Evaluating noun phrase modification as a function of object trait variability

Cassandra Jacobs
Term paper for CS 598JHM

May 8, 2012

Abstract

Knowledge about the world constrains how people feel inclined to talk about what they encounter. In particular, redundant traits are often thought to be active during language production but these are seldom encountered (e.g. *red stop sign* is a perfectly valid potential description of a typical *stop sign*, but is uninformative or redundant without further context). In order to better models of natural language understanding and natural language generation, it is important to properly model the attributes of objects in such a way that redundancy can be minimized and informatively maximized. This paper examines current models of attributes in the computer vision literature as well as models of referring expression generation and proposes one method for examining the relationship between the two. The results suggest a potential relationship between linguistic modification of nouns and visual variability of an object category, though the work is not conclusive.

Background

Current research looking at models of attributes often focuses on semantic knowledge representation by propositions such as **HAS_FEATHERS** for *sparrow* or *penguin*. Models exist that can identify the unusual attributes of an object or scene [5, 2], and some models attempt to identify the factors that people consider when generating referring expressions to generate text automatically using computers [15]. In general, the process of learning the attributes of objects, especially from image data, results in increasingly good understanding of what these properties' semantics actually are. Unfortunately, this is not sufficient for producing natural referring expressions for objects, given that there are many other psychological factors at work.

The problem is two-fold: because people tend to over specify and are extremely sensitive to the discourse context in which a referring expression is produced, it is important to identify the types of objects and properties that are most consistently modified using an adjective or relative clause. However, while referring expression generation can be guided by corpus statistics [16], these do not necessarily generalize to larger sets of objects with unpredictable features, and the concept of *common ground* can be difficult to define objectively, especially in the case of prototypically gradable adjectives like *big* or *cold* [3].

Another aspect of the problem stems from the fact that the properties of objects cannot be learned from

corpus statistics alone; while there are many distributional similarities that could be inferred (e.g. part of speech), it is not clear what the difference is between *green banana* and *green car* with respect to the informativity of these phrases. It is therefore important to assess the attributes of the objects in as non-linguistic of an environment as possible, though one which is still semantically (i.e. propositionally) grounded.

Human referring expression production and comprehension

Humans are mostly sensitive to the discourse context they are in. When asked to describe a scene to another person, participants in an experiment will generally provide a more specific name for an object if it is disambiguating. When given the choice between two identical images of different sizes, participants almost universally chose to modify the referents appropriately to be informative about the salient differences between the two objects [7].

However, people tend to over-describe the things they talk about in their current discourse [4], which means that the most obvious measures of informativeness (e.g. Gricean maxims) in language production are not sufficient to describe the decisions that must be made in generating referring expressions. Work by both Engelhardt et al. (2006) and Sedivy (2003) suggests that while people are somewhat confused by the content of redundant or overly informative referring expressions and may even temporarily refuse to accept a redundant, non-contrastive referring expression [4, 13], they do not necessarily judge such expressions as being of a lower quality [4]. There also tend to be a lot of differences between individuals on the task, meaning that while one referring expression is sufficient to describe an object, there are likely many others [16], any of which could say something different about an object.

Currently much of the work that examines how people calculate the salience of the traits of objects centers on formal semantics and poorly-specified common ground. In general, the concept of a *big mouse* or a *green mouse* is taken, perhaps correctly, to be a function of the natural traits of that object. In general, it is only worth mentioning the size of an object if it is in fact *big* or *small*, at least within the understanding of communicative efficiency [9]. Models that attempt to capture the fact that mice are small, and therefore that a big mouse is worth mentioning, are not strictly successful at generating referring expressions for these types of objects, instead setting hard-and-fast rules to try to capture the gradability of human judgments [3].

Corpus statistics can indeed inform about what sorts of things are talked about, but they cannot tell us why, or even necessarily what makes it worth talking about in the first place. While it is not really a new concept in psychology that experiential and linguistic information work in complementary ways [1, 10], it is worthwhile to examine what can be learned automatically using machine learning in computer vision. Incorporating computer vision techniques with psycholinguistic and corpus data has the potential to enhance the naturalness of automatically generated referring expressions.

Techniques for supervised and unsupervised learning of object features and attributes

There has been substantial research on pattern recognition and the validity of descriptions in computer vision in artificial intelligence. In particular, researchers in computer vision are interested in object identification and the acquisition of the meanings of properties [12], such as identifying which one of the men in a hypothetical picture is wearing green pants, and what colors actually count as *green* in that picture.

Some attempts to identify whether an object is unusual given its environment have been demonstrated with context models [2], which make use of graphical models to allow for multi-correlated structures, such as grass and trees, or airplanes and wings. In the model outlined in [2], the metrics used to evaluate context are not strictly probabilistically dependent on surrounding objects; relations and visual expectations (such as size or position) also play into the successful identification of out-of-context objects.

While not strictly applicable to the study of referring expression generation, the model they use, which is a latent tree model taking into account object co-occurrences. The latent variables, which are the contexts, are constructed from labeled data, which constitute the objects within those contexts. This allows for the model to learn concepts that line up well with *kitchen* or *park*, for example.

Models of this sort for objects instead of scenes can be made analogously by considering a set of properties that might make up an object. Such research has been conducted [5] such that semantic relationships encompassing what were separately co-occurrence and support trees in [2], may be succinctly expressed with propositions like the earlier **HAS_FEATHERS**. One strength of the model they construct to learn such relationships is that the classifiers work within objects first, and then generalize across classes. The ability to generalize while also allowing for individual differences enhances the validity of the constructs. The model of Farhadi et al. (2009) successfully identifies many unusual features, with unusual being the presence or the absence of some trait that is associated with the object.

Importantly, the work in [5] shows that strictly assessing the correlations between objects results in poor annotation and feature identification, while it may have been sufficient in identifying atypicality in a scene as in [2]. While this research does provide a substantial amount of information about unusual visual scenes and can even provide the reasons why the scenes are unusual, it is not clear what criteria people use when talking about these unusual circumstances, even if the ease of comprehending statements about them seems to directly mirror the informativity of a statement [4, 13].

Datasets

The goal of the project was to identify the contribution of visual uncertainty within object categories. In order to do this, an image dataset with sufficiently common object categories with highly frequent noun equivalents containing feature annotations was needed. Ideally, the dataset would be annotated for linguistically relevant features, especially common adjective categories (colors, shapes, etc.).

Attribute Datasets

The dataset described in Farhadi et al. (2009) served as the image and feature dataset for visual complexity. One of the main advantages of this dataset was the size, containing some 10,000 images across 32 object categories, all with at least 150 exemplars.[1] Each image in both of these datasets has been annotated for 64 image features, ranging from *3DBoxy* to *Hand* [5], broadly spanning shapes, parts, and materials. This means that the dataset has an enormous expressive advantage in terms of the ability to describe an object category's complexity.

[1] One potential drawback of this dataset is also its size. Because each of the object categories has an unequal representation (e.g. *people* contains approximately 6,000 images), it cannot be said that some of the variance in any potential relationship could not be attributable to the number of items, even if we are interested in feature proportions, as larger sets of images for an object class may result in smaller variances, such that *person* may be more detailed than *centaur*.

a-PASCAL

The a-PASCAL dataset contains 20 common object classes of 10,000 pictures total. The object classes are: people, bird, cat, cow, dog, horse, sheep, airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, and TV/monitor. For the purposes of finding linguistic parallels using our linguistic datasets, items from this dataset that were initially labeled as compounds were removed.[2] This is due also to the fact that prenominal modification of compounds could potentially be unnatural relative to noun phrases containing only a single noun.

a-Yahoo

The a-Yahoo dataset, part of the same set of experiments in [5] contains an additional 12 object categories, which are similarly common, but selected to be structurally similar to objects in the a-PASCAL dataset. These object categories are: wolf, zebra, goat, donkey, monkey, statue of people, centaur, bag, building, jet ski, carriage, and mug.[3]

This dataset may be more appropriate for determining the degree to which objects are themselves variable, given that the terms themselves do not align very well with the terminology that is very natural to use when describing those objects. For instance, *3DBoxy* is not a very natural trait to mention and has a very infrequent linguistic counterpart *3D*. The lack of appropriate terms that align with actual adjectives, in particular color and size terms, is a limitation of the dataset that cannot be easily addressed through object recognition techniques and would require extensive human annotation.

[2]The "compounds" that were removed included *motorbike*, *dining table*, *potted plant*, and *TV/monitor*.

[3]As in the a-PASCAL dataset, compounds were removed for the ease of n-gram searches. This resulted in the exclusion of jet ski.

Linguistic Surprisal

Multiple sources of linguistic entropy exist. For the purposes of this proposal, comparing the output of the generated referring expressions will be important. In order to gain more precise counts of the probabilities of any given modification for an object, linguistic entropies calculated from backward transitional probabilities between adjectives and the nouns extracted from the PASCAL image dataset (from [6]) can provide an understanding of how likely that object is to be modified prenominally overall.

These linguistic probabilities would be calculated from relatively unambiguous 3- and 4-grams in the Google n-gram corpus, which would allow for a semi-accurate calculation of the degree to which an item is likely to be modified relative to its probability of occurring alone.

Surprisal of Generic Prenominal Modification

In general, we would like to iterate over all possible unambiguous constructions. To do so, we searched for phrases of the type **[determiner] [modifier] [noun] [is]** for the 4-grams, and **[determiner] [noun] [is]** for the 3-grams.[4]

The surprisal of a noun being modified is proportional to the number of times it is modified (the 4-gram frequencies) relative to the number of times it is not (the 3-gram frequencies):

$$-\log_2\left(\frac{\sum_{i=1} freq(modifier_i + noun)}{\sum_{j=1} freq(determiner_j + noun)}\right)$$

Higher values are intuitively more surprising, meaning that nouns with a very small proportion of modified noun phrases to bare noun phrases has a high surprisal value.

[4]The determiners selected were *the*, *this*, and *a*. Generics (e.g. *centaurs*) were not searched partly due to the expected redundancy of the copular constructions for generic noun phrases.

Surprisal of Adjectival Modification

There are various reasons to suspect that adjectives may behave differently from nouns or participles, though the difference between these constructions is blurred [8]. In particular, adjectives denote properties rather than kinds of objects. So, there may be a difference between *big cat* and *house cat*.

We constructed a subset of the 4-gram dataset using WordNet, which contains part-of-speech information that can be used to disambiguate a word [11], to select all adjectives from the set of prenominal modifiers. Unfortunately, doing so eliminated color terms, which are generally classified as nouns in WordNet.

The surprisal of adjectival modification is calculated in the same way as for generic prenominal modification. Because the adjectives form a subset of possible prenominal modifiers, surprisal values for adjectival modification of a noun will be universally higher for all objects in our dataset.

$$-\log_2\left(\frac{\sum_{i=1} freq(adjective_i + noun)}{\sum_{j=1} freq(determiner_j + noun)}\right)$$

Two items from the a-PASCAL and a-Yahoo datasets, *centaur* and *sofa*, were not considered in any further analyses because they were found to contain no adjectival modifications, despite having other kinds of modifiers. This was potentially due to the constructions searched, which were naturally more restricted than noun phrases in general by requiring the nouns in question to be subjects of a clause, so it is possible that adjectival modification data could be extracted from the Google corpus in other ways to obtain values for these objects.

Trait Entropy

It is important to be able to generalize from both object attributes and corpus statistics. Using visual attribute data corresponding to a word, such as a prenominal modification (most commonly an adjective), the variability in an object's possessing that attribute should reflect the degree to which that attribute is mentioned.

An appropriate measure for trait variability or trait consistency is entropy, which can be calculated from proportions or frequencies of traits for an object class. Entropy has the advantage of weighting probabilities far from 0 and 1 as having greater uncertainty, since proportions near 0 and 1 are very uninformative about what sort of object we are dealing with. This is analogous to cars coming in multiple colors, but stop signs only coming in one.

Entropy for a given object type across the 64 traits is calculated as:

$$-\sum_{i=1}^{64} p(trait_i) \times \log_2(p(trait_i))$$

That is, the proportions of each trait ($p(trait_i)$) are calculated for an object and multiplied by their log probability. In the instances where $p(trait_i) = 0$, the value of $\log_2(p(trait_i))$ is set to 0 as well, given that these are otherwise undefined. These values are then summed up across all the traits (64 in our dataset) for an object. The value obtained is unique to every object in our dataset. Larger values are indicative of a high variability in the presence of the properties of objects. Importantly, these values are blind to the actual features of the objects. Objects can be seen as more or less equally variable despite varying along different dimensions and combinations of features.

Results

Linguistic surprisal values for generic prenominal modification were plotted as a function of trait entropy, or trait consistency. Following our hypotheses about what constitutions maximally informa-

tive trait mention, we expect lower surprisal values (greater odds of modification) for objects whose traits are highly inconsistent or uncertain. That is, there should be a negative relationship between linguistic surprisal and trait consistency.

Analysis

Initial analyses showed that in all subsets of the dataset (a-PASCAL training, a-PASCAL test, and a-Yahoo), the items *boat* and *airplane* were clear outliers. The reasons for this are unclear, since this could have been driven by abnormal linguistic surprisal values or abnormal trait entropies. This could have been due to the presence of people in the images, which would result in the identification of several traits which did not necessarily apply to boats or airplanes themselves. It is also possible that the terms *boat* and *airplane* do not represent the most typical ways people refer to these objects. For instance, *boat* is frequently used in compounds (e.g. *sailboat*) and *airplane* is generally reduced to *plane*. For this reason, these items were not considered in any analysis, and no further outliers were removed in subsequent analyses.

Simple linear regression was the primary form of analysis, specifically identifying whether there was a significant slope coefficient which could characterize the probability of using a prenominal modifier for the nouns of interest. Significant slope coefficients are taken as evidence that there is a linear or near-linear relationship between these two variables.

Generic Prenominal Modification

There was a significant linear relationship between linguistic surprisal of generic prenominal modification and the variability or consistency of an object's traits in the image dataset, $R^2 = .18$, $t(20) = -2.102, p < .05$. This suggests that there is a negative relationship between whether an object will be modified and the degree to which its features are uncertain. This result supports the hypothesis that people are informative when they need to be, specifying more for objects that are more variable in the real world. The relationship between these two variables is presented in Figure 1.

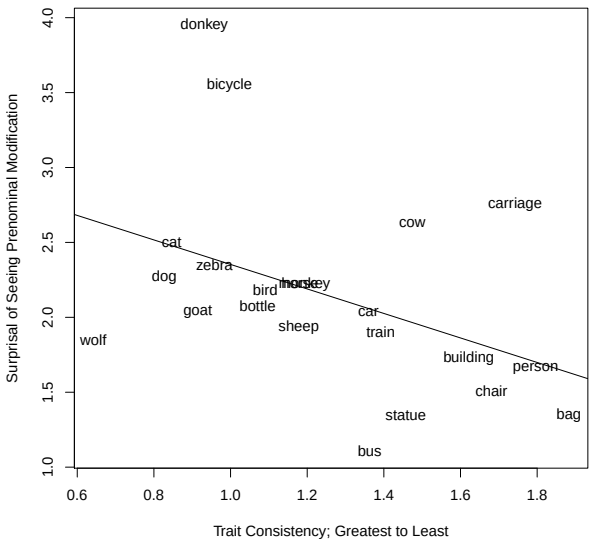

Figure 1: Generic linguistic surprisal as a function of trait consistency. There is a strong relationship between prenominal modification and the variability of an object's features.

Adjectival Modification

Adjectives are potentially informative in different ways, given that they often denote properties of an object and not its kind or what it is a result of, in the case of nominal and participial prenominal modification. In our dataset, the surprisal scores for adjectival modification are strongly correlated with the

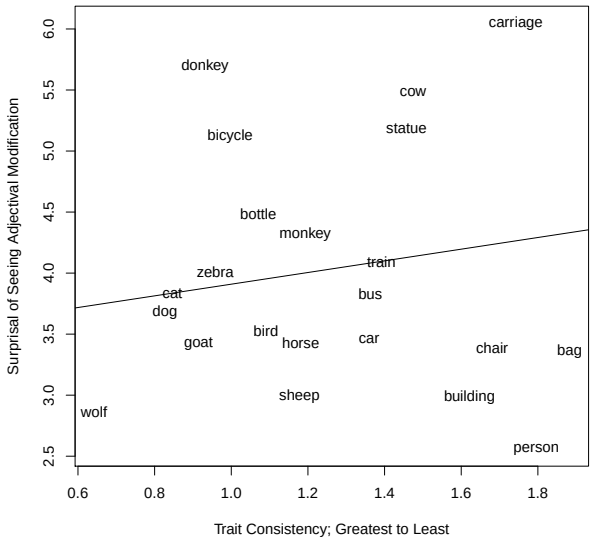

Figure 2: Adjectival linguistic surprisal as a function of trait consistency. There is no clear relationship between prenominal adjective use and the variability of an object's features.

surprisal scores for generic prenominal modification, ($R^2 = .66$, $t(22) = 4.1302, p < .001$), but it is possible that whether an adjective is expected or not differs from whether any kind of modification is expected. Adjectives may, for instance, be less sensitive to the structural variability of an object, given that adjectives most often describe different sorts of traits, most commonly dimensions, qualities, colors, and age [14].

Using the same type of analysis as for generic prenominal modification, we found that there was no significant relationship between the probability of modifying an object with an adjective and how variable those items are in the real world ($R^2 = .05$, $t(22) = 1.093, p = .286$). This result is reported in Figure 2.

Discussion

These results of the adjective analysis are different from those in the overall analysis. This suggests that there is something functionally different between the full dataset and the adjective-only dataset. The n-gram data are themselves not perfect. It is also the case that the 4-gram and 3-gram data are only a subset of all of the instances of these objects in the Google corpus. These words may occur in many more contexts than copular constructions with our three constraining singular determiners; those constructions were avoided because of their potential ambiguity. The adjective-only dataset also does not contain color terms, which are among the most salient of modifications, which are often over-specified by individuals when asked to describe a scene [16].

This could be due to the constructions in the full dataset. It is possible that the nominal modifications, for instance, constitute collocations which could be identified by searching for them as children of the bare nouns in the WordNet database. The fact that the constructions are seen as equivalent in certain theories of syntax [8] suggests that it may be difficult to disambiguate nominal, participial, and adjectival prenominal modification, even if it were possible to identify all the nominal collocations. Furthermore, it is not clear why nominal collocations would be less informative when needing to specify what an object is, considering the need for terms such as *zebra* to describe a particularly exotic, striped horse. So, while there are probably phrases like *school bus* in our full dataset, it is not clear whether such a phrase is less informative than *yellow bus*, unless we assume that adjectives are informative in different ways than nouns are.

Nevertheless, the marked difference between the informativity of all prenominal constructions indicates that there is more to be done to understand the

contribution of trait variability to adjectival modification. It may still be the case that there exists a relationship between the number of ways an object can be talked about and its ability to vary in the real world; there is no particular reason to abandon the hypothesis that people do modify the entities they talk about in a more or less rational way just because the current data seem to not support the idea that adjective use is related to structural variability.

Future Directions

Because there are clear limitations to the linguistic and feature datasets outlined in this paper, it will be important in future work to develop a more sophisticated model of the behavior we are wishing to emulate in generating referring expressions. While it may be tempting to avoid ambiguity [16], it may be even more important to capture the knowledge that people have and attempt to emulate the n-gram counts that are available for prenominal modification in general.

It may be the case that obtaining a more relevant set of traits for object classes, say, those that are built entirely on terms that are actually employed when people describe entities, could result in a model of adjective usage that looks more like Figure 1 than Figure 2. If overspecified annotations of objects, such as color, size, quality, age, or shapes really do point to the salient properties of those objects, then redundant expressions can inform us about the properties that are distributed over an object class.

In order to do this, these traits must be extracted from utterances that are overspecified for the entities that are being described. In particular, instead of a phrase like *The dog is biting the girl*, an overinformative utterance would specify the properties of the objects in that phrase (e.g. *black dog* or *young girl*). By extracting the co-occurrence statistics between named entities and their named properties, support contexts as in [2] can be constructed. The model in [2] effectively determined whether items were out of their typical context, similar in spirit to [5]. Being able to do this with linguistically relevant features should result in a model that can effectively generate noun phrases such as *the black dog* or *the small elephant*, if a feature for a given object is sufficiently surprising. It is possible that such a model can even take into account the same assumptions that were verified in the first analysis, which is that the odds of modification, though not necessarily adjectival modification, increase linearly with the uncertainty of the physical features of an object category, or perhaps following the algorithms in [5] and [2].

References

[1] Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498, Jan 2009.

[2] M Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, Jan 2012.

[3] K Van Deemter. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222, 2006.

[4] Paul E Engelhardt, Karl G D Bailey, and Fernanda Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54:554–573, Mar 2006.

[5] A Farhadi, I Endres, D Hoiem, and D Forsyth. Describing objects by their attributes. *Computer Vision and Pattern Recognition, 2009.*

CVPR 2009. IEEE Conference on, pages 1778–1785, 2009.

[6] A Farhadi, M Hejrati, M Sadeghi, P Young, C Rashtchian, J Hockenmaier, and D Forsyth. Every picture tells a story: Generating sentences from images. *Computer Vision–ECCV 2010*, pages 15–29, 2010.

[7] V.S Ferreira, L.R Slevc, and E.S Rogers. How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3):263–284, 2005.

[8] A.E Goldberg. Constructions: a new theoretical approach to language. *Trends Cogn Sci (Regul Ed)*, 7(5):219–224, 2003.

[9] H P Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, Jan 1975.

[10] Brendan T Johns and Michael N Jones. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, Jan 2012.

[11] George A Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, Mar 1995.

[12] O Russakovsky and L Fei-Fei. Attribute learning in large-scale datasets. *ECCV 2010 Workshop on Parts and Attributes*, 2010.

[13] J.C Sedivy. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1):3–23, 2003.

[14] A Spencer. Adjective classes: A cross-linguistic typology (review). *Language*, 84(2):407–409, 2008.

[15] Henriette Anna Elisabeth Viethen. The generation of natural descriptions. *Unpublished doctoral dissertation*, pages 1–254, Mar 2011.

[16] Jette Viethen, Robert Dale, Emiel Krahmer, Mariët Theune, and Pascal Touset. Controlling redundancy in referring expressions. 2008.