10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Using Prosody to Improve Dependency Parsing

Hussein Ghaly[1], *Michael Mandel*[1,2]

[1] Linguistics PhD Program, Graduate Center, City University of New York
[2] Computer and Information Science, Brooklyn College
hmghaly@gmail.com, mim@sci.brooklyn.cuny.edu

Abstract

The goal of the present study is to use prosodic information to improve automatic syntactic parsing of conversational speech in the Switchboard Corpus. To achieve this, an ensemble classifier, based on a Recurrent Neural Network, is developed to predict the parse with the highest Unlabelled Attachment Score (UAS) from the outputs of multiple dependency parsers, based on syntactic and prosodic features. The main syntactic features proposed, which we refer to as "dependency configurations," represent the relative dependency location of each of a pair of consecutive words. Empirical analysis indicates that configurations with a direct dependency between consecutive words are less likely to be associated with major prosodic breaks. Using syntactic features alone, the system achieved an improvement of 1.1% of UAS on the test set, above the best parser in the ensemble, while using syntactic features combined with prosodic features (pauses and normalized duration) led to a further improvement of 0.4%. Both empirical analysis of dependency configurations and parsing improvement suggest a relationship between prosody and direct dependency relationships between consecutive words.

Index Terms: parsing, prosody, computational linguistics

1. Introduction

Prosody has strong ties with syntax. It can be used to resolve some syntactic ambiguities (Cutler et al., 1997) and prosodic breaks are associated with certain syntactic constituents such as parentheticals, tag questions, vocatives, and major syntactic boundaries (Shattuck-Hufnagel and Turk, 1996). In addition, it has been indicated that prosodic structure is related to the syntactic structure (Selkirk, 2011).

Automatic parsing has been shown to be negatively affected by syntactic ambiguities, such as prepositional phrase attachment and relative clause attachment (Kummerfeld et al., 2012). In addition, speech repairs in spontaneous speech also negatively affect parsing (Charniak and Johnson, 2001). Since both ambiguities and speech repairs can be associated with prosodic cues, there is reason to believe that prosodic information can help improve parsing.

This paper focuses on prosodic phrasing, reflecting the grouping and segmentation of speech into prosodic units, and the boundaries between such units (referred to as prosodic breaks or prosodic boundaries). It therefore does not include intonation, stress or rhythm. A common system for annotating prosody is ToBI (Tones and Break Indexes) (Silverman et al., 1992), which represents prosodic breaks as break indexes on a scale of 0-4 , reflecting the perceived disjuncture between words (i.e. 0 reflects no disjuncture, 1 is the default disjuncture associated with a word boundary, 3 is a strong disjuncture associated with an intermediate phrase boundary, while 4 is the strongest disjuncture associated with an intonational phrase boundary). In addition, prosodic breaks are associated with the presence and absence of silent pauses, and increased durations of syllables preceding a boundary, as well as other acoustic cues (Shattuck-Hufnagel and Turk, 1996).

A number of previous approaches have attempted to use prosody (both the perceptual ToBI breaks and the measured acoustic cues) to improve parsing. An early approach in this area was by Gregory et al. (2004), who treated prosody as a kind of punctuation, but this did not lead to any improvement of F score of the parse output on the Switchboard corpus using prosody, compared with the parse output without using prosody. Kahn et al. (2005) used prosody as features for re-ranking and achieved an improvement of 0.2% in F score on the Switchboard corpus using prosodic information alone and 0.6% when using prosodic features combined with syntactic features. Dreyer and Shafran (2007) used prosodic breaks as latent annotations to words in syntactic trees, with an improvement of 0.2% in F score on the Switchboard corpus. Huang and Harper (2010) developed a number of approaches for attaching prosodic breaks to parts of the syntactic trees, with results ranging from 0.6-0.8% improvement in F score on a mixture of the Switchboard and Fisher corpora. Tran et al. (2017) developed an approach based on neural networks, where they reported an improvement of 0.5% in F score on the Switchboard corpus when using acoustic features.

These approaches were mainly based on constituency parsing, where the syntactic representation is assumed to be phrase structure grammar. However, dependency parsing, which assumes the structure as a set of dependent-head relationships between words, is becoming the norm in Natural Language Processing. The main advantages of dependency parsing is its speed, scalability to new languages, and representation of semantic relationships between words (Choi et al., 2015). In addition, it represents the same syntactic information contained in the phrase structure grammar, as it is possible to convert between dependency and constituency structures (Xia and Palmer, 2001). Therefore, this study attempts to improve dependency parsing using prosodic information. Dependency structure has not been utilized much in prosody research. One notable exception is Pate and Goldwater (2013), who demonstrated that prosody can be used to improve dependency parsing. Their scope was limited to shorter sentences (10 words or less). Using word durations improved the directed attachment score (equivalent to our Unlabeled attachment score, UAS) of their unsupervised generative approach from

10.21437/SpeechProsody.2020-207

36.4% to 40.3% on the Switchboard corpus. Our model applies similar ideas to improve upon modern supervised dependency parsers that achieve much higher UAS scores.

This current study investigates two hypotheses:

- There is a correspondence between prosody and syntax that can be extracted from syntactic structure.
- These correspondences along with prosodic information can select the most appropriate parse for an utterance from among a set of candidate parses.

2. Data

This study uses the Switchboard corpus of spontaneous conversational speech (Godfrey et al, 1992). The sentences in this corpus are annotated in the NXT format (Calhoun et al, 2010). The syntactic information was converted to CoNLL dependency format by Honnibal and Johnson (2014). In the corpus, there are a total of 1285 recordings and the corresponding transcripts and annotations for each recording, with each recording corresponding to a side of a recorded telephone conversation. A subset of this corpus contains conversational sides that have ToBI annotations (150 sides, 11,743 sentences). The full set includes 100,729 sentences, which are divided according to the scheme of Honnibal and Johnson (2014) into a development set of 5,416, a test set of 5,456, and a training set containing the remainder.

3. Proposed features

An important part of this investigation is to identify which elements of syntactic structure, and more specifically, dependency structure, are related to prosody. We propose building on a theoretical foundation from phonology and psycholinguistics, which can provide a starting point for analyzing dependency structure.

One particular category of psycholinguistic research is of interest in this regard, referred to as “algorithmic approaches” (Ferreria, 2007), which attempt to predict the likelihood of prosodic breaks from syntactic information. Two notable approaches were pursued by Ferreira (1988) and Watson and Gibson (2004). In Ferreira’s approach, the likelihood of a prosodic break between two syntactic constituents depends on the height of the path connecting them in the syntactic tree, where higher indicates less semantic relatedness. In Watson and Gibson’s approach, the likelihood of a prosodic break between two constituents depends on the size of each constituent, with a lower likelihood if there is a dependency relationship between them. Both predicted that the highest likelihood of prosodic break comes between words when there is the least amount of semantic coherence or dependence. These approaches are based on Selkirk’s (1984) Sense Unit Condition, which suggested that the immediate constituents of an intonational phrase must be semantically related.

Taking this further to the dependency structure, we investigate here how to characterize this semantic dependence or coherence. Since prosodic breaks occur between two words, our unit of analysis is the word pair, and we identify the corresponding syntactic information for each word in the pair. Dependency structure identifies the “head” of each word, i.e., the word on which it depends. Therefore, for each word in a given pair, its head can be only one of the following:

- the other word in the pair - partner (P)
- to the right of the word pair - excluding partner (R)
- to the left of the word pair - excluding partner (L)
- nothing (the word is the root of the sentence) (O)

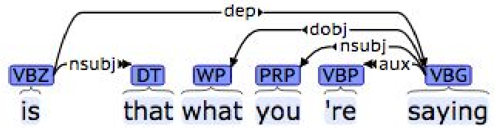

Figure 1- Illustration of a dependency structure for a sentence

Figure 1 shows the dependency structure of the sentence “Is that what you’re saying?” which is taken from the Switchboard corpus. When analyzing the word pair (“that”, “what”) in the sentence, the first word “that” depends on “is” to its left, while the second word “what” depends on “saying” to its right, and hence the configuration is (L, R). However, for the pair (“is”, “that”), “is” depends on “saying” to the right, while “that” depends on “is”, hence the configuration is (R,P). In the pair (“what”, “you”), both words depend on “saying” to the right, and the configuration is (R,R).

We analyzed the subset of Switchboard that was hand-annotated with ToBI breaks, identifying the configurations for each pairs of words and the corresponding ToBI break that was observed between them. Table 1 shows the count of each combination of observed configuration and ToBI break index. It also groups the configurations into four categories:

- (L, R): the first word in the pair depends on a word earlier in the sentence and the second word depends on a word later in the sentence.
- “No direct dependence”: Other situations in which the two words do not depend on one another.
- “Direct dependence”: One of the words depends on the other.
- “Two roots”: an unexpected situation that occurs occasionally because this is conversational speech.

		ToBI Break Index				
category	*config*	**1**	**2**	**3**	**4**	**Total**
(L, R)	(L, R)	6125	557	726	1456	8864
No Direct Dep	(R, R)	15657	1550	602	860	18669
	(L, L)	3268	334	465	829	4896
	(O, R)	2645	130	174	154	3103
	(L, O)	117	11	27	62	217
Direct Dep	(P, L)	10328	402	203	135	11068
	(L, P)	6062	281	308	358	7009
	(P, O)	6032	233	114	94	6473
	(P, R)	3308	102	42	30	3482
	(O, P)	2802	130	120	115	3167
	(R, P)	1011	25	31	22	1089
2 roots	(O, O)	74	30	2	7	113
	Total	**57429**	**3785**	**2814**	**4122**	**68150**

Table 1- Counts of word-pair dependency configurations categorized by manually annotated ToBI break indexes.

	ToBI Break Index			
Configuration Category	**1**	**2**	**3**	**4**
(L, R)	10.7%	14.7%	25.8%	35.3%
No Direct Dependency	37.8%	53.5%	45.1%	46.2%
Direct Dependency	51.4%	31.0%	29.1%	18.3%
2 roots	0.1%	0.8%	0.1%	0.2%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 2- Conditional Probability of configuration categories given observed ToBI break index - P(config | tobi).

Based on the Sense Unit Condition, we expect the (L, R) configuration to most often represent an intonational phrase boundary and thus have the highest likelihood of containing a ToBI break of index 3 or 4. The other configurations represent higher semantic coherence between the words in the pair, so we expect to see fewer strong breaks between them.

Table 1 shows the counts of each combination of configuration and ToBI break index. The most frequent configuration is (R, R), followed by (P, L), and then (L, R). However, the distribution of ToBI breaks for each configuration varies. The main focus here is on ToBI break indexes 1 (default index, no prosodic boundary, which is the most frequent index), 3 (intermediate phrase boundary), and 4 (Intonational Phrase boundary).

Table 2 shows these same data converted into conditional probabilities of configuration category given observed ToBI break index. We find that the configuration (L, R) accounts for 35% of ToBI break 4 and 26% of ToBI break 3, but only 11% of ToBI break 1. Other configurations involving no direct dependency (namely (R, R), (L, L), (O, R), and (L, O)) account for 45% of both ToBI breaks 3 and 4, and 38% of ToBI break 1. However, configurations involving direct dependencies between the two words (those that include “P”), account for 18% of ToBI break 4 and 29% of ToBI 3 (which can be due to disfluencies or planning factors related to spontaneous speech), but 51% of ToBI break 1. The configuration with 2 roots is negligible.

These trends are shown more clearly in Table 3, which shows the conditional probability of observing each ToBI break index given a dependency configuration category. It can be seen that there is a higher probability of 4 and 3 break indexes in the configuration (L,R) than in other configurations with no direct dependency. Configurations with direct dependencies have an even lower probability of ToBI break indexes 3 and 4, and higher probability for ToBI break 1.

	ToBI Break Index			
Configuration Category	1	2	3	4
(L, R)	69.1%	6.3%	8.2%	16.4%
No direct dependency	80.7%	7.5%	4.7%	7.1%
Direct dependency	91.5%	3.6%	2.5%	2.3%

Table 3- Conditional Probability of observed ToBI break index given the configuration category of each word pair P(tobi | config)

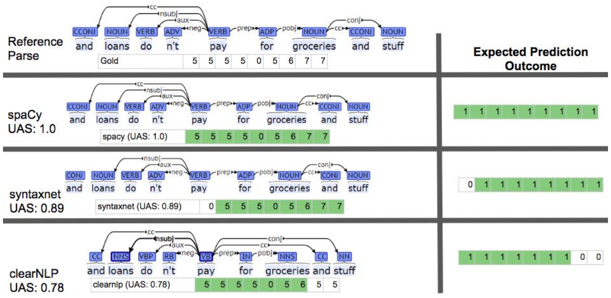
Reference Parse
CCONJ NOUN VERB ADV VERB ADP NOUN CCONJ NOUN
and loans do n't pay for groceries and stuff
cc nsubj aux neg prep pobj conj cc
Gold 5 5 5 5 0 5 6 7 7
Expected Prediction Outcome
spaCy
UAS: 1.0
spacy (UAS: 1.0) 5 5 5 5 0 5 6 7 7
1 1 1 1 1 1 1 1 1
syntaxnet
UAS: 0.89
CONJ NOUN VERB ADV VERB ADP NOUN CONJ NOUN
syntaxnet (UAS: 0.89) 0 5 5 5 0 5 6 7 7
0 1 1 1 1 1 1 1 1
clearNLP
UAS: 0.78
CC NNS VBP RB VB IN NNS CC NN
clearnlp (UAS: 0.78) 5 5 5 5 0 5 6 5 5
1 1 1 1 1 1 1 0 0

Figure 2- Prediction outcome for the heads correctly identified in each parse hypothesis

Therefore, these configurations appear to have a strong relationship with prosodic breaks, confirming our first hypothesis, that there is a correspondence between prosody and syntax that can be extracted from the syntactic structure. This correlation also makes dependency configurations a good candidate to be used as features in the parsing improvement system, which we do in the next section.

4. Prosody-informed parsing ensemble

Based on the above features, we developed an ensemble classifier using Long Short Term Memory (LSTM) Recurrent Neural Networks to select from among several candidate parses for a given sentence. The network takes as input these dependency configurations together with prosodic information and predicts the unlabeled attachment score of several candidate parses. This in effect ensembles together several parsers, in this experiment we use three: spaCy (Honnibal and Johnson, 2015), clearNLP (Choi and McCallum, 2013), and Syntaxnet (Andor et al., 2016).

Recurrent Neural Networks allow a single model to map between sequences of variable length. They also permit the use of both continuous and categorical features.

The outcome predicted by the network is the likelihood of correctness of the head of each word in a dependency parse hypotheses generated by one of the parsers, as shown in the example in figure 2.

In the example, the left column shows the ground truth parse along with each parse hypothesis produced by the three parsers for the same sentence. Each hypothesis is annotated with its UAS, the proportion of correctly identified heads. Correct identification of heads is with reference to the ground truth parse. The heads correctly identified by the parser are shown in green under the dependency structure. This information about head correctness is shown in the right column, by either correct (green) or incorrect (white). This head correctness information is the prediction target of the system.

After training, the system predicts the correctness of the heads in a given parse hypothesis, and the parse with the highest sum of prediction scores is selected from the parse hypotheses.

These predictions are based on two categories of features: text features and prosodic features. The text features consisted of part-of-speech tags, dependency configurations, and word embeddings of both the sentence words and their corresponding heads in the parse hypothesis. The prosodic features were only based on timing information, which can be easily measured and calculated. They consisted of the following: pauses (values of silent intervals between words), and normalized word durations, calculated as follows:

$$\text{Normalized word duration} = \frac{\textit{Actual word duration}}{\textit{Expected word duration}}$$

$$\text{Expected word duration} = \sum_{k=1}^{n} avgdur(ph_k)$$

Where n is the number of phonemes in the word, ph_k is the k-th phoneme in the word, and avgdur is the average duration of this phoneme across all utterances by the current speaker. In the current data, all utterances from a given speaker come from a single side in a single conversation. The expected duration is the sum of these average values for all the phonemes in the current word. Therefore, values greater than 1 indicate lengthening.

The RNN was implemented in Pytorch, with the experiments conducted on Google Colaboratory. The hyperparameters used are the following:

- Number of hidden neurons = 64; hidden layers = 1
- Learning rate = 0.0001; Adam optimizer
- Maximum number of epochs = 15 (results reported reflect early stopping based on UAS performance on the development set)

An oracle system is also computed as an upper bound on the performance of an ensemble. This system selects the parse from the candidates with the highest UAS using knowledge of the ground truth parse.

5. Results

Experiments were conducted to evaluate the performance of the ensemble classifier system relative to the individual parsers, and to evaluate whether using prosodic features improves ensembling performance. Results of these experiments are shown in Table 4.

System	Text Features	Prosodic features	UAS Dev	UAS Test
clearNLP			79.76	79.59
spaCy			79.06	78.91
Syntaxnet			72.54	72.81
Ensemble	POS, configs		80.69	80.73
Ensemble	POS, configs	Dur, dur log, pause	81.21	81.17
Ensemble	POS, configs, embeddings		83.47	83.36
Ensemble	POS, configs, embeddings	Dur, pause	83.51	83.39
Oracle			85.93	85.89

Table 4 - Results of parse ensembling experiments.

The most successful individual system was clearNLP (Dev UAS: 79.76%, Test UAS: 79.59%). The text-only ensemble baseline improved UAS above this best parser by around 1% absolute (Dev UAS: 80.69%, Test UAS 80.73%) using Part-of-Speech tags and Dependency Configurations. Adding prosodic information in the form of the timing features described above led to a further improvement of approximately 0.5% absolute (Dev UAS: 81.21%, Test UAS: 81.17%). This improvement on the test set is statistically significant according to a paired t-test t(5455)=-3.3, p = .0008. This statistical test is conducted by comparing two sets of UAS values achieved by the ensemble for sentences in the test set, where the first consists of the UAS values achieved without prosodic features, and the second with prosodic features.

Thus, the parsing ensemble using only textual features showed better parsing outcomes than the best individual parser, and adding prosodic information (word durations and pauses) improved parsing further. These results confirm our second hypothesis that prosody will permit us to select the most appropriate parse for an utterance from among a set of candidate parses.

Adding word embeddings led to a much higher UAS, close to the oracle UAS. This result is interesting in that it shows the potential of the ensembling approach. Adding prosodic features did not improve performance further, possibly because it was already so good using the word embeddings.

6. Conclusion

These results indicate that dependency configurations can be used as features to improve parsing performance. Furthermore, they have a meaningful relationship with prosody, as shown both by the analysis of ToBI break probabilities, and the improvement achieved by combining them with timing information. They also show promise for future analyses of prosody and dependency structure.

It should be noted that there are other important prosodic features, such as pitch and intensity, that can be used in future work. In addition, factors such as disfluencies and speech repairs, as well as lengths of syntactic constituents, have not been explicitly addressed in this study, despite being marked by prosodic cues. These are further possibilities for parsing improvements. Word embeddings are a promising feature for improving parsing in general, but are not helped by our prosodic features.

7. Acknowledgements

We would like to thank Janet Dean Fodor and Jason Bishop, for their useful conversations and feedback. We would also like to thank Andrew Caine, for providing the Switchboard syntactic data converted to CoNLL format.

8. References

[1] Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., ... & Collins, M. (2016). Globally normalized transition-based neural networks. arXiv preprint arXiv:1603.06042.

[2] Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation, 44(4), 387-419.

[3] Charniak, E., & Johnson, M. (2001, June). Edit detection and parsing for transcribed speech. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-9). Association for Computational Linguistics.

[4] Choi, J. D., & McCallum, A. (2013, August). Transition-based dependency parsing with selectional branching. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1052-1062).

[5] Choi, J. D., Tetreault, J., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Vol. 1, pp. 387-396).

[6] Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. Language and speech, 40(2), 141-201.

[7] Dreyer, M., & Shafran, I. (2007). Exploiting prosody for PCFGs with latent annotations. In INTERSPEECH (pp. 450-453).

[8] Ferreira, M. F. (1988). Planning and timing in sentence production: The syntax-to-phonology conversion.

[9] Ferreira, F. (2007). Prosody and performance in language production. Language and Cognitive Processes, 22(8), 1151-1177.

[10] Godfrey, J.J., Holliman, E.C. and McDaniel, J., 1992, March. SWITCHBOARD: Telephone speech corpus for research and development. In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92. (Vol. 1, pp. 517-520). IEEE.

[11] Gregory, M. L., Johnson, M., & Charniak, E. (2004). Sentence-Internal Prosody Does not Help Parsing the Way Punctuation Does. In HLT-NAACL (pp. 81-88).

[12] Honnibal, M., & Johnson, M. (2014). Joint incremental disfluency detection and dependency parsing. Transactions of the Association for Computational Linguistics, 2, 131-142.

[13] Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

[14] Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. (2005, October). Effective use of prosody in parsing conversational speech. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 233-240). Association for Computational Linguistics.

[15] Kummerfeld, J. K., Hall, D., Curran, J. R., & Klein, D. (2012, July). Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1048-1059). Association for Computational Linguistics.

[16] Pate, J. K., & Goldwater, S. (2013). Unsupervised dependency parsing with acoustic cues. Transactions of the Association for Computational Linguistics, 1, 63-74.

[17] Selkirk, E. (1984) The Relation Between Sound and Structure. Cambridge: MIT Press.

[18] Selkirk, E. (2011) - (2009). The syntax-phonology interface. The handbook of phonological theory, 2, 435-483.

[19] Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. Journal of psycholinguistic research, 25(2), 193-247.

[20] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., 1992. ToBI: A standard for labeling English prosody. In Second International Conference on Spoken Language Processing.

[21] Tran, T., Toshniwal, S., Bansal, M., Gimpel, K., Livescu, K., & Ostendorf, M. (2017). Joint Modeling of Text and Acoustic-Prosodic Cues for Neural Parsing. arXiv preprint arXiv:1704.07287.

[22] Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. Language and cognitive processes, 19(6), 713-755.

[23] Xia, F., & Palmer, M. (2001, March). Converting dependency structures to phrase structures. In Proceedings of the first international conference on Human language technology research (pp. 1-5). Association for Computational Linguistics.