Mapping Syntax and Prosody

Tae-Jin Yoon
Department of Linguistics
University of Illinois at Urbana-Champaign
tyoon@uiuc.edu

The relationship between prosodic structure and syntactic structure has remained a controversial and unresolved area, partly due to the lack of rich corpora of natural speech, and partly due to the complexity involved in both syntax and prosody. Chomsky & Halle (1968, p. 372) state that "although there is a substantial literature on intonational and prosodic features in English, it is largely restricted to citation of examples, and we cannot draw on it for any significant insight into processes of a general nature." More recently, Ladd (1996, p. 334) adds that "in the standard theory, the correspondence between syntactic constituent types and prosodic ones is highly variable, since the make-up of the prosodic constituents is influenced by a variety of essentially linear factors." Despite the *status quo*, different views on the mapping from syntax to prosody have been proposed: (1) Syntax alone determines most of prosodic structure (Cooper & Paccia-Cooper 1980, Langendoen 1975, Downing 1970, Inkelas 1989, Taglich 1998, Truckenbrodt 1999, Steedman 2000); (2) speakers use prosody that signals the syntactic information only when ambiguity is involved (Snedeker & Truswell 2003, Allbritton et al. 1996); (3) Many linguistic and para-linguistic factors along with syntax determine a prosodic structure (Bachenko & Fitzpatrick 1990, Gee & Grosjean 1983, Schafter 2004, Watson & Gibson 2004). Besides the linguistic and psycholinguistic studies, machine learning approaches have been employed to improve the performance of TTS or ASR either using hand-built rules (Ostendorf & Veilleux 1994, Wang & Hirschberg 1992, Bachenko & Fitzpatrick 1990, Erwin 2001) or using stochastic classification algorithms (Taylor & Black 1998, Cohen 2004, Ingulfsen 2004).

Much linguistic and psycholinguistic research is limited in that it has often relied on data from intuition, or small collections of recorded speech. As for machine learning approaches using syntactic parser, a concern is reflected by Taylor & Black (1998) who argue that "[a]lthough we argued ... against using syntactic parsers for phrase break assignment, our reason stem from the basic inaccuracy of these parsers, not because syntactic parsers themselves are unhelpful." Thus, one way to overcome the limitation of dataset and the inaccuracy of the full syntactic parser is to employ machine learning approaches on a large corpus of natural speech with features that are more accurate than those of a full syntactic parser, and at the same time that contains richer syntactic structural information than part of speech. The outcome of the parser is, then, similar to the "flattened syntactic structure" (Chomsky & Halle 1968; Langendoen 1975).

This paper presents an experimental prediction of prosodic information (pitch accents and boundary tones) based on shallow syntactic structure and grammatical relations, together with part of speech, basic syllable information and constituent length. The working hypothesis is that even though there is no one-to-one correspondence between syntax and prosody, the two grammatical components are highly correlated, such that grammatical information is a good predictor of prosodic structure.

A subset of Boston Radio Speech Corpus (BRSC) is used for the experiment. The BRSC is a corpus of speech recorded by professional FM Radio News

Extracted Features
Part of Speech (POS)
Syntactic phrase chunk
Grammatical Relation
Number of words within a phrase
Number of words within a sentence
Number of syllables of the word
Number of phones of the word

announcers. The corpus is prosodically labeled using ToBI (Tones and Break Indices) (Silverman et al., 1992). The total words used for the experiment are about 10,000 and the number of sentences is about 600. Note that that since the speakers produced the same scripts, the type frequency of words is quite limited (about 900 word types). Below are features extracted for these experiments. [1]

Two machine learning algorithms are used for the experiment: (1) CART using Wagon and (2) Memory Based Learning using TiMBL. Previous and following n words, where $n = 1, 2$, are used for contextual information. The dataset is divided into training data (90%) and test data (10%). Word information is excluded from the feature set, because the type frequency of words is quite limited (about 900 words), and thus the result without word information would be more robust to changes in the dataset. Accuracies for the prediction of types of pitch accent (H*, !H*, L*, No Pitch Accent) and types of boundary tones (L-, H-, L-L%, L-H%, H-L%, H-H%, and No Boundary Tone) are reported below. In general, Memory based Learning results in better accuracy than CART-based Wagon for both pitch accent and boundary tone prediction. Chance performance for pitch accent labeling is 48.04 % (490/1020 of words carry an H* accent). The confusion matrices for both Wagon and TiMBL show that the accuracy for predicting presence vs. absence of pitch accent is quite high (85.2%), and that even the four-class pitch accent labeling task achieves an accuracy of 75.18 %. Chance performance for boundary tone labeling in the test data is 72.6% (741/1020 of words carry no boundary tone). The best accuracy for predicting types of boundary tone is 81.86%. As with pitch accents, the accuracy of predicting the presence vs. absence of boundary tone is high (90.2%), but predicting the type of boundary tone is more difficult, in part because some of the possible boundary type labels are extremely infrequent. The accuracy obtained from this experiment is favorable compared to previous studies. For example, the best score for predicting the presence or absence of phrase break using 6-gram part of speech tagging in Taylor and Black (1998) is 86.6%, as opposed to 90.2% in this experiment. Using a full automatic context-free parse of the same data for a slightly different task, Cohen (2004) achieved 89.8% accuracy in the automatic detection of intonational phrase boundary (a subset of the boundaries considered in this paper).

The paper concludes with the discussion of possible methods for improvement. The prediction of types of pitch accent will be more accurate if acoustic information is utilized. For example, the confusion between H* and !H* will be reduced with acoustic information available. Despite current arguments against the categorical status of !H* in American English (Dainora 2001), linear regression analysis reliably discriminates !H* from H*. Basic semantic information such as information content of words and named entity tagggin will further reduce the confusion among types of pitch accents and boundary tones.

[1] POS, syntactic phrase chunk, and grammatical relations are tagged using shallow syntactic parser availble at ILK. (<http://ilk.kub.nl>)

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

@inproceedingsSilBec92, author=”K. Silverman and M. Beckman and J. Pitrelli and M. Ostendorf and C. Wightman and P. Price and J. Pierrehumbert and J.

Hirschberg", booktitle=icslp, title="TOBI: A standard for labeling English prosody", year="1992"

@mastersthesisCoh04, advisor="Mark Hasegawa-Johnson", author="Aaron Cohen", school="University of Illinois at Urbana-Champaign", title="A Survey of Machine Learning Methods for Predicting Prosody in Radio Speech", year="2004"