Treating Similarity with Respect: How to Evaluate Models of Meaning?

Dmitrijs Milajevs
Queen Mary University of London
London, UK
d.milajevs@qmul.ac.uk

Sascha Griffiths
Queen Mary University of London
London, UK
s.griffiths@qmul.ac.uk

Abstract

Similarity is a core notion that is used in psychology, theoretical and computational linguistics. The similarity datasets that come from the two fields differ in design: psychological datasets are focused around a certain topic such as fruit names; linguistic datasets contain words from various categories. The later makes humans assign low similarity scores to the words that have nothing in common and to the words that have contrast in meaning, making similarity scores ambiguous. In this work we discuss the similarity collection procedure for a multi-category dataset that avoids score ambiguity and suggest changes to the evaluation procedure to reflect the insights of psychological literature for word, phrase and sentence similarity. We suggest to ask humans to provide a list of commonalities and differences instead of numerical similarity scores and employ the structure of human judgements beyond pairwise similarity. We believe that the proposed approach will give rise to datasets that test meaning representation models more thoroughly with respect to the human treatment of similarity.

1 Introduction

Similarity is the degree of resemblance between two objects or events (Hahn, 2014) and plays a crucial role in psychological theories of knowledge and behaviour, where it is used to explain such phenomena as classification and conceptualisation. *Fruit* is a *category* because it is a practical generalisation. Fruits are sweet and constitute deserts, so when one is presented with an unseen fruit, one can hypothesise that it is served toward the end of a dinner.

Generalisations are extremely powerful in describing a language as well. The verb *runs* requires its subject to be singular. *Verb*, *subject* and *singular* are categories that are used to describe English grammar. When one encounters an unknown word and is told that it is a verb, one will immediately have an idea about how to use it assuming that it is used similarly to other English verbs.

The semantic formalisation of similarity is based on two ideas. The occurrence pattern of a word *defines* its meaning (Firth, 1957), while the difference in occurrence between two words *quantifies* the difference in their meaning (Harris, 1970). From a computational perspective, this motivates and guides development of similarity components that are embedded into natural language processing systems that deal with tasks such as word sense disambiguation (Schütze, 1998), information retrieval (Salton et al., 1975; Milajevs et al., 2015), machine translation (Dagan et al., 1993), dependency parsing (Hermann and Blunsom, 2013; Andreas and Klein, 2014), and dialogue act tagging (Kalchbrenner and Blunsom, 2013; Milajevs and Purver, 2014).

Because it is difficult to measure performance of a single (similarity) component in a pipeline, datasets that focus on similarity are popular among computational linguists. Apart from a pragmatic attempt to alleviate the problems of evaluating similarity components, these datasets serve as an empirical test of the hypotheses of Firth and Harris, bringing together our understanding of human mind, language and technology.

Two datasets, namely MEN (Bruni et al., 2012) and SimLex-999 (Hill et al., 2015), are currently widely used. They are designed especially for meaning representation evaluation and surpass datasets stemming from psychology (Tversky and Hutchinson, 1986), information retrieval (Finkelstein et al., 2002) and computational linguistics (Rubenstein and Goodenough, 1965) in quantity by having more entries and, in case of SimLex-999, attention to the evaluated relation by distinguishing similarity from relatedness. The datasets provide similarity (relatedness) scores between word pairs.

In contrast to linguistic datasets which contain randomly paired words from a broad selection, datasets that come from psychology contain entries that belong to a single category such as *verbs of judging* (Fillenbaum and Rapoport, 1974) or *animal terms* (Henley, 1969). The reason for category oriented similarity studies is that "stimuli can only be compared in so far as they have already been categorised as identical, alike, or equivalent at some higher level of abstraction" (Turner et al., 1987). Moreover, because of the *extension effect* (Medin et al., 1993), the similarity of two entries in a context is less than the similarity between the same entries when the context is extended. "For example, *black* and *white* received a similarity rating of 2.2 when presented by themselves; this rating increased to 4.0 when *black* was simultaneously compared with *white* and *red* (*red* only increased 4.2 to 4.9)" (Medin et al., 1993). In the first case *black* and *white* are more dissimilar because they are located on the extremes of the greyscale, but in the presence of *red* they become more similar because they are both monochromes.

Both MEN and SimLex-999 provide pairs that do not share any similarity to control for false positives, and they do not control for the comparison scale. This makes similarity judgements ambiguous as it is not clear what low similarity values mean: incompatible notions or contrast in meaning. SimLex-999 assigns low similarity scores to the incompatible pairs (0.48, *trick* and *size*) and to antonymy (0.55, *smart* and *dumb*), but *smart* and *dumb* have relatively much more in common than *trick* and *size*!

The present contribution investigates how a similarity dataset with multiple categories should be built and considers what sentence similarity means in this context.

2 Dataset Construction

Human similarity judgements To build a similarity dataset which contains non-overlapping categories, one needs to avoid comparison of incompatible pairs. However, that itself requires a priori knowledge of item similarity or belongingness to a category, making the problem circular. To get out of this vicious circle, one might erroneously refer to an already existing taxonomy such as WordNet (Miller, 1995). But in case of similarity, as Turney (2012) points out, categories that emerge from similarity judgements are different from taxonomies. For example, *traffic* and *water* might be considered to be similar because of a functional similarity exploited in hydrodynamic models of traffic, but their lowest common ancestor in WordNet is *entity*.

Since there is no way of deciding upfront whether there is a similarity relation between two words, the data collection procedure needs to test for both: relation existence and its strength. Numerical values, as has been shown in the introduction, do not fit this role, because of ambiguity. One way of avoiding the issue is to avoid asking humans for numerical similarity judgements, but instead to ask them to list commonalities and differences between the objects. As one might expect, similarity scores correlate with the number of listed commonalities (Markman and Gentner, 1991; Markman and Gentner, 1996; Medin et al., 1993). For incompatible pairs, the commonality list should be empty, but the differences will enumerate properties that belong to one entity, but not another (Markman and Gentner, 1991; Medin et al., 1993).

The entries in the dataset So far we have proposed a similarity judgement collection method that is robust to incompatible pairings. It also naturally gives rise to categories, because the absence of a relation between two entries means the absence of a common category. It still needs to be decided what words to include to the dataset.

To get a list of words that constitute the dataset, one might think of categories such as *sports*, *fruits*, *vegetables*, *judging verbs*, *countries*, *colours* and so on. Note, that at this point its acceptable to think of categories, because later the arbitrary category assignments will be reevaluated. Once the list of categories is ready, each of them is populated with category instances, e.g. *plum*, *banana* and *lemon* are all *fruits*.

When the data is prepared, humans are asked to provide commonalities and differences between all pairs of every group. First, all expected similarities are judged, producing a dataset that can be seen as a merged version of category specific datasets. At this point, a good similarity model should provide meaning representation that are easily split to clusters: *fruit* members and *sport* members have to be separable.

Intra-category comparisons should be also performed, but because it is impractical to collect all possible pairwise judgements between the number of words of magnitude of hundreds, a reasonable sample should be taken. The intra-category comparisons will lead to unexpected category pairings, such as *food* that contains *vegetables* and *fruits*, so the sampling procedure might be directed by the discovery of compatible pairs: when a *banana* and *potato* are said to be similar, *fruits* and *vegetables* members should be more likely to be assessed.

Evaluation beyond proximity Human judgements validate the initial category assignment of items and provide new ones. If a category contains a superordinate, similarity judgements arrange category members around it (Tversky and Hutchinson, 1986). For example, similarity judgements given by humans arrange fruit names around the word *fruit* in such a way that it is their nearest neighbour, making *fruit* the *focal point* of the category of *fruits*.

As an additional evaluation method, the model should be able to retrieve focal points. Therefore, a precaution should be taken before human judgement collection. If possible, categories should contain a subordinate.

Similarity evaluation needs to focus on how well a model is able to recover human similarity intuitions expressed as groupings, possibly around their focal points. We propose to treat it as a soft multi-class clustering problem (White et al., 2015), where two entities belong to a same class if there is a similarity judgement for them (e.g. *apple* and *banana* are similar because they are *fruits*) and the strength is proportional to the number of such judgements, so we could express that *apple* is more a *fruit* than a *company*. Models also should arrange subordinates to focal points giving labels to the clusters.

3 Similarity in Context

In the previous section we focused on the properties of similarity, illustrating it with word-word examples. This section is dedicated to the nuances of similarity measurement between words, phrases and sentences. Here, we assume that similarity requires the agreement of the grammatical roles.

A noun phrase can be similar to a noun as in *female lion* and *lioness*, and to another noun phrases as in *yellow car* and *cheap taxi*. The same similarity principle can be applied to phrases as to words. In this case, similarity is measured in *context*, but is still comparison of the phrases' head words which meaning is modified by arguments they appear with (Kintsch, 2001; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Dinu and Lapata, 2010; Baroni and Zamparelli, 2010; Thater et al., 2011; Séaghdha and Korhonen, 2011). With verbs this idea can be applied to compare transitive verbs with intransitive. For example, *to cycle* is similar to *ride a bicycle*.

Sentential similarity might be treated as the similarity of the heads in the contexts. That is, the similarity between *sees* and *notices* in *John* ***sees*** *Mary* and *John* ***notices*** *a woman*. This approach abstracts away grammatical difference between the sentences and concentrates on semantics and fits the proposed model as the respect for the head, which is a lexical entity, has to be found.

4 Constructions and Language Sub-systems

The question of sentence similarity is more complex because sentences in many ways are different entities than words. Or are they? Linguistics has recently often pointed toward a continuum which exists between words and sentences (Jackendoff, 2012). Jackendoff and Pinker (2005), for example, point out that there is good evidence that "human memory must store linguistic expressions of all sizes." These linguistic expressions of variable size are often called *constructions*. Several computational approaches to constructions have been proposed (Gaspers et al., 2011; Chang et al., 2012), but to the authors' best knowledge they do not yet feature prominently in natural language processing.

Words, sentences and other constructions draw attention to states of affairs around us. This is both pragmatically and semantically speaking the case. As Tomasello (2009) points out, speakers of English can make sense of phrases like *"X floosed Y the Z."* and *"X was floosed by Y."* This is due to their similarity to sentences such as *"John gave Mary the book."* and *"Mary was kissed by John."* respectively. The point here is that some sentences are similar to others with respect to the functions they perform. For example, *"X floosed Y the Z."* is clearly a *"transfer of possession"* or *dative* (Bresnan et al., 2007).

Thus, the amount in which sentences are similar, at least to a certain extent, corresponds to the function of a given sentence. Tomasello (1998) points out that sentence-level constructions show prototype effects similar to those discussed above for lexical systems (e.g. colours). Consider the following sentences:

- *"John gave Mary the book."* is a example of an *Agent Causes Transfer* construction. These usually are build around words such as *give, pass, hand, toss, bring, etc.*

- *“John promised Mary the book.”* is a example of an *Conditional transfer* construction. These usually are build around words such as *promise, guarantee, owe, etc.*

As soon as one has such a prototype network, one can actually decide sentence similarity as one can say with respect to what sentences and utterances are similar.

However, these categories work on the semantic-grammatical level, and might be still handled by similarity in context as described in previous section. What about pragmatics? As Steels (2008) points out, sentences and words draw attention and do not always directly point or refer to entities and actions in the world. For example, he points to the fact that if a person asks another person to *“pass the wine”* they are actually asking for the *bottle*. The speaker just attracts attention to an object of perception in a given situation. There are several ways in which that sentence can both be *grammaticalized* and *lexicalized*. Dialogue act tags is another way of utterance categorisation, refer to the work of Kalchbrenner and Blunsom (2013) and Milajevs and Purver (2014).

If one conceptualises sentence similarity with respect to a discourse, then one might ask how different sentences fit in to such a discourse. Griffiths et al. (2015) tried to construct to versions of the same dialogue using a bottom-up method. They deconstructed a certain dialogue in a given domain—a receptionist scenario—into *greetings*, *directions* and *farewells*. They used a small custom made corpus for this purpose and created the two dialogues by having people rate the individual utterances by friendliness. The resulting two dialogues were surprisingly uneven. The dialogue was supposed to give instructions to a certain location within a building. The “friendly version” was very elaborated and consisted of several sentences:

(1) The questionnaire is located in room Q2-102. That is on the second floor. If you turn to your right and walk down the hallway. At the end of the floor you will find the stairs. Just walk up the stairs to the top floor and go through the fire door. The room is then straight ahead.

The sentence which served the same purpose in the neutral version was a fairly simple sentence:

(2) The questionnaire is located in Q2-102.

Often the same functions in dialogue can be performed by as little as one word or several phrases or even a complete story.

Steels (2010) introduces the idea of language sub-systems and language strategies. A language subsystem is a means of expressing certain related or similar meanings. Examples of such subsystems include:

- Lexical systems which express colours.
- Morphological devices to encode tenses.
- Usage of word order to express relations between agent and patient.

The later is an illustration of a language strategy. In English agent-patient relations are mainly encoded by syntax whereas German would use intonation and a combination or articles and case to convey the same information. Russian in contrast will use morphological devices for the same purpose.

Also, sentences are different from words because they are part of the discourse and it is easy to come up with a pair of sentences that are similar in a certain situation but do not share grammatical structure at all, for example *No* and *I’ve seen John eating them* are similar sentences if they are answers to the question *Do we have cookies?* Questions the sentences answer are valid respects for similarity explanation, as well as entailment, paraphrase (White et al., 2015) or spatial categories (Ritter et al., 2015). One approach to treat sentences on their own would be to encode the meaning of a sentence into a vector, in such a way that similar sentences are clustered together (Coecke et al., 2010; Baroni et al., 2014; Socher et al., 2012; Wieting et al., 2015; Hill et al., 2016).

5 Conclusion

In this contribution we discussed the notion of “similarity” from an interdisciplinary perspective. We contrasted properties of the similarity relation described in the field of psychology with the characteristics of similarity datasets used in computational linguistics. This lead to the recommendations on how to improve the later by removing low score ambiguity in a multi-category similarity dataset.

In the future, a multi-category similarity dataset should be build that allows evaluation of vector space models of meaning by not only measuring proximity between the points, but also their arrangement to clusters. The same ideas can be used to build phrase and sentence level datasets.

On a broader perspective, this work highlights psychological phenomena that being incorporated into the models of meaning are expected to improve their performance.

References

Jacob Andreas and Dan Klein. 2014. How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 822–827, Baltimore, Maryland, June. Association for Computational Linguistics.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. Predicting the dative alternation. pages 69–94. Royal Netherlands Academy of Arts and Sciences, Amsterdam.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 136–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nancy C Chang, Joachim De Beule, and Vanessa Micelli. 2012. Computational Construction Grammar: Comparing ECG and FCG. In Luc Steels, editor, *Computational Issues in Fluid Construction Grammar*, pages 259–288. Springer Verlag, Berlin.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 164–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1162–1172, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samuel Fillenbaum and Amnon Rapoport. 1974. Verbs of judging, judged: A case study. *Journal of Verbal Learning and Verbal Behavior*, 13(1):54 – 62.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, January.

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

Judith Gaspers, Philipp Cimiano, Sascha S Griffiths, and Britta Wrede. 2011. An unsupervised algorithm for the induction of constructions. In *2011 IEEE International Conference on Development and Learning (ICDL)*, pages 1–6. IEEE, aug.

Sascha Griffiths, Friederike Eyssel, Anja Philippsen, Christian Pietsch, and Sven Wachsmuth. 2015. Perception of Artificial Agents and Utterance Friendliness in Dialogue. In Maha Salem, Astrid Weiss, Paul Baxter, and Kerstin Dautenhahn, editors, *Proceedings of the Fourth Symposium on "New Frontiers in Human-Robot Interaction"*, pages 46 – 53, Canterbury, UK. AISB.

Ulrike Hahn. 2014. Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3):271–280.

Zellig S. Harris, 1970. *Papers in Structural and Transformational Linguistics*, chapter Distributional Structure, pages 775–794. Springer Netherlands, Dordrecht.

Nancy M. Henley. 1969. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, 8(2):176 – 184.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 894–904, Sofia, Bulgaria, August. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Comput. Linguist.*, 41(4):665–695, December.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *CoRR*, abs/1602.03483.

Ray Jackendoff and Steven Pinker. 2005. The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, 97(2):211–225.

Ray Jackendoff. 2012. Language. In Keith Frankish and William Ramsey, editors, *The Cambridge Handbook of Cognitive Science*. Cambridge University Press, Cambridge, MA.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, Sofia, Bulgaria, August. Association for Computational Linguistics.

Walter Kintsch. 2001. Predication. *Cognitive Science*, 25(2):173 – 202.

Arthur B. Markman and Dedre Gentner. 1991. Commonalities, differences and the alignment of conceptual frames during similarity judgments. In *Proceedings of the 13th Annual Meeting of the Cognitive Science Society, USA*, pages 287–292.

Arthur B. Markman and Dedre Gentner. 1996. Commonalities and differences in similarity comparisons. *Memory & Cognition*, 24(2):235–249.

Douglas L Medin, Robert L Goldstone, and Dedre Gentner. 1993. Respects for similarity. *Psychological review*, 100(2):254.

Dmitrijs Milajevs and Matthew Purver. 2014. Investigating the contribution of distributional semantic information for dialogue act classification. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 40–47, Gothenburg, Sweden, April. Association for Computational Linguistics.

Dmitrijs Milajevs, Mehrnoosh Sadrzadeh, and Thomas Roelleke. 2015. IR Meets NLP: On the Semantic Similarity Between Subject-Verb-Object Phrases. In *Proceedings of the 2015 International Conference on Theory of Information Retrieval*, ICTIR '15, pages 231–240, New York, NY, USA. ACM.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Samuel Ritter, Cotie Long, Denis Paperno, Marco Baroni, Matthew Botvinick, and Adele Goldberg. 2015. Leveraging preposition ambiguity to assess compositional distributional models of semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 199–204, Denver, Colorado, June. Association for Computational Linguistics.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, March.

Diarmuid Ó Séaghdha and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1047–1057, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1201–1211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luc Steels. 2008. The symbol grounding problem has been solved, so what's next? In Manuel de Vega, Arthur M Glenberg, and Arthur C Graesser, editors, *Symbols and Embodiment: Debates on Meaning and Cognition*, pages 223–244. Oxford University Press, Oxford.

Luc Steels. 2010. Can Evolutionary Linguistics Become a Science ? *Journal for Evolutionary Linguistics*, 1(1).

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Michael Tomasello. 1998. The return of constructions. *Journal of Child Language*, 25(02):431–442.

Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard University Press, Cambridge, MA.

John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585.

Amos Tversky and J. Wesley Hutchinson. 1986. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3 – 22.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2015. How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 9:1–9:8, New York, NY, USA. ACM.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *arXiv preprint arXiv:1506.03487*.