Initial Steps for Building a Lexicon of Adjectives with Scalemates

Bryan Wilkinson
Dept. of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
bryan.wilkinson@umbc.edu

Abstract

This paper describes work in progress to use clustering to create a lexicon of words that engage in the lexico-semantic relationship known as grading. While other resources like thesauri and taxonomies exist detailing relationships such as synonymy, antonymy, and hyponymy, we do not know of any thorough resource for grading. This work focuses on identifying the words that may participate in this relationship, paving the way for the creation of a true grading lexicon later.

1 Introduction

Many common adjectives, like *small* and *tiny*, can be defined in terms of intensities of other adjectives. These relations, known as grading, intensification, magnification and others, are hypothesized to be one of the more important types in a lexicon (Evens, 1980). This type of relationship has applications in question answering and ontological representations (de Marneffe et al., 2010; Raskin and Nirenburg, 1996).

While the existence of this relationship is widely agreed upon, the study of it has fallen far behind that of synonymy, antonymy, and hyponymy, especially in the computational linguistics community. Recent work has brought renewed attention to this area of research, but there is still no large resource of words that participate in this relationship (van Miltenburg, 2015; Ruppenhofer et al., 2014).

The phenomenon of grading is not the same as gradability, although there is significant overlap among the adjectives that have it. Gradability refers to an adjective's ability to be combined with adverbs like *very* or be used in comparative expressions. It is possible that words like *lukewarm*, which are not considered gradable by most linguists, still have the lexico-semantic relation of grading. Similarly, a word like *spotted*, which is gradable, and in fact can be viewed on its own scale, does not express the relationship of grading with any other words in English.

There is no agreement on what types of adjectives express this relationship. Paradis and Kennedy & McNally propose two similar views that were influential to this work. Kennedy and McNally (2005) focus on the structure of scales, whether they are open at both ends (*tall, short*), closed at both ends (*visible, invisible*), or a combination of the two (*bent, straight* and *safe, dangerous*). Paradis (1997) on the other hand, defines three classes of gradable adjectives, limit adjectives, extreme adjectives, and scalar adjectives. For her, *dead* and *alive* are gradable adjectives but of the limit variety, meaning there is a definite boundary between the two. Extreme and scalar adjectives, such as *terrible* and *good* respectively, are both conceptualized as being on a scale, although extreme adjectives share some properties with limit adjectives as well. Paradis also points out that many adjectives can easily have a scalar interpretation, such as someone being *very Swedish.*

The study of grading has focused on a small number of adjectives (van Tiel et al., 2014). Many previous approaches of automatically learning the relation have relied on existing ontologies such as WordNet and FrameNet to choose which words occur on scales (Sheinman et al., 2013; Ruppenhofer et al.,

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 57–63,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

2014). The issues with using ontologies like these as starting points are pointed out by Van Miltenburg (2015). He notes that words like *difficult* and *impossible* are not grouped together and that limiting scales to WordNet networks prevents ad-hoc scales as introduced by Hirschenberg (1985) from being studied. To this we can add our own observation that many times an ontology can be too broad, including *puffy*, *rangy*, and *large-mouthed* under *size* alongside expected senses of *big*, *small*, and others. Westney investigated what might be necessary for a word to be on a scale while recent work in cognitive science has focused on the acquisition of scalar implicatures in children (Westney, 1986; Verbuk, 2007).

We demonstrate work in progress to cluster adjectives into those that participate in grading and those that do not. While our metrics do not currently match the supervised solution of (Hatzivassiloglou and Wiebe, 2000), the lack of large amounts of training data encourages us to continue to pursue the unsupervised approach. Clustering the adjectives is a critical first step to support further research into semantic intensities of adjectives, which is outlined in section 2.

1.1 Adverb Types

As shown above, adverbs can play a large role in the study of adjectives. Many types of adverbs have been recognized in the literature, with many studies being derived from the classification of Quirk (1985). Many of these studies have been done with an emphasis on adverbs' interactions with verbs. Moryzcki (2008) has noted that at least the subject oriented class (*deliberately, purposely, willfully, etc.*) and what he terms "remarkably adverbs" (*astoundingly, disappointingly, remarkably, etc.*) occur with adjectives as well.

The group of adverbs that have received the most attention in regards to their combinations with adjectives has been degree adverbs. In addition to Kennedy and McNally's use of co-occurrence with degree adverbs to arrive at the scale structures mentioned earlier, Paradis (1997) performed detailed research on this class of adverbs. She found that certain adverbs combine only with certain types of gradable adjectives. Adverbs she terms scalar modifiers (*fairly, slightly, very, etc.*) combine only with scalar adjectives while maximizer adverbs like *absolutely* combine with extreme adjectives.

	rather	pretty
high	175929.0	42533.0
long	141152.0	31229.0
low	161944.0	22953.0
odd	55147.0	3424.0
short	119977.0	8251.0
bad	30308.0	127592.0
funny	13350.0	19563.0
good	79737.0	817421.0
hard	87502.0	110704.0
tough	9620.0	37633.0

Table 1: Co-occurrence matrix from Google syntactic ngrams corpus

This type of pattern of co-occurrence has not only been observed between the classes of adjectives and adverbs but also within them. Desaguilier (2014) showed that *rather* combined more often with words like *long* and *high* while *pretty* combined more often with words like *good* and *stupid*, yet both are considered not only scalar modifiers, but a subtype known as moderators according to (Paradis, 1997). This effect can be seen in the co-occurrence matrix shown in Table 1.

2 Related Work

While this is the first attempt we know of to create a general lexicon of adjectives that participate in grading, several related studies have occurred. We first discuss work on defining gradable and non-gradable adjectives and then discuss several recent works on automatically ordering adjectives.

Using the intuition that gradability is a good indicator of subjectivity Hatzivassiloglou and Wiebe (2000) use the co-occurrence of adjectives with adverbs as well as a word's ability to be inflected for gradability in a classification task. They classified all adjectives that occurred more than 300 times in the 1987 WSJ corpus as gradable or non-gradable, for a total of 496 adjectives. When counting the co-occurrence with adverbs, they used only two features, the number of times an adjective occurred with any of the degree modifiers from a manually created list of 73, and the number of times it occurred with any other type of adverb. The classifier was trained on 100 randomly selected subsets of 300 adjectives and tested on randomly selected subsets of 100 adjectives.

Since Hatzivassiloglou and Wiebe was published, a great number of corpora have been produced. One issue we now face is that the class of degree adverbs is generally agreed to be a closed class in English, while other adverbs are not. This means we can reasonably expect the number of non-modifier adverbs would dominate the other features in an unsupervised situation. Additionally, while the degree adverb class is considered closed, we have not found a comprehensive list of all of them, leading to further reservations about simply counting adverbs as degree modifying and non degree modifying based on a list.

Several works have looked at automatically ordering a group of adjectives by intensity given that they occur on the same scale. Van Miltenburg (van Miltenburg, 2015) uses patterns to find scalemates from a large corpus. He is particularly interested in pairs of words for use in reasoning about scalar implicatures. The candidate pairs generated by the patterns are then validated by using various similarity measures, such as LSA or being under the same attribute in WordNet. This pattern based approach has also been taken by Sheinman (Sheinman et al., 2013), although she starts out with the words on a scale from WordNet and uses the patterns to order the words. As pointed out by (Ruppenhofer et al., 2014), pattern based approaches do not have wide applicability, a fact backed up by the results of van Miltenburg. Out of 32470 pairs identified, only 121 occur in 4 or more of the 6 patterns used.

Ruppenhoffer (2014) has also investigated the automatic ordering of adjectives on a scale. Using adjectives taken from FrameNet, they compare the occurrence of adjectives with 3 “end-of-scale” modifiers and 3 “normal” modifiers, using (Kennedy and McNally, 2005) as a guide. They achieve good correlations to human standards on the 4 scales they chose to investigate using this method, though it should be noted that once these co-occurrence metrics were computed, the scale was constructed manually.

Shivade, et al. (2015) use a combination of clustering and patterns in their approach to ordering not only adjectives, but adverbs as well. To determine scale membership, they cluster 256 adjectives known to occur on scales by their co-occurrence with nouns. They then match patterns of parse trees rather than at string level to derive features for ordering. The order is computed using Mixed Linear Integer Programming as done in (de Melo and Bansal, 2013). Our contribution can be seen as a precursor to their pipeline, providing a list of adjectives that are known to participate in grading to the clustering algorithm.

3 Methodology

While the group of gradable adjectives and those that participate in grading do not entirely overlap, it is a good starting point to build a lexicon of graded adjectives. There are rare cases, like *lukewarm*, but it is not believed there are many other words that would be missed by this assumption.

For a given set of adjectives that we wish to derive a lexicon from, we first build a co-occurrence matrix using the Google syntactic ngrams to select adverbs that are dependent on adjectives (Goldberg and Orwant, 2013). We used the arc relations in this dataset that represent a direct dependency between two words. The adverbs were required to participate in the advmod dependency with the adjective. To ensure a wide representation of adverbs, we use the degree modifiers discussed by Paradis (1997), the remarkably adverbs discussed by Moryzcki (2008), the subject oriented adverbs discussed by Moryzcki and enumerated by Quirk (1985), and the viewpoint and time adverbs from Quirk as our features. This gives us a total of 84 features, which we call the Manual feature set in Table 2. We also produce a variation of the feature set with only five features, where the adjectives are grouped together by type as defined above, denoted by Manual Collapsed in Table 2. A third feature set we investigated was the 1000 most frequent adverbs in the corpus, regardless of their occurrence with adjectives, denoted by Top 1000 Advs.

The matrix is weighted with PPMI as implemented in DISSECT (Dinu et al., 2013). We then run k-means(k=2) clustering to split the adjectives into a group of gradable adjectives and a group of non-gradable adjectives.

As previously discussed, being gradable does not guarantee an adjective participates in the grading lexico-semantic relation. As an approximation of finding only adjectives that occur on the same scale as others, we run anomaly detection on the adjectives which were clustered into the gradable group. We used local outlier factor (LOF) due to its ability to find anomalies locally, rather than on a global scale, better approximating adjectives without scalemates (Breunig et al., 2000).

4 Evaluation

As Hatzivassiloglou and Wiebe did, we use the Collins COBUILD Dictionary for our evaluation (Sinclair et al., 1987). The dictionary classifies adjectives as either classifying or qualitative which correspond approximately to non-gradable and gradable. The distinction here is the narrow sense of gradable, meaning the adjectives can be modified by only scalar modifiers, not maximizers or approximators. This is the best resource we know of at this time however, and it allows comparisons to earlier work. We follow Hatzivassiloglou and Wiebe in removing adjectives from the dataset that we could not reliably label as classifying or qualitative when different senses had conflicting labels.

We ran the clustering and anomaly detection on the 500 and 1000 most common adjectives in the Google syntactic ngrams corpus, removing any that were not labeled as an adjective by COBUILD. This gives of datasets of length 427 (237 gradable and 190 non-gradable) and 838 (461 gradable and 377 non-gradable) respectively. Due to many of the words having conflicting senses, we ran another dataset consisting of only the words for which all senses unanimously chose the same classification.

The results of evaluating the clustering can be seen in Table 2. The data set that should be compared to (Hatzivassiloglou and Wiebe, 2000) who report a precision of .9355, recall of .8224, and accuracy of .8797, is the 500 most frequent adjectives. While we don't achieve as high a precision, our recall is much higher. Partial reasons for this could be that using COBUILD is a flawed choice, as it assigns words like *far* to the classifying class of adjectives in all senses, even though it can be inflected as *farther* and *farthest*. The words that were labeled by COBUILD as non-gradable but clustered as gradable by our method from the 500 words dataset using the 1000 most frequent adverbs are shown in figure 1. While some of the words are true errors, words like *dead* and *alive* are commonly discussed in linguistic literature, with many considering them gradable (Kennedy and McNally, 2005). Other words that were misclustered can easily be placed on a scale, such as *silent* or *everyday*. Ultimately we are using a broader definition of gradable than COBUILD. Additionally it is more likely for a word not traditionally viewed as gradable to appear in gradable context rather than vice-versa. This leads to a high recall due to the fact that the gradable adjectives rarely appear in non-gradable contexts.

*able above absolute actual additional alive
available average based central chief chronic
comprehensive constant contemporary
continuous corresponding criminal current
dead dear double east entire equivalent
eternal everyday extreme facial far fatal
fellow few fewer free front fundamental
future gay giant global horizontal identical
illegal induced inevitable intermediate
known lateral left like logical natural neutral
objective occasional ongoing operational
overall parallel particular past positive
possible potential present previous principal
proper pure ready real related responsible
right same separate silent single solid special
specific subject subsequent sufficient
temporary top total traditional ultimate
unable unique universal unknown up usual
various vertical very whole*

Figure 1: Words labeled by COBUILD as non-gradable, but clustered with gradable words in our data

The most interesting outcome is that the use of manual features does not provide an advantage. This is promising for future work, especially for applications in other languages. Constructing manual features requires the existence of detailed descriptive grammars for the language in question.

Testing against only the words that were assigned one label in the dictionary performed the worst under all conditions. This may be because the distribution of these terms is heavily skewed towards the

Data Set	Feature Set	Precision	Recall	F_1	Accuracy
1000	Manual	.7061	.9696	.8171	.7613
	Manual Collapsed	.7154	.9652	**.8217**	.7697
	Top 1000 Advs	.6931	.9848	.8136	.7517
500	Manual	.7030	.9789	.8183	7587
	Manual Collapsed	.7285	.8945	.8030	.7564
	Top 1000 Advs	.7005	.9873	**.8196**	.7587
Unanimous	Manual	.6493	.9765	.78	.7417
	Manual Collapsed	.6445	.9843	.7789	.7380
	Top 1000 Advs	.6791	.9921	**.8063**	.7765
(Hatzivassiloglou and Wiebe, 2000)	Custom Features	.9355	.8224	**.8753**	.8797

Table 2: Evaluation against COBUILD classifications

less frequent words of the top 1000, rather than any effect from the classification itself.

One group of words that is reliably identified as not having any scalemates are demonyms like *American* and *Swedish*. As another heuristic on our algorithm, we use the list of denonymic names from Wikipedia [1]. We found that 100% of these were correctly excluded from the final list for all feature sets.

While we have no evaluation for the effectiveness of the anomaly detection, the words with the 10 highest LOF are shown in Table 3. Of these, *able* and *logical* are identified by COBUILD as classifying adjectives. If we assume that the synonyms and antonyms given by COBOUILD could be scalemates for these words, we find that only *consistent* and *historic* do not have scalemates in the dataset. This suggests that at least LOF is not a good estimate of words sharing a scale, and possibly anomaly detection in general.

word	LOF
able	34.78
consistent	4.98
realistic	3.42
loyal	2.92
better	2.57
historic	2.56
hungry	2.50
logical	2.46
attractive	2.43
extensive	2.41

Table 3: Top 10 Highest LOF

5 Future Work

There are many areas for improvement. In the methodology, we feel that there is currently too much manual selection of the features. This includes both the selection of adverbs that apply to a wider range of adjectives as well as the ability to automatically group the adverbs into classes similar to those defined in section 2.1.

While using more semantically related feature sets revealed no large improvement, we still believe this could be a productive avenue of further work. One possible source of inspiration for this would be biclustering often used in biology. This works on the assumption that the underlying data has a checkerboard pattern. The problem with this assumption is that this may actually separate the related adjectives and adverbs more. The idea of of grouping the adverbs and adjectives simultaneously is an attractive one however.

Once the adjectives have been placed into preliminary groupings, we need to determine which of the words to not have any scalemates. It was shown above that LOF does not appear to be a viable solution. Several promising solutions to this are still available for exploration. Hypernym identification as performed in (Lenci and Benotto, 2012) has traditionally been used on nouns to build taxonomies, but may have some applications to adjective taxonomies

[1] http://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations

as well. Additionally, (Kanzaki et al., 2004) have exploited the relationship between abstract nouns and adjectives to build a hierarchy of adjectives in Japanese.

Another area of improvement is the need for a better evaluation. In addition to the issue of COBUILD using a narrower version of gradability than us, there is no resource to reliably check if the words produced do in fact have scalemates. Work by (van Miltenburg, 2015) on finding pairs of scalemates used in scalar implicature is a possible solution but notes that their techniques also face evaluation issues.

The relationship between gradability, subjectivity and the lexical relationship we investigate in this paper needs to be further explored. While we do not believe they are the same, they may serve as resources for both the creation of our lexicon as well as evaluation.

Beyond the creation of the lexicon, it will have many potential uses once created. For linguists, it will provide new data on which to test theoretical models of scales, scale structures, and gradability. For the NLP community, it will serve as a resource in investigations into scalar implicature as well as the automatic ordering of adjectives.

6 Conclusion

In this paper we discuss a method to automatically build a lexicon of words that appear on a scale. Our clustering step achieved F_1 scores between .78 and .82. While these are not as high as the those achieved by (Hatzivassiloglou and Wiebe, 2000), we have demonstrated that using an unsupervised method comes close to a supervised one. In addition, we have pointed out many potential flaws with the current evaluation, and provided several future directions on which to further improve the lexicon.

References

Markus Breunig, Hans-Peter Kriegel, Raymond Ng, and Jörg Sander. 2000. LOF: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2010. Was it good? it was provocative. learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176. Association for Computational Linguistics.

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Guillaume Desagulier. 2014. Visualizing distances in a set of near-synonyms. *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, 43:145.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. Dissect: Distributional semantics composition toolkit. In *Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, East Stroudsburg PA. Association for Computational Linguistics.

Martha W Evens. 1980. *Lexical-semantic relations : a comparative survey*. Linguistic Research, Carbondale [Ill.].

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, pages 241–247.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics*, volume 1 of *COLING '00*, pages 299–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julia Hirschberg. 1985. *A theory of scalar implicature*. Ph.D. thesis, University of Pennsylvania.

Kyoko Kanzaki, Eiko Yamamoto, Hitoshi Isahara, and Qing Ma. 2004. Construction of an objective hierarchy of abstract concepts via directional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1147. Association for Computational Linguistics.

Chris Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 75–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcin Morzycki. 2008. Adverbial modification in AP: Evaluatives and a little beyond. In Johannes Dölling and Tatjana Heyde-Zybatow, editors, *Event Structures in Linguistic Form and Interpretation*, pages 103–126. Walter de Gruyter, Berlin.

Carita Paradis. 1997. *Degree modifiers of adjectives in spoken British English*, volume 92 of *Lund studies in English*. Lund University Press.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language (General Grammar)*. Longman, 2nd revised edition edition.

Victor Raskin and Sergei Nirenburg. 1996. Adjectival modification in text meaning representation. In *Proceedings of the 16th conference on Computational Linguistics*, volume 2, pages 842–847. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 117–122. Association for Computational Linguistics.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation*, 47(3):797–816, 1 September.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Folser-Lussier, and Albert Lai. 2015. Corpus-based discovery of semantic intensity scales. In *In Proceedings of NAACL-HTL 2015*, Denver, CO. Association for Computational Linguistics.

John Sinclair, Patrick Hanks, Gwyneth Fox, Rosamund Moon, Penny Stock, et al. 1987. *Collins COBUILD English language dictionary*. Collins London.

Emiel van Miltenburg. 2015. Detecting and ordering adjectival scalemates. In *MAPLEX*, Yamagata, Japan.

Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2014. Scalar diversity. *Journal of Semantics*, 23 December.

Anna Verbuk. 2007. *Acquisition of scalar implicatures*. Ph.D. thesis, University of Massachusetts Amherst.

Paul Westney. 1986. Notes on scales. *Lingua*, 69(4):333–354, August.