An Empirical Study on Compositionality in Compound Nouns

Siva Reddy[1,2], Diana McCarthy[2], **Suresh Manandhar**[1]

[1]Artificial Intelligence Group,
Department of Computer Science,
University of York, UK

[2]Lexical Computing Ltd., UK

IJCNLP 2011 Main Conference
Chiang Mai, Thailand
9th November 2011

THE UNIVERSITY *of York*

Compositionality

The Principle of Semantic Compositionality (Partee, 1995)

The meaning of a complex expression is determined by the meanings of its constituents and its structure

Example

Compound Noun	swimming pool
Adjective Noun	blue sky
Subject Verb	flies fly
Verb Object	lose keys
Verb Particle	*climb up* the hill

Non-Compositionality

Non-Compositional Expressions

Not all the expressions in language have compositional meaning. The applicability of principle of semantic compositionality is widely debated (Pelletier, 1994)

Example

Compound Noun	cloud nine
Adjective Noun	red tape
Verb Object	spill beans
Verb Particle	*carry out* a meeting

THE UNIVERSITY *of York*

What is this paper about?

Part 1: Study on human judgments from Mechanical Turk

- A unique dataset for Compositionality judgments
- Contribution of each constituent to the semantics of a phrase
- Compositionality judgment of phrase
- Relation between constituent contribution to the phrase compositionality

Part II: Two Computational Models for Compositionality

- Constituent based model built upon above conclusions
- Composition function based model

THE UNIVERSITY of York

Existing Datasets

Resource	Phrase Types	# Annotators	# Phrases	Sample Data
McCarthy et al. (2003)	Verb Particle	4	117	step down: 9/10
Bannard et al. (2003)	Verb Particle	28	40	run > run up: False; up > run up: True; no_scores
Venkatapathy and Joshi (2005)	Verb Object	2	800	spill beans: 1/6
Biemann and Giesbrecht (2011)	Verb Object, Verb Subject, Adj Noun	20	145	blue chip: 11/100
Korkontzelos and Manandhar (2009)	Noun Noun	1	38	no_scores

- All these datasets (except Bannard et al.) have
  - *Only* phrase compositionality judgments
  - Constituent contributions are absent
  - Bannard et al. does not have any quantitative scores

THE UNIVERSITY *of York*

Motivation for our dataset preparation

- We work on Compound Nouns containing two words
  - No existing datasets for Compound Nouns
  - Relatively simple than other constructions since no morphological or syntactic variations
- Constituent contribution scores along with phrase level compositionality scores
  - Possible to study the relation between constituents and phrase compositionality
- Aim to prepare a non-skewed data. Most datasets are skewed towards compositional phrases
- Observe the continuum of compositionality, if at all it exists.

Compositionality as literality

Bannard et al. (2003)

"the overall semantics of the multi-word expression (here compound) can be composed from the simplex semantics of its parts, as described (explicitly or implicitly) in a finite lexicon"

Our assumption: Compositionality as literality problem

A compound is compositional if its meaning can be understood from the literal (simplex) meaning of its parts

Similar assumption used by other researchers but not explicitly mentioned

- Lin (1999); Katz and Giesbrecht (2006)
- DisCo 2011 Shared Task Data (Biemann and Giesbrecht, 2011)

Compound Noun Set

We aimed at including data from four different classes

1. Both words are literal
   - swimming pool
2. First word is literal and second is non-literal
   - night owl
3. First word is non-literal and second literal
   - zebra crossing
4. Both words non-literal
   - smoking gun

Classes 2 and 4 are found to be hard to collect

- Heuristics based on WordNet and Wiktionary
- Authors have chosen 90 compounds for final annotation

Experimental Setup

Three tasks per compound

1. How literal the phrase is?
2. How literal the use of first constituent in the given phrase?
3. How literal the use of second constituent in the given phrase?

- Each task annotated by 30 random annotators out of 151 annotators
- Lower chance of bias to any annotator
- Total 8100 annotations (90 * 3 * 30 = 8100)
- 5 random examples from ukWaC (Ferraresi et al., 2008)
  - To capture behavior of most frequent sense of a compound
  - Partially remove subjective differences

How literal is this phrase?

Sample examples at http://tinyurl.com/is-it-lit

web site:

Definitions:

1. a computer connected to the internet that maintains a series of web pages on the World Wide Web

Examples:

1. can simply update the firmware and modem drivers by downloading patches from the modem manufacturers **web site** . It may be best to contact the manufacturers of your modem in the first
2. up with the Government position here (mainly pro-badger killing) , visit the DEFRA **web site** , and use the search function to trace papers about badgers and tuberculosis . Action
3. of galaxy formation and evolution and of the enrichment of the intergalactic medium . This **web site** is part of a research project by Graham Thurgood who is a senior lecturer .
4. of use represent the complete and only statement of the terms of use of this **web site** . 4 . My Portfolio within the Financial Organiser Friends Provident receives its data feed
5. Courts . If you require to contact us in regard to the content of this **web site** or with a view to obtaining consent from the University to use the material contained

Note: Please select the answers below carefully based on the definition which occurs frequently in the examples

Step 1: score of 0-5 for how literal is the use of **"web"** in the phrase **"web site"**

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Please provide any comments in case you want to tell us about your judgement or any other queries/suggestions! Not Mandatory but helpful.

Annotators: Amazon Mechanical Turk

Our Quality Control

- (Snow et al., 2008) demonstrated as the turkers number increase, the quality surpasses expert judgment
- Every turker took online training and a qualification test
- Qualified turkers annotate
- Spammers and Outliers: *Catch me if you can?*

Additional Check

- If Spearman Correlation of a turker averaged over all turkers > 0.6:
  - Accept the annotation
- else if a task's annotation is closer to the task's mean
  - Accept the annotation
- else: reject

Annotation

No. of turkers participated	260
No. of them qualified	151
Spammers $\rho <= 0$	21
Turkers with $\rho >= 0.6$	81
No. of annotations rejected	383
Avg. submit time (sec) per task	30.4

Table: Amazon Mechanical Turk statistics

Annotation

No. of turkers participated	260
No. of them qualified	151
Spammers $\rho <= 0$	21
Turkers with $\rho >= 0.6$	81
No. of annotations rejected	383
Avg. submit time (sec) per task	30.4

Table: Amazon Mechanical Turk statistics

Compound	**Word1**	**Word2**	**Phrase**
swimming pool	4.80±0.40	4.90±0.30	4.87±0.34
fashion plate	4.41±1.07	3.31±2.07	3.90±1.42
face value	1.39±1.11	4.64±0.81	3.04±0.88
blame game	4.61±0.67	2.00±1.28	2.72±0.92
sitting duck	1.48±1.48	0.41±0.67	0.96±1.04

Table: Compounds with their constituent and phrase level mean±deviation scores

Agreement - Disagreement

	highest ρ	avg. ρ
ρ for phrase compositionality	0.741	0.522
ρ for first word's literality	0.758	0.570
ρ for second word's literality	0.812	0.616
ρ for over all three task types	0.788	0.589

Table: Overall Agreements

- For 15 compounds, deviation > $|\pm 1.5|$
- Some compounds are ambiguous
  - Occur frequently with both Compositional and Non-Compositional Senses
  - e.g. silver screen, brass ring
- Some due to subjective differences

Interesting Observation: Continuum of Compositionality

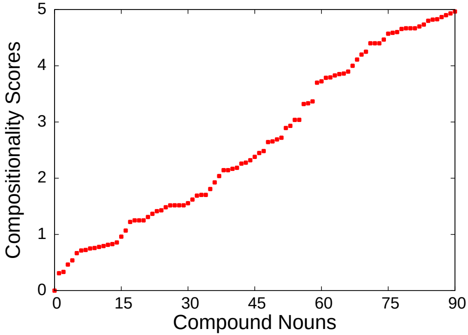

Figure: Mean Phrase Compositionality Score of all compounds

- Continuum of Compositionality is observed in the data
- Dataset is not skewed

Relation between Constituent and Phrase Compositionality Scores

- Existing methods in compositionality detection use constituent word level semantics to compose the semantics of the phrase (Baldwin et al., 2003; Katz and Giesbrecht, 2006; Sporleder and Li, 2009; Biemann and Giesbrecht, 2011)
- Evaluation datasets are not particularly suitable to study the contribution of each constituent word to the semantics of the phrase

Our dataset allows us to examine the relationship between the constituents contribution and phrase compositionality score rather than assume the nature of relationship

THE UNIVERSITY of York

Relation between Constituent and Phrase Compositionality Scores

We tried various function fittings over the human judgments

- ADD: $a.s1 + b.s2 = s3$
- MULT: $a.s1.s2 = s3$
- COMB: $a.s1 + b.s2 + c.s1.s2 = s3$
- WORD1: $a.s1 = s3$
- WORD2: $a.s2 = s3$
    - s1 and s2: Contributions from first and second constituent resp.
    - s3: phrase compositionality score

- 3-fold cross validation to evaluate the above functions (two training samples and one testing sample at each iteration)
- The coefficients of the functions are estimated using least square linear regression technique over the training samples

THE UNIVERSITY of York

Study on human judgments

Function f	ρ	$\mathbf{R^2}$
ADD	0.966	0.937
MULT	0.965	0.904
COMB	0.971	0.955
WORD1	0.767	0.609
WORD2	0.720	0.508

Table: Spearman Correlation ρ and Best-fit R^2 values between functions and phrase compositionality scores

THE UNIVERSITY of York

Study on human judgments

Function f	ρ	$\mathbf{R^2}$
ADD	0.966	0.937
MULT	0.965	0.904
COMB	0.971	0.955
WORD1	0.767	0.609
WORD2	0.720	0.508

Table: Spearman Correlation ρ and Best-fit R^2 values between functions and phrase compositionality scores

Conclusions

- Both the words decide compositionality
  - Earlier methods like (Bannard et al., 2003; Korkontzelos and Manandhar, 2009) used only one of the words (mostly head word)
- Phrase compositionality score can be predicted from constituents literality scores

Computational Models for Compositionality

We experiment with two different models

Constituent based models

- Based on the above study
- First determines the literality of each constituent
- Using the literality score of each constituent, we predict phrase compositionality score

Composition function based models

- Composition function models first build a compositional meaning of a phrase using its constituents
- Difference between the composed meaning and actual meaning is used to decide phrase compositionality score

Distributional Model: Meaning as a distributional vector

Distributional Hypothesis (Harris, 1954)

Words that occur in similar contexts tend to have similar meanings i.e. meaning of a word can be defined in terms of its context.

Word Space Model (WSM)

Meaning of a word is represented as a co-occurrence vector built from a corpus

	police-n	photon-n	speed-n	car-n	soul-n
v1 Traffic	142	0	293	347	1
v2 Light	41	29	222	198	50

THE UNIVERSITY of York

Distributional Model: Meaning as a distributional vector

Distributional Hypothesis (Harris, 1954)

Words that occur in similar contexts tend to have similar meanings i.e. meaning of a word can be defined in terms of its context.

Word Space Model (WSM)

Meaning of a word is represented as a co-occurrence vector built from a corpus

	police-n	photon-n	speed-n	car-n	soul-n
v1 Traffic	142	0	293	347	1
v2 Light	41	29	222	198	50
v3 TrafficLight	5	0	13	48	0

Constituent Based Models

If a constituent word is used literally in a given compound it is highly likely that the compound and the constituent share common co-occurrences e.g. swimming in swimming pool.

Literality of a Constituent

- s1= sim(v1, v3); s2= sim(v2, v3)
- sim is Cosine Similarity.
- i.e. if the number of common co-occurrences between constituent and compound are numerous, it is more likely the constituent has a literal meaning in the compound

	first constituent	second constituent
s1	0.616	–
s2	–	0.707

Table: Constituent level correlations with constituent human judgments

Constituent Based Models

Phrase Compositionality Score

- Study on human data revealed that if literality scores of constituents are known, phrase compositionality scores can be estimated
- s3 = f(s1, s2)

Model	ρ	R^2
ADD	0.686	0.613
MULT	0.670	0.428
COMB	0.682	0.615
WORD1	0.669	0.548
WORD2	0.515	0.410

Table: Phrase level correlations with human phrase compositionality judgments

Composition Function based models

- Several semantic composition functions are proposed to compose meaning of a phrase from its constituents (Mitchell and Lapata, 2008; Widdows, 2008; Erk and Padó, 2008)
- e.g. **Traffic**$\oplus$**Light** is the meaning composed from **Traffic** and **Light**
- $\oplus$ is the composition function
- Most successful ones are simple addition and simple multiplication (Mitchell and Lapata, 2008; Vecchi et al., 2011)

Example

	police-n	photon-n	speed-n	car-n	soul-n
v1 Traffic	142	0	293	347	1
v2 Light	41	29	222	198	50
v3 TrafficLight	5	0	13	48	0
a**Traffic** + b**Light**	183	29	515	545	51
Traffic * Light	5822	0	65046	68706	50

Composition Function based models

Phrase Compositionality Score

- Similarity between composed (compositional) meaning and the true distributional meaning
- s3= sim(v1 $\oplus$ v2, v3)

Model	ρ	R^2
$a\mathbf{v1} + b\mathbf{v2}$	0.714	0.620
v1v2	0.650	0.501

Table: Phrase level correlations with human phrase compositionality judgments

THE UNIVERSITY of York

Winner

Model	ρ	R^2
Constituent Based Models		
ADD	**0.686**	0.613
MULT	0.670	0.428
COMB	0.682	0.615
WORD1	0.669	0.548
WORD2	0.515	0.410
Composition Function Based Models		
$a\mathbf{v1} + b\mathbf{v2}$	**0.714**	0.620
v1v2	0.650	0.501
RAND	0.002	0.000

Table: Phrase level correlations of compositionality scores

Winner

- Both competitive
- Composition Function based models have slight upper hand

Possible Reasons

- Constituent based models use contextual information of each constituent *independently*
- Composition function models use contexts of both the constituents *simultaneously*
- Contexts salient to both the words are important. Foundations for our DisCo 2011 Shared Task System (Reddy et al., 2011)
  - (Biemann and Giesbrecht, 2011) "... across different scoring mechanisms, **UoY is the most robust of the systems**"

Contributions

- Novel dataset for Compositionality judgments
  - Contains constituent level contributions
  - Continuum of Compositionality
  - Not skewed
- Study of relation between constituent contributions to phrase level contributions
- Comparison between two different models for phrase compositionality expectation

The dataset is freely downloadable from `http://sivareddy.in`

Bibliography 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*, MWE '03, pages 89–96, Stroudsburg, PA, USA. 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*, MWE '03, pages 65–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of DISCo-2011 in conjunction with ACL 2011*.

Bibliography II

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the WAC4 Workshop at LREC 2008*, Marrakesh, Morocco.

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

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*, MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

THE UNIVERSITY of York

Bibliography III

Korkontzelos, I. and Manandhar, S. (2009). Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 65–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, ACL '99, pages 317–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliography IV

Mitchell, J. and Lapata, M. (2008). Vector-based Models of Semantic Composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Partee, B. (1995). Lexical semantics and compositionality. *L. Gleitman and M. Liberman (eds.) Language, which is Volume 1 of D. Osherson (ed.) An Invitation to Cognitive Science (2nd Edition)*, pages 311–360.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13:11–24. 10.1007/BF00763644.

Reddy, S., McCarthy, D., Manandhar, S., and Gella, S. (2011). Exemplar-based word-space model for compositionality detection: shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 54–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliography V

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP 2008*, pages 254–263, Stroudsburg, PA, USA.

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*, EACL '09, pages 754–762, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vecchi, E. M., Baroni, M., and Zamparelli, R. (2011). (linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, Oregon, USA. Association for Computational Linguistics.

THE UNIVERSITY of York

Bibliography VI

Venkatapathy, S. and Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 899–906, Vancouver, B.C., Canada.

Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction, Oxford*.