

# Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics

Angeliki Lazaridou

Marco Marelli

Roberto Zamparelli

Marco Baroni

University of Trento

ACL 2013

# Making sense of the title

Compositional-ly Derived Representations of  
Morphologically Complex Words  
in **Distributional Semantics**

# Distributional Semantic Models (DSMs)

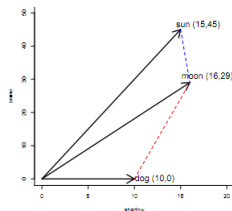
- A representation of meaning based on the *Distributional Hypothesis* ...

he curtains open and the moon shining in on the barely  
 ars and the cold , close moon " . And neither of the w  
 rough the night with the moon shining so brightly , it  
 made in the light of the moon . It all boils down , wr  
 surely under a crescent moon , thrilled by ice-white  
 sun , the seasons of the moon ? Home , alone , Jay pla  
 m is dazzling snow , the moon has risen full and cold  
 un and the temple of the moon , driving out of the hug  
 in the dark and now the moon rises , full and amber a  
 bird on the shape of the moon over the trees in front  
 But I could n't see the moon or the stars , only the  
 rning , with a sliver of moon hanging among the stars  
 they love the sun , the moon and the stars . None of



	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

- ...useful for...



and

Nearest Neighbors (NNs)

star  
 sun  
 Armstrong  
 space

# Distributional Semantic Models (DSMs)

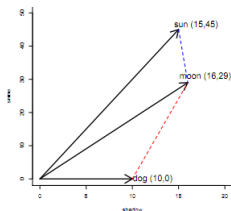
- A representation of meaning based on the *Distributional Hypothesis* ...

he curtains open and the moon shining in on the barely  
 ars and the cold , close moon " . And neither of the w  
 rough the night with the moon shining so brightly , it  
 made in the light of the moon . It all boils down , wr  
 surely under a crescent moon , thrilled by ice-white  
 sun , the seasons of the moon ? Home , alone , Jay pla  
 m is dazzling snow , the moon has risen full and cold  
 un and the temple of the moon , driving out of the hug  
 in the dark and now the moon rises , full and amber a  
 bird on the shape of the moon over the trees in front  
 But I could n't see the moon or the stars , only the  
 rning , with a sliver of moon hanging among the stars  
 they love the sun , the moon and the stars . None of



	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

- ...useful for...



and

Nearest Neighbors (NNs)

star  
 sun  
 Armstrong  
 space

- But **word frequency** affects **vector quality** [Bullinaria and Levy, 2007]

# Making sense of the title

Compositional-ly Derived Representations of  
**Morphologically Complex Words**  
in Distributional Semantics

# Morphologically Complex Words

- high productivity of word derivations
  - complex word = stem and affix
  - e.g. rebuild, colorful, frequently ...
- 55% of the words in CELEX Lexical Database are morphologically complex
- BAD NEWS: **Sparsity**
  - lexicalizable and affixless do not occur **at all** in our large corpus!

# Morphologically Complex Words

- high productivity of word derivations
  - complex word = stem and affix
  - e.g. rebuild, colorful, frequently ...
- 55% of the words in CELEX Lexical Database are morphologically complex
- BAD NEWS: **Sparsity**
  - lexicalizable and affixless do not occur at all in our large corpus!
- **GOOD NEWS:** 80% of times in our large corpus, the stem appears more frequently than the complex word!

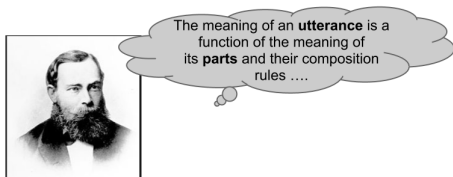
# Making sense of the title

## **Compositional-ly Derived Representations** of Morphologically Complex Words in Distributional Semantics



# Compositional DSMs (cDSMs)

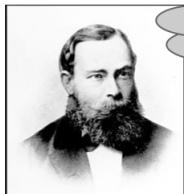
- Increasing need to represent meaning **beyond words** (e.g. paraphrase detection, etc)
- Productivity of syntax causes **sparsity**, which does not allow to collect co-occurrence statistics for whole sentences
- Solution à la Frege...



- ...operationalized in DSM with different **composition functions** on vectors for **phrases**  
 [Mitchell and Lapata, 2008, Socher et al., 2012, Coecke et al., 2010, Baroni and Zamparelli, 2010, Mitchell and Lapata, 2010]

# cDSMs to the rescue for morphology

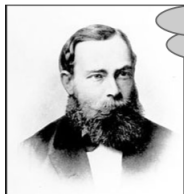
- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?



The meaning of **complex word** is a function of the meaning of its **morphemes** and their composition rules ...

# cDSMs to the rescue for morphology

- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?

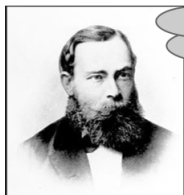


The meaning of **complex word** is a function of the meaning of its **morphemes** and their composition rules ...

- affixless means without affix

# cDSMs to the rescue for morphology

- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?



The meaning of **complex word** is a function of the meaning of its **morphemes** and their composition rules ...

- **affixless** means *without affix*

Our proposal:  $\text{affixless} = f_{\text{composition function}}(\text{affix}, \text{less})$

# Morphological Data

- CELEX English Lexical Database
  - Rich morphological annotations
- Extract stem/derived pairs
  - e.g. valid/invalid, frequent/frequently, run/rerun etc
- 18 derivational affixes
  - 3 prefixes: in, re, un
  - 15 suffixes: al, er, ful, ic, ion, ist, ity, ize, less, ly, ment, ness, ous, y
- Report results on 50 stem/derived pairs (900 in total)

# Morphemes in Distributional Semantics

## Representing corpus-extracted vectors

- **stems** and derived words:
  - collect co-occurrence statistics from 2-word windows around each target item from a large **large**<sup>1</sup> corpus
- **affixes** as vectors
  - accumulate context vectors of derived words with the same **affix**
  - e.g.  $\vec{less} = \vec{effortless} + \vec{colorless} + \dots$

---

<sup>1</sup><http://wacky.sslmit.unibo.it>, <http://en.wikipedia.org>,  
<http://www.natcorp.ox.ac.uk/>

# Compositional Models for Derivational Morphology

Derive representation of complex word by composing its **stem** and **affix**

Model	Composition function	By Who?
<b>weighted additive</b>	$w_1 \vec{stem} + w_2 \vec{affix}$	[Mitchell and Lapata, 2010]
<b>multiplicative</b>	$\vec{stem} \odot \vec{affix}$	[Mitchell and Lapata, 2010]
<b>dilation</b>	$\ \vec{stem}\ _2^2 \vec{affix} + (\lambda - 1) \langle \vec{stem}, \vec{affix} \rangle \vec{stem}$	[Mitchell and Lapata, 2010]
<b>full additive</b>	$W_1 \vec{stem} + W_2 \vec{affix}$	[Guevara, 2010]
<b>lexical function</b>	$A_{affix} \vec{stem}$	[Baroni and Zamparelli, 2010]

# NON-Compositional models for Derivational Morphology

- **corpus**: How problem is currently been handled in DSMs i.e. it isn't!!  
ignore morphology and extract representation of complex word from the corpus
  - frequently 😊
  - affixless 😞
- **stem**: back off to the meaning of the stem
  - build instead of rebuild 😊
  - cycle instead of recycle 😞



# NON-Compositional models for Derivational Morphology

- **corpus**: How problem is currently been handled in DSMs i.e. it isn't!!  
ignore morphology and extract representation of complex word from the corpus
  - frequently 😊
  - affixless 😞
- **stem**: back off to the meaning of the stem
  - build instead of rebuild 😊
  - cycle instead of recycle 😞
- Introduce 2 annotations to dichotomize dataset and test our intuitions

# Dichotomization of dataset based on high OR low vector quality

- Approximate quality of a **corpus**-extracted vector by the quality of its nearest neighbors (NNs)
  - car**: automobile, truck, drive 😊
  - river**: potato, school, t-shirt 😞
- Relatedness judgments<sup>2</sup> to classify items as high or low quality

Affix	Type	Derived form	Neighbors
-ist	High	transcendentalist	mythologist, futurist, theosophist
	Low	florist	Harrod, wholesaler, stockist
in-	High	inaccurate	misleading, incorrect, erroneous
	Low	inoperable	metastasis, colorectal, biopsy
re-	High	recapture	retake, besiege, capture
	Low	rename	defunct, officially, merge

<sup>2</sup><http://crowdfLOWER.com/>

# Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. **in-** (valid-**invalid**)
  - ...simply change the POS-tag of the stem, e.g. **-ly** (sad-**sadly**)
- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems
- Examples:
  - **High**: **compulsory**/compulsorily, **chaos**/chaotic
  - **Low**: **cycle**/recycle, **believable**/**un**believable

# Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. **in-** (valid-**invalid**)
  - ...simply change the POS-tag of the stem, e.g. **-ly** (sad-**sadly**)
- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems
- Examples:
  - **High**: **compulsory**/compulsorily, **chaos**/chaotic
  - **Low**: **cycle**/**recycle**, **believable**/**unbelievable**
- What is causing the low judgements?:
  - opacity

# Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. **in-** (valid-**invalid**)
  - ...simply change the POS-tag of the stem, e.g. **-ly** (sad-**sadly**)
- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems
- Examples:
  - **High**: **compulsory**/compulsorily, **chaos**/chaotic
  - **Low**: cycle/recycle, **believable**/**unbelievable**
- What is causing the low judgements?
  - opacity
  - affixes triggering antonymy

# Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the **model**-generated ones.  
Focus only on **high quality** complex words.

	<i>stem</i>	<i>mult</i>	<i>dil.</i>	<i>wadd</i>	<b>fulladd</b>	<b>lexfunc</b>
<b>All</b>	0.47	0.39	0.48	0.50	<b>0.56</b>	<b>0.54</b>
High Relatedness	0.52	0.43	0.53	0.55	0.61	0.58
Low Relatedness	0.32	0.28	0.33	0.38	0.41	0.42

- Overall composition models apart from **mult** outperform **stem**
  - fulladd** and **lexfunc** significantly outperform **stem**

# Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the **model**-generated ones.  
Focus only on **high quality** complex words.

	<i>stem</i>	<i>mult</i>	<i>dil.</i>	<i>wadd</i>	<b>fulladd</b>	<b>lexfunc</b>
<b>All</b>	0.47	0.39	0.48	0.50	0.56	0.54
<b>High Relatedness</b>	0.52	0.43	0.53	0.55	<b>0.61</b>	<b>0.58</b>
<b>Low Relatedness</b>	0.32	0.28	0.33	0.38	0.41	0.42

- Overall composition models apart from **mult** outperform **stem**
  - **fulladd** and **lexfunc** significantly outperform **stem**
- Surprisingly, even when meanings of stem/derived are very close!!

# Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the **model**-generated ones.  
Focus only on **high quality** complex words.

	<i>stem</i>	<i>mult</i>	<i>dil.</i>	<b>wadd</b>	<b>fulladd</b>	<b>lexfunc</b>
<b>All</b>	0.47	0.39	0.48	0.50	0.56	0.54
<b>High Relatedness</b>	0.52	0.43	0.53	0.55	0.61	0.58
<b>Low Relatedness</b>	0.32	0.28	0.33	<b>0.38</b>	<b>0.41</b>	<b>0.42</b>

- Overall composition models apart from **mult** outperform **stem**
  - **fulladd** and **lexfunc** significantly outperform **stem**
- Surprisingly, even when meanings of stem/derived are very close!!
- **stem** is a particularly bad surrogate when meanings of stem/derived are not close
  - **fulladd**, **lexfunc**, and to a lesser extent **wadd**, can capture **affix-triggered** semantics



## Experiment 2: Focus on high-quality antonym complex words

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the **model**-generated ones.  
Focus only on **high quality** complex words.

	<i>stem</i>	<i>mult</i>	<i>dil</i>	<i>wadd</i>	<i>fulladd</i>	<i>lexfunc</i>
-less	0.22	0.23	0.24	0.30	0.38	<b>0.44</b>
in-	0.39	0.34	0.40	0.45	<b>0.47</b>	0.45
un-	0.33	0.33	0.34	0.41	0.44	<b>0.46</b>

- Encouraging results for modeling negation in DSM and cDSMs [Mohammad et al., 2013, Preller and Sadrzadeh, 2011]
  - modeling **in**operable is close to modeling syntactic negation compositionally, i.e. **not** operable

## Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on **low quality** complex words
- Evaluate quality of **model**-generated vector representations.
  - Higher scores = better quality

	<i>corpus</i>	<i>stem</i>	<i>wadd</i>	<i>fulladd</i>	<i>lexfunc</i>
<b>All</b>	2.28	3.26	<b>4.12</b>	<b>3.99</b>	3.09
<b>HR</b>	2.29	3.56	<b>4.48</b>	<b>4.31</b>	3.31
<b>LR</b>	2.22	2.48	<b>3.14</b>	<b>3.12</b>	2.52

- **lexfunc** is significantly outperformed by **wadd** and **fulladd**.
  - **fulladd** and **wadd** feature almost always the stem as **NN**
- Stemploitation strategy for **fulladd** and **wadd**?

## Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on **low quality** complex words
- Evaluate quality of **model**-generated vector representations.
  - Higher scores = better quality

	<i>corpus</i>	<i>stem</i>	<i>wadd</i>	<i>fulladd</i>	<i>lexfunc</i>
<b>All</b>	2.28	3.26	<b>4.12</b>	<b>3.99</b>	3.09
HR	2.29	3.56	4.48	4.31	3.31
LR	2.22	2.48	3.14	3.12	2.52

- **lexfunc** is significantly outperformed by **wadd** and **fulladd**.
  - **fulladd** and **wadd** feature almost always the stem as **NN**
- Stemploitation strategy for **fulladd** and **wadd**? **NO**
  - Both models over**All** significantly outperform **stem**

## Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on **low quality** complex words
- Evaluate quality of **model**-generated vector representations.
  - Higher scores = better quality

	<i>corpus</i>	<i>stem</i>	<i>wadd</i>	<i>fulladd</i>	<i>lexfunc</i>
<b>All</b>	2.28	3.26	4.12	3.99	3.09
<b>HR</b>	2.29	3.56	4.48	4.31	3.31
<b>LR</b>	2.22	2.48	<b>3.14</b>	<b>3.12</b>	2.52

- **lexfunc** is significantly outperformed by **wadd** and **fulladd**.
  - **fulladd** and **wadd** feature almost always the stem as **NN**
- Stemploitation strategy for **fulladd** and **wadd**? **NO**
  - Both models over**All** significantly outperform **stem**
  - In **LR**, **corpus** is significantly outperformed by both models, but not **stem**

## Experiment 3: Qualitative analysis of NNs

Target	Model	Neighbors
florist	<i>wadd</i>	<b>flora</b> , fauna, ecosystem
	<i>fulladd</i>	<b>flora</b> , fauna, ecologist
	<i>lexfunc</i>	ornithologist, naturalist, botanist
	<i>corpus</i>	Harrod, wholesaler, stockist
inoperable	<i>wadd</i>	<b>operable</b> , palliation, biopsy
	<i>fulladd</i>	<b>operable</b> , inoperative, ventilator
	<i>lexfunc</i>	inoperative, unavoidably, flaw
	<i>corpus</i>	metastasis, colorectal, biopsy
rename	<i>wadd</i>	<b>name</b> , later, namesake
	<i>fulladd</i>	<b>name</b> , namesake, later
	<i>lexfunc</i>	temporarily, reinstate, thereafter
	<i>corpus</i>	defunct, officially, merge

# Conclusions

- First work to test validity of compositionality principle for generating representations of complex words
- Morpheme composition can improve on corpus-based representations when the latter are low-quality (e.g., because of sparsity problems)
- Morpheme composition outperforms **stem** baseline, even when similarity between stem and derived is high
- Finally, new evaluation task for compositional approaches, in which **fulladd** emerged as the best model

# Future Work

- Model jointly morphological induction and composition
- Try to “decompose” frequent complex words to derive the representations of their less frequent stems.
- Test compositionality principle for inflectional morphology
- Test recursive morpheme composition (e.g. **unoperable**), especially useful for agglutinative languages

# Thank you for your attention!

The implementation is based on the DISSECT toolkit,  
demo at the ACL demo session this year!

<http://clic.cimec.unitn.it/composes/toolkit>



# References I



Baroni, M. and Zamparelli, R. (2010).

Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.  
In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.



Coecke, B., Sadrzadeh, M., and Clark, S. (2010).

Mathematical foundations for a compositional distributional model of meaning.  
*Linguistic Analysis*, 36:345–384.



Guevara, E. (2010).

A regression model of adjective-noun compositionality in distributional semantics.  
In *Proceedings of GEMS*, pages 33–37, Uppsala, Sweden.



Mitchell, J. and Lapata, M. (2008).

Vector-based models of semantic composition.  
In *Proceedings of ACL*, pages 236–244, Columbus, OH.



Mitchell, J. and Lapata, M. (2010).

Composition in distributional models of semantics.  
*Cognitive Science*, 34(8):1388–1429.



Mohammad, S., Dorr, B., Hirst, G., and Turney, P. (2013).

Computing lexical contrast.  
*Computational Linguistics*.  
In press.

# References II



Preller, A. and Sadrzadeh, M. (2011).

Bell states and negative sentences in the distributed model of meaning.  
*Electr. Notes Theor. Comput. Sci.*, 270(2):141–153.



Socher, R., Huval, B., Manning, C., and Ng, A. (2012).

Semantic compositionality through recursive matrix-vector spaces.  
In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.