WordNet-based similarity metrics for adjectives

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

Le and Fokkens (2015) recently showed that taxonomy-based approaches are more reliable than corpus-based approaches in estimating human similarity ratings. On the other hand, distributional models provide much better coverage. The lack of an established similarity metric for adjectives in WordNet is a case in point. I present initial work to establish such a metric, and propose ways to move forward by looking at extensions to WordNet. I show that the shortest path distance between derivationally related forms provides a reliable estimate of adjective similarity. Furthermore, I find that a hybrid method combining this measure with vector-based similarity estimations gives us the best of both worlds: more reliable similarity estimations than vectors alone, but with the same coverage as corpus-based methods.

1 Introduction

In this paper I present new WordNet-based (Fellbaum, 1998) measures to provide reliable estimates of human word similarity ratings. Ever since Hill et al. (2014) published their SimLex-999 data set, many people have tried to find a way to determine the similarity of all the word pairs without being affected by the relatedness of the words. Recently, Le and Fokkens (2015) showed that taxonomy-based approaches beat vector-based approaches (Turney et al., 2010) in the estimation of the SimLex data. This is because corpus-based approaches are more affected by association, while taxonomy-based approaches mainly use vertical relations that are well-suited for determining similarity. However, corpus-based approaches do have a big advantage in their coverage. Moreover, Le and Fokkens left adjectives out of consideration, for lack of a good WordNet-similarity measure. My aim was to fill this lacuna, and also to find a way to mitigate the coverage issue. In section 3, I propose three WordNet-based adjective similarity measures, and evaluate them on the SimLex-999 data.[1] Section 4 provides a more thorough discussion of our results. At the same time, we should acknowledge that the representation of the adjectives in WordNet could use some attention. Section 5 proposes future work, looking at some extensions to WordNet that might improve our proposed measures. Section 6 concludes.

2 Evaluation

It is important to note that similarity is a *relative* measure; we do not learn anything from the fact that the similarity between adjectives X and Y is 2.4 unless we also know the similarity between other pairs of adjectives. Only then do we learn whether X and Y are very similar or not similar at all. In other words, being able to *rank* adjective pairs in terms of their similarity is more important than having a specific number for each pair. This is why the Spearman rank correlation is typically used for evaluation. I follow this standard procedure in our general evaluation.

Le and Fokkens (2015) argue for the use of multiple different evaluation methods, since they may lead to different conclusions about the results. They propose to use *ordering accuracy* (an evaluation of the relative ordering between all combinations of pairs, following Agirre et al. (2009)), supplemented with tie correction, i.e. giving a partial score to word pairs having the same similarity score. This levels the playing field, as taxonomy-based similarity values are more prone to yield ties than corpus-based measures (discrete versus real scores). The intuition behind this proposal is that

[1] All the code and data is available for replication at https://github.com/evanmiltenburg/gwc2016-adjective-similarity

overall ranking is more important than arbitrary local differences. Therefore, we should not punish algorithms as much for getting specific pair orderings 'wrong' when they are too close to call. In the discussion (section 4), I will use Le and Fokkens' comparison by group, where *pairs of pairs* of adjectives are grouped by the difference in their similarity scores in the gold standard. This is useful to see how well different models perform at varying levels of granularity.

3 Current possibilities

In this section, I examine distance metrics for adjectives in WordNet. I will first look at two classical measures, *Hso* (Hirst and St-Onge, 1998) and *Lesk* (Lesk, 1986), and show that they perform reasonably well (although not state-of-the-art). Next, I propose a method based on derivationally related forms, that are associated with the adjective lemmas. Though this approach achieves good results, it does suffer from poor coverage. I will then look at an alternative approach using attributes, but conclude that it is not feasible to incorporate them in our distance metric. Finally, to remedy the coverage issue, I propose a hybrid approach using both WordNet and distributional vectors.

3.1 Classical measures

Two classical similarity measures are given by the *Lesk* and the *Hso* methods. The former uses word overlap between glosses as a similarity measure, while the latter uses path distance (with some restrictions on the path). Both are implemented in Perl by Pedersen et al. (2004). Banjade et al. (2015) evaluate these measures on the adjectives in SimLex-999 taking only the first sense in WordNet into account, achieving a Spearman correlation (ρ) of 0.42 for the Lesk measure, and $\rho = 0.236$ for Hso.

Following Resnik (1995), I evaluated these measures using *all* senses for each word form, and taking the highest similarity. Intuitively, this comes closer to what Hill et al.'s participants did during the judgment task: they were already primed to look for similarities, so they were likely to be biased towards selecting the most similar senses. This idea is reinforced by the Lesk results: now this method (taking the maximal Lesk similarity between all synsets) yields a stronger correlation of $\rho = 0.51$. The correlation of the Hso scores with SimLex almost doubled: $\rho = 0.45$.

3.2 Using derivationally related forms

For all adjectives that have derivationally related forms in WordNet, one can use the distance between those related forms as a measure of adjective similarity. This roughly equates to saying that similarity between adjectives is a function of the properties they describe. I again used the 111 adjective pairs in SimLex-999 to evaluate the performance of this measure. To perform the evaluation, I selected all pairs of adjectives for which WordNet 3.0 specifies derivationally related nouns (for at least the first sense of the adjective). This resulted in 88 (out of 111) pairs, consisting of 89 (out of 107) different adjectives. Our distance measure is defined as follows:

1. For both adjectives A and B, get a list of all synsets corresponding to A and B.
2. Then, generate two new lists of derivationally related nouns: DRN_A, DRN_B.
3. The distance between A and B is given by $\mathit{min}(\{\mathit{distance}(x, y) : \langle x, y\rangle \in DRN_A \times DRN_B\})$, where *distance* is the shortest-path distance.[2]

I predicted that there would be a (negative) correlation between the distance between A and B and the similarity between A and B (i.e. items that are further apart in WordNet should be less similar). This expectation is corroborated by the results: our similarity measure has a Spearman correlation (ρ) of -0.64 with the SimLex data, which is near human performance (overall human agreement $\rho = 0.67$). To compare this result, I used the best performing predict-vector from (Baroni et al., 2014)[3] to generate cosine similarities for the same pairs of adjectives, achieving $\rho = -0.59$.

3.3 Using attributes: negative results

A problem with using derivationally related forms is that only 41% of all adjective synsets *have* derivationally related nouns. For better coverage, can we apply a similar technique to measure similarity through each adjective's attributes? The answer seems to be negative. I took two types of

[2] I did not experiment with alternative measures, as performance is not the main goal of this paper.

[3] This model was trained using `word2vec` (Mikolov et al., 2013) on the UkWac corpus, the British National Corpus, and the English Wikipedia. It is available here: `http://clic.cimec.unitn.it/composes/semantic-vectors.html`.

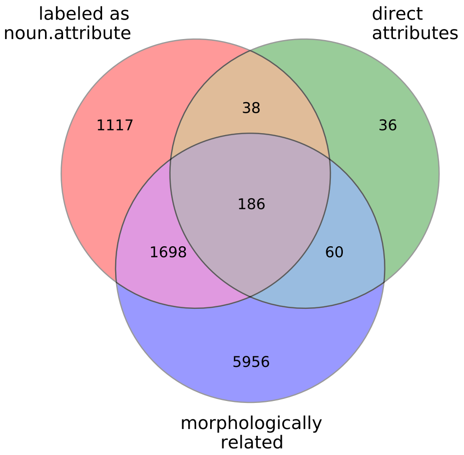

Figure 1: Nouns in WordNet that are, or could potentially be linked to adjectives in WordNet 3.0.

approaches, but neither produced any significant correlation with the SimLex data:

1. Take the shortest path distance between all attributes of the first/all senses of A and B.
2. Use the (relative) size of the overlap between the sets of attributes of A and B.

It is unclear why we get such a different result using attributes instead of derivationally related forms, but it probably has to do with the current status of WordNet attributes. A closer look at the adjectives in WordNet 3.0 teaches us that there are only 620 adjectives that even have attributes, and on average each adjective has 1.03 attributes. Furthermore, only a fraction of nouns that are labeled as `noun.attribute` is actually used as an attribute. Figure 1 provides an illustration of the current situation. In sum: it might be too soon to write off an attribute-based similarity measure, but getting such a measure to work requires a serious effort to link adjectives to all their possible attributes. Fortunately, there is already some work in this direction: Bakhshandeh and Allen (2015) describe a method to automatically learn from WordNet glosses which attributes an adjective can describe.

3.4 Going hybrid: WordNet plus vectors

What we *can* do, is make use of WordNet as much as possible, and only rely on vectors or other techniques if WordNet fails to provide a measure.[4] I used the following general algorithm, substituting Baroni et al's vectors for X:

1. Generate similarity values for all the pairs using WordNet, and other approach X, so that we have two lists of similarity values: L_W and L_X.
2. Sort both lists, so that we get a ranking for all pairs. In L_W, there will typically be many pairs with the same rank (i.e. ties).
3. Create a new output list L_O; initially a copy of L_W. Use the values from L_X as a tie-breaker, so that all pairs in L_O have a unique rank.
4. Iterate over all the pairs p in L_X that do not occur in L_W. The first pair is a special case: if p is the first item of L_X, put it at the start of L_O. Otherwise, treat it like the other pairs: get the pair immediately preceding p in L_X and look up its position in L_O. Insert p immediately after that position in L_O.

The result (L_O) is a sorted list that maintains the structure of L_W, but that also contains all the pairs under consideration. For the SimLex data set, the hybrid approach achieves a correlation of $\rho = -0.62$, compared to $\rho = -0.58$ for Baroni et al.'s vectors alone.

4 Discussion

From the Spearman correlations alone, it seems that we gain precision by involving derivationally related forms (DRF) in the estimation of similarity values. This picture changes when we look at ordering accuracy. I found that the DRF-based and vector-based approaches achieve comparable results. For the subset of 88 pairs where both adjectives have DRFs, I found a slight advantage for the vector-based method compared to the DRF-based method: 70% versus 71%. For the full dataset, this is exactly reversed, with a precision of 71% for the hybrid method and 70% for the vector-based method. That is not to say that both measures encode the same information; indeed we find interesting differences when we compare the pairs on a group-by-group basis.

Table 1 shows the ordering accuracy by group. When differences (in similarity scores) between two word pairs are small, the vector-based approach seems to have the upper hand in determining which is more similar. On the other hand, when differences between pairs are larger it seems that the hybrid approach is better at determining which pair is more similar. As the table shows,

[4] Banjade et al. (2015) also use a hybrid system to estimate similarity scores, but they use many different measures and combine them using a regression model.

Δ	WordNet	Vectors	Hybrid	Vectors
0	52	54	53	54
1	57	68	63	64
2	65	73	66	73
3	89	69	82	74
4	92	91	91	89
	Subset		Full dataset	

Table 1: Ordering accuracy scores by group, for the 88-pair subset from section 3.2 and the full dataset from section 3.4. The Δ-column indicates levels of granularity in the differences between pairs being compared. It runs from 0 (pairs with comparable similarity scores) to 5 (pairs with large differences in their similarity scores).

both effects are more pronounced in the 88-pair subset. Note especially the marked 20 percentage point difference with $\Delta = 3$.

Issues with tie-correction

The fact that with $\Delta \in \{0, 1, 2\}$ we find that vector-based approaches have a better ordering accuracy is interesting, but may also be an artifact of the tie-correction. Consider the way tie correction works: whenever a model predicts a tie, a score of 0.5 is awarded. In groups where the differences are small, the likelihood of a tie using the DRF-based method increases, and so the average score is drawn towards 50%. This is not what we want, as it actively biases the evaluation against coarse-grained measures in first group(s).

When we make the score linearly dependent on the difference between the pairs in SimLex-999 (punish the model for predicting a tie when there is actually a big difference, and reward the model for predicting a tie when there is little-to-no difference at all), the DRF-based method with the 88-pair subset gets an increased overall score of 74% whereas the vector-based method achieves the same score as before (71%).[5] More work is needed to determine whether this is a good way to do tie-correction, and whether it is at all possible to reliably compare fine-grained similarity measures with course-grained ones. But if we just ignore any ties between pairs in either the gold standard or in both of the similarity measures, then we are left with 3299 pairs where the DRF-based method has an accuracy of 74%, versus 73% for the vector-based approach.

5 Future work: extensions to WordNet

There are several projects that add new information to the adjective synsets, which can be used to increase coverage. Below I discuss potential uses and the current limitations of this information.

Adjective hierarchy GermaNet (Hamp and Feldweg, 1997) contains a hierarchy for adjectives, structured using hyponymy relations. This means that it is possible to use any of the available WordNet distance metrics directly on the adjective synsets. Unfortunately, the mapping between GermaNet and Princeton WordNet is still incomplete, and there is no dataset similar to SimLex for German to test this idea.

Add new cross-POS relations In this paper we have used the two types of cross-POS links that are available in WordNet: attributes and derivationally related forms. Other projects have a more diverse set of relations between adjectives and nouns. EuroWordNet (Vossen, 1998) has the *xpos_near_synonym, xpos_has_hyperonym* and *xpos_has_hyponym*-relations that can be used as access points to the noun hierarchy. WordNet.PT (Mendes, 2006) has similar relations. These seem like a good addition to the *'derivationally related to'*-link that we have been using, as they encode very similar information without the requirement of the two words morphologically resembling each other. Adding these relations would give us a much better coverage, while hopefully still providing a good score, but this remains to be tested.

Add domain information a more general approach is WordNet-domains (Magnini and Cavaglia, 2000), where each synset is associated with a particular domain. Examples of domains are: ECONOMY, SPORT, MEDICINE, and so on. Like the *property-of* relation, domain information does not seem to be helpful in the actual ranking procedure, but the knowledge whether two adjectives are associated with the same domain may serve as a useful bias.

6 Conclusion

We have seen several different WordNet-based measures of adjective similarity: the classical

[5] The updated scoring function returns the result of the following function if a tie is predicted (with P as the set of all pairs in the gold standard):

$$\text{score}_{\text{tie}}(p_1, p_2) = 1 - \frac{abs(p_1 - p_2)}{max(\{abs(p_i - p_j) : \langle p_i, p_j \rangle \in P \times P\})}$$

Lesk and Hso measures, and two new measures based on specific cross-POS links and the shortest-path distance between the nouns they are related to. It turns out that the *derivationally related forms*-link can be used to get state-of-the-art results on the SimLex-999 dataset. If coverage is an issue, then the hybrid method from section 3.4 is a better option than using vectors alone (though not by a large margin). We also noted that, on closer inspection, these measures do not seem to capture the same information. Therefore, future research should look at new ways to combine distributional and taxonomy-based measures.

Another way to improve similarity estimations would be to extend WordNet with new information. For example, the *attributes*-relation currently seems unusable for any similarity-related work, but may still be useful if more attribute links are added to WordNet. And looking at the literature, there is a lot of promising work being done with other WordNets, leaving us with many interesting avenues to explore the relation between WordNet and lexical similarity.

Acknowledgments

Thanks to Tommaso Caselli, Antske Fokkens, Minh Le, Hennie van der Vliet, and Piek Vossen for valuable comments on earlier versions of this paper. This research was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of HLT*, pages 19–27. Association for Computational Linguistics.

Omid Bakhshandeh and James F Allen. 2015. From adjective glosses to attribute concepts: Learning different aspects that an adjective can describe. *IWCS 2015*, page 23.

Rajendra Banjade, Nabin Maharjan, Nobal B Niraula, Vasile Rus, and Dipesh Gautam. 2015. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *Computational Linguistics and Intelligent Text Processing*, pages 335–346. Springer.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, volume 1, pages 238–247.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. Cambridge, MA: The MIT Press.

Minh Ngoc Le and Antske Fokkens. 2015. Taxonomy beats corpus in similarity identification, but does it matter? In *Proceedings of Recent Advances in NLP*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into wordnet. In *LREC*.

Sara Mendes. 2006. Adjectives in WordNet.PT. In *Proceedings of the GWA*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004*, pages 38–41. Association for Computational Linguistics.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.