Predicting Relative Prominence in Noun-Noun Compounds

Taniya Mishra
AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932
taniya@research.att.com

Srinivas Bangalore
AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932
srini@research.att.com

Abstract

There are several theories regarding what influences prominence assignment in English noun-noun compounds. We have developed corpus-driven models for automatically predicting prominence assignment in noun-noun compounds using feature sets based on two such theories: the informativeness theory and the semantic composition theory. The evaluation of the prediction models indicate that though both of these theories are relevant, they account for different types of variability in prominence assignment.

1 Introduction

Text-to-speech synthesis (TTS) systems stand to gain in improved intelligibility and naturalness if we have good control of the prosody. Typically, prosodic labels are predicted through text analysis and are used to control the acoustic parameters for a TTS system. An important aspect of prosody prediction is predicting which words should be prosodically *prominent*, i.e., produced with greater energy, higher pitch, and/or longer duration than the neighboring words, in order to indicate the former's greater communicative salience. Appropriate prominence assignment is crucial for listeners' understanding of the intended message. However, the immense prosodic variability found in spoken language makes prominence prediction a challenging problem. A particular sub-problem of prominence prediction that still defies a complete solution is prediction of relative prominence in noun-noun compounds.

Noun-noun compounds such as ***White*** *House, cherry* ***pie****,* ***parking*** *lot, Madison* ***Avenue****,* ***Wall*** *Street,* ***nail*** *polish,* ***french*** *fries, computer* ***programmer****,* ***dog*** *catcher, silk* ***tie****, and self* ***reliance***, occur quite frequently in the English language. In a discourse neutral context, such constructions usually have *leftmost prominence*, i.e., speakers produce the left-hand noun with greater prominence than the right-hand noun. However, a significant portion — about 25% (Liberman and Sproat, 1992) — of them are assigned rightmost prominence (such as cherry pie, Madison Avenue, silk tie, computer programmer, and self reliance from the list above). What factors influence speakers' decision to assign left or right prominence is still an open question.

There are several different theories about relative prominence assignment in noun-noun (henceforth, NN) compounds, such as the structural theory (Bloomfield, 1933; Marchand, 1969; Heinz, 2004), the analogical theory (Schmerling, 1971; Olsen, 2000), the semantic theory (Fudge, 1984; Liberman and Sproat, 1992) and the informativeness theory (Bolinger, 1972; Ladd, 1984).[1] However, in most studies, the different theories are examined and applied in isolation, thus making it difficult to compare them directly. It would be informative and illuminating to apply these theories to the same task and the same dataset.

For this paper, we focus on two particular theories, the informativeness theory and the semantic composition theory. The informativeness theory posits that the relatively more informative and unexpected noun is given greater prominence in the NN compound than the less informative and more predictable noun. The semantic composition theory posits that relative prominence assignment in NN compounds is decided according to the semantic relationship between the two nouns.

We apply these two theories to the task of predicting relative prominence in NN compounds via statistical corpus-driven methods, within the larger context of building a system that can predict appropriate prominence patterns for text-to-speech synthesis. Here we are only focusing on predicting relative prominence of NN compounds in a neutral context, where there are no pragmatic reasons (such as contrastiveness or given/new distinction) for shifting prominence.

[1] In-depth reviews of the different theories can be found in Plag (2006) and Bell and Plag (2010).

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 609–613,
Portland, Oregon, June 19-24, 2011.

2 Informativeness Measures

We used the following five metrics to capture the individual and relative informativeness of nouns in each NN compound:

- Unigram Predictability (UP): Defined as the predictability of a word given a text corpus, it is measured as the log probability of the word in the text corpus. Here, we use the maximum likelihood formulation of this measure.

$$\text{UP} = \log \frac{Freq(w_i)}{\sum_i Freq(w_i)} \quad (1)$$

  This is a very simple measure of word informativeness that has been shown to be effective in a similar task (Pan and McKeown, 1999).

- Bigram Predictability (BP): Defined as the predictability of a word given a previous word, it is measured as the log probability of noun N2 given noun N1.

$$\text{BP} = \log\left(Prob(N2 \mid N1)\right) \quad (2)$$

- Pointwise Mutual Information (PMI): Defined as a measure of how collocated two words are, it is measured as the log of the ratio of probability of the joint event of the two words occurring and the probability of them occurring independent of each other.

$$\text{PMI} = \log \frac{Prob(N1, N2)}{Prob(N1)Prob(N2)} \quad (3)$$

- Dice Coefficient (DC): Dice is another collocation measure used in information retrieval.

$$\text{DC} = \frac{2 \times Prob(N1, N2)}{Prob(N1) + Prob(N2)} \quad (4)$$

- Pointwise Kullback-Leibler Divergence (PKL): In this context, Pointwise Kullback-Leibler divergence (a formulation of relative entropy) measures the degree to which one overapproximates the information content of N2 by failing to take into account the immediately preceding word N1. (PKL values are always negative.) A high absolute value of PKL indicates that there is not much information contained in N2 if N1 is taken into account. We define PKL as

$$Prob(N2 \mid N1) \log \frac{Prob(N2 \mid N1)}{Prob(N2)} \quad (5)$$

  Another way to consider PKL is as PMI normalized by the predictability of N2 given N1.

All except the first the aforementioned five informativeness measures are relative measures. Of these, PMI and Dice Coefficient are symmetric measures while Bigram Predictability and PKL are non-symmetric (unidirectional) measures.

3 Semantic Relationship Modeling

We modeled the semantic relationship between the two nouns in the NN compound as follows. For each of the two nouns in each NN compound, we maintain a semantic category vector of 26 elements. The 26 elements are associated with 26 semantic categories (such as food, event, act, location, artifact, etc.) assigned to nouns in WordNet (Fellbaum, 1998). For each noun, each element of the semantic category vector is assigned a value of 1, if the *lemmatized noun* (i.e., the associated uninflected dictionary entry) is assigned the associated semantic category by WordNet, otherwise, the element is assigned a value of 0. (If a semantic category vector is entirely populated by zeros, then that noun has not been assigned any semantic category information by WordNet.) We expected the cross-product of the semantic category vectors of the two nouns in the NN compound to roughly encode the possible semantic relationships between the two nouns, which — following the semantic composition theory — correlates with prominence assignment to some extent.

4 Semantic Informativeness Features

For each noun in each NN compound, we also maintain three semantic informativeness features: (1) Number of possible synsets associated with the noun. A *synset* is a set of words that have the same sense or meaning. (2) Left positional family size and (3) Right positional family size. *Positional family size* is the number of unique NN compounds that include the particular noun, either on the left or on the right (Bell and Plag, 2010). These features are extracted from WordNet as well.

The intuition behind extracting synset counts and positional family size was, once again, to measure the relative informativeness of the nouns in NN compounds. Smaller synset counts indicate more specific meaning of the noun, and thus perhaps more information content. Larger right (or left) positional family size indicates that the noun is present

in the right (left) position of many possible NN compounds, and thus less likely to receive higher prominence in such compounds.

These features capture type-based informativeness, in contrast to the measures described in Section 2, which capture token-based informativeness.

5 Experimental evaluation

For our evaluation, we used a hand-labeled corpus of 7831 NN compounds randomly selected from the 1990 Associated Press newswire, and hand-tagged for leftmost or rightmost prominence (Sproat, 1994). This corpus contains 64 pairs of NN compounds that differ in terms of capitalization but not in terms of relative prominence assignment. It only contains four pairs of NN compounds that differ in terms of capitalization and in terms of relative prominence assignment. Since there is not enough data in this corpus to consider capitalization as a feature, we removed the case information (by lowercasing the entire corpora), and removed any duplicates. Of the four pairs that differed in terms of capitalization, we only retained the lower-cased NN compounds. By normalizing Sproat's hand-labeled corpus in this way, we created a slightly smaller corpus 7767 utterances that was used for the evaluation.

For each of the NN compounds in this corpus, we computed the three aforementioned feature sets. To compute the informativeness features, we used the LDC English Gigaword corpus. The semantic category vectors and the semantic informativeness features were obtained from Wordnet. Using each of the three feature sets individually as well as combined together, we built automatic relative prominence prediction models using Boostexter, a discriminative classification model based on the boosting family of algorithms, which was first proposed in Freund and Schapire (1996).

Following an experimental methodology similar to Sproat (1994), we used 88% (6835 samples) of the corpus as training data and the remaining 12% (932 samples) as test data. For each test case, the output of the prediction models was either a 0 (indicating that the leftmost noun receive higher prominence) or a 1 (indicating that the rightmost noun receive higher prominence). We estimated the model error of the different prediction models by computing the relative error reduction from the baseline error. The baseline error was obtained by assigning the majority class to all test cases. We avoided overfitting by using 5-fold cross validation.

5.1 Results

The results of the evaluation of the different models are presented in Table 1. In this table, INF denotes informativeness features (Sec. 2), SRF denotes semantic relationship modeling features (Sec. 3) and SIF denotes semantic informativeness features (Sec. 4). We also present the results of building prediction models by combining different features sets.

These results show that each of the prediction models reduces the baseline error, thus indicating that the different types of feature sets are each correlated with prominence assignment in NN compounds to some extent. However, it appears that some feature sets are more predictive. Of the individual feature sets, SRF and INF features appear to be more predictive than the SIF features. Combined together, the three feature sets are most predictive, reducing model error over the baseline error by almost 33% (compared to 16-22% for individual feature sets), though combining INF with SRF features almost achieves the same reduction in baseline error.

Note that none of the three types of feature sets that we have defined contain any direct lexical information such as the nouns themselves or their lemmata. However, considering that the lexical content of the words is a rich source of information that could have substantial predictive power, we included the lemmata associated with the nouns in the NN compounds as additional features to each feature set and rebuilt the prediction models. An evaluation of these lexically-enhanced models is shown in Table 2. Indeed, addition of the lemmatized form of the NN compounds substantially increases the predictive power of all the models. The baseline error is reduced by almost 50% in each of the models — the error reduction being the greatest (53%) for the model built by combining all three feature sets.

6 Discussion and Conclusion

Several other studies have examined the main idea of relative prominence assignment using one or more of the theories that we have focused on in this paper (though the particular tasks and terminology used were different) and found similar results. For example, Pan and Hirschberg (2000) have used some of the same informativeness measures (denoted by INF above) to predict pitch accent placement in word bi-

Feature Sets	Av. baseline error (in %)	Av. model error (in %)	% Error reduction
INF	29.18	22.85	21.69
SRF	28.04	21.84	22.00
SIF	29.22	24.36	16.66
INF-SRF	28.52	19.53	**31.55**
INF-SIF	28.04	21.25	24.33
SRF-SIF	29.74	21.30	28.31
All	28.98	19.61	**32.36**

Table 1: Results of prediction models

Feature Sets	Av. baseline error (in %)	Av. model error (in %)	% Error reduction
INF	28.6	14.67	48.74
SRF	28.34	14.29	49.55
SIF	29.48	14.85	49.49
INF-SRF	28.16	14.81	47.45
INF-SIF	28.38	14.16	50.03
SRF-SIF	29.24	14.51	50.30
All	28.12	13.19	**52.95**

Table 2: Results of lexically-enhanced prediction models

grams. Since pitch accents and perception of prominence are strongly correlated, their conclusion that informativeness measures are a good predictor of pitch accent placement agrees with our conclusion that informativeness measures are useful predictors of relative prominence assignment. However, we cannot compare their results to ours directly, since their corpus and baseline error measurement[2] were different from ours.

Our results are more directly comparable to those shown in Sproat (1994). For the same task as we consider in this study, besides developing a rule-based system, Sproat also developed a statistical corpus-based model. His feature set was developed to model the semantic relationship between the two nouns in the NN compound, and included the lemmata related to the nouns. The model was trained and tested on the same hand-labeled corpus that we used for this study and the baseline error was measured in the same way. So, we can directly compare the results of our lexically-enhanced SRF-based models to Sproat's corpus-driven statistical model. In his work, Sproat reported a baseline error of 30% and a model error of 16%. The reported relative improvement over the baseline error in Sproat's study was 46.6%, while our relative improvement using the lexically enhanced SRF based model was 49.5%, and the relative improvement using the combined model is 52.95%.

Type-based semantic informativeness features of the kind that we grouped as SIF were analyzed in Bell and Plag (2010) as potential predictors of prominence assignment in compound nouns. Like us, they too found such features to be predictive of prominence assignment and that combining them with features that model the semantic relationship in the NN compound makes them more predictive.

7 Conclusion

The goal of the presented work was predicting relative prominence in NN compounds via statistical corpus-driven methods. We constructed automatic prediction models using feature sets based on two different theories about relative prominence assignment in NN compounds: the informativeness theory and the semantic composition theory. In doing so, we were able to compare the two theories.

Our evaluation indicates that each of these theories is relevant, though perhaps to different degrees. This is supported by the observation that the combined model (in Table 1) is substantially more predictive than any of the individual models. This indicates that the different feature sets capture different correlations, and that perhaps each of the theories (on which the feature sets are based) account for different types of variability in prominence assignment.

Our results also highlight the difference between being able to use lexical information in prominence prediction of NN compounds, or not. Using lexical features, we can improve prediction over the default case (i.e., assigning prominence to the left noun in all cases) by over 50%. But if the given input is an out-of-vocabulary NN compound, our non-lexically enhanced best model can still improve prediction over the default by about 33%.

Acknowledgment We would like to thank Richard Sproat for freely providing the dataset on which the developed models were trained and tested. We would also like to thank him for his advice on this topic.

[2]Pan and Hirschberg present error obtained by using a unigram-based predictability model as baseline error. It is unclear what is the error obtained by assigning left prominence to *all* words in their database, which was our baseline error.

References

M. Bell and I. Plag. 2010. Informativeness is a determinant of compound stress in English. Submitted for publication. Obtained from `http://www2.uni-siegen.de/~engspra/publicat.html` on February 12, 2010.

L. Bloomfield. 1933. *Language*, Holt, New York.

D. Bolinger. 1972. Accent is predictable (if you're a mind-reader). *Language* 48.

C. Fellbaum (editor). 1998. *WordNet: An Electronic Lexical Database*, The MIT Press, Boston.

Y. Freund and R. E. Schapire, 1996. Experiments with a new boosting alogrithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148-156.

E. Fudge. 1984. *English Word-Stress*, Allen and Unwin, London and Boston.

H. J. Giegerich. Compound or phrase? English noun-plus-noun constructions and the stress criterion. In *English Language and Linguistics*, 8:1–24.

R. D. Ladd, 1984. English compound stress. In Dafydd Gibbon and Helmut Richter (eds.) *Intonation, Accent and Rhythm: Studies in 1188 Discourse Phonology*, W de Gruyter, Berlin.

M. Liberman and R. Sproat. 1992. The Stress and Structure of Modified Noun Phrases in English. In I. Sag (ed.), *Lexical Matters*, pp. 131–181, CSLI Publications, Chicago, University of Chicago Press.

H. Marchand. *The categories and types of present-day English word-formation*, Beck, Munich.

S. Olsen. 2000. Compounding and stress in English: A closer look at the boundary between morphology and syntax. *Linguistische Berichte*, 181:55–70.

S. Pan and J. Hirschberg. 2000. Modeling local context for pitch accent prediction. *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*, pp. 233-240, Hong Kong. ACL.

S. Pan and K. McKeown. 1999. Word informativeness and automatic pitch accent modeling. *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*, pp. 148–157.

I. Plag. 2006. The variability of compound stress in English: structural, semantic and analogical factors. *English Language and Linguistics*, 10.1, pp. 143–172.

R. Sproat. 1994. English Noun-Phrase Accent Prediction for Text-to-Speech. *Computer Speech and Language*, 8, pp. 79–94.

R.E. Schapire, A brief introduction to boosting. In *Proceedings of IJCAI*, 1999.

S. F. Schmerling. 1971. A stress mess. *Studies in the Linguistic Sciences*, 1:52–65.