Learning Verb-Noun Relations to Improve Parsing

Andi Wu
Microsoft Research
One Microsoft Way, Redmond, WA 98052
andiwu@microsoft.com

Abstract

The verb-noun sequence in Chinese often creates ambiguities in parsing. These ambiguities can usually be resolved if we know in advance whether the verb and the noun tend to be in the verb-object relation or the modifier-head relation. In this paper, we describe a learning procedure whereby such knowledge can be automatically acquired. Using an existing (imperfect) parser with a chart filter and a tree filter, a large corpus, and the log-likelihood-ratio (LLR) algorithm, we were able to acquire verb-noun pairs which typically occur either in verb-object relations or modifier-head relations. The learned pairs are then used in the parsing process for disambiguation. Evaluation shows that the accuracy of the original parser improves significantly with the use of the automatically acquired knowledge.

1 Introduction

Computer analysis of natural language sentences is a challenging task largely because of the ambiguities in natural language syntax. In Chinese, the lack of inflectional morphology often makes the resolution of those ambiguities even more difficult. One type of ambiguity is found in the verb-noun sequence which can appear in at least two different relations, the verb-object relation and the modifier-head relation, as illustrated in the following phrases.

(1) 登记 手续 的 费用
dengji shouxu de feiyong
register procedure DE expense
"the expense of the registration procedure"

(2) 办理 手续 的 费用
banli shouxu de feiyong
handle procedure DE expense
"the expense of going through the procedure"

In (1), the verb-noun sequence "登记 手续" is an example of the modifier-head relation while "办理 手续" in (2) is an example of the verb-object relation. The correct analyses of these two phrases are given in Figure 1 and Figure 2, where "RELCL" stands for "relative clause":

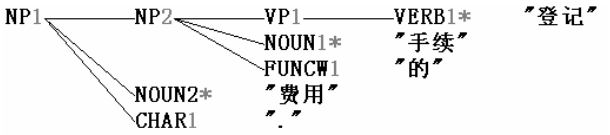

Figure 1. Correct analysis of (1)

Figure 2. Correct analysis of (2)

However, with the set of grammar rules that cover the above phrases and without any semantic or collocational knowledge of the words involved, there is nothing to prevent us from the wrong analyses in Figure 3 and Figure 4.

Figure 3. Wrong analysis of (1)

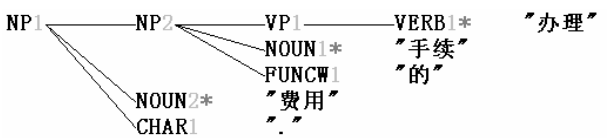

Figure 4. Wrong analysis of (2)

To rule out these wrong parses, we need to know that 登记 is a typical modifier of 手续 while 办理 typically takes 手续 as an object. The question is how to acquire such knowledge automatically. In the rest of this paper, we will present a learning procedure that learns those relations by processing a large corpus with a chart-filter, a tree-filter and an LLR filter. The approach is in the spirit of Smadja (1993) on retrieving collocations from text corpora, but is more integrated with parsing. We will show in the evaluation section how much the learned knowledge can help improve sentence analysis.

2 The Learning Procedure

The syntactic ambiguity associated with the verb-noun sequence can be either local or global. The kind of ambiguity we have observed in (1) and (2) is global in nature, which exists even if this noun phrase is plugged into a larger structure or complete sentence. There are also local ambiguities where the ambiguity disappears once the verb-noun sequence is put into a broader context. In the following examples, the sentences in (3) and (4) can only receive the analyses in Figure 5 and Figure 6 respectively.

(3) 这 是 新 的 登记 手续。
zhe shi xin de dengji shouxu
this be new DE register procedure
"This is a new registration procedure."

(4) 你 不 必 办理 手续。
ni bu bi banli shouxu
you not must handle procedure
"You don't have to go through the procedure."

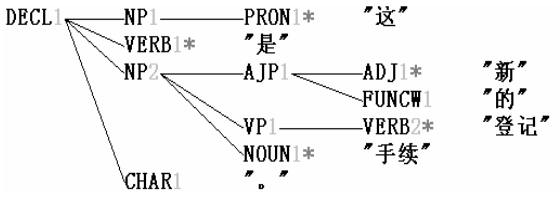

Figure 5. Parse tree of (3)

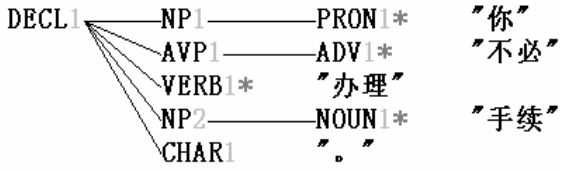

Figure 6. Parse tree of (4)

In the processing of a large corpus, sentences with global ambiguities only have a random chance of being analyzed correctly, but sentences with local ambiguities can often receive correct analyses. Although local ambiguities will create some confusion in the parsing process, increase the size of the parsing chart, and slow down processing, they can be resolved in the end unless we run out of resources (in terms of time and space) before the analysis is complete. Therefore, there should be sufficient number of cases in the corpus where the relationship between the verb and the noun is clear. An obvious strategy we can adopt here is to learn from the clear cases and use the learned knowledge to help resolve the unclear cases. If a verb-noun pair appears predominantly in the verb-object relationship or the modifier head relationship throughout the corpus, we should prefer this relationship everywhere else.

A simple way to learn such knowledge is by using a tree-filter to collect all instances of each verb-noun pair in the parse trees of a corpus, counting the number of times they appear in each relationship, and then comparing their frequencies to decide which relationship is the predominant one for a given pair. Once we have the information that "登记" is typically a modifier of "手续" and "办理" typically takes "手续" as an object, for instance, the sentence in (1) will only receive the analysis in Figure 1 and (2) only the analysis in Figure 2. However, this only works in idealized situations where the parser is doing an almost perfect job, in which case no learning would be necessary. In reality, the parse trees are not always reliable and the relations extracted from the parses can contain a fair amount of noise. It is not hard to imagine that a certain verb-noun pair may occur only a couple of times in the corpus and they are misanalyzed in every instance. If such noise is not filtered out, the knowledge we acquire will mislead us and minimize the benefit we get from this approach.

An obvious solution to this problem is to ignore all the low frequency pairs and keep the high frequency ones only, as wrong analyses tend to be random. But the cut-off point is difficult to set if we are only looking at the raw frequencies, whose range is hard to predict. The cut-off point will be too low for some pairs and too high for others. We need a normalizing factor to turn the raw frequencies into relative frequencies. Instead of asking "which relation is more frequent for a given pair?", the question should be "of all the instances of a given verb-noun pair in the corpus, which relation has a higher percentage of occurrence?". The normalizing factor should then be the total count of a verb-noun pair in the corpus regardless of the syntactic relations between them. The normalized frequency of a relation for a given pair is thus the number of times this pair is assigned this relation in the parses divided by this normalizing factor. For example, if 登记 手续 occurs 10 times in the corpus and is analyzed as verb-object 3 times and modifier-head 7 times, the normalized frequencies for these two relations will be 30% and 70% respectively. What we have now is actually the probability of a given pair occurring in a given relationship. This probability may not be very accurate, given the fact that the parse trees are not always correct, but it should a good approximation, assuming that the corpus is large enough and most of the potential ambiguities in the corpus are local rather than global in nature.

But how do we count the number of verb-noun pairs in a corpus? A simple bigram count will unjustly favor the modifier-head relation. While the verb and the noun are usually adjacent when the verb modifies the noun, they can be far apart when the noun is the object of the verb, as illustrated in (5).

(5) 他们 正在 **办理** 去 台湾 参加
tamen zhengzai banli qu taiwan canjia
they PROG handle go Taiwan participate
第十九届 国际 计算 语言学
dishijiujie guoji jisuan yuyanxue
19th international compute linguistics
会议 的 **手续**。
huiyi de shouxu
conference DE procedure
"They are going through the procedures for going to Taipei for the 19th International Conference on Computational Linguistics."

To get a true normalizing factor, we must count all the potential dependencies, both local and long-distance. This is required also because the tree-filter we use to collect pair relations consider both local and long-distance dependencies as well. Since simple string matching is not able to get the potential long-distance pairs, we resorted to the use of a chart-filter. As the parser we use is a chart parser, all the potential constituents are stored in the chart, though only a small subset of those will end up in the parse tree. Among the constituents created in the chart for the sentence in (5), for instance, we are supposed to find [办理] and [去台湾参加第十九届国际计算语言学会议的手续] which are adjacent to each other. The fact that 手续 is the head of the second phrase then makes 手续 adjacent to 办理. We will therefore be able to get one count of 办理 followed by 手续 from (5) despite the long span of intervening words between them. The use of the chart-filter thus enables us to make our normalizing factor more accurate. The probability of a given verb-noun pair occurring in a given relation is now the total count of this relation in the parse trees throughout the corpus divided by the total count of all the potential relations found in all the charts created during the processing of this corpus.

The cut-off point we finally used is 50%, i.e. a pair+relation will be kept in our knowledge base if the probability obtained this way is more than 50%. This may seem low, but it is higher than we think considering the fact that verb-object and modifier-head are not the only relations that can hold between a verb and a noun. In (6), for example, 办理 is not related to 手续 in either way in spite of their adjacency.

(6) 他们 去 上海 办理 手续 所需
tamen qu shanghai banli shouxu suoxu
they go Shanghai handle procedure need
的 公证 材料。
de gongzheng cailiao
DE notarize material

"They went to Shanghai to handle the notarized material needed for the procedure."

We will still find the 办理 手续 pair in the chart, but it is not expected to appear in either the verb-object relation or modifier-head relation in the parse tree. Therefore, the baseline probability for any pair+relation might be far below 50% and more than 50% is a good indicator that a given pair does typically occur in a given relation. We can also choose to keep all the pairs with their probabilities in the knowledge base and let the probabilities be integrated into the probability of the complete parse tree at the time of parse ranking.

The results we obtained from the above procedure are quite clean, in the sense that most of the pairs that are classified into the two types of relations with a probability greater than 50% are correct. Here are some sample pairs that we learned.

Verb-Object:

检验 真理	test - truth
配置 资源	allocate - recourses
经营 业务	manage - business
奉献 爱心	offer - love
欺骗 行人	cheat - pedestrians

Modifier-Head:

检验 标准	testing - standard
配置 方案	allocation - plan
经营 方式	management - mode
奉献 精神	offering - spirit
欺骗 行为	cheating - behavior

However, there are pairs that are correct but not "typical" enough, especially in the verb-object relations. Here are some examples:

具有 意义	have - meaning
具有 效力	have - impact
具有 色彩	have - color
具有 作用	have - function
具有 功效	have - effect
…	

These are truly verb-object relations, but we may not want to keep them in our knowledge base for the following reasons. First of all, the verbs in such cases usually can take a wide range of objects and the strength of association between the verb and the object is weak. In other words, the objects are not "typical". Secondly, those verbs tend not to occur in the modifier-head relation with a following noun and we gain very little in terms of disambiguation by storing those pairs in the knowledge base. To prune away those pairs, we used the log-likelihood-ratio algorithm (Dunning, 1993) to compute the degree of association between the verb and the noun in each pair. Pairs where there is high "mutual information" between the verb and noun would receive higher scores while pairs where the verb can co-occur with many different nouns would receive lower scores. Pairs with association scores below a certain threshold were then thrown out. This not only makes the remaining pairs more "typical" but helps to clean out more garbage. The resulting knowledge base therefore has higher quality.

3 Evaluation

The knowledge acquired by the method described in the previous section is used in subsequent sentence analysis to prefer those parses where the verb-noun sequence is analyzed in the same way as specified in the knowledge base. When processing a large corpus, what we typically do is analyzing the corpus twice. The first pass is the learning phase where we acquire additional knowledge by parsing the corpus. The knowledge acquired is used in the second pass to get better parses. This is one example of the general approach of "improving parsing by parsing", as described in (Wu *et al* 2002).

To find out how much the learned knowledge contributes to the improvement of parsing, we performed a human evaluation. In the evaluation, we used our existing sentence analyzer (Heidorn 2000, Jensen *et al* 1993, Wu and Jiang 1998) to process a corpus of 271,690 sentences to learn the verb-noun relations. We then parsed the same sentences first without the additional knowledge and then with the acquired knowledge. Comparing the outputs, we found that 16,445 (6%) of the sentences had different analyses in the two passes. We then randomly

selected 500 sentences from those "diff" sentences and presented them to a linguist from an independent agency who, given two different parses of the same sentence, was asked to pick the parse she judged to be more accurate. The order in which the parses were presented was randomized so that the evaluator had no idea as to which tree was from the first pass and which one from the second pass.

The linguist's judgment showed that, with the additional knowledge that we acquired, 350 (70%) of those sentences parsed better with the additional knowledge, 85 (17%) parsed worse, and 65 (13%) had parses that were equally good or bad. In other words, the accuracy of sentence analysis improved significantly with the learning procedure discussed in this paper.

Here is an example where the parse became better when the automatically acquired knowledge is used. Due to space limitation, only the parses of a fraction of the sentence is given here:

(7) 要 遵照 国家 测试 标准
yao zunzhao guojia ceshi biaozhun
want follow nation testing standard
"(You) must follow the national testing standards."

Because of the fact that 遵照 is ambiguous between a verb ("follow") and a preposition ("in accordance with"), this sentence fragment got the parse tree in Figure 7 before the learned knowledge was used, where 标准 was misanalyzed as the object of 测试:

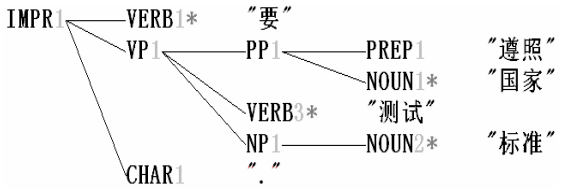

Figure 7: old parse of (7)

During the learning process, we acquired "测试-标准" as a typical pair where the two words are in the modifier-head relationship. Once this pair was added to our knowledge base, we got the correct parse, where 遵照 is analyzed as a verb and 测试 as a modifier of 标准:

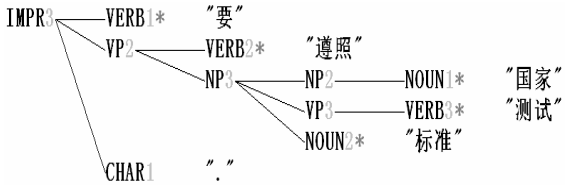

Figure 8: New tree of (7)

We later inspected the sentences where the parses became worse and found two sources for the regressions. The main source was of course errors in the learned results, since they had not been manually checked. The second source was an engineering problem: the use of the acquired knowledge required the use of additional memory and consequently exceeded some system limitations when the sentences were very long.

4 Future work

The approach described in this paper can be applied to the learning of many other typical syntactic relations between words. We have already used it to learn noun-noun pairs where the first noun is a typical modifier of the second noun. This has helped us to rule out incorrect parses where the two nouns were not put into the same constituent. Other relations we have been trying to learn include:

- Noun-noun pairs where the two nouns are in conjunction (e.g. 新郎 新娘 "bride and bridegroom");
- Verb-verb pairs where the two verbs are in conjunction (e.g. 调查 研究 "investigate and study");
- Adjective-adjective pairs where two adjectives are in conjunction (e.g. 年轻 漂亮 "young and beautiful");
- Noun-verb pairs where the noun is a typical subject of the verb.

Knowledge of this kind, once acquired, will benefit not only parsing, but other NLP applications as well, such as machine translation and information retrieval.

In terms of parsing, the benefit we get there is similar to what we get in lexicalized statistical parsing where parsing decisions can be based on

specific lexical items. However, the training of a statistical parser requires a tree bank which is expensive to create while our approach does not. Our approach does require an existing parser, but this parser does not have to be perfect and can be improved as the learning goes on. Once the parser is reasonably good, what we need is just raw text, which is available in large quantities.

5 Conclusion

We have shown in this paper that parsing quality can be improved by using the parser as an automatic learner which acquires new knowledge in the first pass to help analysis in the second pass. We demonstrated this through the learning of typical verb-object and modifier-head relations. With the use of a chart-filter, a tree-filter and the LLR algorithm, we are able to acquire such knowledge with high accuracy. Evaluation shows that the quality of sentence analysis can improve significantly with the help of the automatically acquired knowledge.

References

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61-74.

Heidorn, G. E. 2000. Intelligent writing assistance, in *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, Dale R., Moisl H., and Somers H. eds., Marcel Dekker, New York, pp. 181-207.

Jensen, K., G. Heidorn and S. Richardson. 1993. *Natural Language Processing: the PLNLP Approach".* Kluwer Academic Publishers, Boston.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1): 143-177.

Wu, Andi, J. Pentheroudakis and Z. Jiang, 2002. Dynamic lexical acquisition in Chinese sentence analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 1308-1312, Taipei, Taiwan.

Wu, Andi, J. and Z. Jiang, 1998. Word segmentation in sentence analysis, in *Proceedings of 1998 International Conference on Chinese Information Processing*, pp. 46-51.169-180, Beijing, China.