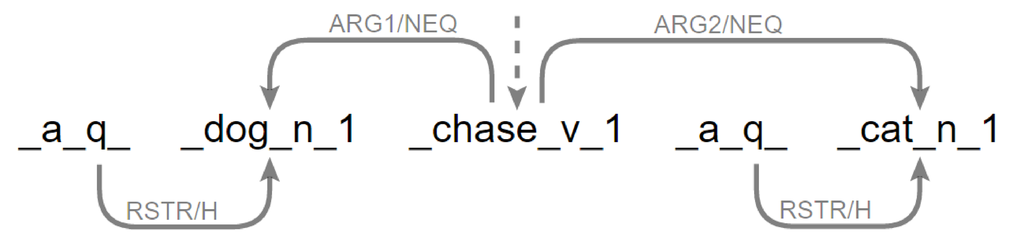
Investigating the effect of controlled context choice in distributional semantics

Alexander Kuhnle, University of Cambridge (aok25@cam.ac.uk)

Dependency Minimal Recursion Semantics

- DMRS semantic graphs as representation of an underlying MRS logical form structure
- Nodes roughly correspond to words and links to relations between them
- The (un)directed neighbour nodes are used as context for distributional semantic analysis



Distributional semantics: Why use DMRS graphs?

- Comparable to using syntactic parses [Padó & Lapata, 2007; Levy & Goldberg, 2014]
- DMRS used before for distributional analysis [Herbelot, 2013]
- Nodes/predicates represent the *semantic atoms* in the compositional structure of a sentence
  - Words are *lemmatised* and inflectional information is kept separately
  - No bijective correspondence to words, which in contrast is typical for syntactic parses
  - Combined representation of surface words in e.g. `_rely_v_on` or `_for+instance_a_1`
  - Words considered semantically empty are not represented, e.g. passive *"by"* or copula *"be"*
- Links represent the argument structure and scopal relationship between predicates
  - Link labels add *shallow semantic information* about the relation, e.g. `ARG1` or `RSTR` (but comparatively unspecific, as compared to `AGENT`, `PATIENT` etc)
  - Node neighbours result naturally in *more (syntacto-)semantic context*, as compared to a noisy word window approximation

Most important here: Link labels allow to control (to some degree) what kind of context is taken into account, e.g. for extracting only "agentive" `ARG1` information about verbs or the attributive/predicative usage of adjectives (see below)
⇒ Allows to analyse subtle distributional/semantic differences in various aspects of a word's usage, which word-window-based methods presumably struggle to detect

Top context similarity measures

Goal: Compare most significant, i.e. highest-valued, indices between context vectors

Average precision (AP)

$$\mathrm{AP}\big(\mathrm{top}, \mathrm{ranking}\big) = \frac{1}{\#(\mathrm{top})} \cdot \sum_i \delta\big(\mathrm{ranking}[i] \in \mathrm{top}\big) \cdot \frac{\#\big(\mathrm{ranking}[1\ldots i] \cap \mathrm{top}\big)}{i}$$

Reciprocal average precision (RAP)

$$\mathrm{RAP}_{k,l}(v, w) = \frac{1}{2} \cdot \Big(\mathrm{AP}\big(v[1\ldots k], w[1\ldots l]\big) + \mathrm{AP}\big(w[1\ldots k], v[1\ldots l]\big)\Big)$$

Jaccard Index (JI)

$$\mathrm{JI}_l(v, w) = \frac{\#\big(v[1\ldots l] \cap w[1\ldots l]\big)}{\#\big(v[1\ldots l] \cup w[1\ldots l]\big)}$$

(Vector entries (context) assumed to be sorted decreasingly w.r.t. context count/PPMI; in experiments: $k = 100$, $l = 1000$)

Data and pre-processing

- WikiWoods, a parsed Wikipedia 2008 snapshot (http://moin.delph-in.net/WikiWoods)
- Basic co-occurrence count / PPMI extraction with neighbour nodes as context (no dim.red.)
- Context: directed neighbours, but undirected neighbours for word similarity datasets
- Skip over 0.5%/2.0% of the least occurring predicates/words, resulting in 84,117/153,354 predicates/words for DMRS-/word-window-based system
- Positive pointwise mutual information (PPMI) with cds=0.75 and neg=3 [Levy et al., 2015]
- Sub-sampling with a threshold value of 10^{-5}
- Skipped context nodes in a DMRS graph are replaced by taking their neighbours instead

Bibliography

Sebastian Padó and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Ann Copestake. 2009. Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go. *Proceedings of the 12th Meeting of the European Chapter of the Association for Computational Linguistics*, 1–9.

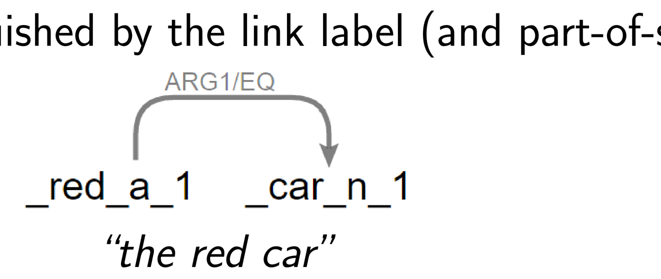
Aurelie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. *Proceedings of the 10th International Conference on Computational Semantics*, 321–327. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2:302–308. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Adjectives: Attributive vs predicative usage

Can be distinguished by the link label (and part-of-speech field of source/target node):



Adjective meaning can vary significantly, depending on its attributive or predicative usage:

- Similar meaning: *"the red car"* vs *"the car is red"*
- Different meaning: *"bad luck"* vs *"luck is bad"*, *"the sore loser"* vs *"the loser is sore"*
- Impossible construction: *"the former president"* vs **"the president is former"*

⇒ **Can we detect a difference in the distributional context of these two usages?**

Distributional difference in attr./pred. context

How similar are the top context entries?

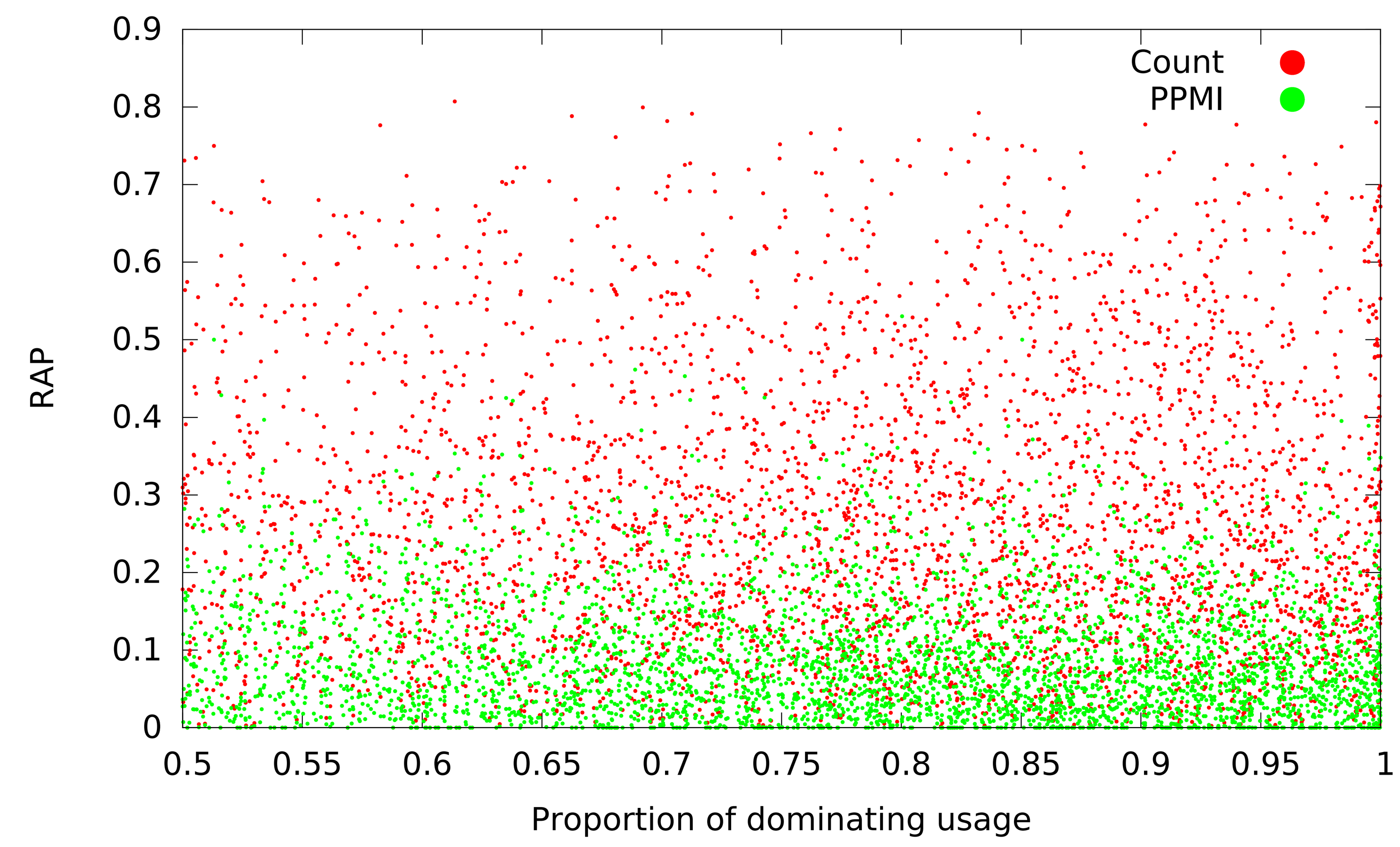
	Count vectors		PPMI vectors	
	Mean RAP	Mean JI	Mean RAP	Mean JI
Attributive vs combined	87.52%	74.91%	83.45%	73.85%
Predicative vs combined	44.13%	31.94%	28.37%	23.74%
Attributive vs predicative	25.37%	19.16%	8.63%	10.70%

⇒ **The contexts are definitely different, with PPMI intensifying the effect!**

Examples of top context entries (for count vectors; shared entries in bold)

- Attributive usage of *"good"*:
  friend player example **result time performance** finish **way award album record work thing condition** place **quality team** deal year luck **man life** film **school relation**
- Predicative usage of *"good"*: (RAP: 60.12%, JI: 36.05%)
  thing performance award quality life player result friend condition man people **record relation team time** song **work school** escape **way** relationship system situation game **album**
- Attributive usage of *"bad"*:
  weather luck guy news boy **thing condition** reputation girl religion company blood habit faith **behavior** idea day taste **time** temper publicity shape man start experience
- Predicative usage of *"bad"*: (RAP: 46.78%, JI: 30.72%)
  thing condition weather situation **time luck** effect deed quality action fortune performance people road business food relationship relation **behavior** life year side result injury film

Effect is independent of imbalanced attributive/predicative distribution



(Overall 4082 adjectives; context vectors evenly sampled according to the observed distribution, to neutralise frequency effects)

Sanity check: Word similarity datasets

Spearman based on cosine/RAP similarity	DMRS-	DMRS+	WW2	Levy (only cosine-based similarity)	LevyPPMI
SimLex-999	**44.19%** / **41.35%**	41.38% / 36.57%	32.90% / 26.37%	43.8%	39.3%
WordSim-353*	63.44% / 64.91%	**70.74%** / **67.85%**	60.47% / 57.56%	-	-
→ Similarity	77.44% / **79.10%**	**78.29%** / 74.80%	71.94% / 68.70%	79.3%	75.5%
→ Relatedness	47.87% / 49.66%	**63.53%** / **60.16%**	53.27% / 46.07%	69.7%	69.7%
MEN*	54.13% / 54.26%	**75.85%** / **72.56%**	74.91% / 70.21%	77.8%	74.5%
MTurk-287*	50.65% / 50.26%	63.31% / 54.55%	**65.57%** / **57.83%**	69.3%	68.6%
MTurk-771*	50.60% / 49.30%	61.89% / **54.37%**	**62.43%** / 53.22%	-	-
Rare-Words*	**42.25%** / **41.29%**	40.65% / 41.01%	36.53% / 34.82%	51.4%	46.2%
YP-130	**60.06%** / **59.04%**	54.72% / 45.02%	51.79% / 47.42%	-	-
Verb-143*	34.31% / 27.97%	**36.24%** / **32.53%**	28.77% / -3.64%	-	-

- `DMRS-/+`: Context as undirected neighbour predicates in DMRS graph (plus sub-sampling)
- `WW2`: Context as words in a 2-word-window within a surface sentence
- `Levy(PPMI)`: Best (PPMI-)system in [Levy et al., 2015] (trained on 2013 Wikipedia dump)

(Necessary pre-processing: Extraction of predicates corresponding to words in datasets; *: including manual lemmatisation)