Intensionality of Adjectives

Mark Kowarsky, Neha Nayak MARKAK@STANFORD.EDU, NAYAKNE@STANFORD.EDU

1. Introduction

Given the information *"The young president has been travelling for eight weeks"*, we can correctly conclude that *"The president has been travelling for eight weeks"* However, if it is known that *"The former president has been travelling for eight weeks"*, it would be incorrect to conclude that *"The president has been travelling for eight weeks"*, since the former president is not the president. As is obvious to a native English speaker, the adjective "former" somehow modifies the noun "president" such that we conclude that it no longer belongs to *the set of things that are presidents* by virtue of being *former*. Linguists categorize adjectives with this effect as being *intensional*. Distinguishing between intensional and extensional (non-intensional) adjectives is an important part of understanding textual entailment, essential to the domain of Natural Language Understanding (1).

Related work

In Boleda et al.(3), the distributional representations of various intensional and extensional adjectives were examined.Different composition functions were used to model adjectival modification, and the predicted properties of the compositional representations were evaluated. Intensional adjective-noun vectors were most similar to those of the unmodified noun using cosine similarity. Boleda et al.(4) conducted an extension to their earlier work, using an enhanced list of intensional adjectives, as well as a more varied set of extensional adjectives. Both concluded that modification by intensional and extensional adjectives were both modelled equivalently well with existing composition functions.

This work

This final project for CS224N and CS229, will attempt to classify whether an unseen adjective is intensional based on a linguistic model and to expand the list of intensional adjectives known in the literature. It increases the number of known intensional adjectives by 120%. It does this using adjective-noun co-occurrences (section 4) and grammatical and contextual information (section 5) and the machinery of support vector machines. Adverbs associated with adjectives were also considered, but the results were uninformative.

2. Data

The three data sets we extracted for this report all derive from dumps of English Wikipedia articles. One database which had undergone tokenization, part-of-speech tagging and sentence splitting (carried out by the NLP group at Stanford) was used for adjective-noun co-occurrences. In total, 19GB of uncompressed text was used, covering 2019 different types of adjective (188 075 unique), 26 728 types of noun (172 090 unique) and total co-occurrences of 16 996 971. The other database had dependency and token contexts annotated. This allowed adverbs (1428 unique) modifying adjectives (1302 unique) to be extracted for a total of 644 052 co-occurrences. Finally, 15 077 adjective-noun pairs were extracted with 371 582 different contexts, for a total of 16 027 099 co-occurrences.

The list of known intensional adjectives (see Appendix A) was curated by reading the literature (4), and expanded by adding synonyms as well as false positives found during early testing of the classifier[1].

3. Methods

Labelled data This problem is a supervised learning classification problem. The prior knowledge of only 30 known intensional adjectives had the following repercussions:

Incorrect example labels All other adjectives were assumed (incorrectly) to be extensional. This affected classification as was seen by relabelling discovered intensional adjectives in the training set.

Train-Test division We chose to split the known intensional adjectives into a set of 10 for testing and a set of 20 for testing. Up to 1000 and 2000 'extensional' adjectives were added to these. Learning algorithms were applied using stratified k-fold cross-validation to maintain the ratio of positive:negative examples.

Classifier A linear support vector machine (SVM) was used (6; 5) to classify the tokens as being either intensional or extensional. Input data was scaled to have zero mean and unit variance, which increased precision and recall and sped up the algorithm. The penalty for errors for each class was set to be inversely proportional to their frequency in

[1] Examples found by early classifiers were strictly restricted to the training set in the classifiers used for evaluation.

the training data. Earlier in the project Naive Bayes and Logistic Regression models were also used, but they did not work as well.

Features Co-occurrences are frequencies, which do not adequately represent the significance of a particular co-occurrence. We applied some standard transformations (LMI, PMI, PMI^2) to the data before classification, with the objective of filtering out less meaningful co-occurrences. The following transformation was used:

$$\text{PMI}^2(A,N) = \log(\frac{\Pr(A \cap N)^2}{\Pr(A) \cdot \Pr(N)})$$

Feature selection To minimize the effect of over-fitting with too many features and to provide a meaningful subset of features for understanding how intensional adjectives are use, the following methods. We first used a thresholding scheme to select nouns who co-occured with some at least some number of intensional adjectives. This proved better than using only frequency of co-occurence, but was not as effective as using differential expression.

Differential Expression Another approach to categorising and building models that predict the class of adjective has been to appropriate techniques used in bioinformatics to find nouns (genes) that are called "differentially expressed" between different adjectives (tissues/samples)(7).

Briefly, it works by performing statistical tests between different classes for a given feature, taking into account the variance both within and between classes to measure the likelihood that a feature is expressed higher or lower (p-value$<$0.05, after Benjamini-Hochberg corrections for multiple testing(2)). The other method is to sort features by their "log fold change". That is, after suitable normalizations, how much are they expressed in one class compared to another. It was observed that picking features using the log fold-change method performed best.

In an attempt to improve the list of features again, an automatic recursive feature elimination with cross-validation algorithm was also employed to minimize the set, using the

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

as the metric of the quality of the fit.

4. Identifying nouns indicative of intensionality

Our baseline implements a simple hypothesis: that intensional and extensional adjectives differ in the nouns they can modify. Using adjective-noun bigrams, we constructed a vector space of adjectives, using their co-occurrence frequencies (above a threshold) with nouns as features.

apartheid {anti-,former}
contradiction {apparent, seeming}
lack {alleged, apparent, former, likely, past, possible, potential, probable, seeming, virtual}

Table 1. Some nouns associated with the intensional class - selected by DE

The best results were achieved by using nouns (features) chosen by log fold-change as determined by differential expression and counts modified by PMI^2. Using the recursive feature elimination with cross-validation increased the number of false positives and false negatives, so was not used.

We updated the training set by incorporating false positive adjectives that turned out to be mislabelled to attempt to better train the learning algorithm.

4.1. Results

We trained classifiers on sets different ratios of extensional and intensional adjectives ($m_{ext} : m_{int}$), using randomly selected subsets of the training set. Using a different number of features selected by DE, and different $m_{ext} : m_{int}$ ratios, we produced learning curves of the average precision, recall and F_1 scores over the stratified test folds. The solid lines in the plots below is for the original training set, the dashed line when we relabelled false positives that were in fact intensional adjectives.

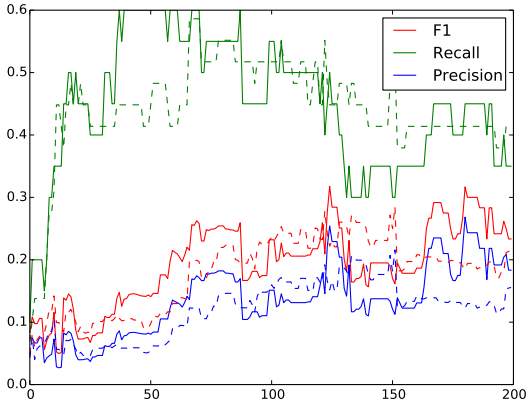

Figure 1. Learning curve for 100:1 ratio of intensional:extensional adjectives

The confusion matrices from Table 2 to Table 3 show

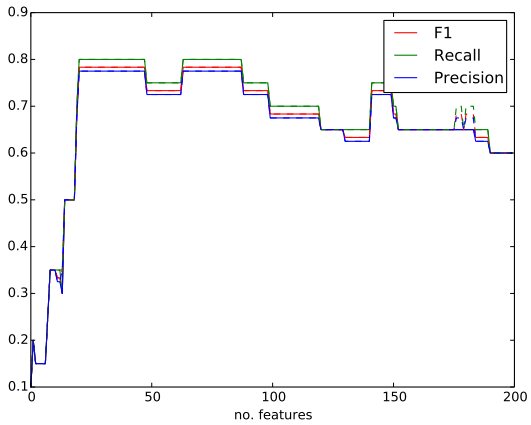

Figure 2. Learning curve for 1:1 ratio of intensional:extensional adjectives

more detailed results of the classifiers. Each cell contains the average of the corresponding cells in the confusion matrix of the folds of k-fold cross validation. The matrices labelled *With 'Bootstrapping'* correspond to the classifiers in which we manually flipped the labels on the false positives that were actually intensional.

The first classifier, with an $m_{ext} : m_{int}$ ratio of 100:1, represents a model of the 'real-life' distribution of adjectives between the classes. In this situation, we aimed to increase recall while maintaining high precision. We were not able to beat the majority baseline of 99% in this scenario.

The third classifier, using an $m_{ext} : m_{int}$ ratio of 1:1, represents an attempt to characterise fundamental differences between intensional and extensional adjectives. With an accuracy of 85%, this classifier was able to beat the majority baseline of 50%. The constant values post-'bootstrapping' occurred because the training examples whose labels were flipped did not occur in the training set for this classifier.

Initial		Predicted Class	
		Ext	Int
True Class	Ext	96.85	3.10
	Int	0.4	0.6
With 'Bootstrapping'		Predicted Class	
		Ext	Int
True Class	Ext	65.28	3.34
	Int	0.45	0.55

Table 2. Confusion matrix - Best classifier for 100:1. 82 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	0.9	0.1
	Int	0.2	0.8
With 'Bootstrapping'		Predicted Class	
		Ext	Int
True Class	Ext	0.9	0.1
	Int	0.2	0.8

Table 3. Confusion matrix - Best classifier for 1:1. 64 nouns selected by DE

4.2. Analysis

False negatives Polysemous adjectives with extensional meanings but labelled 'intensional' in our data were frequently mislabelled. For example, this occurred with 'theoretical' (which often modifies 'physics') and 'artificial' (which often modifies 'intelligence' and 'insemination'). Adjectives such as 'likely' and 'probable' - with exactly one meaning, which is intensional - tended to fare better, and were classified as extensional less frequently. Our classifiers also misclassified 'phony' and 'erstwhile' as extensional, which may be due to the lack of data, stemming from their low frequencies.

False Positives Manually sifting through the false positives revealed many intensional adjectives that we had not considered in our seed list. Two conclusions can be drawn from this. First, the accuracy of our system may be higher than the confusion matrices show. Second, our system recovers intensional adjectives at a rate higher than chance. Some examples include : *"historic", "faulty", "uncertain", "plausible", "erroneous", "illegitimate", "simulated"*.

Polysemous words Certain adjectives in our seed set which were labelled as intensional were actually only intensional when used in a certain sense. For example, assumed culprit may not actually be a culprit, but an assumed name is, in fact, a name. This led us to the inference that intensionality is the characteristic of a particular instance of adjectival modification, and informed the classifiers in 5.

Idioms and common collocations While the word 'potential' in the context of 'potential solution' is intensional, 'potential' occurs most commonly in the corpus as part of the collocation 'potential energy', which is, in fact, subsective. Also, the idioms such as 'false alarm' occur with high frequencies and are not examples of intensional modification. This is further justification for instance-based classification.

Antonyms Many antonyms of privative adjectives were wrongly classified as intensional. These are adjectives that affirm the membership of a noun to its class, and appear in similar context to privative adjectives. It was not expected that instance-based classification could correctly classify these instances, since the ambiguity was inherent in the contexts. However, it was worth noting that these adjectives might cause persistent errors. Some examples include: *"true", "complete", "obvious", "factual"*.

5. Instance-based classification

Method

We modified our hypothesis, claiming that intensionality is a characteristic of particular instances of adjectival modification. In the instance-based classifier, each example corresponds to an adjective-noun pair. One advantage of this model was that we were no longer restricted to a maximum of 30 positive examples, since a positive example could be constructed out of any occurrence of an intensional adjective. For each adjective in the training and test sets, we selected five distinct nouns most frequently modified by it. All adjective-noun pairs in which the adjective was usually intensional were labelled as intensional, although this contradicts the idea behind instance-based classification. We hoped to find reliable false positives that would allow bootstrapping, which would address this problem.

We hypothesised that since a noun N modified by an intensional adjective is no longer 'an N', the contexts in which the modified and unmodified noun occur would be distinct. For every pair of modified noun and context (A-N, y) that co-occurred in the corpus, we calculated

$$f(\text{A-}N, y) = \frac{\Pr(y \mid \text{A-}N)}{\Pr(y \mid \text{unmodified-}N)}.$$

Using add-one smoothing for the probabilities of the unmodified contexts (since is it possible that they might be zero), this can be expressed as

$$f(\text{A-}N, y) = \frac{c(y \cap \text{A-}N) \cdot (c(\text{unmodified-}N) + |Y|)}{c(\text{A-}N) \cdot c(\text{unmodified-}N \cap y)},$$

where c is a counting function. We also calculated,

$$g(\text{A-}N, y) = \frac{c(y \cap \text{A-}N) - c(y \cap \text{unmodified-}N)}{c(y)}$$

The experiments were conducted using token-based contexts, with a window of up to two tokens on either side of the target.

To reduce the number of features used by the classifier, we considered only the top 5000 context features by frequency. Since these were not always informative, we also considered the top 5000 context features after weighting with PMI^2. We also modified the experiment to include a threshold for the number of times a context must co-occur with a noun or adjective-noun instance to be considered.

Results

The instance-based classifiers performed very poorly, achieving accuracy below the majority baseline for all ratios of $m_{extensional} : m_{intensional}$. Without further investigation, this could be due to the assumption made about our training data, that if the adjective is marked as intensional then the adjective-noun pair is as well. A more nuanced training set may improve the results and reduce the large number of false positives and negatives we receive. Results for a representative subset of classifiers are shown in Table 4.

	F1	R	P
f, top 5k contexts	4.05%	29.5%	2.21%
g, top 5k contexts	7.98 %	23.87%	4.89%
f, strict threshold	10.3%	20.45%	7.14%
g, strict threshold	6.34%	18.18%	3.90%
f, PMI^2 contexts	0.20%	9.09%	0.10%
g, PMI^2 contexts	0.20%	9.09%	0.10%

Table 4. Results for instance-based classification.

Analysis

The frequency distribution of the contexts observed was similar to that predicted by Zipf's law. The patterns that occurred with high frequencies were relatively uninformative. Frequencies then dropped off sharply, resulting in high data sparsity. This could be amended by using generalisations. In the noun co-occurrence experiments, we used transformations to ensure that more meaningful co-occurrences were given a higher weight. A more effective measure by which to weight pair-context co-occurrences would contribute to more effective feature selection.

6. Evaluation

As only the adjective-noun co-occurrence classifier performed reasonably and features from the other classifiers were uninformative, we chose to use only the adjective-noun co-occurrence data in our final training set. The linear SVM was trained against all the training data using the appropriate number of features as found maximized the number of correct predictions in the training set.

These confusion matrices show that we often only identify 10% of the intensional adjectives in the test set. Increasing the number of features we trained upon increases this

Initial		Predicted Class	
		Ext	Int
True Class	Ext	921	79
	Int	9	1

Table 5. Confusion matrix - Final evaluation for 100:1. 82 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	87	13
	Int	9	1

Table 6. Confusion matrix - Final evaluation for 10:1. 67 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	9	1
	Int	7	3

Table 7. Confusion matrix - Final evaluation for 1:1. 64 nouns selected by DE

percentage for the lower extensional:intensional ratio samples, indicating that it may just be doing an improved job at classifying those adjectives as being intensional against the *particular* extensional adjectives, rather than the class of all of them.

Further work could divide the intensional adjectives into a few subclasses based on their semantic properties (for example, separating privative and plain non-subsective adjectives) and try to train on each of those. This would hopefully better capture the variety of ways different classes of adjectives are used. Although the instance based features did not work well here, finding a good choice and training with appropriately labelled training data, data that does not assume every instance of an intensional adjective is used in an intensional manner, still feels like it may be useful.

Even though we were not highly successful in classifying unseen adjectives, we did manage to increase the number of known intensional adjectives by 120%.

Acknowledgments

We would like to thank Gabor Angeli for the original project idea of detecting intensionality and for the numerous discussions we had with him during the course of this project. He was always good to bounce ideas off of and also provided us with some of the scripts needed to extract the data from the corpora.

A. Intensional adjectives

From literature: possible, potential, apparent, likely, theoretical, alleged, hypothetical, probable, presumed, putative, former, future, past, false, artificial, impossible, mock, fake, counterfeit, fictitious, ostensible
From synonyms: ex-, phony, virtual, vice, adjunct, unlikely, unnecessary
From false positives: anti-, assumed, mistaken, erstwhile, fictional, seeming, deputy, associate, historic, faulty, uncertain, erroneous, plausible, suspicious, unsuccessful, illegitimate, simulated

References

[1] G. Angeli and C. D. Manning. Philosophers are mortal: Inferring the truth of unseen facts. *CoNLL-2013*, page 133, 2013.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. (B)*, pages 289–300, 1995.

[3] G. Boleda, , et al. First-order vs. higher-order modification in distributional semantics. In *Proceedings of EMNLP*, pages 1223–1233. Association for Computational Linguistics, 2012.

[4] G. Boleda et al. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pages 35–46, 2013.

[5] R.-E. Fan et al. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] M. D. Robinson et al. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.