Iterative Integration of Unsupervised Features for Chinese Dependency Parsing

Te Luo, Yujie Zhang(✉), Jinan Xu, and Yufeng Chen

School of Computer and Information Technology,
Beijing Jiaotong University, Beijing, China
{14120472,yjzhang}@bjtu.edu.cn

Abstract. Since Chinese dependency parsing is lack of a large amount of manually annotated dependency treebank. Some unsupervised methods of using large-scale unannotated data are proposed and inevitably introduce too much noise from automatic annotation. In order to solve this problem, this paper proposes an approach of iteratively integrating unsupervised features for training Chinese dependency parsing model. Considering that more errors occurred in parsing longer sentences, this paper divide raw data according to sentence length and then iteratively train model. The model trained on shorter sentences will be used in the next iteration to analyze longer sentences. This paper adopts a character-based dependency model for joint word segmentation, POS tagging and dependency parsing in Chinese. The advantage of the joint model is that one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks. The higher accuracy of the three tasks on shorter sentences can bring about higher accuracy of the whole model. This paper verified the proposed approach on the Penn Chinese Treebank and two raw corpora. The experimental results show that F1-scores of the three tasks were improved at each iteration, and F1-score of the dependency parsing was increased by 0.33%, compared with the conventional method.

Keywords: Chinese dependency parsing · Iteration · Unsupervised learning · Joint model

1 Introduction

Dependency parsing, which attempts to build dependency arcs between words in a sentence, is widely used in Machine Translation and automatic question answering system. Many methods of Chinese dependency parsing are proposed recently. Whether graph-based method [1, 2] or transition-based method [3, 4] are belong to supervised learning method, therefore, the model accuracy is limited by the scale and quality of the manual annotated treebank. Due to the difficulty of manual annotation of dependency treebank, there are almost no large-scale Chinese dependency treebank, and most of the dependency treebank used in research are automatically converted from the phrase treebank. On the other hand, large-scale raw corpus are relatively easy to obtain, and many researchers have proposed some unsupervised learning methods [5–9] using raw corpus to improve the accuracy of dependency parsing.

C.-Y. Lin et al. (Eds.): NLPCC-ICCPOL 2016, LNAI 10102, pp. 531–540, 2016.
DOI: 10.1007/978-3-319-50496-4_46

Unsupervised learning methods usually extract the feature from the results of automatic annotation [8, 10]. The main problem is that the errors in automatic annotation resulted in a large amount of noise in the feature extraction. There are less errors in the short sentences than the long sentences owning to the simple structure. Conventional methods do not pay attention to this difference and use the raw corpus without any discrimination on sentence length. Considering the less errors in automatic annotation for short sentences, we propose an approach of iteratively exploring unsupervised features for training Chinese dependency parsing model. We prefer to use shorter sentences of raw data to train model firstly, and then the trained model will be used in the next iteration to analyze longer sentences. Particularly, we adopt a character-based dependency model for joint word segmentation, POS tagging and dependency parsing in Chinese. The advantage of the joint model is that one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks. The higher accuracy of each task on short sentences can bring about higher accuracy of the whole model.

2 Previous Work

There are usually two ways using raw corpus in unsupervised learning method. The one is to directly use the automatic annotation as training data. Zhu [5] applied a high-accuracy parser (such as the Berkeley parser) to automatically analyze raw corpus, and then the new annotated treebank was applied as additional training data to build a shift-reduce parser.

Another one is to extract statistical features from raw corpus. Zhou exploited the feature of web-data to improve the supervised statistical dependency parsing [6]. Chen extracted the short dependency relations from the results of automatic annotation, and then map to different categories based on their frequency. Finally, they train the dependency parser by using the information as features [7]. Chen calculated the scores of dependency language model from the results of automatic annotation, and then map the scores to different categories, and integrate them in the decoding algorithm directly using beam-search [9].

Although the two ways improved the accuracy of the dependency parsing by large-scale raw corpus, lots of noise from automatic annotation remained because of long sentences. In this paper, we focus on the more effective unsupervised learning method by preferring the short sentences.

3 Joint Word Segmentation, POS Tagging and Dependency Parsing Model

In this paper, we adopt the shift-reduce frame, combine the three task, word segmentation, POS tagging and dependency parsing, into a joint model [10–13]. We use the online perceptron algorithm with early-update [14] for global learning and beam search algorithm for decoding [15]. The advantage of the joint model is that one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks.

3.1 Character-Based Joint Model

Word-based dependency tree is for build dependency arcs between words in a sentence. Because a sentence can be divided into different numbers of words in word segmentation, the number of dependency arc is also different. Character-based dependency tree is for build dependency arcs between characters in a sentence. For a sentence with L characters, the number of dependency arcs is N-1 for character-based dependency tree.

The analysis of a sentence is divided into several transition actions in shift-reduce joint model. In order to improve the search efficiency, for candidate results with the same number of transition actions, we only keep the top N results. Therefore, the model requires the candidate with the same number of transition actions is comparable, which requires all candidate results just experience the same number of transition actions from the initial state to the ending state. Thus, character-based dependency tree meets the requirements. Zhang [10] manually annotate the structures of words that occur in CTB5. We transform the word-based dependency tree into the character-based dependency tree by this way.

In a shift-reduce parser, an input sentence is processed in a linear left-to-right pass, and the output is constructed by a state-transition process. Every transition state includes a stack and a queue, where stack contains a sequence of partially-parsed dependency trees, and the queue consists of unprocessed input characters. There are two transition actions, shift and reduce, in word-based joint model. In this paper, we adjust the two transition actions in character-based joint model. The shift action is divided into four types, which are shift_S (the character is a single word), shift_B (the character is the first character of a word), shift_M (the character is the middle character of a word) and shirt_E (the character is the tail character of a word). The reduce action is divided into two types, which are the construction of inter-word dependency arc and intra-word dependency arc. Based on the above transfer strategy, a sentence with L characters requires 2L-1 transition actions from the initial state to the terminal state. In this paper, we use the same feature template with Guo [12] and Zhang [13].

3.2 Unsupervised Feature Using in Joint Model

In this paper, we extract two kinds of unsupervised feature, 2-gram string feature [16] and 2-gram dependency subtree feature [8], from large-scale raw corpus.

2-gram string feature is added to the joint model in the following way. In a sentence, each character c_i is labeled with a tag t_i after automatically word segmentation. In other word, the output of automatically word segmentation is a sequence $\{(c_i, t_i)\}_{i=1}^{L}$, L is the length of the sentence. Then, we can extract all two consecutive characters and its label (g, seg) from the segmented data, g is $c_i c_{i+1}$, and seg is $t_i t_{i+1}$. Next, we can extract a list of {g, seg, f(g, seg)} from the segmented data. Here, f(g, seg) is the frequency of the cases where 2-gram g is segmented with the segmentation profile seg. In order to alleviate the sparseness of the data, we group all the (g, seg) into three sets: high-frequency(HF), middle-frequency(MF), and low-frequency(LF). The grouping way are defined as follows: if the f(g, seg) is one of the top 10% of all the f(g, seg), the label of (g, seg) is represented as HF; if it is between top 10% and 30%, it is represented as MF, otherwise it is represented as LF. Finally 2-gram $\{g, seg, label\}$ lists are produced. When transition actions for word segmentation and POS tagging are being formed, we extract the 2-gram string about the character and get label from the $\{g, seg, label\}$ lists. We combine the 2-gram string and the label as the 2-gram string feature of the character.

2-gram dependency subtree feature is added to the joint model in the following way. We extract subtrees containing two words from the automatically parsed dependency trees express as st = w1_w2_R/L. Here, w1 and w2 are words, and the order of w1 and w2 corresponds to the sequence of them in the original sentence, R and L is right dependency arc and left dependency arc respectively. Then, we can extract a list of $\{(st, f(st))\}$ from the parsed data. Here, $f(st)$ is the frequency of st appeared in the whole corpus. Next, we group all the $f(st)$ into three sets: high-frequency(HF), middle-frequency(MF), and low-frequency(LF). The grouping way is same with the previous paragraph. Finally we get the subtree lists $\{st, label\}$. When we judge whether the top two nodes S1 and S0 on stack have dependency relationship, we get labels for all kinds of subtree between S1 and S0 as features using the subtree lists.

4 Iterative Exploring of Unsupervised Features for Chinese Dependency Parsing

In word segmentation, POS tagging and dependency parsing joint model, we propose a more effective unsupervised learning method in which the shorter sentences of raw corpus are preferred and iterative training is conducted. In this way, less noise will be introduced in the feature extraction. At first, we investigate the relationship between the accuracy of the dependency parsing and the length of sentences.

4.1 Preliminary Investigation

Given the sentence of length L, the number of possible dependency tree can be calculated by the following formula (1) [17].

$$\frac{1}{L}{}_{2(\mathrm{L}-1)}C_{(L-1)} \tag{1}$$

The number of possible dependency trees grows rapidly as the length becomes larger. The longer the sentence, the higher the complexity of the dependency parsing. The accuracy will be decreased unavoidably.

We conducted the following preliminary experiment related to the sentence length. First, we trained a joint model using CTB5 and then conducted a closed test. We calculated the average F1-score on different sentence length for word segmentation, POS tagging and dependency parsing. The average numbers of sentences of different length are about 200. The longest sentence consists of 418 characters and 240 words. The relationship between the length of sentence and F1-score of the three tasks on the sentences is shown in Fig. 1. From Fig. 1, we can see that with the increase of the

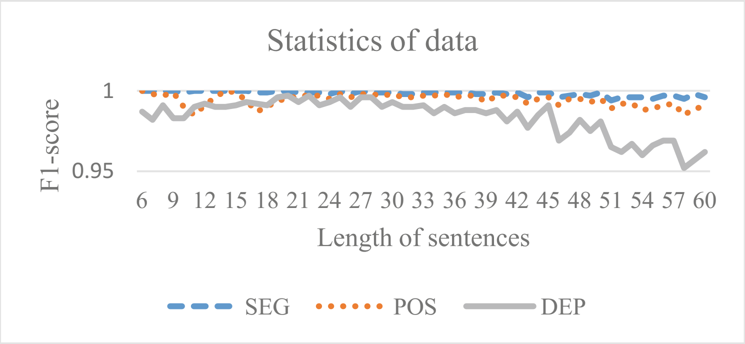


Fig. 1. The relationship between the sentence length and F1-score on the three tasks

length of sentence, the F1-score of the three tasks got decreased. Particularly, the decrease on the accuracy of dependency parsing is obvious. In POS tagging, F1-score at sentence length of 11 and 17 have greatly decreased. This is because that the word "新华社" appears several times, but is annotated POS with "NN" here and "NR" there.

The investigation result further prove the above expectation we obtained based on the formula (1) and provide the support for our proposed method.

4.2 Iterative Exploring of Unsupervised Feature

We propose an approach of iterative exploring of unsupervised features for training Chinese dependency parsing model. The framework is shown in Fig. 2.

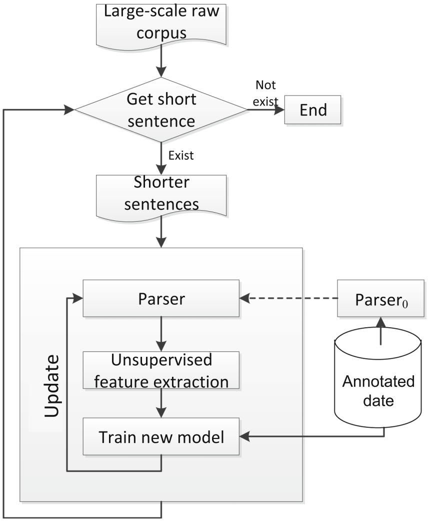


Fig. 2. Framework of iterative training of dependency parsing

The steps of iterative training model are as follows:

I. Obtain the initial dependency parser Parser_0 by using annotated treebank, set current parser Parser = Parser_0, start iteration.
II. Extract shorter sentences from raw corpus.
III. Use the current parser to analyze the shorter sentences and extract unsupervised feature according to the method of the 2.2 section.
IV. Re-train the model using the unsupervised feature and replace the current parser with the new parser, jump to II.

In Fig. 2, the process of II, III and IV are shown in one box to emphasize the iterative updating of model based on unsupervised feature.

Owning to using shorter sentences of raw corpus, the higher accuracy of automatic annotation of three tasks, word segmentation, POS tagging and dependency parsing, are expected to obtain. Since in the adopted joint model, one task can be promoted by exploring the available internal results from the other tasks during processing, the higher accuracy of three tasks on short sentences can bring about higher accuracy to the whole model. As a result, the higher accuracy of automatic annotation will be achieved.

5 Experiments

5.1 Experimental Settings

We use Chinese Tree Bank (CTB5) as annotated corpus, and it was separated into several parts: Training data set (chapter: 1–270, 400–931 and 1001–1151), development data set (chapter: 301–325) and test data set (chapter 271–300) [10]. As the names described, training data used for training joint model, development data was used for tuning parameters, and test data used for evaluation. We adopted Penn2Malt to transfer phrase structure tree to dependency tree. The People's Daily corpus (the first half of 1998 year) and Sogou Web News corpus were regard as large-scale raw corpus, which the People's Daily corpus belongs to a more standardized corpus. Statistics of datasets are shown in Table 1.

Table 1. Statistics of datasets

	Training	Development	Test	People's daily	Sogou web news
Number of sentences	18 K	350	348	295 K	18 M
Average length	44.4	38.2	39.5	40.5	51.3

In order to compare the accuracy in each iteration, and compare with the conventional method. The experimental setting is as follow, we have four experiments on The People's Daily corpus and Sogou Web News corpus respectively. (1) We extract the sentences with length of 1 to 10 from the raw corpus, and then extract unsupervised feature using Parser_0. Finally, we get Parser_1 through the first iteration. (2) We extract the sentences with length of 11 to 20 from the raw corpus, and then extract unsupervised feature using Parser_1. Finally, we get Parser_2 through the second iteration. (3) We extract the sentences with length of 21 to 30 from the raw corpus, and then extract unsupervised feature using Parser$_2$. Finally, we get Parser$_3$ through the third iteration. (4) We merge the raw corpus extracted in (1), (2) and (3), and we get Parser$_{mix}$ by using the mix raw corpus to train with conventional method. The beam size of joint model is set as 64 in this paper.

In this paper, we used F1-score as the accuracy metric to measure the performance of word segmentation, POS tagging and dependency parsing. Note that a dependency relationship is correct only when the two related words are all recalled in word segmentation and the head direction is correct. Following conventions, the relationships containing and punctuation are ignored.

5.2 Experimental Result and Analyses

The F1-score of four models' evaluation results on The People's Daily corpus are shown in Table 2. From Table 2, we can see that Parser$_3$ achieved higher F1-score than Parser$_2$ and Parser$_2$ achieved higher F1-score than Parser$_1$ in word segmentation, POS tagging and dependency parsing respectively, demonstrating the effectiveness of iterative training the model. We speculate that with the increase of the number of iterations, the accuracy of the three tasks will continue to improve. Parser$_{mix}$ achieved higher F1-score than Parser$_1$ in the three tasks, demonstrating the increase of the scale of raw corpus will increase the accuracy of the model. Parser$_{mix}$ achieved the same F1-score with Parser$_2$ in word segmentation, slightly lower F1-score than Parser$_2$ in POS tagging, and slightly higher F1-score than Parser$_2$ in dependency parsing. However, the corpus' scale of Parser$_2$ is smaller than Parser$_{mix}$, which indicates that the iterative method is significant in the three tasks. With the same scale of raw corpus, Parser$_3$ achieved higher F1-score than Parser$_{mix}$, demonstrating iteratively using raw corpus is better than conventional method. The F1-score of four models' evaluation results on Sogou Web News corpus are shown in Table 2. The difference in accuracy between the three models is similar to Table 3, which further proves the effectiveness of our method.

Table 2. The people's daily corpus

Model	SEG	POS	DEP
Parser$_1$	97.71	94.19	80.10
Parser$_2$	97.80	94.40	80.36
Parser$_3$	**97.95**	**94.53**	**80.71**
Parser$_{mix}$	97.80	94.36	80.38

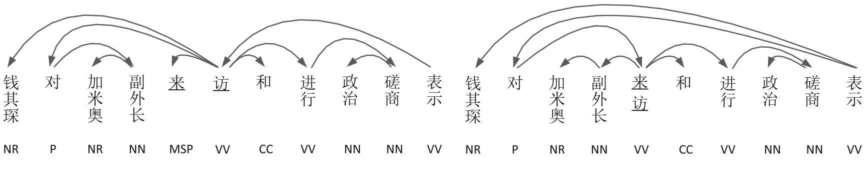
Table 3. Sogou web news corpus

Model	SEG	POS	DEP
Parser$_1$	97.87	94.26	80.05
Parser$_2$	97.90	94.40	80.21
Parser$_3$	**97.98**	**94.49**	**80.51**
Parser$_{mix}$	97.92	94.38	80.25

The accuracy of iterative exploring of unsupervised feature in three tasks is better than the conventional method. By comparing the results of the two methods, we find some errors in conventional method is correct in our method. This situation includes the following three types:

The first one is the dependency error caused by the word segmentation error. For example: (1) 在会见乌拉圭客人时，钱其琛对加米奥副外长来访和进行政治磋商表示欢迎。

The partial result of conventional method and our method are shown in Fig. 3(a) and 3(b). Figure 3(a) incorrectly divided the “来访(visit)” into two words “来(come)” and “访(visit)”, shown by the underline. “来(come)” become a verb, resulting in “钱其琛(Qian Qichen)” and “对(treat)” modified “访(visit)”. As our method get the less noise of 2-gram string feature from the shorter sentences, it correct the word segmentation of “来访(visit)”, and the three tasks is promoted by each other, and the dependency error is also corrected.

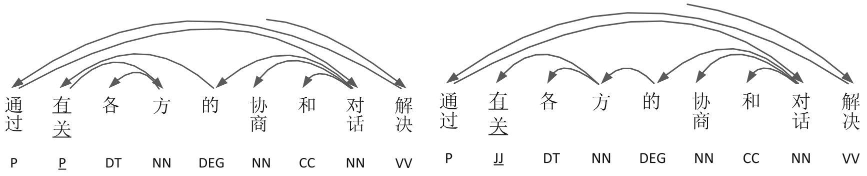


(a) (b)

Fig. 3. Dependency results of example (1)

The second one is the dependency error caused by the POS tagging error. For example: (2) 中方主张应通过有关各方的协商和对话解决朝鲜半岛的有关问题。

The partial result of conventional method and our method are shown in Fig. 4(a) and 4(b) (Because the sentence is too long to display all, the incomplete part of line indicates that the modified word is not in the figure). Figure 4(a) incorrectly labeled the POS tagging of “有关(concerned)” as preposition (P), leading the dependency relationship error of “有关(concerned)”, “方(parties)” and “的(of)”. As our method get the less noise of 2-gram subtree feature from the short sentences, it correct the POS tagging error, and the three tasks is promoted by each other, the dependency error is also corrected.



(a) (b)

Fig. 4. Dependency results of example (2)

The third one is the dependency error with the correct word segmentation and POS tagging. For example: (3) 尼克松先生是一位具有战略远见和政治勇气的政治家。

The partial result of conventional method and our method are shown in Fig. 5(a) and 5(b). Figure 5(a) shows “战略(strategic)”, “远见(vision)” and “和(and)” all incorrectly modified “政治(political)”. As our method get the less noise 2-gram subtree feature from the short sentences, it correct the dependency error directly.

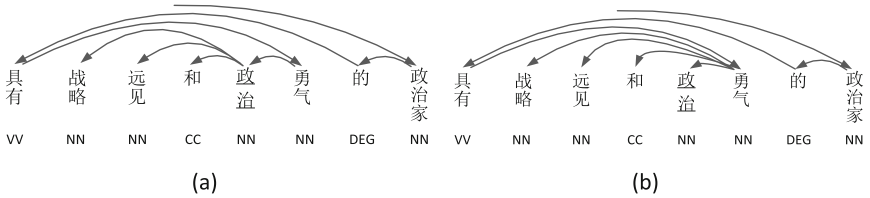


Fig. 5. Dependency results of example (3)

6 Conclusions

This paper proposes an approach of iterative exploring of unsupervised features for training Chinese dependency parsing model. Considering more errors are resulted in long sentence, we prefer to use shorter sentences as raw data first. The model trained on short sentences will be used in the next iteration to analyze longer sentences, and so on. We use a character-based dependency model for joint word segmentation, POS tagging and dependency parsing in Chinese. The advantage of the joint model is that one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks. The higher accuracy of three task on short sentences can bring about higher accuracy of the whole model. We verified the approach on the Penn Chinese Treebank. The experimental results show that F1-scores of three tasks were improved at each iteration, and F1-score of dependency parsing was increased by 0.33%, compared with the conventional method.

Acknowledgments. The authors are supported by National Nature Science Foundation of China (Contract 61370130 and 61473294), and the Fundamental Research Funds of the Central Universities (2014RC040).

References

1. Koo, T., Collins, M.: Efficient third-order dependency parsers. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1–11 (2010)
2. McDonald, R., Crammer, K., Pereira, F.: Online large-margin training of dependency parsers. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 91–98. Association for Computational Linguistics (2005)

3. Yamada, H., Matsumoto, Y.: Statistical dependency analysis with support vector machines. In: Proceedings of IWPT, vol. 3 (2003)
4. Nivre, J.: Algorithms for deterministic incremental dependency parsing. Comput. Linguist. **34**(4), 513–553 (2008)
5. 朱慕华, 王会珍, 朱靖波, 等. 向上学习方法改进移进-归约中文句法分析. 中文信息学报 29(2), 33–39 (2015)
6. Zhou, G., Zhao, J., Liu, K., et al.: Exploiting web-derived selectional preference to improve statistical dependency parsing. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 1556–1565. Association for Computational Linguistics (2011)
7. Chen, W., Kawahara, D., Uchimoto, K., et al.: Dependency parsing with short dependency relations in unlabeled data. In: IJCNLP, pp. 88–94 (2008)
8. Chen, W., Kazama, J., Uchimoto, K., et al.: Improving dependency parsing with subtrees from auto-parsed data. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 2, pp. 570–579. Association for Computational Linguistics (2009)
9. Chen, W., Zhang, M., Li, H.: Utilizing dependency language models for graph-based dependency parsing models. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Long Papers-Volume 1, pp. 213–222. Association for Computational Linguistics (2012)
10. Zhang, M., Zhang, Y., Che, W., et al.: Chinese Parsing Exploiting Characters. Proceedings of the 51st Annual meeting of the Association for Computational Linguistics, Long Papers-volume 1. Association for Computational Linguistics, pp. 125–134 (2013)
11. Hatori, J., Matsuzaki, T., Miyao, Y., et al.: Incremental joint approach to word segmentation, pos tagging, and dependency parsing in Chinese. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Long Papers-Volume 1. Association for Computational Linguistics, pp. 1045–1053 (2012)
12. Guo, Z., Zhang, Y., et al.: Character-level dependency model for joint word segmentation, POS tagging, and dependency parsing in Chinese. IEICE TRANS. Inf. Syst. **99**, 257–264 (2016)
13. Zhang, M., Zhang, Y., Che, W., et al.: Character-level chinese dependency parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1326–1336 (2014)
14. Collins, M., Roark, B.: Incremental parsing with the perceptron algorithm. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 111. Association for Computational Linguistics (2004)
15. Zhang, Y., Nivre, J.: Analyzing the effect of global learning and beam-search on transition-based dependency parsing. In: Proceedings of the COLING (Posters), pp. 1391–1400 (2012)
16. Wang, Y., Jun'ichi Kazama Y.T., Tsuruoka Y., et al.: Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In: IJCNLP, pp. 309–317 (2011)
17. Ozeki, K.: A multi-stage decision algorithm to select optimum bunsetsu sequences based on degree of Kakariuke-dependency. IEICE Trans. Inf. Syst. **70**, 601–609 (1987)