Exemplar-based Word-Space Model for Compositionality Detection

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

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

[2]Lexical Computing Ltd., UK

DisCo Workshop, ACL, Portland
Jun 24 2011

THE UNIVERSITY of York

Notation

We use DH to mean:

DH-based cooccurrence vector captures the actual meaning of the compound reflected in the corpus. e.g. the actual vector for **TrafficLight**

We use Traffic⊕Light to mean:

the *computed* compositional meaning of the compound

THE UNIVERSITY *of York*

Compositionality Detection - One key idea

Relation between DH and $\oplus$

If the compound is compositional, DH- and $\oplus$-based vectors are *identical*.

Main idea

Similarity(DH-based meaning, $\oplus$-based meaning)

$=$

Degree of compositionality

THE UNIVERSITY *of York*

Current methods

Existing work Schone and Jurafsky (2001); Baldwin et al. (2003); Katz and Giesbrecht (2006); Giesbrecht (2009)

- if sim($\mathbf{V}^{w_1 w_2}$, $\mathbf{V}^{w_1} \oplus \mathbf{V}^{w_2}$) > γ, MWE is compositional
- Thus for compositional **RiverBank** :
  expect sim(**RiverBank**, **River**$\oplus$**Bank**) to be ***high***
- Similarly for non-compositional **SmokingGun** :
  expect sim(**SmokingGun**, **Smoking**$\oplus$**Gun**) to be ***low***

THE UNIVERSITY *of York*

Problems with current methods

- $\text{sim}(\mathbf{V}^{w_1 w_2}, \mathbf{V}^{w_1} \oplus \mathbf{V}^{w_2}) = \gamma$
- One common observation: γ **varies highly**
- Instead, with *noisy* vectors, we get:
  sim(**RiverBank**, **River**$\oplus$**Bank**) < sim(**SmokingGun**, **Smoking**$\oplus$**Gun**)

THE UNIVERSITY of York

Problems with current methods

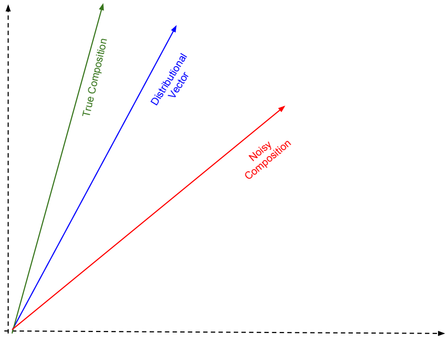
- Most current methods are based on using static prototype vectors
- Why static prototype vectors do not work?
- Noise due to *polysemy*

Reason: Polysemy

	police-n	photon-n	speed-n	car-n	soul-n
Traffic	142	0	293	347	1
Light	41	29	222	198	50
TrafficLight	5	0	13	48	0
a**Traffic** + b**Light**	5	0.8	14	15	1.4
Traffic * Light	5	0	56	59	0

THE UNIVERSITY *of York*

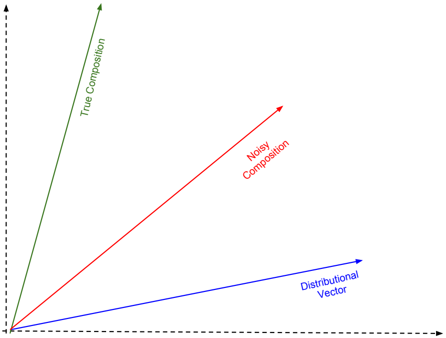
Why static prototype vectors do not work?



Compositional Multiword

THE UNIVERSITY of York

Why static prototype vectors do not work?



Non-compositional Multiword

THE UNIVERSITY of York

Problem: Polysemy

Due to polysemy of the constituent words, *compositionality functions* compose a *noisy vector* away from the *true compositional vector*.

Problem: Polysemy

Prototype Vectors are the problem

Currently most methods represent each word as a single vector i.e. a prototype vector for each word ***irrespective*** of its sense.

Problem: Polysemy

Prototype Vectors are the problem

Currently most methods represent each word as a single vector i.e. a prototype vector for each word ***irrespective*** of its sense.

Light occur in many contexts

- like quantum theory, optics, bulbs and traffic domain
- Not all contexts of *light* are relevant for *traffic light*
- **Light** is noisy $\Rightarrow$ **Traffic** $\oplus$ **Light** is noisy

THE UNIVERSITY *of York*

Concordance of *light*

the idea of a much worse prison ; where No	**light**	, but rather darkness visible , Served
tabernacle-work over the stalls carved in a	**light**	and elegant manner . St. John 's , which
There was a general , unrelieved , dull	**light**	; so that , unless when looking at your
He half opened one of them , and as the	**light**	poured in , looked round with mournful
beauty heightened by the aid of brilliant	**lights**	, of costly jewels , and all the pride
use the cycle paths and have good bright	**lights**	, then you should have no problems . Bus
. I think it puts business in a very bad	**light**	. Alan Sugar does everyone a great disservice
framework , Tati became an influential guiding	**light**	for the generations of comedians and filmmakers
morning - it 's night It 's dark - it 's	**light**	It 's raining - it 's sunny life 's serious
or feeling low M Baird 167 The Northern	**lights**	and Mackie 's means home sweet home to
This investigation is intended to bring to	**light**	some reasons for connecting the notion
24 hours a day and the propritators keep	**light**	security , perticulary a local rent a cop
form with the only pleasantness being the	**light**	white fluffy foam of the recently sumped
I thought his material had all seen the	**light**	of day . TK 's mentor , Henry Stone sent
1973 . I am amazed this has n't seen the	**light**	of day . It is wonderful , and definitely
legal issues , that it would never see the	**light**	of day . Frank has taken the reigns , as
quite rightly so . Bright and breezy and	**lit**	up a few dancefloors as well as receiving
requested . â ¢ Ideal as a clip-on book	**light**	â ¢ Reaches places other torches can
cover at the end of the arm . The Flexi	**Light**	requires two AAA batteries (not included
and clips in for compact storage . Flexi	**Light**	FAQ 's : Q) Hi , what bulb should I use

THE UNIVERSITY *of York*

Solution: Dynamic Prototypes using Exemplar based Models

Static prototype vectors are noisy

A need for a better representation of meaning

Solution: Dynamic Prototypes using Exemplar based Models

Static prototype vectors are noisy
A need for a better representation of meaning

Exemplar-based Word Space Model

- Select (examples) exemplars of *light* which have similar context to *traffic*
- Prune out the irrelevant exemplars
- Use selected exemplars to build the **Dynamic Prototype** $\textbf{Light}_{\textbf{Traffic}}$
- Exemplar based Models (Smith and Medin, 1981; Erk and Padó, 2010)

THE UNIVERSITY of York

Solution: Dynamic Prototypes using Exemplar based Models

Static prototype vectors are noisy
A need for a better representation of meaning

Exemplar-based Word Space Model

- Select (examples) exemplars of *light* which have similar context to *traffic*
- Prune out the irrelevant exemplars
- Use selected exemplars to build the **Dynamic Prototype** $\mathbf{Light_{Traffic}}$
- Exemplar based Models (Smith and Medin, 1981; Erk and Padó, 2010)

Dynamic Prototypes

- $\mathbf{Light_{Traffic}}$ represents dynamic vector of *light* relative to *traffic*
- $\mathbf{Traffic_{Light}} \oplus \mathbf{Light_{Traffic}}$ is closer to true compositional meaning than $\mathbf{Traffic} \oplus \mathbf{Light}$
- Others: Static Multi Prototypes (Reisinger and Mooney, 2010; Korkontzelos and Manandhar, 2009)

Building $\textbf{Light}_{\textbf{Traffic}}$

the idea of a much worse prison ; where No	light	, but rather darkness visible , Served
tabernacle-work over the stalls carved in a	light	and elegant manner . St. John 's , which
There was a general , unrelieved , dull	light	; so that , unless when looking at your
He half opened one of them , and as the	light	poured in , looked round with mournful
beauty heightened by the aid of brilliant	lights	, of costly jewels , and all the pride
use the cycle paths and have good bright	lights	, then you should have no problems . Bus
. I think it puts business in a very bad	light	. Alan Sugar does everyone a great disservice
framework , Tati became an influential guiding	light	for the generations of comedians and filmmakers
morning - it 's night It 's dark - it 's	light	It 's raining - it 's sunny life 's serious
or feeling low M Baird 167 The Northern	lights	and Mackie 's means home sweet home to
This investigation is intended to bring to	light	some reasons for connecting the notion
24 hours a day and the propritators keep	light	security , perticulary a local rent a cop
form with the only pleasantness being the	light	white fluffy foam of the recently sumped
I thought his material had all seen the	light	of day . TK 's mentor , Henry Stone sent
1973 . I am amazed this has n't seen the	light	of day . It is wonderful , and definitely
legal issues , that it would never see the	light	of day . Frank has taken the reigns , as
quite rightly so . Bright and breezy and	lit	up a few dancefloors as well as receiving
requested . â ¢ Ideal as a clip-on book	light	â ¢ Reaches places other torches can
cover at the end of the arm . The Flexi	Light	requires two AAA batteries (not included
and clips in for compact storage . Flexi	Light	FAQ 's : Q) Hi , what bulb should I use

for each e in E_{light}: $score(e|traffic) = e \cdot c + e \cdot s$

- E_{light} are the set of exemplars of light
- **e** is the exemplar of *light*
- **c** is the (static) co-occurrence vector of *traffic*
- **s** is the distributional similar neighbours of *traffic*

Cooccurrences of *traffic*

object_of	18950	1.4
divert	198	7.78
slow	212	7.58
block	319	7.5
generate	711	7.34
speed	149	7.25
motorise	93	7.17
encrypt	83	6.79
rout	78	6.76
direct	218	6.65
calm	75	6.5
congest	55	6.43
wheel	55	6.28
redirect	52	6.24
monitor	260	6.21
increase	985	6.18
drive	407	6.12
stop	330	6.1
reduce	698	6.08
induce	83	6.01

and/or	10005	0.8
pedestrian	157	7.56
congestion	134	6.92
pollution	168	6.42
noise	188	5.8
parking	173	5.57
freight	39	5.45
roadwork	14	5.26
traffic	190	5.17
transportation	32	5.09
highway	38	5.05
passenger	99	4.87
fume	14	4.73
pedestrianisation	8	4.63
lorry	20	4.53
commuter	13	4.49
motorway	20	4.38
transport	130	4.33
TCP	11	4.33
road	240	4.15

pp_along-i	95	5.2
road	25	0.92
route	14	0.83
street	9	0.62

pp_onto-i	64	4.6
road	25	0.92
route	9	0.2

pp_off-i	54	4.4
road	31	1.23

pp_through-i	305	2.8
firewall	10	4.35
village	46	2.74
port	18	2.7
tunnel	8	2.69

n_modifier	23201	2.5
freight	512	8.63
road	5520	8.63
air	2248	8.09
passenger	708	7.54
rush	211	7.45
commuter	134	7.09
motor	360	6.94
rail	240	6.27
network	874	6.0
motorway	85	5.96
good	350	5.85
lorry	73	5.85
Internet	492	5.83
coal	117	5.75
multicast	37	5.63
tourist	106	5.48
barge	40	5.27
data	143	5.25
container	72	5.23

- cooccurrence vector of *traffic* is computed using logDice Curran (2003)
- can substitute your favourite method

Distributionally similar words to *traffic*

Lemma	Score	Freq
transport	0.362	134717
road	0.339	324641
train	0.336	114514
vehicle	0.331	160671
bus	0.322	131884
route	0.312	168121
network	0.311	262162
trade	0.308	165216
market	0.307	379176
travel	0.306	115459
communication	0.3	171501
flow	0.299	77846
station	0.295	175788
operation	0.295	198053
car	0.293	419404
access	0.289	376109
business	0.287	700710
speed	0.286	146544
sale	0.285	239489

- Not only context words of *traffic* but also words distributionally similar to *traffic* are useful
- Computed using method described in Rychlý and Kilgarriff (2007)
- Can use another method

THE UNIVERSITY *of York*

Constructing Dynamic Prototype Vector for **Light**$_{\textbf{Traffic}}$

Ranked exemplars of *light*

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0
'bright-j': 1.0, 'day-n': 1.0

THE UNIVERSITY *of York*

Constructing Dynamic Prototype Vector for **Light**$_{\textbf{Traffic}}$

Ranked exemplars of *light*

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0
'bright-j': 1.0, 'day-n': 1.0

- **Light**$_{\textbf{Traffic}}$ is built by from the top n % exemplars of *light*
- Single prototype vector for **Light**$_{\textbf{Traffic}}$
- Re-weight features using $\frac{p(f|w)}{p(f)}$
- Similarly **Traffic**$_{\textbf{Light}}$ is built

THE UNIVERSITY *of York*

Constructing Dynamic Prototype Vector for **Light$_{\textbf{Traffic}}$**

Ranked exemplars of *light*

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0
'bright-j': 1.0, 'day-n': 1.0

- **Light$_{\textbf{Traffic}}$** is built by from the top n % exemplars of *light*
- Single prototype vector for **Light$_{\textbf{Traffic}}$**
- Re-weight features using $\frac{p(f|w)}{p(f)}$
- Similarly **Traffic$_{\textbf{Light}}$** is built

THE UNIVERSITY of York

DisCo 2011 Shared Task (Biemann and Giesbrecht, 2011)

- Phrases consist of two lemmas and come in three grammatical relations:
  - ADJ_NN: adjective modifying a noun
  - V_SUBJ: noun as a subject of a verb
  - V_OBJ: noun as an object of a verb
- For each phrase, 4 Amazon Mechanical Turkers annotate the data
  - Each score in the range 0-10 for compositionality
  - 4-5 random sentences are presented to the annotator
- Final compositionality score is averaged over all the workers
- 0-25 as non-compositional, 38-62 as medium and >75 as compositional
- 40% training, 10% validation and 50% test

THE UNIVERSITY of York

DisCo 2011 Shared Task (Biemann and Giesbrecht, 2011)

Distribution within Coarse grained evaluation

- Training set (107 total)
  - low: 10
  - medium: 47
  - high: 76
- Test set (118 total)
  - low: 7
  - medium: 42
  - high: 69
- 58.5% (69/118) were “highly compositional”
- Always choosing ’high’ will give you 58.5% score
- Only one system was able to achieve this baseline

Computing coarse-grained values

0-25 : non-compositional, 38-62 : medium, >75 : compositional

- ADJ_NN:
  - blue chip: 11, non
  - great deal: 40, medium
  - stainless steel: 92, high
- V_SUBJ:
  - interest lie: 40, medium
  - women want: 81, high
- V_OBJ:
  - reinvent wheel: 5, non
  - put pressure: 44, medium
  - give advice: 86, high

Compositionality Score

$$
\begin{aligned}
Score\ \ \alpha(V^{w_1}, V^{w_2}) = a_0 &+ a_1.sim(V^{w_1 w_2}, V^{w_1}) \\
&+ a_2.sim(V^{w_1 w_2}, V^{w_2}) \\
&+ a_3.sim(V^{w_1 w_2}, V^{w_1} + V^{w_2}) \\
&+ a_4.sim(V^{w_1 w_2}, V^{w_1} * V^{w_2})
\end{aligned}
$$

- Use linear regression to estimate all a_i
- Estimate a_i's separately for each of ADJ_NN, V_SUBJ, V_OBJ
- Only a_3 and a_4 involve compositionality operators

THE UNIVERSITY of York

Our Shared Task System: Exm-Best

V_OBJ

- $\alpha(V_{OBJ}, OBJ_V)$
- Both the constituent words help each other in disambiguation

V_SUBJ

- $\alpha(V_{SUBJ}, SUBJ_V)$
- It is found a3=0, a4=0 i.e. using $\oplus$ doesn't help.

ADJ_NN

- $\alpha(ADJ_{NN}, NN)$
- Adjective fails in disambiguating the noun
- Hence switch to using static prototype for *NN*

Our Other Systems

Exm

- We use Dynamic prototypes for both the words
- None of the a_i's is taken to be 0
- V_OBJ: $\alpha(V_{OBJ}, OBJ_V)$
- V_SUBJ: $\alpha(V_{SUBJ}, SUBJ_V)$
- NN_ADJ: $\alpha(ADJ_{NN}, NN_{ADJ})$

Pro-Best

- We just use the prototypes (i.e. no exemplar selection)
- V_OBJ: $\alpha(V, OBJ)$
- V_SUBJ: $\alpha(V, SUBJ)$
- NN_ADJ: $\alpha(ADJ, NN)$

Dynamic weights in additive model

In the simple additive model a**Traffic** + b**Light**

- Mitchell and Lapata (2008) use static weights a= (say) 0.2, b= 0.8
- Guevara (2010) also use static weights. But A and B are matrices
- We use Dynamic Weights

  a= $\frac{sim(TrafficLight, Traffic)}{sim(TrafficLight, Traffic) + sim(TrafficLight, Light)}$ and

  b= $\frac{sim(TrafficLight, Light)}{sim(TrafficLight, Traffic) + sim(TrafficLight, Light)}$

- $sim(TrafficLight, Traffic) = 0.54$
- $sim(TrafficLight, Light) = 0.27$
- **Traffic** contributes more towards the meaning of **TrafficLight**

THE UNIVERSITY of York

Dynamic weights in additive model

In the simple additive model a**Traffic** + b**Light**

- Mitchell and Lapata (2008) use static weights a= (say) 0.2, b= 0.8
- Guevara (2010) also use static weights. But A and B are matrices
- We use Dynamic Weights

  a= $\frac{sim(TrafficLight, Traffic)}{sim(TrafficLight, Traffic) + sim(TrafficLight, Light)}$ and

  b= $\frac{sim(TrafficLight, Light)}{sim(TrafficLight, Traffic) + sim(TrafficLight, Light)}$

- $sim(TrafficLight, Traffic) = 0.54$
- $sim(TrafficLight, Light) = 0.27$
- **Traffic** contributes more towards the meaning of **TrafficLight**

- $sim(Student, StudentNurse_{Dist}) = 0.238$
- $sim(Nurse, StudentNurse_{Dist}) = 0.893$

Average Point Difference Scores

	en-all	en-ADJ-NN	en-SUBJ	en-OBJ
Rand-Base	32.82	34.57	29.83	32.34
Zero-Base	23.42	24.67	17.03	25.47
Exm-Best	16.51	15.19	**15.72**	18.6
Pro-Best	16.79	**14.62**	18.89	18.31
Exm	17.28	15.82	18.18	18.6
SharedTaskBest	**16.19**	14.93	21.64	**14.66**

Table: Average Point Difference Scores

Correlation Scores

	TotPrd	Spearman ρ	Kendalls τ
Rand-Base	174	0.02	0.02
Exm-Best	169	**0.35**	**0.24**
Pro-Best	169	0.33	0.23
Exm	169	0.26	0.18
SharedTaskNextBest	174	0.33	0.23

Table: Correlation Scores

Coarse Grained Accuracy

	en-all	en-ADJ-NN	en-SUBJ	en-OBJ
Rand-Base	0.297	0.288	0.308	0.30
Zero-Base	0.356	0.288	0.654	0.25
Most-Freq-Base	**0.585**	0.654	0.346	**0.65**
Exm-Best	0.576	0.692	**0.5**	0.475
Pro-Best	0.567	**0.731**	0.346	0.5
Exm	0.542	0.692	0.346	0.475
SharedTaskBest	**0.585**	0.654	0.385	0.625

Table: Coarse Grained Accuracy

THE UNIVERSITY of York

Final Words

- Biemann and Giesbrecht (2011) referred to our system Exm-Best as the *most robust system* among all the participating systems
- Polysemy is a problem for semantic composition
- Dynamic prototypes provide a mechanism to address polysemy
- However for this task:
  - Results are mixed and incomplete
  - Comparison with static multi-prototypes Korkontzelos and Manandhar (2009)
  - Unsupervised evaluation
  - Evaluation on noun-noun compounds

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.

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*.

Curran, J. R. (2003). From distributional to semantic similarity. Technical report, PhD Thesis, University of Edinburgh.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 92–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliography II

Giesbrecht, E. (2009). In search of semantic compositionality in vector spaces. In *Proceedings of the 17th International Conference on Conceptual Structures: Conceptual Structures: Leveraging Semantic Technologies*, ICCS '09, pages 173–184, Berlin, Heidelberg. Springer-Verlag.

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*, GEMS '10, pages 33–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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.

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.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *HLT-NAACL*, pages 109–117.

Rychlý, P. and Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 41–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliography IV

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *EMNLP '01*.

Smith, E. E. and Medin, D. L. (1981). *Categories and concepts / Edward E. Smith and Douglas L. Medin*. Harvard University Press, Cambridge, Mass. :.