ASRAEL - Acquisition of Semantic RelAtions bEtween Latin nouns

Konstantin Schulz
Humboldt-Universität zu Berlin

XXXabstractXXX

Contents

Introduction

This dissertation deals with different ways of detecting semantic relations in text corpora. As such, its focus is more on methodology than on analytical results. However, it is still important to consider the various implications of, e.g., the precise research question, which necessitates some important restrictions. Since the question of semantic relations in language cannot be exhaustively treated here, I will concentrate only on a subset of all the existing phenomena: polysemy and disambiguation, synonymy and antonymy, hypernymy and hyponymy, homonymy. This choice is motivated by the objective of utilizing the obtained insights for language learning, e.g. as the basis for pedagogical exercises. Thus, the detection of said relations is not an end in itself, but embedded into the context of how humans acquire language. Specifically, I hope that an improved integration of polysemy into the learning process can lead to a higher level of language skills:

> In cognitive linguistics, the word itself with its network of polysemous senses came to be regarded as a category in which the senses of the word (i.e. the members of the category) are related to each other by means of general cognitive principles such as metaphor, metonymy, generalization, specialization, and image-schema transformations. (Nerlich and Clarke 2003, p. 5)

If there is a parallel between polysemy and the way that humans think, it seems reasonable to exploit that connection for learning purposes. Unfortunately, this still leaves us with the lack of a clearly defined research object:

> Another problem arising from polysemy and homonymy is lexical ambiguity, and the precise relationship between polysemy, homonymy, ambiguity and vagueness is still an unresolved issue in lexical semantics. (Nerlich and Clarke 2003, p. 4)

It is therefore important to define the various semantic relations and distinguish between them as clearly as possible. In the case of, e.g., synonymy, a good operationalization is needed to retrieve the relevant instances from a corpus of text data:

> Polysemy requires the researcher to determine whether two usage events are identical or sufficiently similar to be considered a single sense, what the degree of similarity is between different senses, where to connect a sense to others in the network, and which sense(s) to recognize as prototypical one(s). [...] in addition, [linguists] have to decide what the differences are between the near-synonyms as well as what the relation is between semantically similar words in a domain. (Divjak and Gries 2009, p. 274)

In my study, I will follow the approach of distributional semantics, describing specific semantic relations between words in terms of their co-occurrences and typical contexts in a corpus. Previous research has been carried out using, e.g., association measures to detect synonymy:

> The value of distributional features $f_j^D(x,z)$ is determined so that it represents the degree of commonality of context c_j shared by the word pair (x, z). [...] The advantage of this feature construction is that, given the independence assumption between word x and z , the feature value is easily calculated as the simple sum of two corresponding pointwise mutual information weights as: $f_j^D(x,z) = PMI(x,c_j) + PMI(z,c_j)$ [...]. (Hagiwara, Ogawa, and Toyama 2009, p. 566)

One of the critical parts in that view is the term "word pair", which introduces the linguistically ill-defined concept of word into an otherwise seemingly precise mathematical formula. Thus, the usage of bare token-based PMI (Pointwise Mutual Information) for distributional questions may lead to severely skewed results:

> [...] strong collocation effects can influence the measurement of information negatively: it is

> an open question which phrases should be considered ‘words-with-spaces’ when building distributions. (Herbelot and Ganesalingam 2013, p. 444)

Therefore, corpora may have to be tokenized and segmented using more complex separator rules than "whitespace and punctuation" before they can serve as a basis for distributional analyses. Besides, polysemy itself may also become a problem for resolving other issues like synonymy, e.g. if two words are synonymous only in one of their possibly numerous senses:

> Some of the errors we observe may also be related to word senses. For instance, the word medium, to be found in the pair magazine – medium, can be synonymous with middle, clairvoyant or again mode of communication. In the sense of clairvoyant, it is clearly more specific than in the sense intended in the test pair. As distributions do not distinguish between senses, this will have an effect on our results. (Herbelot and Ganesalingam 2013, p. 444)

Since the solution to this problem may be to disambiguate word senses for every single token in a corpus, polysemy will be treated in this study not just as one of the analytical targets, but also as a methodological aspect that needs to be addressed before other kinds of analyses can be performed. Finally, some of the second-tier relations may have to be redefined when looked at from a distributional perspective, e.g. hypernymy: “Although beverage is an umbrella word for many various types of drinks, speakers of English use it in very particular contexts. So, distributionally, it is not a ‘general word’.” (Herbelot and Ganesalingam 2013, p. 443) In this view, other existing taxonomies that may have acted as models of hypernymy need to be replaced by a definition that reflects how actual usage contexts can limit our selection of possible candidates for hypernymy:

> [The Distributional Inclusion Hypothesis] states that more specific terms appear in a subset of the distributional contexts in which more general terms appear. So, animal can occur in all the contexts in which dog can occur, plus some contexts in which dog cannot – for instance, rights can be a typical cooccurrence for animal (e.g. “animal rights”), but not so much for dog (e.g. #“dog rights”). (Roller, Erk, and Boleda 2014, p. 443)

Unfortunately, this view may clash with the observations of Herbelot and Ganesalingam 2013:

- Hypernyms can have very specific usage contexts.
- Collocations may lead to unexpected results in the distribution of words.

Similar adjustments have to be made in treating other kinds of semantic relations. Therefore, many researchers apply preprocessing to their corpora before performing the actual analysis on them:

> It is worth noting that the skipgram model uses subsampling of common words, which is an optimization introduced to compensate for the power law distribution in common vocabularies. Also, the skipgram model controls for collocations by dampening the impact of frequent collocations. (Gyllensten, Ekgren, and Sahlgren 2019, p. 56f.)

However, this increases the danger of circularity, revealing only those results that we were already expecting in the first place. As an alternative, we should look for a model that is robust enough to detect semantic relations despite the confounding factors. The objective should be to fit the model to the base text, not the other way round.

Agile and Open Methodology

To achieve that, this study will follow an approach which is borrowed from computer science, i.e. the agile methodology for software development:

> [...] agile processes assume and encourage the alteration of requirements while the code is being written. As such, design cannot be a purely up-front activity to be completed before construction. Instead, design is a continuous activity that’s performed throughout the project. (Fowler and Highsmith 2001, p. 32)

Coding will be used to establish a proof of concept for this study by implementing the discussed models and testing them on real-world corpus data. Furthermore, the basic notion of incremental and iterative progress also applies to the way the models themselves are being created: The dissertation will see a range of approaches and considerations being applied for various purposes, which reflects my own thoughts about the detection of semantic relations growing from initial, very basic foundations (e.g. pure contingency tables) to more complex environments of semantic study. The application of agile methodology to linguistic research is motivated not just by the related coding activities, but also by the similarities between dissertations and software products: In this view, a dissertation can be seen as a product to be delivered to customers, i.e. the research community. The analogy is enhanced further by the overlapping quality factor of openness: Dissertations are necessarily products of open science in much the same way as open source software is the product of other communities sharing the desire for unrestricted public access to creative commons.

> We think that Open Knowledge comprises Open Software, Open Content, Open Science and Open Innovation. [...] Open Software owes its deepest roots to Open Access; Open Contents are related to open access to the educative, cultural or divulgative contents that are published under a non restrictive license that allows copy and distribution, but also the right to modify works. Open Science is devoted to the open access to scientific contents, while Open Innovation transfers the Open Access principles to the enterprise production world, which is actually indispensable for the enhancement of University-Enterprises relationships. (García-Penalvo, García de Figuerola, and Merlo 2010, p. 518)

To demonstrate the viability of this transfer, the dissertation is split into various modules, separating, e.g., the bibliography from the rest of the documents and employing cross-references to put it all together for presentational purposes. These modules are, to some extent, directly visible in the sense of multiple source files that may be compiled to form a more complex document. Also, they are all publicly accessible at and subject to version control, thus enabling readers to trace the resulting thoughts back to their very roots.

Base Text

...

References

Divjak, Dagmar and Stefan Th Gries (2009). “Corpus-Based Cognitive Semantics: A Contrastive Study of Phasal Verbs in English and Russian”. In: *Studies in cognitive corpus linguistics*, pp. 273–296.

Fowler, Martin and Jim Highsmith (2001). “The Agile Manifesto”. In: *Software Development* 9.8, pp. 28–35.

García-Penalvo, Francisco J., Carlos García de Figuerola, and Jose A. Merlo (2010). “Open Knowledge Management in Higher Education”. In: *Online Information Review* 34.4, pp. 517–519.

Gyllensten, Amaru Cuba, Ariel Ekgren, and Magnus Sahlgren (2019). “R-Grams: Unsupervised Learning of Semantic Units in Natural Language”. en-us. In: *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pp. 52–62.

Hagiwara, Masato, Yasuhiro Ogawa, and Katsuhiko Toyama (2009). “Supervised Synonym Acquisition Using Distributional Features and Syntactic Patterns”. In: *Information and Media Technologies* 4.2, pp. 558–582.

Herbelot, Aurélie and Mohan Ganesalingam (2013). “Measuring Semantic Content in Distributional Vectors”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 440–445.

Nerlich, Brigitte and David D. Clarke (2003). “Polysemy and Flexibility: Introduction and Overview”. In: *Polysemy. Flexible Patterns of Meaning in Mind and Language*. Ed. by Brigitte Nerlich et al. Berlin: Mouton de Gruyter, pp. 3–30.

Roller, Stephen, Katrin Erk, and Gemma Boleda (2014). “Inclusive yet Selective: Supervised Distributional Hypernymy Detection”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1025–1036.