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 items 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[1]. 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 treshold.

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 Pantel (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.

[1] However, the input is usually at least tokenised, possibly lemmatised (to some degree) and/or filtered.

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[2]) 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. as well as the (underspecified) scopal relationship of quantifiers/adverbs/etc. explicit in its links[3]. Context defined as neighbouring nodes/predicates consequently is expected to capture semantic association between predicates better. In particular, this also allows to restrict oneself to a subset of relations, as I will do in section 3.

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

[2] `http://sweaglesw.org/linguistics/ace/`

[3] The argument structure is not further specified semantically, i.e. links are annotated with `ARG1`/`ARG2`/etc. and not, for instance, `AGENT`/`THEME`/etc. .

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 explicified 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[4] 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; Reichard, 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/EQ` 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[5]) 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[6]. 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*

Predicative usage for *"good"*:
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 for *"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 for *"bad"*:
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

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.

[4] `http://moin.delph-in.net/WikiWoods`

[5] This POS tag also includes a few predicates primarily acting as adverbs, and I so far did not use a more sophisticated filter method (e.g. looking at the relative amount of context verbs to identify adverbial-only usage).

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

References

Emily M. Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

Ann Copestake, Guy Emerson, Michael W. Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszyńska. 2016. Resources for building applications with Dependency Minimal Recursion Semantics. In *To appear at LREC 2016*.

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

John R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.

Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. 2010. Wikiwoods: Syntacto-semantic annotation for english wikipedia. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

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

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

Chris Kennedy. 2012. Adjectives. In *Routledge Companion to Philosophy of Language*, pages 328–341. Routledge.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. 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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Marcin Morzycki. 2015. The lexical semantics of adjectives: more than just scales. In *Modification*, pages 13–87. Cambridge University Press.

Ulrich Reichard. 2013. Inference and grammar: Intersectivity, subsectivity, and phases. In *Microvariation, Minority Languages, Minimalism and Meaning: Proceedings of the Irish Network in Formal Linguistics*, pages 222–244. Cambridge Scholars.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 298–307.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.