Joint International Conference on Artificial Intelligence
Poster session

Integration of Statistical Techniques for Parsing

INUI Kentaro, SHIRAI Kiyoaki, TOKUNAGA Takenobu and TANAKA Hozumi
Department of Computer Science, Tokyo Institute of Technology
2–12–1 O-okayama Meguro Tokyo 152 Japan
{inui,kshirai,take,tanaka}@cs.titech.ac.jp

1 Introduction

In statistical parsing, the following three types of statistics have been empirically proven to be effective: (a) *part-of-speech n-gram*, typically for word segmentation and part-of-speech tagging, (b) *structural preference* such as a probabilistic CFG model, and (c) *lexical association* such as mutual information, or some other collocational statistics, between words. Given this background, we consider the following two basic requirements:

Integration of different types of statistics: Since these types of statistics represent analyses of linguistic distribution from different points of view, language models for statistical parsing are ideally required to capture all of them.

Modularity of statistics types: Since such an integrated model would still only reflect limited aspects of linguistic phenomena, we would need to analyze the model's behavior in order to refine it further. This requires that the total score of a parse derivation can be decomposed into the factors derived from the different types of statistics, which would facilitate analysis in terms of each statistics type.

However, it seems to be the case that no existing framework of language modeling sufficiently satisfies both of these requirements. In this presentation, we propose a new framework of probabilistic language modeling that satisfies both requirements simultaneously by introducing the notions of *lexical contexts* and *lexical dependency parameters*.

2 Integrated language modeling

Given an input word sequence W, we rank its parse derivations according to the joint distribution $P(R, W)$, where R is a parse derivation whose terminal symbols are part-of-speech tags. We first decompose $P(R, W)$ into the syntactic model $P(R)$ and the lexical model $P(W|R)$. The syntactic model is expected to cover both part-of-speech n-gram statistics and structural preference, which can be achieved through, for example, our probabilistic GLR language model [Inui *et al.*, 1997b], whereas the lexical model is expected to reflect lexical association.

Lexical context The lexical model $P(W|R)$ is the product of the probability $P(w_i|l_i, w_1^{i-1}, R)$ of each lexical derivation $l_i \rightarrow w_i$, where l_i is the part-of-speech tag of w_i, and words are assumed to be derived, in principle, in a head-centered derivation order. The key idea for estimating each factor $P(w_i|w_1^{i-1}, R)$ is in assuming that each lexical derivation depends only on a certain small part of its whole context, which we refer to as the lexical context of that derivation. We specify the lexical context of each lexical derivation in a rule-based manner; e.g.,

- If w_i functions as a slot-marker of a certain lexical head h, the lexical context of w_i must include h. For example, the probability of deriving *with* from its part-of-speech tag P in "*she ate spaghetti with chopsticks*" is estimated to be $P(\textit{with}|\mathrm{P}, \textit{eat})$.
- If w_i functions as a filler of a slot s of a head word h, the lexical context of w_i must include h coupled with s. For example, the probability of deriving *chopsticks* in the above sentence is estimated to be $P(\textit{chopsticks}|\mathrm{N}, \textit{eat} : \textit{with})$.

By localizing lexical association statistics into the lexical derivation model as above, our framework can separate this type of statistics from other statistics types.

Lexical dependency parameter A lexical derivation may be associated with more than one lexical head; for example, *spaghetti* in "*she ate the spaghetti I cooked*" functions as a filler of the object slots of both *eat* and *cook*. In such a case, we subdivide the lexical context to reduce the model's complexity as follows:

$$P(w_i|l_i, c_1, c_2) \approx P(w_i|l_i) \cdot \frac{P(w_i|l_i, c_1)}{P(w_i|l_i)} \cdot \frac{P(w_i|l_i, c_2)}{P(w_i|l_i)}$$

where c_1 and c_2 are the lexical context of w_i, and the second and third factors are what we call lexical dependency parameters. This formulation shows that we can decompose the lexical model in a probabilistically well-founded manner even in cases where the input sentence includes relative clauses or coordinate structures. For the details of our modeling, see [Inui *et al.*, 1997a].

3 Implementation and Evaluation

For the syntactic model, we have fully implemented a probabilistic GLR parser, and conducted several preliminary experiments, attaining promising results. As for the lexical model, we are currently conducting preliminary experiments, the target language being Japanese.

References

[Inui *et al.*, 1997a] K. Inui, K. Shirai, H. Tanaka, and T. Tokunaga. Integrated probabilistic language modeling for statistical parsing. Technical Report TR97-0005, Dept. of Computer Science, Tokyo Institute of Technology, 1997 Available from http://www.cs.titech.ac.jp/.

[Inui *et al.*, 1997b] K. Inui, V. Sornlartlamvanich, H. Tanaka, and T. Tokunaga. A new probabilistic LR language model for statistical parsing. Technical Report TR97-0004, Dept. of Computer Science, Tokyo Institute of Technology, 1997.