Investigating the effect of controlled context choice in
distributional semantics
— DSALT 2016 Extended Abstract —

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

1 Introduction

Distributional semantics tries to capture aspects of meaning of a linguistic item by looking at its distributional properties in corpora, i.e. occurrence and co-occurrence with other items. At its heart lies the distributional hypothesis (Harris, 1954; Firth, 1957) – that there is a correspondence between similar meaning and similar distributional properties. Recently, a set of distributional techniques producing low-dimensional, so-called word embeddings (most notably Mikolov et al. (2013a), Mikolov et al. (2013b)) are very successful in various evaluations (Schnabel et al., 2015).

Apart from their often superior results, the appeal of distributional methods is also due to the fact that they are essentially unsupervised learning methods, i.e. they can be learned directly from raw, un-processed text. While it is in principle seen as a virtue that the respective method is supposed to learn to appropriately disregard or discount noise (w.r.t. semantics) originating either from unusual use of content words (e.g. fixed phrases like “with respect to”) or semantically largely vacuous words (passive, infinitive, copula, etc. involve the use of fixed, purely syntactic function words in English), it is common practice to apply in advance certain noise-reducing measures that have proven to significantly boost the performance of distributional techniques (Levy et al., 2015). The most common are discarding rare words, a pre-defined list of stop words considered irrelevant, or, similar to the latter, sub-sampling words with corpus frequency exceeding a certain threshold.

The aforementioned methods of pre-processing data basically are simple attempts to modify the input to exhibit a semantically more appropriate context. Here I argue that a more principled way of choosing appropriate context (as compared to the definition of context as words occurring within a certain word window) is a promising approach to improve word vectors for certain tasks, and in general is a way to control which aspect of meaning word vectors do (and are able to) capture. This idea is inspired by the central role of context in the distributional hypothesis, and different contexts were investigated before (e.g. Turney and Pan etal (2010) emphasise pair-pattern-matrices). Recently, Levy and Goldberg (2014) had some success in using dependency tree parses as the input format to extract context from. My approach uses the semantic graph representations of DMRS in a similar fashion (see also Herbelot (2013)), and postulates that these intrinsically noise-reducing and dependency-emphasising structures allow for a structural and flexible context choice.

2 DMRS semantic graphs

The semantic graphs of Dependency Minimal Recursion Semantics (DMRS) (Copestake, 2009) are a representation of Minimal Recursion Semantics (MRS) structures (Copestake et al., 2005). The English Resource Grammar (Flickinger, 2000) is a general-purpose and wide-coverage grammar for English, and with an appropriate parser (e.g. ACE) enables one to obtain MRS from sentences. More on how to convert to and work with DMRS can be found in Copestake et al. (2016).

DMRS semantic graphs have several properties that are potentially interesting for distributional context extraction (consequently based on (D)MRS predicates instead of words). First of all, a DMRS graph makes the argument structure of verbs/adjectives/etc. explicit in its links (ARG1/ARG2/etc. and not, for instance, AGENT/THEME/etc.).

The predicates in a DMRS graphs do in general not directly map to the words of the surface

1However, the input is usually at least tokenised, possibly lemmatised (to some degree) and/or filtered.
string. Instead, some words are considered semantically vacuous and hence not present as predicate, while other semantic elements are only implicitly present in a sentence but get explicifed in the semantic graph (e.g. implicit quantification for mass nouns, compared to explicit quantification via determiners). Additionally, certain fixed word compounds like “such as” or light verbs like “take on” are represented by a single predicate. DMRS parsing furthermore involves lemmatisation and some limited entity recognition for names, certain temporal/locational phrases, or numbers.

Potential disadvantages of using DMRS are systematic parse errors that will be reflected in the resulting semantic vectors, and the occasional inability to parse a sentence altogether (5–15%, Bender et al. (2015)). Furthermore, obtaining a sufficiently large corpus of parsed data is expensive – however, there is a parsed version of a Wikipedia snapshot of 2008, WikiWoods (Flickinger et al., 2010), which is publicly available[1] and which I will use for my experiments.

3 Adjectives: Attributive vs. predicative usage

For an initial evaluation of this concept that controlling the choice of context can bring out semantic subtleties, I looked at the differences between attributive (“the brown dog”) and predicative (“the dog is brown”) usage of adjectives. Adjective semantics can vary significantly (Kennedy, 2012; Richardson, 2013; Morzycki, 2015), and these two usages can be one source for variation: While the example of the brown dog is semantically equivalent, there can be divergences (“a sore loser” vs. “the loser is sore”), changes in meaning (“bad luck” vs. “luck is bad”), or even impossible constructions (“the former president” vs. “the president is former”).

The two different usages of adjectives can easily be distinguished in DMRS graphs – an adjective in attributive position is linked to the modified noun via an ARG1/NEQ link, whereas the link of a predicatively acting adjective is labeled with ARG1/NEQ. I analysed the 100 top context nouns (n, predicate POS field) of the 1000 most frequent adjectives (a, predicate POS field) in WikiWoods when restricting context to the respectively labeled links. Comparing the context nouns both usages of an adjective have in common, one gets an average overlap of around 54.9% when ranking context w.r.t. co-occurrence counts, which drops to 29.2% when ranking w.r.t. PPMI value[2] Adjectives like “available”, “historical” or “religious” exhibit a rather high overlap of >70%, while e.g. “true”, “certain” or “full” have a much less similar context of <30% between both usages.

Finally, below a list of the 25 most frequent context nouns for the adjectives “good” and “bad”. Note how the lists for “bad” differ much more than for “good” (6 vs. 18 shared words, in italics).

**Attributive usage for “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

**Attributive usage for “bad”:**
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

**Predicative usage for “good”:**
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 for “bad”:**
weather, luck, deer, quality, action, fortune, performance, people, road, business, food, relationship, relation, behavior, life, year, side, result, injury, film

4 Conclusion and future work

Even though this is just a first crude analysis which can be improved in many ways to yield better and more representative results, it reinforces the hypothesis that there are semantic effects in the argument structure of words that can be accounted for in distributional techniques by a more sophisticated context extraction.

On the one hand, I plan to continue the analysis of constructions that have an effect on the semantics of the words involved and how context distributions extracted from DMRS graphs can help to compare them. On the other hand, I want to improve on the method of how to construct and post-process distributional vectors based on DMRS by integrating performance-boosting techniques like sub-sampling of frequent words or singular value decomposition (Levy et al., 2015). One aim is to finally compare the resulting vectors to other word vectors like word2vec (Mikolov et al., 2013a) on standard vector evaluation tasks. However, I believe that the truly interesting aspect of DMRS word vectors lies in the fact that their construction can be controlled to some degree w.r.t. linguistic aspects (as I have shown for attributive/predicative adjectives) and hence capture semantic effects that word-window-based methods presumably struggle to detect.

---


[2]I so far did not account for the known problem of PPMI with rare words.
References


