LETTER

A New Probabilistic Dependency Parsing Model for Head-Final, Free Word Order Languages

Hoojung CHUNG[†], ***Student Member and*** **Hae-Chang RIM**[†], ***Nonmember***

SUMMARY We propose a dependency parsing model for head-final, variable word order languages. Based on the observation that each word has its own preference for its modifying distance and the preferred distance varies according to surrounding context of the word, we define a parsing model that can reflect the preference. The experimental result shows that the parser based on our model outperforms other parsers in terms of precision and recall rate.

key words: *dependency parsing, statistical parsing*

1. Introduction

The dependency grammar has been widely used for analyzing free-word order languages. Assuming independence between each dependency relation in a parse tree t, the statistical dependency parsing model for a given sentence S can be defined as follows:

$$P(t|S) \approx \prod_{i=1}^{|S|-1} P(dep_i|S) \qquad (1)$$

where $|S|$ is a number of words in S, and dep_i is a dependency relation that starts from the ith word, w_i. The probability of a single dependency relation dep_i that links w_i with its head $w_{h(i)}$ can be simply defined as follows:

$$P(dep_i|S) \approx P(Yes|w_i\, w_{h(i)})$$

The probability represents the likelihood of a dependency relation established between two words and it reflects lexical preference between two words. By finding $argmax_t P(t|S)$, a parser can find appropriate heads for every word in S.

However, it is hard to decide correct heads with the simple model. Let's consider an example for deciding a head for the word *'wanjeonhi'* (completely) in the Korean sentence (Fig. 1). The word *wanjeonhi* has two alternative head candidates, which are *silpaeha-n* (failed) and *deureona-go* (was revealed). Because lexical preference of *wanjeonhi* to both head candidates are similar (See the lexical dependency probabilities to two head candidates in Fig. 2), it is uneasy to decide the correct head for the word *wanjeonhi* with the simple model.

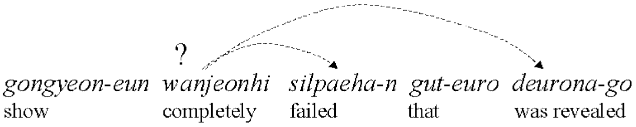

Fig. 1 The word 'wanjeonhi' has two alternative head candidates. (The sentence can be interpreted as "It was revealed that the show was completely failed" in English.)

To solve this problem, the distance between two depending words was considered in many previous statistical works. Some approaches can be described as follows:

- [1] distinguished the dependency probability between adjacent words from other dependency probability in its probabilistic dependency parsing model for Korean.
- [2] proposed a statistical parsing model for English, based on bigram lexical dependencies. It used a number of features (Δ) to consider the distance between the two depending words.

  $$P(dep_i|S) = P(Yes|w_i, w_j, \Delta_{i,j})$$

  [3]–[5] proposed similar models for parsing Korean and Japanese.
- [6] and [7] utilized hand-crafted HPSG for dependency analysis of Japanese. HPSG is used to find three alternative head candidates: the nearest, the second nearest, and the farthest candidates from a certain modifier. Then, the probabilistic model chooses an appropriate head among three candidates.

However, these approaches have following problems:

- The usage of distance features in the conditional part of the probability equation assumes that the dependency relation of a certain length is different from dependency relations with different lengths [1]–[5]. This assumption may cause sparse data problems in estimating lexical preferences between two words, especially for the languages allowing variable word order.
- The parsing model used in [6] and [7] requires a hand-crafted grammar. And the parsing model considers only three head candidates at most. The restriction is set based on the statistics from Japanese corpus. So it may fit for parsing Japanese but would cause a problem when the parsing model is used for other languages.

Manuscript received June 19, 2003.
Manuscript revised July 22, 2003.

[†] The authors are with the Department of Computer Science, Korea University, Seoul, Korea.

In this paper, we propose a new probabilistic model for head-final, free word order languages. In order to avoid the above problems, our model is designed to meet following requirements:

- We distinguish lexical dependency probability from modifying distance probability. It alleviates sparse data problem in estimating lexical dependencies.
- Our model does not ignore any grammatically correct head candidates, while [6] and [7] ignore less likely grammatical head candidates.
- Our model does not depend on language specific features. It does not require any language-specific, manual rules – such as heuristic constraints or HPSG – too. So it can be adapted to other languages easily.

2. Dependency Parsing Model

We define a parsing model based on the following observation: *a word selects its head based on its lexical preference and modification distance preference to the other word*. We observed that lexical dependency is influenced by two words that are linked by the dependency, while the modification distance of a word depends on a modifying word only.

To model the observation formally, we define a dependency relation dep_i as follows:

$$dep_i = \begin{bmatrix} dependent & i \\ head & h(i) \\ length & \Psi(h(i)-i) \\ exist & Yes \end{bmatrix}$$

The dependency relation dep_i contains a head position $h(i)$ and a dependent position i that the dependency relation links. The length of the dependency (which is modification distance) is expressed with the following function Ψ:

$$\Psi(x) = \begin{cases} x & \text{if} \quad x < k \\ long & \text{if} \quad x \geq k \end{cases}$$

The attribute *exist* expresses whether dep_i exists in the dependency tree. The value of *exist* is always Yes because we consider only dependency relations existing in a tree. Considering the attribute structure of dep_i, the Eq. (1) can be rewritten as follows:

$$\begin{aligned} P(t|S) &\approx \prod_{i<|S|} P(dep_i|S) \\ &= \prod_{i<|S|} P(i, h(i), \Psi(h(i)-i), Yes|S) \\ &\approx \prod_{i<|S|} P(\Psi(h(i)-i), Yes|w_i\, w_{h(i)}\, \Phi_i\, \Phi_{h(i)}) \end{aligned} \tag{2}$$

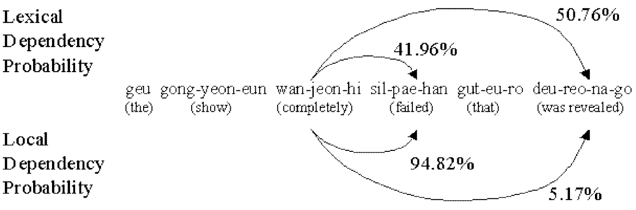

Fig. 2 Lexical and local dependency probability of *wanjeonhi*.

where Φ_x is a context† surrounding w_x, which is the xth word in a sentence.

According to our observation, an event of dependency relation existence depends on two words (modifier and modifyee), while the modification distance depends on the modifier only. So we need to decompose the Eq. (2) to differentiate depending features for two events.

$$\begin{aligned} P(t|S) \approx \prod_{i<|S|} & P(Yes|w_i\, w_{h(i)}\, \Phi_i\, \Phi_{h(i)}) \\ & \cdot P(\Psi(h(i)-i)|w_i\, w_{h(i)}\, \Phi_i\, \Phi_{h(i)} Yes) \end{aligned} \tag{3}$$

As mentioned in [8], the decomposition has several additional merits too. It facilitates parameter estimation and makes analyzing model's behavior easier.

Under the assumption that the existence of a dependency relation depends only on the two words that are linked by the dependency relation and that the length of a dependency relation depends on a modifier only, not on a head, (3) becomes as follows:

$$\begin{aligned} P(t|S) &\approx \prod_{i<|S|} P(Yes|w_i\, w_{h(i)}) \cdot P(\Psi(h(i)-i)|w_i\, \Phi_i) \\ &= \prod_{i<|S|} P(Yes|w_i\, w_{h(i)}) \\ &\qquad \cdot P(\Psi(h(i)-i)|w_i\, p(w_{i-n}) \ldots p(w_{i+m})) \end{aligned} \tag{4}$$

where $p(x)$ is a part-of-speech tag of a word x. We can observe that the statistical parsing model becomes a product of following two probabilities:

1. A probability of a dependency relation for the given two words.
2. A probability of a certain modifying distance for a given word and its context.

We name them *lexical dependency probability* and *local dependency probability*. The two probabilities formalize our observation on the head decision process. By composing the two probabilities, Fig. 2 shows two probabilities from the word *wanjeonhi* in the same sentence represented in Fig. 1. As shown in the figure, local dependency probability from *wanjeonhi* helps the word

[†] We used part-of-speech tag lists as contextual information rather than lexical word squences to avoid sparse data problem.

to select its correct head, *silpaeha-n*.

3. Experimental Results

We have performed an experiment to evaluate our parsing model and to compare the proposed model with other parsing models considering the distance between a modifier and a head. We select Korean as our target language for our experiments. Korean , like Japanese, is a verb- or adjective-final and free word order language.

We have implemented a parser that uses the proposed parsing model. The parser was trained on 27,694 sentences and tested on heldout 3,386 sentences of dependency tagged sections of KAIST Language Resources [9]. The average lengths of sentences in the training and test set are 11.13 and 11.31 words, respectively. We used the maximum likelihood estimation for estimating parameter for Eq. (4) and deleted interpolation to alleviate data sparseness. All sentences are part-of-speech tagged, and this information as well as lexical information was used as an input for the parser.

To evaluate the performance of the parser, we used arc based precision, recall, and F-measure. To select appropriate values for m, n (in Eq. (4)) and k (in Ψ function) in the parsing model, we performed a preliminary experiment and selected 0, 2, and 2 for those variables, respectively. We have also implemented three other parsing models for comparison.

Model1 Lexical dependency probability model using heuristic constraints on distance (similar to [10]).

Model2 Lexical dependency probability model that distinguishes dependencies between adjacent words from other lexical dependencies (similar to [1]).

Model3 Lexical dependency probability model combined with triplet/quadruplet head candidate decision model[†] (similar to [6]).

Table 1 shows experimental results. Our model does not perform well in the training set. The Model1 and Model2 show almost 100% arc precision and recall in the training set. It is because they are highly lexicalized models. The Model3 shows a worse result than our model for the training set. The triplet/quadruplet model used in the Model3 assumes that a head of a word is one among the nearest, second nearest, or the last head candidates from the dependent word. However, only 91.48% of heads are among the three head candidates for the training data and this means that the model is useful for only 91.48% of words. This makes the performance of the Model3 poor.

For parsing the heldout data, our model outperformed all other models in the experiment. And the improvement (+0.65 F score from Model3) was statistically meaningful. By analyzing the parsing result (Fig. 3), we found out that our model was more effective for finding the short-distance head, compared to Model3. However, our model shows a worse performance for finding the long-distance head. If the advantage of Model3 can be combined with our model, the proposed model would have a better performance for the long dependency relations too.

Table 1 Experimental results for various parsing models.

		Model1	Model2	Model3	Ours
Training	Arc Prec.	99.35	99.55	90.96	98.50
	Arc Recall	99.33	99.52	90.84	98.39
	Arc F	99.34	99.54	90.89	98.45
Testing	Arc Prec	79.40	83.09	84.34	85.00
	Arc Recall	79.02	82.41	83.87	84.51
	Arc F	79.21	82.04	84.11	84.76

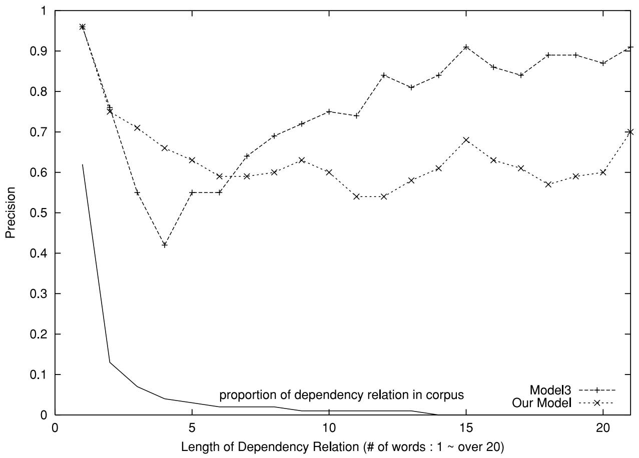

Fig. 3 Precision versus length of dependency relations figures for Model3 and our model in the test data.

4. Conclusion

We have proposed a new probabilistic model for parsing head final, free word order languages. The model consits of two probabilities. The lexical dependency probability reflects selectional restrictions, while the local dependency probability reflects the preferred length of a dependency relation from a certain dependent word. The proposed model formally considers the preference on modifying distance without ignoring grammatically correct dependency relations. It does not require any manually constructed rules such as HPSG or heuristic constraints. Furthermore, our model is more robust for the data sparseness problem than other models that separate lexical probability distribution based on the length of dependency relation. Since it does not use any language dependent feature, the proposed parsing model can be used for various languages. In future work, we are planning to carry out experiments with other languages such as Japanese.

[†] [6]'s model requires hand-crafted grammar (HPSG). Instead of HPSG, we used a set of dependency rules whose frequency is more than one in the training corpus as the grammar.

Acknowledgement

This work was supported by Korean Research Foundation Grant (KRF-2002-042-D00485).

References

[1] H. Kim, Korean Syntactic Analysis with Probabilistic Dependency Grammar, Master's Thesis, Dept. of Computer Science, KAIST, 1994.

[2] M.J. Collins, "New statistical parser based on bigram lexical dependencies," Proc. 34th Annual Meeting of the ACL, pp.223–230, 1996.

[3] H. Kim and J. Seo, "A statistical Korean parser based on lexical dependencies," Spring Proceedings of Conference on Korea AI Society, pp.332–338, 1997.

[4] M. Haruno, S. Shirai, and Y. Ooyama, "Using decision trees to construct a practical parser," Proc. COLING-ACL 98, pp.505–512, 1998.

[5] K. Uchimoto, S. Sekine, and H. Isahara, "Japanese dependency structure analysis based on maximum entropy models," Proc. 13th EACL, pp.163–203, 1998.

[6] H. Kanayama, K. Torisawa, Y. Mitsuichi, and J. Tsujii, "Statistical dependency analysis with an HPSG-based Japanese grammar," Proc. 5th Natural Language Processing Pacific Rim Symposium, pp.138–143, 1999.

[7] H. Kanayama, K. Torisawa, Y. Mitsuichi, and J. Tsujii, "A hybrid Japanese parser with a hand-crafted grammar and statistics," Proc. COLING 2000, pp.411–417, 2000.

[8] K. Shirai, K. Inui, T. Tokunaga, and H. Tanaka, "An empirical evaluation on statistical parsing of Japanese sentences using lexical association statistics," Proc. 3rd Conference on Empirical Methods in Natural Language Processing, pp.505–512, 1998.

[9] K.S. Choi, "KAIST language resources v.2001," Result of Core Software Project from Ministry of Science and Technology, Korea (http://kibs.kaist.ac.kr), 2001.

[10] S.W. Jung, E.K. Park, D.Y. Ra, and J.T. Yoon, "A study on Korean dependency parser using case relation and mutual information," Proc. Hangul and Korean Information Processing Conference, pp.450–456, 2001.