Modelling Free Associations with Co-occurrence Data

Stefan Evert[1], Gabriella Lapesa[2]

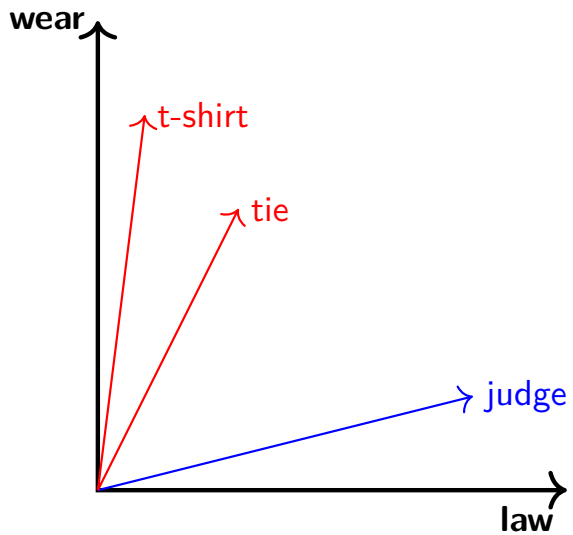
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]University of Stuttgart, Germany

24 March 2017

Distributional semantics

- **Distributional Hypothesis** (Harris 1954; Firth 1957)
  difference in meaning $\Longleftrightarrow$ difference in distribution
- **Distributional Semantic Models** (DSM)
  meaning of w = collocational profile (co-occurring words)

	law	wear
judge	8	2
t-shirt	1	8
tie	3	6



Distance between word vectors $\Longleftrightarrow$ semantic similarity

empirical correlate of the amount of shared meaning

Syntagmatic vs. paradigmatic relations

Definitions and general assumptions

- **Syntagmatic** $\Longleftrightarrow$ contiguity
  - Examples: {*dog*, *barks*}, {*dog*, *bone*}
  - Words appear together: 1st-order co-occurrence
  - Found in: collocational profiles, DSM dimensions

- **Paradigmatic** $\Longleftrightarrow$ interchangeability
  - Examples: {*book*, *volume*}, {*dog*, *animal*}
  - Words appear in similar contexts: 2nd-order co-occurrence
  - Usually semantically related
  - Found in: DSM nearest neighbours

However ...

DSM neighbourhoods include syntagmatically related words (collocates) if certain parameters are properly set, in particular if the context window is large enough (Lapesa *et al.* 2014).

DSM Evaluation

Standards and problems

A number of linguistic tasks targeting semantic relations:

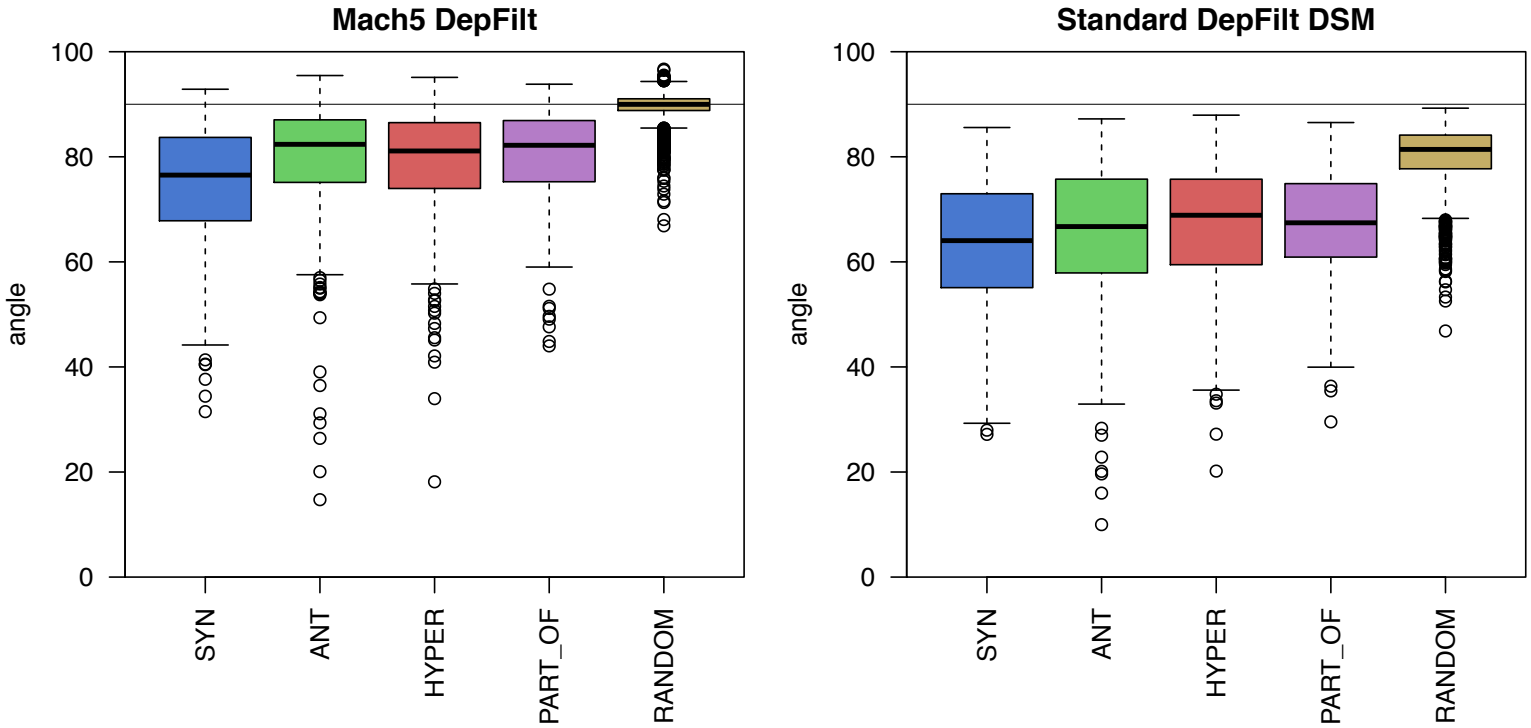
- Multiple-choice task (e.g. TOEFL synonymy)
- Classification of semantic relations
- Concept clustering (in practice, it targets co-hyponymy)

Two major problems:

- DSMs may exploit contingent properties of the task
  - random fillers as distractors ("controls")
    ➥ recognize random word pairs rather than semantic relations
  - clearly separated categories in noun clustering
  - typical superordinate-level words in hypernym detection
- Dataset size too small
  - e.g. 97.5% accuracy on 80 TOEFL items

DSM Evaluation

A disturbing result from the CogALex-V Shared Task on semantic relation identification



Mach5 DSM: $F_1 = 77.88\%$ for related vs. unrelated (Evert 2016)

- but weird parameters (best features: ranks 100k–150k)
- nearest neighbours often unintuitive

☞ DSM has learned to recognize random word pairs (at 90°)

Today . . .

- ☞ . . . we explore an alternative DSM evaluation task, based on **free association norms**
- ☞ . . . discuss the concrete steps in the construction of such a task
- ☞ . . . look at preliminary modelling results, discuss problems and further steps

Free associations

. . . a cue into the organization of the mental lexicon?

Which words come to your mind if you hear . . .

- whisky → gin, drink, scotch, bottle, soda
- giraffe → neck, animal, zoo, long, tall

- Hypotheses concerning the nature of the underlying process:
  - Result of learning-by-contiguity (James 1890)
    ☞ syntagmatic (1st-order)
  - Result of symbolic processes which make use of complex semantic structures (Clark 1970) ☞ paradigmatic (2nd-order)

Free associations

. . . a cue into the organization of the mental lexicon?

Which words come to your mind if you hear . . .

- whisky → gin, drink, scotch, bottle, soda
- giraffe → neck, animal, zoo, long, tall

- Hypotheses concerning the nature of the underlying process:
  - Result of learning-by-contiguity (James 1890)
    ☞ syntagmatic (1st-order)
  - Result of symbolic processes which make use of complex semantic structures (Clark 1970) ☞ paradigmatic (2nd-order)
- Large collections available
  - Edinburgh Associative Thesaurus (**EAT**)
    8210 stimuli, 100 subjects (Kiss *et al.* 1973)
  - University of South Florida Free Association Norms (**USF**)
    5019 stimuli, 6000 subjects (Nelson *et al.* 2004)

Why are free associations promising for DSM evaluation?

- Cognitively motivated, hence semantically plausible
- Large collections allow for robust evaluation
- Large collections allow for better selection of fillers, making the task more difficult but also more interesting:
  - words that are produced only once in response to a target stimulus can be used as plausible distractors
  - we use plausible first responses (taken from other stimuli) as random distractors
  - large pool of filler candidates allows controlling for frequency (marked differences in frequency bias vector similarities)

What do free association norms contain?

Syntagmatic vs. paradigmatic associates

- Brown and Berko (1960):
  - 74% of the responses of adults are paradigmatic
  - 72% of the responses of first grade children are syntagmatic
- Fitzpatrick (2007): 100 English stimuli (middle-frequency), 60 subjects, manual classification of free associates
  - Consecutive xy collocations: *significant–other*
  - Defining synonyms: *significant–important*
  - Conceptual associations: *coordination–driving*
  - Consecutive yz collocations: *instance–first*
  - Context-dependent synonyms: *label–name*
  - Lexical set: *pet–dog*

Mixture of syntagmatic and paradigmatic associates

. . . can we find both relations in corpus data?

Free associations & co-occurrence data

Previous work

- Wettler *et al.* (2005)
  - Data: subset of EAT (100 stimuli)
  - Task: prediction of the most frequent free associate
  - Model: first-order model, BNC, large window (20 words)
  - Result: human associative responses can be predicted from contiguities between words in language use (collocations)
    - ★ syntagmatic associates are captured with larger windows
- ESSLLI 2008 Shared Task
  - Data: subset of EAT (100 stimuli)
  - Tasks:
    - ★ discrimination btw. the first associates and hapaxes/random
    - ★ prediction of the most frequent associates
  - Result: first order models (collocations) are better than DSMs

Our new data sets

Preprocessing

- Annotate items in EAT and USF with part of speech information (most frequent POS in Web corpus ENCOW)
  - publicly available 10-billion-word Web corpus ➜ replicability
- Lemmatize with `morpha`, a robust morphological analyser
  - `http://users.sussex.ac.uk/~johnca/morph.html`
  - lemmatization of unknown words based on POS tag
- After having lemmatized ENCOW with `morpha`, annotate stimuli and responses with ENCOW lemma frequency

Our new data sets

Item selection

For each stimulus in EAT (8210) and USF (5019) select a:

- FIRST: the most frequent associate response
- HAPAX: a response generated for the target once
  - or twice for USF (hapax responses are omitted there)
  - if more HAPAX candidates are available, pick the one whose lemma frequency matches more closely that of FIRST
- RANDOM, by randomly picking a word which was among the top 25% associates *of another stimulus* (and produced at least 5 times). If possible:
  - match lemma frequency of RANDOM and FIRST
  - try to use each RANDOM only once

Multiwords, numbers, closed-class words, and other words that do not occur in ENCOW were discarded.

Experiments

Corpus based models (ENCOW):

- Context window: L2/R2 vs. L10/R10
- 1st-order models: $P(w_2|w_1)$ vs. PPMI vs. MI^2
- 2nd-order models: parameters from Lapesa and Evert (2014)

Tasks

- Task 1: multiple-choice
  - given a stimulus and a <FIRST, HAPAX, RANDOM> triple, determine which of the three candidates is FIRST.
  - baseline accuracy: 33.3%
- Task 2: open-vocabulary lexical access
  - given only a stimulus, predict the FIRST associate
  - problem: may learn to recognize "typical" free associates

First results

model	span	$n = 1973$ train	$n = 3863$ test
DSM	2	82.4%	81.8%
$\text{DSM}_{P=0}$	2	79.5%	78.6%
$P(w_2\|w_1)$	2	80.0%	80.3%
PPMI	2	77.6%	76.1%
MI^2	2	79.7%	79.9%
Combined	2	85.4%	84.6%
DSM	10	81.7%	82.5%
$\text{DSM}_{P=0}$	10	79.1%	79.8%
$P(w_2\|w_1)$	10	83.7%	84.0%
PPMI	10	81.9%	81.2%
MI^2	10	84.4%	83.9%
Combined	10	85.8%	85.4%
Combined	mix	86.4%	85.7%

- "training" data (1973 items) for exploration
- results validated on test set (3863 items)
- combine 1st- and 2nd-order models: $\text{DSM}_{P=0} + \text{MI}^2$
- with L10/R10 cooc span, 1st-order models are better
- only small improvement from combined model
- but cross-combination helps

Problems

The new EAT task isn't perfect either . . . yet

- Guessing POS from corpus doesn't always work
  - e.g. *fit*$_{\text{VERB}}$ → *epileptic*$_{\text{ADJ}}$, *aristocracy*$_{\text{NOUN}}$ → *lords*$_{\text{NAME}}$
  - but very few lemmatization errors
- Colloquialisms and British slang
  - e.g. *bod*$_{\text{NOUN}}$ → *person*$_{\text{NOUN}}$ (rare in written corpus)
  - but Web corpus has Welsh *bod* 'to be' mistagged as noun
  - DSM neighbours: *yn*, *hynny*, *mewn*, *hwn*, *gyfer*, . . . , 49. *bloke*, *techy*$_{\text{NOUN}}$, *nus*, *hon*, . . . , 60. *guy*, *mai*, *geezer*, . . .
  - another example is *mellow*$_{\text{ADJ}}$ → *yellow*$_{\text{ADJ}}$
- Model performs well $\not\Rightarrow$ neighbours = human associations
  - e.g. *firmness*$_{\text{NOUN}}$ → *hard*$_{\text{ADJ}}$ (MI^2 correct, DSM wrong)
  - but rank 897 vs. 1026 for *material*$_{\text{NOUN}}$ (DSM: 4797 vs. 772)

First conclusions

1^{st}-order = syntagmatic vs. 2^{nd}-order = paradigmatic?

- 1^{st}- and 2^{nd}-order models less complementary than expected
  ➥ relatively small benefit from combination
- But intuition not completely wrong (L2/R2):
  - DSM: *duckling* → *piglet, chick, duck, cygnet, hatchling, . . .*
  - MI^2: *duckling* → *ugly, chick, duck, swan, fluffy, roast, . . .*

- Possible explanation for the overlap under (many) simplifying assumptions (sentence span, raw cooc freqs, . . .)

term-sentence matrix: $\mathbf{F} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$

term-term matrix: $\mathbf{M} = \mathbf{F}\mathbf{F}^T = \mathbf{U} \cdot \mathbf{\Sigma}^2 \cdot \mathbf{V}^T$

☞ cosine similarity in $\mathbf{F} \approx 1^{st}$-order association MI^2

An alternative task based on association norms

The Cogalex-IV shared task (Rapp and Zock 2014)

Reverse multiword free association

- wheel, driver, bus, drive, lorry $\rightarrow$?
- away, minded, gone, present, ill $\rightarrow$?

- Data: subset of EAT (2000 stimuli training/test)
- Very challenging (best: 35% accuracy)
  - open-ended vocabulary (including inflected surface forms!)
  - need for integrating predictions of different stimuli
- And the winner was . . .
  - A system using first-order statistics to re-rank the output of a "standard" DSM (Ghosh *et al.* 2015)
- Our submission: several 1^{st}-order vs. 2^{nd}-order models
  - best 1^{st}-order: 27.7% / best 2^{nd}-order: 14.0%

References I

Brown, R. and Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, **31**, 1–14.

Clark, H.H. (1970). Word associations and linguistic theory. In J. Lyons (ed.), *New horizons in linguistics*. Harmondsworth: Penguin.

Evert, Stefan (2016). CogALex-V shared task: Mach5 – a traditional DSM approach to semantic relatedness. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 92–97, Osaka, Japan.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford.

Fitzpatrick, Tess (2007). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics*, **17**(3).

Ghosh, Urmi; Jain, Sambhav; Paul, Soma (2015). A two-stage approach for computing associative responses to a set of stimulus words. In Z. (eds.) (ed.), *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon,*.

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

James, W (1890). *The principles of psychology*. New York: Dover.

References II

Kiss, G.R; Armstrong, C.; Milroy; Piper, J. (1973). An associative thesaurus of english and its computer analysis. In R. B. Aitken and N. Hamilton-Smith (eds.), *The computer and literary studies*. Edinburgh University Pres.

Lapesa, Gabriella and Evert, Stefan (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.

Lapesa, Gabriella; Evert, Stefan; Schulte im Walde, Sabine (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland.

Nelson, Douglas L.; McEvoy, Cathy L.; Schreiber, Thomas A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*.

Rapp, Reinhard and Zock, Michael (2014). The cogalex-iv shared task on the lexical access problem. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon,*, pages 1–14. Zock/Rapp/Huang (eds.).

Wettler, Manfred; Rapp, Reinhard; Sedlmeier, Peter (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, **12**(2–3), 111–122.