Assign Stress for Interrogative Sentences via Syntax Structure Mapping

Ya Li[1], Xuefei Liu[1,2], Xiaoying Xu[1,2], Jianhua Tao[1]

[1] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Department of Chinese Language & Literature, Beijing Normal University, Beijing, China
{yli,jhtao}@nlpr.ia.ac.cn, liuxuefei1234@163.com, xuxiaoying2000@bnu.edu.cn

Abstract

Compared to the prosody study in statement-like sentence, little research has been conducted in interrogative sentence, which will definitely improve the naturalness of spoken dialog system. This paper proposes a computational model to assign stress in interrogative sentences using linguistic knowledge. We firstly investigate the interrogative sentences perceptually to label their stress locations and then summarize a series of syntax-stress mapping rules. The dependency syntax is adopted in this work and automatically obtained by Stanford Parser. All the dependency relations in Stanford Parser are studied and stable stress patterns are assigned. Syntax depth is considered as a factor to weight each dependency relation. Finally, the linear superposition of stress score calculated from each dependency relation is assigned to prosodic words. Experiments show that the weighted accuracy of syntax-stress mapping is 74%. A detailed analysis of results is also provided.

Index Terms: stress, speech prosody, syntax, interrogative sentence

1. Introduction

Spoken dialog system (SDS) has been achieved a great success with the development of statistical machine learning approach, however, the application of SDS still limited in information providing, such as weather information retrieval and tickets ordering system [1]. The main task of these SDSs is answering what people ask and in this case, the speech synthesis module only needs to synthesize statement-like utterances. Nevertheless, in real communication, people do not always answer, they can also ask for further information to clarify the question, etc. To provide a natural SDS and expand its application, SDS system should have the capability to generate natural and expressive interrogative sentence [2].

However, generating natural interrogative sentence is a complicate task. The current Text-to-Speech (TTS) system focus on statement-like utterance synthesis and rare work has been done on other sentence forms such as interrogative sentence due to the weakness of their prosodic structure research. The explicit prosodic difference between statement and interrogative sentence is their pitch contour patterns. Statement has a relative stable falling-down pitch contour, while interrogative sentence has various patterns according to their question types, for instance, Yes/No question has a sharp rising pitch contour, whereas Wh-question starts high and falls down rapidly on the last syllable of the question word. Among the three components in prosody structure, which are rhythm, stress and intonation, the latter two are the major factors to form the overall pitch pattern. Stress is the perceptual prominence within words or utterances. This paper attempts to study the default/normal stress in interrogative sentence. We admit that the stress location varies among people and the discourse context. To simplify the work, we do not discuss it in this paper.

The relation between syntax and prosody has been discussed for a long time. Although they are not consistent with each other, they do have a close relation, especially in interrogative sentence [3-11]. For example, in a Subject-Predicate phrase, the predicate is often stressed [3-4]. Feng S. [5] sums up some nuclear stress rules for Mandarin and he argues that prosody does constrain syntax. Veilleux N. M. [6] constructs a series of computational models, such as Decision Tree (DT) model, Hierarchical DT model for prosody/syntax mapping for spoken language systems.

Although the clear relation between syntax and prosody has not been discovered yet, many related research has already introduced syntax features to improve performance. Fitzpatrick E. and Bachenko J. [7] argue that it is possible to generate natural-sounding prosodic phrasing by some syntax information. Atterer M. [8] presents a rule-based syntax-prosody mapping approach to predict phrase breaks. The experiments of Wen M. et al. [9] show that syntax features could improve Mandarin segmental duration prediction. Nevertheless, only a few of work has been done on interrogative sentence. Kitagawa Y. [10] introduces the correlation between prosody and Wh-question in Japanese. He suggests that prosody and syntax should be considered together in understanding and analyzing Wh-questions. Li A. and Wang H. [12] analyze the stress in declarative and interrogative sentences acoustically in Mandarin; however, their work do not involves stress assignment in interrogative sentence.

The ultimate goal of this research is to generate natural speech, which will definitely benefit to the research of SDS, etc. We have done some work on expressive speech analysis and generation with statements [13-14] using machine learning techniques. This paper serves as a preliminary research on interrogative sentence study. We firstly analyze interrogative sentences perceptually to find the relation between syntax structure and stress. Then sum up a syntax-stress mapping strategy and finally use it to assign the default stress location in a sentence.

The rest of the paper is organized as follows. Section 2 analyzes the stress in interrogative sentences in Mandarin. Section 3 introduces the syntax-stress mapping strategy. The evaluation of the mapping strategy is presented in Section 4. Finally, Section 5 concludes the paper and puts forward our future research.

2. Interrogative sentence in Chinese

Communication is the progress that exchanging information between participates, therefore, statement and interrogative sentences which are typically used for providing and asking

information become the dominate sentence patterns in people's communication. This paper focuses on interrogative sentence study, especially, the interrogative sentence with question particles, such as *ma* and *ne*.

2.1. Corpus and stress annotation

The original corpus contains 600 independent interrogative sentences with wav file which are recorded in a professional studio by a male speaker. The speaking style of this corpus is natural speech without act or exaggeration which is shown in Figure 1. The natural or spontaneous speech brings about much difficult in stress perception and processing. We remove the sentences that not end up with question particles because most of them are Positive/Negative questions, for instance, "ni3 qu4 hai2 shi4 bu2 qu4? (Do you go or not?)". This kind of question sounds rude and will not be used in SDS. Finally 400 sentences are selected to construct the final corpus and used as the prompt in corpus recording.

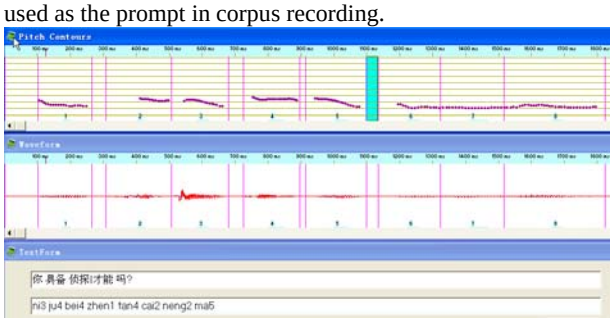

Figure 1: *A sample utterance of "ni3 ju4 bei4 zhen1 tan4 cai2 neng2 ma5 ? (Do you have the detective ability ?)".*

In Figure 1, the first line of TextForm window is Chinese Characters. Four prosodic layers are adopted here, and in this example, the space, "|" and "$" represent the prosodic word, prosodic phrase and intonation phrase boundaries respectively. The second line is the PINYIN transcript. All the word boundaries are manually checked.

Before the syntax stress mapping, we firstly construct the golden data for the subsequent evaluation. The stress annotation is obtained by perception and no syntax structure is taken into consideration. The aim of this experiment is to find the relative "correct" stress location because stress is the perpetually prominence. In the training stage, two assistants are asked to label the prominence word separately and discuss for the difference to achieve a higher consistency. In the labeling stage, only one assistant worked for this task.

2.2. Syntactic analysis

The syntax features we used are the dependency relation and the depth of the syntax tree. Stanford Parser [15] is utilized to automatically get syntactic features, which is one of the best open source parser for Chinese. The other reason for choosing Stanford parser is it provides dependency output as well as phrase structure tree. The dependency represents grammatical relation between words in a sentence. Stanford dependency (SD) are triplets: name of the relation, governor and dependent, e.g., *acomp(looks, beautiful)*, in which, *acomp* denotes an adjectival complement.

Table 1 shows the syntax parser output of example "ni3 ju4 bei4 zhen1 tan4 cai2 neng2 ma5? (Do you have the detective ability?)", and Figure 2 shows the syntax tree for easy reading. In the perception experiment, the annotator label "*侦探(detective)*" as stressed word. This could be explained by syntax analysis. In this example, *VP* is more likely to be stressed than *(NP (PN 你-you))*, in addition, *(NP (NN 侦探-detective) (NN 才能- ability)))* is a noun compound modifier, in which the modifier is more likely to be stressed than the noun. So the *(NN 侦探-detective)* obtained the final sentential stress. This example shows that the syntax and stress are relevant to a certain extent and it is possible to get the stress location from syntactic information.

Table 1. *The syntax analysis output of Stanford Parser (English translations are added).*

Syntax tree
(ROOT (IP (NP (PN 你-you)) (IP (VP (VV 具备-have) (NP (NN 侦探-detective) (NN 才能- ability))) (SP 吗-question particle, ma))))
Typed Dependencies
nsubj(具备-2, 你-1) nn(才能-4, 侦探-3) dobj(具备-2, 才能-4) dep(具备-2, 吗-5)

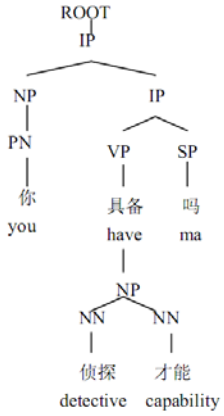

Figure 2: *Syntax tree of "ni3 ju4 bei4 zhen1 tan4 cai2 neng2 ma5?".*

3. Syntax-stress mapping

3.1. Stress in dependency relations

With the above evidence, we further attempt to find a comprehensive relation between syntax and stress. This is also obtained by perception which is a gradual deepening progress. The explored relations are checked in the further perception and more relations are added in the next procedure. In this experiment, the assistant did not refer to the golden data, and she just tried to find the stable or dominate stress pattern of each dependency relation and verified it in the whole corpus. Until all the sentences in the corpus were listened for several

times, the stable stress patterns of each dependency relation could be determined.

Table 2 shows the stable stress pattern, the unstable ones and the dependency relations without anyone stressed are not listed. In this table, the number in last column indicates the index of stressed word in dependency relation ("3" indicates that both words are stressed). For instance, in relation *nsubj*, the first word is more likely to be stressed, which means, *跳舞-dance* is stressed in *nsubj (跳舞-dance,我们-we)*. It should note that the stress pattern in this table is the most likely pattern, but it is not always true in various contexts.

Table 2. *The stress pattern of each dependency relation (English translations are added).*

Dependency relations and samples		Stress word index
nsubj	(跳舞-dance, 我们-we)	1
pobj	(在-at, 家-home)	2
aux	(打开-open, 可以-can)	2
tmod	(去-go, 今夜-tonight)	2
dobj	(看见-see, 钢琴-piano)	2
conj	(高兴-happy, 悲伤- sad)	3
csubj	(越好-better, 吃-eat)	1
poss	(客人们-guests, 我的-my)	1
det	(才能- ability, 侦探- detective)	2
cop	(医生-doctor, 是-is)	1
advmod	(安排-anrangement, 都-all)	2
prep	(谈起-talk, 关于-about)	2
mmod	(死-dead, 打-hit)	1
nummod	(女孩-girl, 一个-one)	2
rcmod	(消息-news, 查利- Charlie)	2
nn	(小姐-lady, 护士-nurse)	2
appos	(模范-model, 李四- Li Si)	1
purpcl	(叫-ask, 买票-buy tickets)	2
infmod	(什么-what, 毁灭-destroy)	1
acomp	(看起来-looks, 好看-beautiful)	2

3.2. Mapping strategy

With the relative stable stress pattern of each dependency relation, we propose a computational model to mapping the stress location from syntax structure.

The mapping strategy is similar to the syntax analysis mentioned in subsection 2.2. Firstly, the given sentence is segmented into words either manually or automatically, then we use Stanford Parser to get the syntax tree and dependency relation $DR(A,B)$, in which A and B are the governor and dependent words respectively. The stress score of governor and dependent words are defined as S_A and S_B.

$$S_A = \begin{cases} 1, if\ A\ is\ stressed\ \text{in}\ DR(A,B) \\ 0, else \end{cases} \quad (1)$$

$$S_B = \begin{cases} 1, if\ B\ is\ stressed\ \text{in}\ DR(A,B) \\ 0, else \end{cases} \quad (2)$$

The summation of the stress score represents the overall prominence degree in a sentence. However, in the pervious syntactic analysis, we find that the stress assignment in interrogative sentence with question particles is a gradually deepening progress. The dependency relation in deeper level of the syntax tree has more probability to be stressed, and the final stress will be transmitted to the leaf node of the syntax tree. Therefore, the syntax depth is also considered in this work, which is calculated as,

$$D_A = \alpha \times Depth_A \quad (3)$$

in which , α is a weight factor.

The final stress score of each word is defined as

$$Score_A = \sum_{A \subset (DR(A,B) \cap DR(B,A))} S_A \times D_A \quad (4)$$

In (4), the governor and dependent words are treated equally. Then we need to determine which words are stressed in a sentence. We use a simple binarization to categorize the scores into 1 and 0, which means stressed or not.

$$i^* = \beta \times Word\ Amount \quad (5)$$

$$Stress_A = \begin{cases} 1, if\ Stress_A > Score_{i^*} \\ 0, else \end{cases} \quad (6)$$

Firstly, the stress scores are sorted in descending order, and then the $i^{*\text{ th}}$ score is chosen as the threshold. In (5), β is the weight factor representing the percentage of stressed word in a sentence. We set $\beta = 0.2$, which is the percentage of the stressed words in whole corpus. Though this mechanism, the overall stress distribution keeps the same as the golden data.

In the final binarization, we introduce some heuristic rules to solve the conflict of same score. For the example mentioned in Section 2, there are 5 words in the sentence, which means only 1 word will carry the stress in our mapping strategy, however, the words "*侦探*" and "*才能*" both obtain the maximum stress score, 5. It happens that these two words are in one dependency relation *nn(才能-4, 侦探-3))*, and the second word is stressed in relation *nn*. So "*侦探*" obtains the final stress. If there are still more words than we expected and no other heuristic rules can be applied, these words are all considered to be stressed.

It should note that the effect of α is eventually offset by (6), however, the idea of weighing or normalization of syntax depth should be considered in this work and we will try other forms for applying this information.

4. Experiments and results

4.1. Experiment preprocessing

The corpus used for evaluation is the golden data described in Section 2, which contains 400 interrogative sentences. Before the syntax parsing, Maximum Entropy (ME) model is used to segment words automatically. The segmentation accuracy is 96.7%. Then Stanford Parser is used to get the syntax tree and the dependency relation. The syntax depth is automatically calculated and the depth of ROOT is set as 0.

4.2. Results and discussion

The TestSet I includes all the 400 manually labeled data and the weighted mapping accuracy is 74%, in which the precision, recall and F-score of the stressed word are only 37%, 35% and 36%. The F-score is calculated as follows:

$$F - score = \frac{2}{1/\ precision + 1/\ recall} \quad (7)$$

It is not a very promising result. By analyzing the results, we find that some mapping results are not entirely wrong or not acceptable by listening. Therefore, we selected the first 50 sentences to construct the TestSet II to check the acceptance rate of the mapping result. Two assistants are asked to check whether the automatic mapped stress location is acceptable by listening to the original wav corpus. Assistant A is the lady who labeled the corpus which means that she listened to the audio file for many times and assistant B had never heard the audio corpus. The acceptance rates are listed below.

Table 3. *The acceptance rate of syntax-stress mapping.*

Assistant	Acceptance
A (who listened to audio corpus for many times)	76%
B (who had never heard the audio corpus)	80%

Additionally, since the syntax structure of long sentence is always complicated, we further select 30 sentences which have more than 10 words (TestSet III) and 30 sentences with less than 8 words (TestSet IV) to evaluate the effect of the sentence length. The acceptance rates on these two test sets are 54% and 78% respectively. It is clear that long sentence is difficult to map stress.

Besides the sentence length, there are several reasons for the low weighted mapping accuracy. Firstly, the error caused by Stanford Parser may results in mapping error. Secondly, the final stress assignment mechanism is not perfect. Sometimes there are not enough heuristic rules for determining the final stress with two words that have the same stress score. However, these kinds of errors are often acceptable in the acceptance test. The third kind of error is that some words are prominent in linguistic point of view, but the tone and phrase boundary make it sounds less prominent than the surrounding words. This is because Chinese is a tonal language and thus the tone, rhythm and intonation are combined to affect the stress perception [16]. The latter two kinds of errors may not be serious in synthesizing interrogative speech. The future synthetic experiment will testify this point.

5. Conclusion and future work

The latest Text-to-Speech system still fails to produce natural interrogative sentence, which limits the application of TTS and some related work, specifically, the spoken dialog system which relies much on the interrogative sentence synthesis. This aim of this work is exploring a way to assign stress from raw text, which is the first step of a Text-to-Speech system. By perceptually analyzing the audio corpus, we propose a syntax-stress mapping method. All the dependency relations in Stanford Parser are assigned with a relative stable stress pattern. Then the stress weighted by their syntax depth is summarized and assigned to each prosodic word. Although the mapping accuracy is not as high as we expect, the acceptance rate shows the promise of this work. We also analyze the reasons comprehensively.

The future work includes natural interrogative sentence synthesis. The more effective stress assignment approach will also be investigated.

6. Acknowledgements

The work was supported by the National Science Foundation of China (No. 60873160, 61011140075 and 90820303) and partly supported by China-Singapore Institute of Digital Media (CSIDM), Beijing Normal University (004-127028), the Fundamental Research Funds for the Central University (2010105565004GK) and the State Language Commission 12th Five-Year language research programme(YB125-41).

7. References

[1] Zue V., et. al., "JUPITER: A Telephone-Based Conversational Interface for Weather Information", IEEE Trans. Speech and Audio Proc., 8(1):85-96, 2000.

[2] Kiriyama S., Hirose K., and Minematsu, N., "Control of Prosodic Focuses for Reply Speech Generation in a Spoken Dialogue System of Information Retrieval on Academic Documents", IEEE Workshop on Speech Synthesis, 139-142, 2002.

[3] Wang Y., Chu M. and He L., "An experimental study on the distribution of the focus-related and semantic accent in Chinese", Chinese Teaching in The World, (2), 86-98, 2006.

[4] Wang D., Cheng Z., Zheng B. and Yang Y., "Study on the rules of default stress distribution in Mandarin Chinese", Applied Acoustics, 26(1):46-54, 2007.

[5] Feng S., "Prosodically constrained postverbal PPs in Mandarin Chinese", Linguistics, 41(6):1085–1122, 2003.

[6] Veilleux N. M., "Computational models of the prosody/syntax mapping for spoken language systems", Thesis of Boston University, 1994.

[7] Fitzpatrick, E. and Bachenko, J., "Parsing for prosody: what a text-to-speech system needs from syntax", AI Systems in Government Conference, 188-194, 1989.

[8] Atterer M., "Assigning Prosodic Structure for Speech Synthesis via Syntax-Prosody Mapping", thesis of University of Edinburgh, 2000.

[9] Wen M. Wang, M., Hirose, K. and Minematsu N., "Improving Mandarin Segmental Duration Prediction with Automatically Extracted Syntax Features", INTERSPEECH 2010, 2178-2181.

[10] Kitagawa Y., "Prosody, Syntax and Pragmatics of Wh-questions in Japanese", English Linguistics 22(2):302-346, 2005.

[11] Selkirk E. O., "Sentence Prosody: Intonation, Stress and Phrasing", Goldsmith J., Ed., Handbook of Phonological Theory. Oxford, U. K: Blackwell, 1994.

[12] Li A. and Wang H., "Friendly Speech Analysis and Perception in Standard Chinese", INTERSPEECH 2004, 1-4.

[13] Li Y., Pan S. and Tao J., "HMM-based Expressive Speech Synthesis with a Flexible Mandarin Stress Adaptation Model", ICSP 2010, Beijing, 625-628.

[14] Li Y., Tao J. and Xu X., "Hierarchical Stress Modeling in Mandarin Text-to-Speech", INTERSPEECH 2011, Florence, 2013-2016.

[15] Chang P. C., Tseng H., Jurafsky D. and Manning D. C., "Discriminative Reordering with Chinese Grammatical Relations Features". In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, 2009. Online: http://nlp.stanford.edu/software/lex-parser.shtml.

[16] Wang Y., et al. "Stress perception of Chinese disyllabic words in utterance". Chinese Journal of Acoustic, 28(6):534-539, 2003.