* and *boat*
  - exceptions exits, e.g. non-intersective adjectives: *former president*
- Compositionality test: if DS similarity between a MWE and its constituent words is sufficiently high, then MWE is compositional
- Experiments on noun-noun and verb-particles compounds
- WordNet hyponymy-based evaluation: MWE endocentric if it is a hyponym of its head
- Results: moderate correlation between LSA similarities and occurrences of hyponymy (problems with polysemy of high-frequency items, WordNet inconsistencies)

Katz and Giesbrecht (2006)

- Compare MWE vector $\vec{ab}$ against combined vector $\vec{c} = \vec{a} + \vec{b}$
- If vectors are dissimilar, MWE is probably non-compositional



- LSA vectors, cosine similarity, supervised threshold optimization
- Similar idea in (Schone and Jurafsky, 2001) for MWE re-ranking

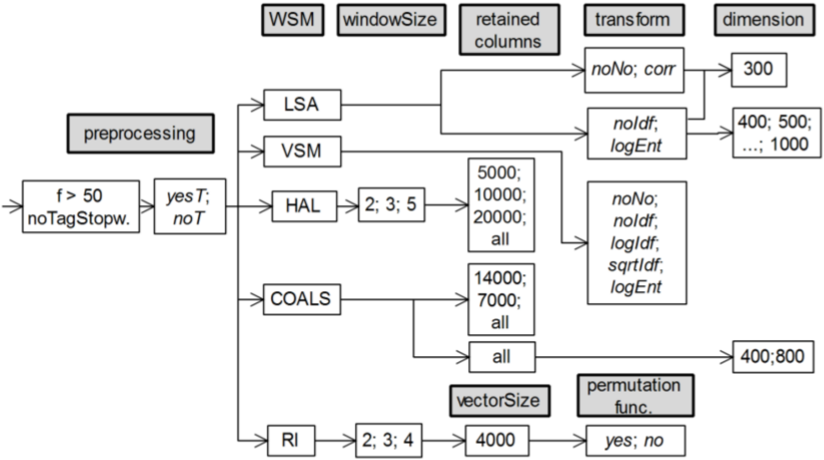
Biemann and Giesbrecht (2011)

- Shared task at DiSCo 2011 (Distributional Semantics and Compositionality): extracting non-compositional MWEs from corpora
- Graded compositionality judgments obtained by crowdsourcing
  - English & German datasets
  - in-context annotations, later averaged over contexts and annotators
- Seven teams participated, with various (1) lexical association measures, (2) DS models, (3) supervised models on top
- Results:
  - no clear winner on the English dataset
  - DS models performed slightly better
- Corpus-based acquisition of graded compositionality is a hard task

Krčmář *et al.* (2013)

- A very systematic evaluation of several DS models and DS-based compositionality measures
- Models: VSM, LSA, HAL, COALS, RI
- Measures:
  - Substitutability-based measure (SU)
    hot dog vs. *warm dog*
  - Endocentricity-based measure (EN)
    hot dog vs. *dog*
  - Compositionality-based measure (CO)
    hot dog vs. *hot*$\odot$*dog*
  - Neighbors-in-common-based measure (NE)
    hot dog$\rightarrow$*food,chips* vs. *dog*$\rightarrow$*cat,bark*
- Spearman correlation on 400+ manually annotated MWEs (DiSCo + Reddy dataset)

Krčmář *et al.* (2013)



Krčmář *et al.* (2013)

WSM	**Measure**	**wAvg(of ρ)**	**ρAN-VO-SV**	**ρAN**	**ρVO**	**ρSV**	**ρNN**
VSM$_1$	SU$_1$	0.28	0.03	0.01	**0.51**	**0.04**	**0.62**
VSM$_2$	EN$_1$	0.26	0.19	0.08	0.29	**0.04**	**0.69**
VSM$_3$	CO$_1$	0.32	0.26	0.24	0.23	**0.25**	**0.65**
VSM$_1$	NE$_1$	0.32	0.19	**0.36**	0.25	-0.13	**0.73**
LSA$_1$	SU$_2$	0.31	0.06	0.05	**0.50**	**0.20**	**0.59**
LSA$_2$	EN$_2$	0.50	**0.40**	**0.39**	**0.55**	**0.32**	**0.78**
LSA$_3$	CO$_1$	0.48	**0.36**	**0.29**	**0.60**	**0.42**	**0.69**
LSA$_2$	NE$_2$	0.44	**0.33**	**0.34**	0.40	**0.44**	**0.67**
HAL$_1$	SU$_3$	0.29	0.16	0.09	0.32	**0.34**	**0.56**
HAL$_2$	EN$_3$	0.36	0.28	**0.33**	0.35	**0.26**	**0.53**
HAL$_3$	CO$_1$	0.24	0.22	0.25	0.16	**0.15**	0.42
HAL$_4$	NE$_3$	0.21	0.14	0.02	0.33	**0.06**	0.47
COALS$_1$	SU$_4$	0.42	0.28	0.28	**0.54**	**0.30**	**0.59**
COALS$_2$	EN$_2$	0.49	**0.44**	**0.52**	**0.51**	**0.07**	**0.72**
COALS$_2$	CO$_1$	0.47	**0.40**	**0.47**	**0.51**	**0.07**	**0.74**
COALS$_2$	NE$_4$	0.52	**0.48**	**0.55**	**0.50**	**0.21**	**0.74**
RI$_1$	SU$_5$	0.30	0.14	0.14	0.29	**0.12**	**0.72**
RI$_2$	EN$_3$	0.44	**0.34**	**0.37**	**0.54**	**0.20**	**0.63**
RI$_3$	CO$_1$	0.23	0.23	**0.29**	0.17	**0.17**	0.26
RI$_2$	NE$_5$	0.31	0.26	0.26	0.42	**0.04**	0.44

Token-based idiom classification

- Many MWEs are used regularly in both their idiomatic and in their literal senses (*green light*)
  - Katz and Giesbrecht (2006): about 1/3 of the uses of the MWE *ins Wasser fallen* in their corpus are literal uses
  - Cook *et al.* (2007): 20% of idioms are used literally
- Literal usage can even dominate in some domain (*drop the ball*)
- Token-based idiom classification
  - Katz and Giesbrecht (2006)
  - Cook *et al.* (2007)
  - Sporleder and Li (2009)

Outline

Why composition?

Idea: explicitly construct a composed representation in vector space

1. Distributional semantic representation of compositional MWEs
   - accounts for productivity of language
   - accounts for sparsity problem
2. Detecting non-compositionality using composition-based methods
   - good semantic composition models for detecting lack of compositionality

Mitchell and Lapata (2008)

- Implemented and tested a number of vector composition models

(1) **(Weighted) additive model**: $\vec{p} = \alpha\vec{u} + \alpha\vec{v}$

(2) **Multiplicative model**: $\vec{p} = \vec{u} \odot \vec{v}, \quad p_i = u_i \cdot v_i$

(3) **Tensor (outer) product**: $\mathbf{P} = \vec{u} \otimes \vec{v}$

(4) **Dilation**: $\mathbf{p} = (1-\lambda)(\vec{u} \cdot \vec{v})\vec{u} + (\vec{u}\vec{u})\vec{v}$
(stretching $\vec{v}$ in the direction of $\vec{u}$)

- Evaluated on phrase similarity task (e.g., *vast amount* vs. *large quantity*)
- Dilatation performs consistently well, multiplicative model is good for simple spaces, additive model for LDA
- In much subsequent work multiplicative model proved to work well and is widely used

Mitchell and Lapata (2008)

	music	solution	economy	craft	reasonable
practical	0	6	2	10	4
difficulty	1	8	4	4	0
practical + difficulty	1	14	6	14	4
practical $\odot$ difficulty	0	48	8	40	0

Mitchell and Lapata (2008)

- Mitchell & Lapata models do composition via vector averaging
- Some problematic cases:
  - *the boy*
  - *red face* vs. *red boy*
  - *cat eats mouse* vs. *mouse eat cat*
  - *the valley of the moon* vs. *the valley and the moon*

Baroni and Zamparelli (2010)

- Adjectives in attributive position are functions (linear maps) from one noun meaning to another

$$\vec{n}' = f(\vec{n}) = \mathbf{A}\vec{n}$$

- Each adjective has its own specific matrix $\mathbf{A}$, modeling its meaning

OLD	runs	barks
runs	0.5	0
barks	0.3	1

$\times$

	dog
runs	1
barks	5

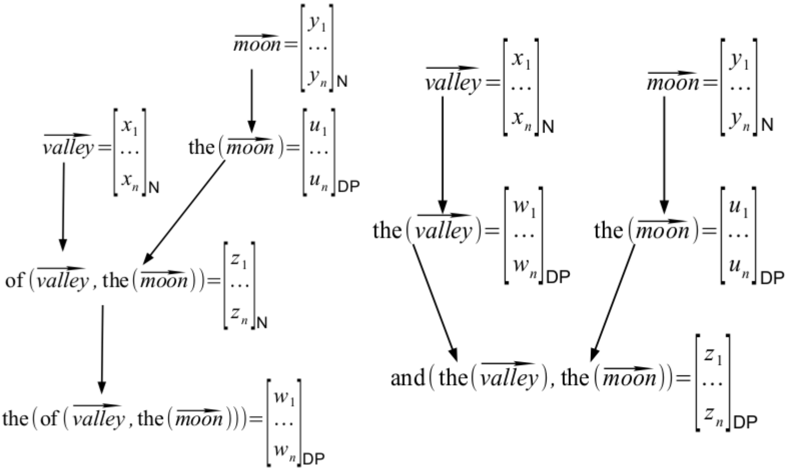
$=$

	OLD(dog)
runs	$(0.5 \times 1) + (0 \times 5)$ $= 0.5$
barks	$(0.3 \times 1) + (5 \times 1)$ $= 5.3$

- Matrix weights can be trained obtained from corpus using regression, as proposed by Guevara (2010) for generic DSC
- Distributional functions map from vectors to vectors

Baroni *et al.* (2012)

- Scale up to represent the meaning of longer phrases and sentences
- Syntactic analysis guides the semantic composition of vectors
- Type-logical syntax-semantic interface based on categorial grammar
- Categories define linear algebraic objects (vectors, matrices, tensors)
  - nouns, determiner phrases, and sentences are still represented as vectors
  - adjectives, verbs, determiners, prepositions, conjunctions, etc. are modeled by distributional functions


$\overrightarrow{moon} = [y_1 \ldots y_n]_N$
$\overrightarrow{valley} = [x_1 \ldots x_n]_N$
$\text{the}(\overrightarrow{moon}) = [u_1 \ldots u_n]_{DP}$
$\text{of}(\overrightarrow{valley}, \text{the}(\overrightarrow{moon})) = [z_1 \ldots z_n]_N$
$\text{the}(\text{of}(\overrightarrow{valley}, \text{the}(\overrightarrow{moon}))) = [w_1 \ldots w_n]_{DP}$
$\overrightarrow{valley} = [x_1 \ldots x_n]_N$
$\overrightarrow{moon} = [y_1 \ldots y_n]_N$
$\text{the}(\overrightarrow{valley}) = [w_1 \ldots w_n]_{DP}$
$\text{the}(\overrightarrow{moon}) = [u_1 \ldots u_n]_{DP}$
$\text{and}(\text{the}(\overrightarrow{valley}), \text{the}(\overrightarrow{moon})) = [z_1 \ldots z_n]_{DP}$

Scaling up to sentences?

- In previous models, the result of a composition is a vector (or matrices in, case of tensor product)
- Can the meaning of a whole sentences be represented as a vector (matrix), regardless of sentence length?

Grefenstette *et al.* (2010)

- A mathematical framework for a compositional distributional model of meaning, consisting of
  - formalism for type logical-syntax: Lambek's Pregroup Grammars
  - formalism for vector space semantics: tensor mathematics
  - syntax-semantics interface formalized via category theory
- SVO constructions:
  - noun type n is assigned vector space $\mathbf{N}$ (ordinary vector space)
  - sentence type $n^r s n^l$ is assigned tensor space $\mathbf{N} \otimes \mathbf{S} \otimes \mathbf{N}$
  - intransitive verbs $\Rightarrow$ vectors, transitive verbs $\Rightarrow$ matrices, ditransitive verbs $\Rightarrow$ rank-3 tensors
  - rank increases with meaning complexity
  - but simpler sentences can be embedded in higher-rank space
  - sentences are comparable, regardless of their length

Towards combining DS and formal semantics

- Previous models deal with composition in vector (tensor) spaces
- An alternative is to combine formal semantics and distributional semantics to exploit their complementarity
    - Copestake and Herbelot (2012)
    - Erk (2013)

Conclusion & Perspectives

- DS used for non-compositionality detection, but this is far from being a solved problem
- Various DSC models around, with a trend towards (1) more structured distributional representations and/or (2) combining formal and distributional semantics

Perspectives within PARSEME:

- (Multilingual) DS representations in MWE dictionaries?
- DS for improved parsing of MWE?
- MWE representations that rely on more structured semantic spaces (including syntax) or on a combination of formal and distributional semantics?
- ???

References I

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Bannard, C., Baldwin, T., and Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 65–72. Association for Computational Linguistics.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Baroni, M., Bernardi, R., and Zamparelli, R. (2012). Frege in space: A program for compositional distributional semantics.

References II

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. Association for Computational Linguistics.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, **44**(3), 890–907.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48. Association for Computational Linguistics.

Copestake, A. and Herbelot, A. (2012). Lexicalised compositionality. *Unpublished draft*.

Erk, K. (2013). Towards a semantics for distributional representations. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*.

References III

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., and Pulman, S. (2010). Concrete sentence spaces for compositional distributional models of meaning. *arXiv preprint arXiv:1101.0309*.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, **10**(23), 146–162.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics.

Krčmář, L., Ježek, K., and Pecina, P. (2013). Determining compositionality of word expressions using word space models. *NAACL HLT 2013*, **13**, 42.

Lapesa, G., Evert, S., and Erlangen-Nürnberg, F. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming.

References IV

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. V. (2010). Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, **44**(1), 1–5.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108.

Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.