Corpus-based discovery of semantic intensity scales

Chaitanya Shivade[†], Marie-Catherine de Marneffe[§], Eric Fosler-Lussier[†], Albert M. Lai[*]
[†]Department of Computer Science and Engineering,
[§]Department of Linguistics,
[*]Department of Biomedical Informatics,
The Ohio State University, Columbus OH 43210, USA.
shivade@cse.ohio-state.edu, mcdm@ling.ohio-state.edu
fosler@cse.ohio-state.edu, albert.lai@osumc.edu

Abstract

Gradable terms such as *brief*, *lengthy* and *extended* illustrate varying degrees of a scale and can therefore participate in comparative constructs. Knowing the set of words that can be compared on the same scale and the associated ordering between them (*brief* < *lengthy* < *extended*) is very useful for a variety of lexical semantic tasks. Current techniques to derive such an ordering rely on WordNet to determine which words belong on the same scale and are limited to adjectives. Here we describe an extension to recent work: we investigate a fully automated pipeline to extract gradable terms from a corpus, group them into clusters reflecting the same scale and establish an ordering among them. This methodology reduces the amount of required handcrafted knowledge, and can infer gradability of words independent of their part of speech. Our approach infers an ordering for adjectives with comparable performance to previous work, but also for adverbs with an accuracy of 71%. We find that the technique is useful for inferring such rankings among words across different domains, and present an example using biomedical text.

1 Introduction

Gradability (Sapir, 1944) is a property of words that identifies different degrees of the quality the word denotes. For example, adjectives such as *large*, *huge* and *gigantic* present different degrees of size or volume. Similarly, adverbs such as *approximately*, *almost* and *roughly* present different degrees of how accurate a measurement is. Thus, one of the characteristics of gradable terms is that they participate in a scale and can be ordered along that scale: for example, *good* < *great* < *excellent* (Kennedy, 2007). Another characteristic is that gradable terms can appear in comparative constructions, e.g., "A is larger than B". Such comparative judgments are a psychological process that precedes judgments of counting, e.g., "A is twice as large as B" (Sapir, 1944).

Modern NLP systems face the challenge of interpreting language as close to human perception as possible. Modeling gradable terms as well as their associated meaning and ordering is an important aspect of this challenge. Such information can be very useful for a variety of inference tasks, such as sentiment analysis (Pang and Lee, 2008) and textual inference (Dagan et al., 2006). However, current lexical resources, like WordNet (Fellbaum, 1998), lack annotations capturing the gradability of words. This weakens the notion of similarity: although words such as *small* and *minuscule* illustrate varying degrees of size, they are listed as synonyms in WordNet.

Recently, there has been a lot of interest in exploring different approaches to derive an ordering among gradable adjectives based on their semantics (Ruppenhofer et al., 2014; Sheinman et al., 2013; Schulam and Fellbaum, 2010). de Melo and Bansal (2013) propose a novel Mixed Integer Linear Programming (MILP) based approach, publish a gold standard dataset and report the best performance on ordering scalar adjectives on this dataset. However, these approaches are limited in two ways. First, they depend on a manually created resource, such

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 483–493,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

as WordNet or FrameNet (Baker et al., 1998). Lexical patterns (e.g., 'not just x but y') are used both to extract words that belong to the same scale and to determine the direction of the ordering (e.g., in the above pattern, x is weaker than y). However, this extraction process gives noisy results that require filtering using an electronic thesaurus. The domain of application is thus restricted to words that exist in an electronic thesaurus. Second, previous work is limited to the study of adjectives.

In this paper, we propose a fully automated pipeline that uses structural patterns to extract gradable terms from a corpus, cluster them into groups that reflect the same semantic scale of comparison, and finally rank them using de Melo and Bansal's MILP technique to establish an ordering among them. We also explore how the technique fares on domain-specific (biomedical) text, deriving scales for domain-specific terms that might not exist in thesauri. Our approach achieves a comparable performance to previous studies on scalar adjectives, and can be reliably extended to adverbs.

2 Related work

Hatzivassiloglou and McKeown (1993) present the first work on automatically clustering adjectives that belong to the same scale, identifying scalar adjectives based on the intuition that similar nouns are modified by similar adjectives. They use a hierarchical clustering algorithm on a newswire corpus for grouping similar adjectives, but do not provide ranking among a given cluster of related adjectives.

Assuming a pair of related adjectives, de Marneffe et al. (2010) use reviews from the Internet Movie Database and their associated ratings to infer an ordering in the adjective pair. Kim and de Marneffe (2013) also obtain an ordering given a pair of adjectives, using distributional word vectors derived from a recursive neural network.

Sheinman et al. (2013) and de Melo and Bansal (2013) present similar approaches, which make use of WordNet *dumbbells* to determine words that belong to the same scale as proposed in Sheinman et al. (2012). A WordNet *dumbbell* is a representation involving an antonym pair (e.g., *small* and *large*) as two ends of a semantic scale with semantically similar adjectives arranged in a radial fashion around each adjective. The antonym acting as a centroid and its synonyms as members of a cluster represent words that most likely participate in the same scale. For example, the antonym pair (*small, large*) results in the dumbbell with clusters (*small, tiny, pocket-size, smallish*) and (*large, gigantic, monstrous, huge*) at the two ends. It should be noted that even with such a representation, there can be words that fall into the same WordNet synset but do not participate in the scalar relationship (e.g., *violent* with respect to *supernatural* and *affected*). This is primarily because of polysemy and semantic drift (de Melo and Bansal, 2013).

Sheinman et al. (2013) present a two-step approach for establishing an ordering among scalar adjectives. They extract adjectives from the Web using lexical patterns indicative of the direction of the scalar relationship between a pair of adjectives. Two sets of patterns are defined: *mild* patterns in which participating words are such that the first word has a weaker semantic intensity than the second word (e.g., '$*$ but not $*$' – *good but not great*); and *intense* patterns, in which the first word has a stronger semantic intensity than the second word (e.g., 'not $*$ but still $*$' – *not freezing but still cold*). In the first step, they assign a positive score to an adjective if it is seen as a part of the intense pattern and a negative score if seen as part of the mild pattern. In the second step, they use these scores to partition the adjectives into two subsets one representing *mild* and the other representing *intense* adjectives. They perform this partitioning recursively to obtain a complete ordering for a given cluster of adjectives from a WordNet dumbbell.

de Melo and Bansal (2013) improve upon Sheinman et al. (2013) by refining their lexical patterns, and refer to them as "strong-weak" and "weak-strong" patterns. Using frequencies of occurrence for a pair of adjectives across the strong-weak and weak-strong patterns in a corpus, they define an overall weak-strong score. They optimize for this score using MILP. The constraints of the MILP model two types of strength relationships: the strength relationships between two adjectives in a pair with a possible third adjective, and synonymy relationship between two adjectives based on information from an external resource. Given a cluster of terms, the MILP produces an ordering of the cluster members using frequency counts of instances where these members are found in strong-weak and weak-strong patterns. To evaluate their approach, de Melo and Bansal construct a manually curated gold standard of 88 clusters, each with a cardinality of three or more adjectives. These 88 clusters are randomly drawn from all possible clusters that are either half of a WordNet dumbbell. Two annotators manually examined these clusters to remove words that did not belong to the same scale. Further, all pairs within these clusters were annotated for scalar relationship: is the adjective in a pair weaker than the other, stronger than the other, or of equivalent intensity. The output of the MILP was tested on these 88 clusters (569 word pairs). They achieve a pairwise accuracy of 78.2%.

3 Our approach

3.1 Extraction using structural patterns

As observed by Ruppenhofer et al. (2014), lexical pattern-based approaches suffer from a coverage issue. This is because these patterns consist of longer n-grams, which are sparsely found in a small dataset. Therefore, Sheinman et al. (2013) use the Web as their corpus, and de Melo & Bansal use Google N-grams (Brants and Franz, 2006). However, this results in a large number of instances where satisfied lexical patterns do not correspond to adjectives (e.g., *sometimes but not always*). Moreover, since the Google N-grams corpus is limited to 5-grams, adjective pairs of interest beyond a five-word window are lost.

To deal with these shortcomings, we use *Tregex* (Levy and Andrew, 2006), which enables pattern matching on parse trees based on syntactic relationships and regular expression matches on nodes. Using Tregex, we transform de Melo and Bansal's weak-strong and strong-weak lexical patterns into structural patterns. For example, one way of expanding the lexical pattern '$*$ but not $*$' into a structural Tregex pattern for adjectives is 'ADJP< ((ADJP<JJ) $ (CC<but)$(RB<not)$ (ADJP<JJ)).' Similarly, a structural pattern for adverbs can be written as 'ADVP< ((ADVP<RB) $ (CC<but)$(RB<not)$ (ADVP<RB)).' These patterns are available for download[1].

Introducing tree patterns requires parsing a corpus: while this additional step in the pipeline might lead to error propagation, the advantages of the structural patterns are that (i) they are more robust than the lexical ones and (ii) restricting results to a desired part-of-speech comes for free. In the experiments reported here, we use the Stanford parser v3.3.1 (Klein and Manning, 2003).

3.2 Automatic clustering

In order to determine a ranking of words based on their semantic intensity, the first step is to determine words that belong to the same scale of meaning. As pointed out earlier, previous work (de Melo and Bansal, 2013; Sheinman et al., 2013) use WordNet *dumbbells*, and this restricts the utility of these approaches to the scope of a manually created lexical resource. We overcome this limitation by automatically clustering words that belong to the same scale. As the clustering algorithm, we use the Matlab (2014) implementation of K-means++ (Arthur and Vassilvitskii, 2007), a hard clustering algorithm[2] with cosine similarity as a distance metric. Following Hatzivassiloglou and McKeown (1993), we use context vectors to represent the words to cluster. They make use of standard context vectors for clustering adjectives, where context for every adjective comprises of nouns it modifies across all sentences in a corpus.

However, recent work shows promise for context vectors embedded in a compressed semantic space that are derived using neural networks: Baroni et al. (2014) compare standard context vectors with embedded vectors for a wide range of lexical semantic tasks and found embedded vectors to yield better results. We therefore generate context vectors and compare the utility of both skip-gram and continuous bag of words (CBOW) representations using the `word2vec` tool (Mikolov et al., 2013) for our task. These two representations have demonstrated varying degrees of success in different NLP tasks (Baroni et al., 2014; Bansal et al., 2014). Given a

[1] http://web.cse.ohio-state.edu/˜shivade/naacl2015

[2] The choice of a hard-clustering algorithm was mostly for implementational convenience, but carries with it the issue that polysemous words can only appear in one semantic cluster. We leave the issue of deriving a soft clustering approach that works with context vectors, a separate research problem in its own right, to future work.

window size w, the CBOW model predicts the current word given the neighboring words as context. In contrast, the skip-gram model predicts the neighboring words given the current word. We used $w = 5$ and found CBOW to yield better results for our task. Thus the terms extracted from a corpus by the structural patterns are automatically clustered, and these clusters are used as an input to the ranking algorithm.

3.3 Ranking based on semantic intensity

Once the terms have been clustered, the second step is to provide a ranking between the cluster members. To do so, we use the MILP implementation provided by de Melo and Bansal (2013). This method computes an overall weak-strong score for a pair of adjectives based on the frequency of that pair in the matches for weak-strong and strong-weak patterns. The MILP then uses these scores among all relevant pairs of adjectives belonging to the same scale, capturing complex interactions to infer an ordering among them.

3.4 Data: PubMed corpus

In this work, we want to provide an approach that can infer scalar orderings for any domain-specific terms. Such terms might be absent from existing thesauri. Our approach is thus corpus-based as outlined above. We chose to test the robustness of our technique on PubMed, a large domain-specific corpus of biomedical texts. It is a free resource developed and maintained by the National Center for Biotechnology Information at the National Library of Medicine. It provides access to scientific abstracts, full text articles and associated resources. We used $10,875,982$ freely available abstracts (not full text articles) from PubMed as our corpus. This corresponds to $88,303,272$ sentences in total, where the average length of a sentence is 28 words (including punctuations). We used this corpus to find instances of the structural strong-weak and weak-strong patterns, both for adjectives and adverbs.

4 Comparison with the gold standard of de Melo & Bansal

To evaluate our approach, we need to establish how good the clustering step is, as well as how good the ranking step is. Each step is evaluated separately using annotations obtained from Amazon Mechanical Turk.

4.1 Clustering

In order to evaluate the automatic clustering procedure that uses K-means++ and word vectors, we start with the gold standard provided by de Melo and Bansal (2013): as mentioned above, their data set has 88 gold standard clusters, corresponding to 346 adjectives, annotated by humans for scale ordering. One problem with evaluating a hard clustering algorithm is that the same word may appear in multiple WordNet synsets, corresponding to multiple clusters (soft clustering). We therefore made a “hard cluster version” of the de Melo & Bansal dataset by removing any adjectives that occur in multiple clusters, and then eliminating any singleton clusters. This resulted in a gold standard set of 256 adjectives belonging to 84 clusters.

We clustered the 256 adjectives from the gold standard data subset into 84 clusters: the representation for each adjective was a neural embedding derived using the `word2vec` tool trained on our PubMed data. We experimented with both the skip-gram and continuous bag of words (CBOW) models to derive vectors of dimension sizes varying from 200 to 800 in increments of 100. To choose the right dimensionality and the best model, we evaluated the quality of the automatically derived 84 clusters against the gold standard. As a metric of evaluation for cluster quality, we follow Hatzivassiloglou and McKeown (1993) and use F1 calculated by comparing equivalence relations generated by the clusters (as implemented in LingPipe (2008)). We found that the CBOW model gave clusters closer to the gold standard than the skip-gram model. We found that a dimension size of 600 for the vectors yielded clusters with a maximum F1 score of 57%. Thus, we were able to fix the parameters for our clustering task. Figure 1 summarizes the results of this experiment.

In their study, Hatzivassiloglou and McKeown (1993) evaluate the results of their clustering on a small set of 21 adjectives. They presented the 21 adjectives to 9 annotators and asked them to partition these adjectives such that each partition contain adjectives that belong to the same scale. They re-

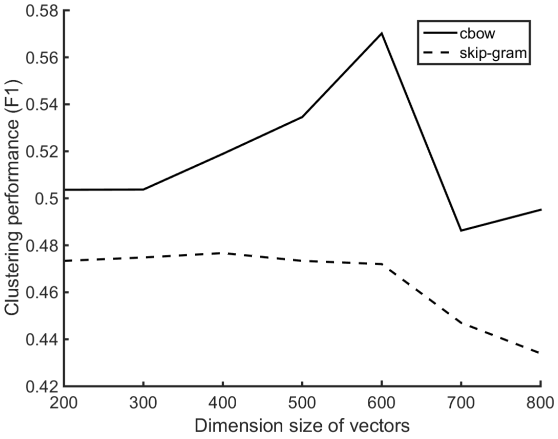

Figure 1: Comparison of CBOW and skip-gram for clustering of adjectives.

port a best F1 score of 48% but note that such score does not seem to reflect the quality of the clustering. We observe the same problem: our automatically derived clusters have a different organization for words that belong to the same cluster than the gold standard clusters, but in a way that seems intuitive. Some differences between our automatically derived clusters and the gold standard clusters are illustrated in Table 1. For example, the adjectives *false* and *misleading* belong to the same cluster in both the gold standard as well as the automatic clustering output. However, the automatic clustering groups the adjectives *false* and *misleading* together with *unreliable* and *wrong*, whereas the gold standard groups *false* and *misleading* with *deceptive* and *fraudulent*. Both clusterings are plausible, though. The adjectives *fraudulent* and *deceptive* become part of new clusters in our automatic clustering. It could be argued that the gold standard cluster "deceptive, false, fraudulent, misleading" represents different degrees of "trickery," whereas the automatic cluster "false, misleading, unreliable, wrong" represent different degrees of "wrongness." Thus, although both clusters contain different adjectives, they group adjectives that are on the same scale of a different meaning.

Therefore, to evaluate the quality of the automatic clustering, we sampled 50 clusters containing three or more adjectives (corresponding to a total of 190 adjectives) from all the generated clusters and obtained annotations using Amazon Mechanical Turk (AMT), a crowdsourcing platform that has been shown useful for a number of NLP tasks (Snow et al., 2008). Annotators (workers, in AMT parlance) were presented with 15 clusters in each worker session, whose members were each associated with a checkbox. For each cluster, workers had to uncheck the adjectives that did not belong to the same scale. The nature of the annotation task does involve inherent subjectivity which cannot be avoided. We tried to minimize this by giving detailed instructions with accompanying examples to achieve coherent annotations. To make sure workers were paying attention to the task, 2 clusters among the 15 clusters they saw were clusters for which we a priori knew which adjectives should be removed (e.g., *beautiful, pretty,* and *rainy* where *rainy* had to be unchecked). Most workers did the task well: we only had to discard annotations from 4 worker sessions (out of 140). We ended up with annotations from 8 to 10 workers per cluster. To create a gold standard, we retained in each cluster only those words that were ascertained to be in the same cluster by 6 or more annotators.

For each cluster, we calculated an accuracy score equivalent to the number of correct adjectives (determined to be on the same scale by the annotators) divided by the total number of adjectives in the generated cluster. This accuracy was averaged across all 50 clusters, and yielded a final micro-averaged accuracy of 74.36% as seen in Table 2.

4.2 Ranking

Since our end goal is to establish an ordering among scalar adjectives, we use the automatically derived clusters (rather than the WordNet dumbbells) as input to the MILP algorithm. To determine the performance of the ranking produced by the MILP algorithm, we use AMT to obtain pairwise ranking annotations for all unique adjective pairs within a cluster. Workers were presented with 15 word pairs in each worker session. For each pair $(a1, a2)$, the worker had to pick one of four options: (1) $a1$ is stronger than $a2$, (2) $a2$ is stronger than $a1$, (3) both are equally strong, and (4) $a1$ and $a2$ are not comparable. Option (4) was present because our clusters possibly contained adjectives that are not on the same scale. As in the previous task for getting annotations for clusters, we inserted two items with a clear ranking (e.g., *hot, hotter*) for every set of 15

Gold standard clusters	**Automatic clusters**
deceptive, false, fraudulent, misleading	false, misleading, unreliable, wrong
evil, immoral, sinful, wrong	desperate, humiliated, immoral, insane, sinful
dangerous, risky, suicidal, unreliable	dangerous, harmful, toxic

Table 1: Example comparison of automatically derived clusters against gold standard clusters from WordNet.

Data	**Corpus for strength counts in MILP ranking**	**Clustering Accuracy**	**Ranking Pairwise Accuracy**
Clusters automatically derived from non-polysemous WordNet adjectives	Google N-grams	74.36	84.74
	PubMed	74.36	69.23
PubMed-derived clustering:			
Regular adjectives	PubMed	86.26	70.37
Domain-specific adjectives	PubMed	64.30	–
PubMed-derived clustering:			
Regular adverbs	PubMed	89.36	71.00
Domain-specific adverbs	PubMed	53.80	–

Table 2: AMT-based evaluations of cluster accuracy and pairwise ranking accuracy of systems that vary in the source of clustering data, source of strength counts, and part of speech. For comparison, the approach used by de Melo and Bansal (2013) achieves a pairwise ranking accuracy of 76.1% on the non-polysemous WordNet clusters.

pairs to avoid random annotations. Each set was annotated by 10 workers. All workers passed all the checks and we did not discard any annotations for this task. To create a gold standard we assigned each pair one of four labels, *weaker, stronger, equal,* or *not comparable*. A value was assigned based on a majority vote. In case of a tie, the pair was assigned a label of being *equal*.

In order to compute a ranking, the MILP needs two inputs: 1) the cluster of terms that are on the same scale, and 2) the counts for how many times all pairs of adjectives in that cluster satisfied the weak-strong and strong-weak patterns (henceforth referred to as "strength counts"). In the first experiment of ranking adjectives, we ran the full pipeline used by (de Melo and Bansal, 2013) on the 256 adjective (84 hard cluster) subset of their gold standard (see Section 4.1). Thus, this experiment uses hand-corrected WordNet dumbbells to determine adjectives on the same scale of semantic intensity, followed by the MILP using strength counts from the Google N-gram corpus, to determine the ranking. Their pipeline resulted in a pairwise accuracy of 76.1% which serves as a baseline for comparison. In the second experiment of ranking adjectives, we used the 50 automatically derived adjective clusters described in Section 4.1 as an input for the MILP. Since these adjectives originate from WordNet dumbbells, we refer to them as "WordNet adjective clusters." We determined the ranking for adjectives within these clusters using strength counts obtained from our PubMed corpus. We obtained an accuracy of 69.23% across 105 pairs. The strength counts for all adjectives in these clusters, from Google N-grams corpus, used in the experiments of (de Melo and Bansal, 2013) were also available to us by the authors. We repeated the previous experiment by substituting strength counts from PubMed corpus with these strength counts from the Google N-grams corpus and obtained an accuracy of 84.74% across 119 pairs.[3] It appears from our experiment that pattern counts from a general corpus

[3] The MILP does not produce a strength relationship between a pair of adjectives if there are no strength counts for this pair. Hence, we observe a difference in the number of pairs for which accuracy is determined in the two ranking experiments.

is a better match for determining the adjective ordering than a more-limited domain corpus, despite the limitation of Google N-grams being restricted to 5-word sequences. We think this is because of two reasons: First, Google N-grams is a very large corpus compared to the one we use. Second, our corpus consists of abstracts and not full text of scientific articles from PubMed. Hence there is less variety in the language used; capturing fewer comparative constructs than Google N-grams. However, it is interesting that we can still extract patterns from domain-specific corpora to act as constraints for the MILP process.

5 Rankings for adjectives extracted from PubMed

We also desired to see how well our approach does on terms that are not specifically in WordNet, but present in a domain-specific corpus such as PubMed. We therefore also evaluate the clustering and ranking steps on a set of adjectives extracted from the PubMed data using structural patterns.

5.1 Clustering

Since there was no gold standard reflecting ideal clustering of data, we explored heuristic measures to choose parameters for our clustering step. We used CBOW vectors over skip-gram vectors since these were more effective in the previous experiment. Since the true value for number of clusters k was unknown, we chose k such that the average cardinality of a cluster was three. The value of k was found to be the same ($k = 375$) for all clustering experiments conducted using vector dimension sizes varying from 200 to 800 in increments of 100. To choose the right dimension size d of the CBOW vectors for this fixed value of k, we obtained clusters for incremental values of d from 200 to 800 in increments of 100. We determined the number of identical clusters obtained using a particular value of d with its next increment. The lowest value of d which resulted in a maximum number of identical clusters with its next increment was chosen: $d = 400$.

Using vectors of 400 dimensions, we obtained 375 adjective clusters with cardinality varying from 1 to 9. Since these clusters were derived from our biomedical dataset, they comprised of domain-specific adjectives, which are quite unfamiliar even to native English speakers. We manually partitioned the clusters into two sets: (i) containing domain-specific words, and (ii) containing words used in day-to-day English (henceforth referred to as "regular" terms). Examples of clusters from both sets are summarized in Table 3. The clusters we obtain look reasonable, grouping together adjectives that pertain to the same scale. The first cluster of domain-specific adjectives qualifies the nouns corresponding to different types of protein with varying degree of specificity, the second cluster contains different qualifications of a tumor, and adjectives in the third cluster qualify different parts of a living cell. For the regular adjective clusters, the clusters look intuitive too, except for the first cluster. The adjectives *male* and *female* are not scalar, but match the structural patterns, and are grouped together with adjectives describing age qualifications, due to a strong context overlap in which these words are used.

Clusters of domain-specific adjectives
cytokine, gm-csf, ifn-gamma, il-10, il-12, il-2
benign, malignant, metastatic, neoplastic, squamous
mitochondrial, nuclear, ribosomal

Clusters of regular adjectives
female, male, middle-aged, older, young, younger
accurate, precise, reliable, reproducible, robust
additive, insignificant, negligible

Table 3: Examples of automatically derived adjective clusters from PubMed abstracts.

We randomly sampled 25 clusters from each set, "regular adjectives" and "domain-specific adjectives", for our evaluation. We evaluated the clustering quality of the regular adjectives using the exact same approach as described in Section 4.1. We obtained a clustering accuracy of 86.26% for 25 clusters across 101 regular adjectives. This is substantially better than the performance of clustering in the previous experiment. We believe that this is due to the fact that the adjectives in the dataset used in the previous experiment originate from WordNet and contain many words (e.g., *handsome*, *crazy*, *spicy*), which are less likely to be found in scientific abstracts. Therefore, the context vectors learnt for these words are possibly less accurate compromising the clustering quality.

For the domain-specific adjectives, the annotations require specialized skills. PubMed hosts scientific articles from different disciplines of biological sciences. We obtained annotations from three annotators specializing in disciplines of Biomedical Informatics, Biochemistry, and Nursing. To create the gold standard, a word was retained or discarded from a cluster if two or more annotators agreed on it. We obtained a clustering accuracy of 64.3% for 25 clusters across 101 domain-specific adjectives.

5.2 Ranking

We obtained gold standard annotations for ranking using AMT for these 25 "regular adjective" clusters derived from the PubMed corpus using the exact same methodology as described in Section 4.2. The strength counts for these adjectives were also derived from the PubMed corpus. We obtained an accuracy of 70.37% across 109 pairs, indicating a similar level of performance to WordNet-based clusters.

Our expert annotators for the domain-specific adjectives faced problems in assessing an ordering between adjectives in a cluster. They report that for majority of the clusters, the ordering of the words would vary given the context. For example, consider the following modifications of the domain specific adjectives of the third cluster in Table 3: ribosomal particles, mitochondiral compartments and nuclear compartments, representing different parts of a living cell. If we consider "number of" as the relation in context, we get (*ribosomal* $>$ *mitochondrial* $>$ *nuclear*) as an ordering since number of ribosomal particles is greater than number of mitochondrial compartments, and number of mitochondrial compartments is greater than number of nuclear compartments. However, if we consider "size of" as the context, the ordering is reversed.

6 Extension to adverbs

A novelty of our approach is that we can also apply the technique to other parts of speech (e.g., adverbs). The structural patterns we describe in Section 3.1 also enable us to extract candidate scalar adverbs. We follow a similar approach to adjectives: extract adverbs, derive strength counts and rank them using the MILP.

Clusters of domain-specific adverbs
anteriorly, caudally, distally, proximally
chromosomally, clonally, genetically, phenotypically
neonatally, prenatally, postnatally

Clusters of regular adverbs
always, certainly, inevitably, invariably, universally
marginally, modestly, slightly, somewhat
excessively, inappropriately, overly

Table 4: Examples of automatically derived adverb clusters from PubMed abstracts.

6.1 Clustering

We used CBOW vectors to perform clustering and derived $k = 300$ and $d = 250$ using the approach described in Section 5. As with the adjective clusters, we found that there were also domain-specific adverbs, illustrated in Table 4. Again, the clusters obtained look reasonable. The first cluster of domain-specific adverbs describes relative position of a body part, the second cluster corresponds to adverbs describing identity of a gene that may have an observable effect, the third cluster represents temporal descriptions that relate an event to child birth. The clustering of regular adverbs is accurate, except for the third cluster where *inappropriately* was found to be an outlier based on our annotations. We followed a similar approach to the adjective experiment, creating two partitions for domain-specific and regular adverbs and sampling 25 clusters from each. Annotations for regular adverbs were obtained from AMT while annotations for domain-specific adverbs were obtained from 3 domain experts. The annotation process for both clusters of adverbs was identical to that of adjectives. We obtained a micro-averaged accuracy of 89.36% for 25 clusters across 104 regular adverbs and a 53.8% for 25 clusters across 89 domain-specific adverbs.

Accuracy	POS	Examples
Good	Adj	serious < life-threatening < fatal
Good	Adv	considerably < significantly < dramatically
Average	Adj	common < frequent = prevalent
Average	Adv	slightly < modestly < marginally
Bad	Adj	useful < helpful
Bad	Adv	continuously = regularly

Table 5: Example rankings for adjectives and adverbs from PubMed data.

6.2 Ranking

As in the case of adjectives, our annotators for domain-specific adverbs faced a challenge in ranking adverbs due to lack of context. Therefor we do not report results on ranking of adverbs. We obtained gold standard annotations for ranking using AMT for 25 clusters of regular adverbs derived from the PubMed corpus, using the exact same methodology as described in Section 4.2. The strength counts for these adverbs were also derived from the PubMed corpus. We obtained an accuracy of 71.00% across 38 pairs – a performance similar to the adjectives. However, we observe that there are a large number of pairs for which there are no strength counts, and the MILP does not generate a ranking. Table 5 shows sample results for ranking adjectives and adverbs from the PubMed data.

7 Limitations and future work

We present an approach to gradable modifier ordering that replaces WordNet-based clusters with automatically derived word clusters, replaces lexical patterns with structural patterns, and show that the approach has utility for not only discovering adjective patterns but also adverb patterns in biomedical text. We observe that while automatic ranking based on semantic intensity can be successful established between regular terms, doing so for domain-specific terms requires knowledge of context.

We plan to expand the structure patterns derived from the lexical patterns of de Melo and Bansal (2013), looking for new patterns that could be more suited for adverbs. We also plan to investigate soft clustering algorithms such as (Pereira et al., 1993) that may allow us to model polysemous words better. Furthermore, recent studies have compared traditional vectors against embedded vectors (such as the CBOW vectors used in this study) for different lexical semantic tasks (Levy and Goldberg, 2014; Baroni et al., 2014), which suggests that such a comparison for our clustering task could be insightful.

Our experimental results show that automatic clustering of gradable words produces promising results. However, we also observe that with domain-specific words, context is important for establishing a ranking between words that is based on semantic intensity. Thus, rather than clustering adjectives or adverbs in isolation, a joint with the clustering of nouns or verbs with which they occur is a possible direction of research. Finally, studies deriving a ranking based on semantic intensities are limited to unigrams belonging to different parts of speech. Our future work would focus on performing a similar task on bigrams consisting of adverb-adjective pairs (e.g., *somewhat unclear* < *quite hard* < *very difficult*) that exhibit properties of gradability.

Acknowledgements

We would like to thank Mohit Bansal and Gerard de Melo for providing the implementation and results of the experiments from their previous work. Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011116. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

David Arthur and Sergei Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Sympo-*

sium on Discrete Algorithms, SODA '07, pages 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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 the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. The Google Web 1T 5-gram Version 1.1.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence DAlché-Buc, editors, *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190, Berlin, Heidelberg, April.

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, July.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inferences of Semantic Intensities. *Transactions of the Association of Computational Linguistics*, 1(July):279–290.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales. In *Proceedings of the 31st Annual meeting on Association for Computational Linguistics*, pages 172–182, Morristown, NJ, USA, June.

Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, March.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving Adjectival Scales from Continuous Space Word Representations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer.

Omer Levy and Yoav Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Language Learning*, Batimore, Maryland USA. Association for Computational Linguistics.

2008. LingPipe 4.1.0.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 183–190, Stroudsburg, PA, USA.

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

Edward Sapir. 1944. Grading, A Study in Semantics. *Philosophy of Science*, 11(2):93–116.

Peter F Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of German adjectives. *Semantic Approaches to Natural Language Proceedings, Saarbruecken, Germany*, page 163.

Vera Sheinman, Takenobu Tokunaga, Isaac Julien, Peter Schulam, and Christiane Fellbaum. 2012. Refining WordNet adjective dumbbells using intensity relations. In *Sixth International Global Wordnet Conference*, pages 330–337, Matsue, Japan.

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, January.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, USA, October.