Measuring the compositionality of collocations via word co-occurrence vectors

Shared task system description

Alfredo Maldonado-Guerra   Martin Emms

School of Computer Science and Statistics
Trinity College Dublin
Ireland

DISCo 2011: Distributional Semantics and Compositionality
Workshop at ACL/HLT 2011

Outline

Introduction

- A shared task system that measures the compositionality of bigrams

Basic intuition

A highly compositional bigram would tend to have a considerable semantic overlap with its constituents whereas a bigram with low compositionality would share little semantic content with its constituents.

- Intuition operationalised via three configurations that exploit cosine similarity measures to detect the semantic overlap between the bigram and its constituents
- Fully unsupervised system that could be employed for any language, including under-resourced languages

Introduction

This work uses vectors as defined by Schütze (1998):

- Word (co-occurrence) vector $\mathbf{W}(w)$: types
  - Counts words that co-occur with target word w in corpus
  - 20 word window centred at target word
- Second-order context vector $\mathbf{C}^2(p)$: tokens
  - Sum of word vectors of words co-occurring with target word at position p in corpus.
  - 20 word window centred at target word

In these vectors the simplest approach possible was used: no normalisation, no weighting, etc. $\rightarrow$ just counts

Introduction

We assume each bigram is made up of a **headword** and a **modifier**

Type	Headword	Modifier
A-N	N	A
S-V	V	S
V-O	V	O

System description

Three configurations:

- Two configurations that use cosine similarity measures in two different ways (configurations 1 and 2)
- One configuration that attempts to address the issue of polysemy (configuration 3)

Conf 1: Average of cosine similarity measures

Build word vectors for:

- Modifier $\mathbf{W}(x)$
- Headword $\mathbf{W}(y)$
- Bigram $\mathbf{W}(x\,y)$

Compositionality score for Configuration 1

$$c_1 = \frac{1}{2}\left[\begin{array}{c} \cos\left(\mathbf{W}\left(x\,y\right), \mathbf{W}\left(x\right)\right) \\ + \cos\left(\mathbf{W}\left(x\,y\right), \mathbf{W}\left(y\right)\right) \end{array}\right] \tag{1}$$

Conf 2: Headword in bigram vs not in bigram

In this configuration we want to look at:

- Contexts where the modifier and the headword form a bigram
- Contexts where the headword occurs but does not form a bigram with the modifier

red herring	Indeed, reflexive practice in the arts is a **red herring**, not because it doesn't exist, but because all practice is inherently reflexive.
$\overline{red}$ *herring*	Peterhead enjoys an increasingly important role in the trade of pelagic species of **herring** and mackerel, particularly with the processing plant at Albert Quay.

Sentences taken from the UK WaC corpus.

Conf 2: Headword in bigram vs not in bigram

Build word vectors for:

- Headword y when forming a bigram with modifier x: $\mathbf{W}^{\mathbf{x}}(y)$
- Headword y when **not** forming a bigram with modifier x: $\mathbf{W}^{\bar{\mathbf{x}}}(y)$

Compositionality score for Configuration 2

$$c_2 = \cos\left(\mathbf{W}^{\mathbf{x}}(y), \mathbf{W}^{\bar{\mathbf{x}}}(y)\right) \tag{2}$$

Conf 3: Cluster potential bigram senses

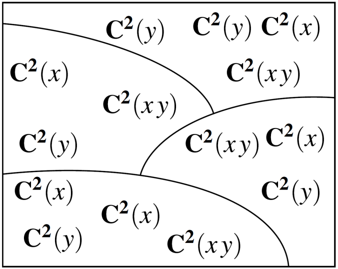
Intuition

Different senses of a bigram might have different degrees of compositionality. E.g.:

1. Two cans of soup for the price of one is such a **great deal**!
2. The tsunami caused a **great deal** of damage to the country's infrastructure.

Conf 3: Cluster potential bigram senses

- Cluster occurrences of headword y, modifier x and bigram $x\,y$ via second-order context vectors



- Each cluster could represent a different sense of the bigram
- If we knew what cluster represents the bigram sense seen by human annotators, we could compute compositionality score from the sub-corpus represented by that cluster only.

Conf 3: Cluster potential bigram senses

- But since we do not know what sense is used, we choose to compute the compositionality score as a weighted average from each cluster → a *polysemy-enhanced* version of Conf 1
- For each cluster k build the word vectors:
  - $\mathbf{W}_k(x\ y)$ for the bigram
  - $\mathbf{W}_k(x)$ for the modifier
  - $\mathbf{W}_k(y)$ for the headword

Compositionality score for Configuration 3

$$c_3 = \sum_{k=1}^{K} \frac{\|k\|}{N} \frac{1}{2} \begin{bmatrix} \cos(\mathbf{W}_k(x\ y), \mathbf{W}_k(x)) \\ +\cos(\mathbf{W}_k(x\ y), \mathbf{W}_k(y)) \end{bmatrix} \quad (3)$$

where $\|k\|$ is the number of contexts in cluster k and N is the total number of contexts across all clusters.

Results and discussion

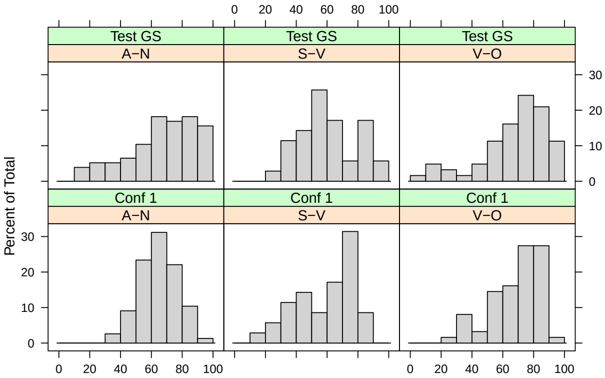
C	Average diffs (numeric)				Precision (coarse)			
	ALL	**A-N**	**S-V**	**V-O**	**ALL**	**A-N**	**S-V**	**V-O**
1	**17.95**	**18.56**	20.80	**15.58**	53.4	**63.5**	19.2	62.5
2	18.35	19.62	**20.20**	15.73	**54.2**	**63.5**	19.2	**65.0**
3	25.59	24.16	32.04	23.73	44.9	40.4	**42.3**	52.5
R	32.82	34.57	29.83	32.34	29.7	28.8	30.0	30.8

- Conf1 and Conf 2 show very similar performance
- Disappointingly, Conf 3 —the polysemy enhanced version of conf 1— did much **worse**
- S-V came out worse than A-N and V-O

Results and discussion

Gold Standard

System Conf 1



	ALL	A-N	S-V	V-O	ALL	A-N	S-V	V-O
A	16.86	17.73	15.54	16.52	58.5	65.4	34.6	65.0

Future work

- Investigate effects of weighting schemes (IDF and others)
- Similarity measures other than cosine
- Further research into the role played by context in determining the compositionality of a bigram
  - In configuration 2, involve modifier in computation of compositionality score
  - In configuration 3, create separate clustering spaces for bigram, headword and modifier
  - Explore other ways of clustering

Thank you for your attention! Questions?

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

References I

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta.
2009.
The WaCky wide web: a collection of very large linguistically processed web-crawled corpora.
Language Resources and Evaluation, 43(3):209–226, February.

Chris Biemann and Eugenie Giesbrecht.
2011.
Distributional Semantics and Compositionality 2011: Shared Task Description and Results.
In *Proceedings of the Distributional Semantics and Compositionality workshop (DISCo 2011) in conjunction with ACL 2011*, Portland, Oregon.

References II

Diana McCarthy, Bill Keller, and John Carroll.
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*, pages 73–80, Sapporo. Association for Computational Linguistics.

Ted Pedersen.
2010.
Duluth-WSI: SenseClusters applied to the sense induction task of SemEval-2.

References III

In *Proceedings of the 5th International Workshop on Semantic Evaluation*, number July, pages 363–366, Uppsala, Sweden. Association for Computational Linguistics.

Amruta Purandare and Ted Pedersen.
2004.
Word sense discrimination by clustering contexts in vector and similarity spaces.
Proceedings of the Conference on Computational Natural Language Learning, pages 41–48.

Hinrich Schütze.
1998.
Automatic word sense discrimination.
Computational Linguistics, 24(1):97–123.

Appendix A: Preliminary definitions

Definitions

First-order context vector

$$\mathbf{C}^1(p)(w) = \sum_{\substack{p' \neq p \\ p-10 \leq p' \\ p' \leq p+10}} (1 \text{ if } w = \mathrm{doc}(p'), \text{ else } 0) \tag{4}$$

Word (co-occurrence) vector

$$\mathbf{W}(w) = \sum_{p} (1 \text{ if } w = \mathrm{doc}(p), \text{ else } 0) \cdot \mathbf{C}^1(p) \tag{5}$$

Appendix A: Preliminary definitions

Definitions

Second-order context vector

$$\mathbf{C}^2(p) = \sum_{\substack{p' \neq p \\ p-10 \leq p' \\ p' \leq p+10}} \mathbf{W}(\mathrm{doc}(p)) \tag{6}$$

Vectors based on work by Schütze (1998)

Appendix A: Preliminary definitions

Generalisation to MWEs:

Single token: *make*

They	will	*make*	a	decision	based	on	...
$p-2$	$p-1$	p	$p+1$	$p+2$	$p+3$	$p+4$	...

MWE: *make decision*

They	will	*make a decision*	based	on	...
$p-2$	$p-1$	p	$p+1$	$p+2$	...

- Up to 3 intervening words allowed.

Appendix A: Preliminary definitions

- Similarity measure between vectors done via cosine, defined in the standard way:

Definition

$$\cos(\mathbf{v},\mathbf{w}) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \tag{7}$$

Appendix A: Conf 2 Word vector definitions

Definitions

Headword vector forming a bigram with x:

$$\mathbf{W}^{x}(y) = \sum_{p} (1 \,\mathrm{if}\, \begin{smallmatrix} \mathrm{doc}(p) = y \\ \mathrm{coll}(p, x) \end{smallmatrix}, \mathrm{else}\, 0) \cdot \mathbf{C}^{1}(p) \tag{8}$$

Headword vector not forming a bigram with x:

$$\mathbf{W}^{\bar{x}}(y) = \sum_{p} (1 \,\mathrm{if}\, \begin{smallmatrix} \mathrm{doc}(p) = y \\ \neg \mathrm{coll}(p, x) \end{smallmatrix}, \mathrm{else}\, 0) \cdot \mathbf{C}^{1}(p) \tag{9}$$

where y is the headword and $\mathrm{coll}(p)$ is a Boolean function that determines whether the word at position p forms a bigram with modifier x.

Appendix B: Results and conclusion

	A-N	S-V	V-O
Instances	177,254	11,092	121,317
Avg intervening	0.0684	0.3867	0.4612

Table: *Some corpus statistics: the number of matched bigrams per subtype (**Instances**) and the average number of intervening words per subtype (**Avg intervening**).*

A-N	S-V	V-O
digital radio	future lie	add value
small island	government intend	address issue
hard copy	business need	help children
black hole	event occur	raise bar

Table: *A few bigram examples provided by organisers.*