CS224U Literature Review

Kevin Leung
Stanford University
kkleung@cs.stanford.edu

1 Introduction

This paper reviews several papers relevant to to finding structure in gradable adjectives from a semi-supervised approach combining corpus data and human judgment. With that in mind, the papers being discussed cover a broad range of topics: instead of discussing different approaches to a similar problem, they discuss various steps and components of a possible project. Kennedy and McNally presents the linguistic theory being tested. Syrett and Lidz presents the use of corpus to test this theory and some basic results. Maas et al. presents an improved method for learning vector-based word representations from corpora using expert data.

2 Kennedy and McNally 2005

This paper presents a linguistic theory on the types of adverbs that can modify gradable adjectives. Unlike nongradable adjectives, gradable adjectives can hold to its subject at differing degrees, depending on the modification. This theory argues that gradable adjectives accept degree modification according to the scale structure that they fit. A scale is a total ordering over degrees for the concept being expressed by an adjective. These scales come in various types depending on whether they have minimum and maximum values. Thus, a scale can be open (no minimum or maximum), closed (minimum and maximum), lower closed (only minimum), and upper closed (only maximum). These correspond to different possible degree modifiers that they accept. For example, proportional modifiers work with adjectives with closed scales.

1. Chris thought the football stadium was half full.
2. ?? Bill bought a half expensive shirt.

Another example of a scale-constrained modifier is *slightly*, which relates to a minimum value.

1. Chris rode his bike on the slightly wet road.
2. ?? Bill made many slightly accurate comments during class.

These scale structures can also be thought to vary on the standard of comparison. Some adjectives, such as *expensive* depend on the context, while others, such as *full* do not. This paper argues that context-dependent (relative) adjectives correlate to completely open scales, whereas context-independent (absolute) adjectives correlate to closed scales.

Moreover, the paper makes the claim that scale structures are not only a helpful way for us to conceptualize language, but are actually the cognitive basis for how we think about language. Thus, scale structures are important in human reasoning, and this theory has consequences on how language shapes thought.

Most of the details of this paper are more specific to the theory and beyond the scope I'm interested in exploring now. Even so, they do put together a compelling reason for verifying this theory because of the consequences of proving it true or false. Although this paper proposes an interesting theory and some examples to provide intuition, it also lacks empirical evidence for it beyond one small chart in

the introduction showing some properties of usage[1]. The most testable part of these claims is the correlation of proportional modifiers with non-relative gradable adjectives.

3 Syrett and Lidz 2010

This paper argues that children are sensitive to the distribution of adverbs used with a given adjective (and vice versa), and this ability helps them learn novel adjectives more quickly. The analysis mainly focuses upon gradable adjectives and argues for a structure based on that presented in the paper above. The paper both presents experimental evidence for this sensitivity and children as well as corpus analysis to show that such structure in gradable adjectives exists for children to learn. Although interesting, the developmental research is less relevant to the project at hand, and this review will mainly focus on the corpus analysis.

They searched the British National Corpus for co-occurrences of 10 adverbs (5 proportional modifiers, 5 intensifiers) with any adjective and 10 adjectives (5 maximal gradable adjectives and 5 relative gradable adjectives) with any adverb. Each of these instances was coded and reviewed by 3 people on the same attributes: proportional modifier or not, and maximal adjective or not. These were compiled, and they found that proportional modifiers were significantly more likely to appear with maximal adjectives, and intensifiers are significantly more likely to appear with nonmaximal adjectives. This is also confirmed by looking at the conditional probabilities of co-occurrences.

Although these results are fairly striking, I have a few concerns with this analysis. First, this approach doesn't appear to be particularly efficient: if I interpret the method correctly, the first author (as well as 2 other reviewers[2]) was responsible for coding the properties of tens of thousands of instances. This process is not generally practical for doing similar research on other data.

Second, the data chosen is very narrow and likely biased towards positive results. The paper notes that "we targeted canonical exemplars from each set of lexical items that are frequently cited in discussions of GAs in the semantics literature." Imagining an entire co-occurrence matrix, however, their analysis only covers 10 rows and 10 columns of data that may potentially span 2 orders of magnitude more data points, and many of those likely don't have as significant characteristics as the words chosen for this analysis. The results presented here are promising evidence of structure in gradable adjectives but are not generally conclusive.

4 Maas et al. 2011

This paper tries to improve the accuracy of vector-based word representations by augmenting the data with sentiment during unsupervised learning. Traditional approaches to these representations depended on largely semantic features, such as nearby words in the sentences where the word appeared. The drawback is that words of varying polarity can appear very similar. For example, "delicious" and "disgusting" will likely have close representations since they often appear as descriptions of food. By adding sentiment annotations from the document as a whole, this approach attempts to improve representations on polarity. For example, "delicious" and "disgusting" would likely appear in restaurant reviews that are "thumbs up" and "thumbs down" respectively.

The basic model used for unsupervised learning is based on topic models. Roughly, the model assumes there are an unspecified, unseen number of topics that involve each document, and words can be represented by a vector for the strength of association with the topics. The word representations are found by training the model to maximize the likelihood of the documents (and words within).

To include sentiment, the model also tries to predict the sentiment label (mapped onto the interval [0, 1]) for each document based on the words within. The word representation vector is passed through logistic regression, and these are aggregated for all of the words in a document as a prediction for the sentiment label. This is also learned by including it in the objective function along with that from the basic model above. The authors suggest that although this objective function is non-convex, it is still tractable

[1] I have learned that many linguists appear loathe to provide empirical evidence

[2] The paper says "reviewed by at least two other linguists with a background in semantics": I suspect they found students taking "Intro to Semantics" to review.

to estimate as the work is parallelizable over the documents.

The model was trained on 25,000 movie reviews from IMDB, which includes both the text and a star rating for the review. This approach is tested on a sentiment classification task for a previous set of 2,000 movie reviews, where the model performs well against other vector space models. This model continued to perform well on a larger dataset of 50,000 intended to correct for some biases in the original training set.

The approach in this paper is particularly interesting because it uses a probabilistic approach to discovering these representations instead of applying specific transformations to a raw data matrix. This flexibility creates a framework for introducing better features into the learning process. Future work may extend this semi-supervised approach into other domains. Although the primary data remains the same, expert data (being the sentiment labels in this case) can also help to guide learning.

I would also be interested to see how this model performs on general word representation tasks. Although the model is tuned for sentiment, one would hope that the training on sentiment would have aided in general word meaning and not at the expense of it. The paper provides a few examples of general improvement but doesn't provide any results from extensive testing.

5 Boleda Torrent and Alonso i Alemany 2003

This paper tries to find adjective classes using clustering on corpus data. Specifically, they looked at adjectives in Catalan, which are believed to have 3 main classes, with differences in syntax, semantics, denotation, and morphology between them. One of the large obstacles for learning how to classify these adjectives was polysemy as specific adjectives may have two meaning that fall into different classes.

They used a corpus of 8.5 million words, marked with the lemma, part of speech, basic syntactic function, and other morphological details. For learning, they used the basic features just mentioned, a sliding window of parts of speech, and several handcrafted, linguistically motivated features that were easily mapped into combinations of the known features. No additional, external processing was done on the data. They used CLUTO as their clustering algorithm.

For evaluation, they compared the results both to clusters determined by human judges and by an metric of internal consistency. For the human evaluation, they found relatively poor agreement among judges (kappa = .52), as well as having far too few data points. They expanded their judgements to the entire set using "derivational morphology"[3]. Although the results of the clustering weren't great (kappa = .45), it was close to human agreement. This result was also fairly robust with different settings.

The authors ultimately resign to the problem of polysemy, though their approach does show some promise. They were able to elicit some sense of classes using fairly simple features. Their evaluation using an expanded version of human judgements also demonstrates one method for testing results against what is presumably the "ground truth" of the language.

One aspect that makes this approach somewhat easy for them is the large differences between the classes. Since these classes have differences along 4 different dimensions (syntax, semantics, denotation, morphology), evidence from all of these properties reinforced the differences that they found. This consistency, however, doesn't appear in many other classes and structure. For example, gradable adjectives in English don't appear to have consistent denotation within the Kennedy and McNally theory.

6 Mitchell 2009

This relatively recent paper tries to learn prenominal modifier orderings by using a somewhat simple unsupervised algorithm for learning classes from corpus counts. Traditional approaches to this problem have included general rules, such as Behaghel's First Law that "What belongs together mentally is placed close together syntactically". Results in learning these orderings before this paper had tried a variety of counting and clustering methods but generally performed poorly when tested on different domains. These approaches had depended on word-to-word ("comfortable red" or "red comfortable") re-

[3]This is not well described, and I don't entirely understand it

lationships, which were sparse and didn't indicate any level of confidence. This paper proposes that words be classified directly on position (1 word before, 2 words before, etc), which effectively deals with sparsity of co-occurrences.

For training, they extracted noun phrases from The Penn Treebank-3, the Wall Street Journal, the Brown corpus, and the Switchboard corpus. The possible classes for modifiers were being in one of four positions before the noun, and 2 and 3 consecutive position subsequences (1-2, 2-3, 3-4, 1-2-3, 2-3-4) for 9 possible classes. Words were assigned to single position classes based strictly on counts, filtered for appearing $>$ .25 of the time (less is considered a baseline since there are 4 positions), and this is further distributed over the subsequences for weighted preferences for each position. Overall, the algorithm is somewhat convoluted and appears to depend on situationally motivated properties instead of a general approach.

The model was evaluated using 10-fold cross validation by determining the learned order of a set of modifiers appearing before a noun. The results are somewhat confusing and not directly comparable to previous results. Even so, they report high (90%) precision.

One of the criticisms that this paper offers of previous models is the poor generalization to other domains. This paper sidesteps the issue by gathering data from 4 different corpora. Although generalization doesn't appear to be a problem, it's not clear whether this was a property of the model or from training.

Overall, I was not particularly impressed by the methods of this paper: the algorithm was not well-motivated, and results weren't directly compared to existing models. Even so, there are some aspects I find interesting about it. First, the paper directly challenges existing linguistic theory by suggesting that the word-to-word relationships are less important than the general position of modifiers. The results show that this shallow approach depending on large-scale statistical properties can perform better than intuition. It suggests that similar approaches may be helpful in evaluating other linguistic theories. Second, it assigns words to classes only to a probability instead of doing hard clustering. They leverage this particular property to improve performance by using it not only for flexibility but also as a proxy for confidence.

7 General Discussion

As mentioned, these papers cover several different types and levels of analyses. In general, they all revolve around this issue of learning properties of words from large corpora. These range from relatively simple methods to fairly complex models.

Boleda Torrent and Alonso i Alemany (2003) differs from the other two papers doing classification in that it relies on an opaque, hard clustering algorithm to learn about the structure of words. It's hard to tell how this choice affected results, but soft clustering can only give more information and seems like a more sound method of evaluation, even if the results are more difficult to interpret. Even so, they all depend on vector-based representations of words, though the meaning of the values varies widely. This is somewhat controlled by the task demands (semantic meaning may not be terribly important to prenominal positioning), but it reflects different ways that corpus data can be boiled down into a representation.

Given this diversity, it seems particularly important to choose a representation that emphasizes the properties of interest. Syrett and Lidz (2010) did this by selecting their data very narrowly, which brings into question the generalization of their results. Boleda Torrent and Alonso i Alemany (2003) hand-crafted some of the linguistically-motivated features to include. I would have preferred to see an approach like that in Maas et al. (2011) which implicitly learns the features of interest. Hopefully, this approach could also be run in reverse to develop features for some evaluation, then interpret the features into a theory.

Both Boleda Torrent and Alonso i Alemany (2003) and Maas et al. (2011) inject judgement into their models, albeit at different times. Boleda Torrent and Alonso i Alemany (2003) use it as a method for evaluating their model, whereas Maas et al. (2011) use it to guide their learning. Mitchell (2009) even mentions that their approach may benefit from bootstrapping. A general theme we can gather from this is that although large corpora can provide rich data, it may also require some guid-

ance from humans. I would be interested to see how Boleda Torrent and Alonso i Alemany (2003) may have benefitted from using the judgements they gathered as a part of their model. The comparable kappas suggest a similarity that may be explained if these align in training.

The method of evaluation varied widely between the papers, likely due to the very different goals that each one had. Syrett and Lidz (2010) did a direct statistical test to find a difference in the distribution between the 2 classes. Maas et al. (2011) had a separate task for external evaluation. Boleda Torrent and Alonso i Alemany (2003) tried both human judgement and internal evaluation, and Mitchell (2009) tested the generalizability of the model to unseen examples. Even within this general approach of learning properties of words from corpora, there appear to be many different tasks and applications one can find.

The future work from these papers fits together as an approach for a semi-supervised method for learning linguistic structure from corpora. Of the approaches, Maas et al. (2011) seems the most general and applicable to multiple domains, even though it isn't interested in modifiers as the rest of them are. Taking inspiration from these previous approaches, however, it seems promising to apply a probabilistic model for unsupervised learning augmented with human judgement to tune the model towards the task at hand instead of depending on hand-coded features and design.

References

G. Boleda Torrent and L. Alonso i Alemany. Clustering adjectives for class acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 9–16, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

C. Kennedy and L. McNally. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2):345–381, 2005.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

M. Mitchell. Class-based ordering of prenominal modifiers. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 50–57, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

K. Syrett and J. Lidz. 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development*, 6(4):258–282, 2010.