Integrating Multiple Layers of Concept Information into N-gram Modeling for Spoken Language Understanding

Nick J.C. Wang[1, 2]

Graduate Institute of Communication Engineering, National Taiwan University[1] &
Department of Man-Machine Interface, R&D Center, Delta-Electronics Inc.[2]
Taipei, Taiwan, R.O.C.
nick.jc.wang@delta.com.tw

Abstract

The paper presents a novel approach, integrating multi-layer concept information into the trigram language model, to improve the understanding accuracy for spoken dialogue systems. With this approach, both the recognition accuracy and out-of-grammar problem can be largely improved. In the experiment using a real-world air-ticket information spoken dialogue system for Mandarin Chinese, a relative concept error rate reduction of 33% is achieved.

1. Introduction

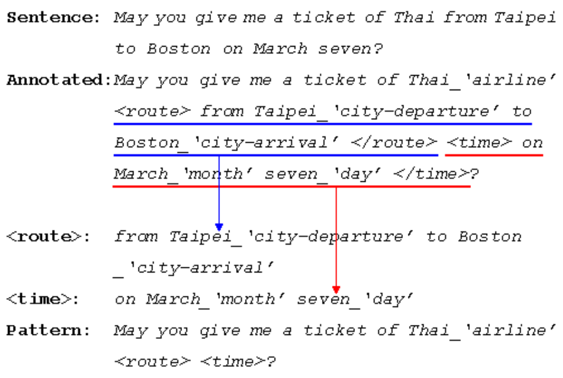
Spoken language understanding is composed of speech recognition and language understanding. Consequently, its performance would depend on both components, as well as their interface [1]. Our speech group in Delta Electronics Inc. has been working on research and development of Mandarin spoken-language technologies for years and cooperating with MIT Spoken Language Systems group. The study of the paper was conducted on a telephony Mandarin real-time flight-schedule inquiry and booking dialogue system, *Mandarin Mercury*, based on the *Galaxy* architecture [2]. The system interacts with the user over the phone through a natural conversation and delivers flight schedules and pricing information. Its vocabulary includes over 200 major city names worldwide and 23 major airline names. It was designed as a way of mix-initiative interactions between man and machine; hence, natural speaking in Mandarin could be understood. However, in our experience, longer utterances are still not easy to be understood, which causes the major problem on using the system.

Conventionally speech recognition and language understanding are interfaced by n-best word sequences or word graphs [3]. A long sentence with speech recognition errors or out-of-grammar expressions would cause parsing failures. A partial parsing strategy may help with this, if only the errors occur beyond the target concept phrases. However, the partial parsing sacrifices completeness and depth of analysis [4][5]. Our proposed approach integrates multiple layers of concept information into the N-gram model for speech recognition. Therefore, the speech recognizer not only can output the word sequence, but also some additional information of the concepts. For instance, the recognizer will output *"I would like <route>to go to Taipei_'city-arrival'</route>"*, where *"<route>"* and *"</route>"* show the beginning and the end of a concept phrase—chunk [6], and "*Taipei_'city-arrival'*" shows a concept attribute-value pair. The chunk tags are about concept information of multiple connected words in a phrase, where the lexical attributes are that of one word or several words for one entity name. Our previous publication presented the use of chunk [7]. The study was continued in the paper, extending to the use of lexical attributes. Experiments showed them largely reducing parsing failures. In our approach, N-gram modeling of using the chunk phrases was constructed like a two-layer Stochastic Context-Free Grammar (SCFG): a 'sentence-pattern' layer using words and chunk tags as the basic units and a 'chunk phrase' layer using words in the chunk as the units. The two-layered organization of chunk and sentence corpora for N-gram modeling was shown to be able to deal with data sparseness, so as to significantly reduce the error.

The following section will explain our proposed approaches. Section 3 shows our experimental setup and results. Conclusion is made in the end.

2. Multiple concept layers

A layer of chunk phrase concept information and a layer of lexical attribute concept information were added onto the words in the training sentences in our experiments. Two major chunk phrase categories, <time> and <route>, were defined in our flight-schedule inquiry application. Within each chunk category or sentence pattern category, several concept attributes were defined, including 'city-departure', 'city-arrival', 'airport-departure', 'airport-arrival', 'depart', and 'arrive' attributes in <route> chunk, and 'day', 'month', 'year', and 'weekday' attributes in <time> chunk, and 'airline', 'confirm', and 'deny' in sentence pattern.


```
Sentence:  May you give me a ticket of Thai from Taipei
           to Boston on March seven?
Annotated: May you give me a ticket of Thai_'airline'
           <route> from Taipei_'city-departure' to
           Boston_'city-arrival' </route> <time> on
           March_'month' seven_'day' </time>?

<route>:   from Taipei_'city-departure' to Boston
           _'city-arrival'
<time>:    on March_'month' seven_'day'
Pattern:   May you give me a ticket of Thai_'airline'
           <route> <time>?
```

Figure 1. Decomposing of chunks and sent-patterns

Sentences for language training were annotated like in Figure 1. The example sentence contains a <time> chunk *"on March seven"* and a <route> chunk *"from Taipei to Boston"*, with additional lexical attributes such as 'city-departure', 'city-arrival', 'day' and 'month'. Its sentence pattern is *'May* *you please give me a ticket of Thai_'airline' <route> <time>?'*

A word in the vocabulary may have lexical variations due to different chunks or attributes. Given the above lexical terms, our speech recognizer generates words with additional information about chunk and attribute, which were utilized in our language understanding process. We adopted category-based N-gram modeling approach, which shows advantages on dealing with data sparseness [8]. The overall process based on the layered concept information within the N-gram models showed advantages on understanding accuracy, comparing to the traditional way. However, on the contrary, the attachment of labels slightly enlarges the vocabulary of the recognizer, which may decrease recognition accuracy mainly due to the unavoidable data sparseness for almost all applications.

2.1. Two-layer structural approach of trigram modeling

The multi-layer stochastic approach is popular in natural language understanding as in [9] and [10]. In the paper, we experiment the use of multi-layer stochastic approach in constructing the N-gram models. A merged N-gram model for the recognizer is computed via three different sub-models, which were computed separately by the corpora of <route> and <time> chunks and of the sentence pattern, respectively.

```
Sentence pattern corpus:
S→   I would like a ticket of <route> <time>?
S→   China airline_'airline' <route>, please.
S→   …

<route> chunk corpus:
<route>→   from Taipei_'city-departure' to
           Boston_'city-arrival'
<route>→   to fly to Paris_'city- arrival'
<route>→   …

<time> chunk corpus:
<time>→    on March_'month' seven_'day'
<time>→    in the morning_'period_of_day'
<time>→    …
```

Figure 2. Sentence pattern and chunk corpora

The three corpora of <route> and <time> chunks and the sentence pattern work conceptually as a form of a two-layer SCFG, as illustrated in Figure 2. The chunk labels are saved for the space in the illustration. An attributed word like "Boston_'city-arrival'<route>" was treated different from "Boston_'city-departure'<route>".

In the first layer, there are rules from the start symbol *S* leading to all abstractive forms. In the second layer, there are rules from the only two non-terminals "<route>" and "<time>" leading to all <route> and <time> phrases, respectively. A large number of grammatical sentences can be derived from the above rules, as illustrated in Figure 3. These sentences are then used to train the N-gram model. The resulted N-gram model inherits the property of two-layer statistics and embraces longer distance dependency.

The computing on unifying the three N-gram count-trees into a merged N-gram count-tree is by replacing the chunk unit <X> with all possible candidates of unit sequence in the chunk tree of <X>. First of all, a count number of a triplet in the sentence pattern, $n_S(X, p, q)$, containing chunk <X> as its first element, would be updated by the following three conditions: (1) length-one extension, as in Equation 1: replacement by single words x_i for those with $n_X(x_i, e)$ observed in <X> chunk count-tree, where *e* denoting phrase end; (2) length-two extension, as in Equation 2: replacement by word pairs (x_i, x_j) for those with $n_X(x_i, x_j, e)$ observed; and (3) length-three extension, as in Equation 3: replacement by word triplets (x_i, x_j, x_k) for those with $n_X(x_i, x_j, x_k)$ observed.

$$n(x_i, p, q) = n_S(X, p, q) \cdot n_X(x_i, e) / n_X \tag{1}$$

$$n(x_i, x_j, p) = n_S(X, p, q) \cdot n_X(x_i, x_j, e) / n_X \tag{2}$$

$$n(x_i, x_j, x_k) = n_S(X, p, q) \cdot n_X(x_i, x_j, x_k) / n_X \tag{3}$$

p and *q* denote either a word or a non-terminal node (such as chunk <route> or <time> in our case) and n_X the total number of <X> chunk phrases.

Secondly, a count number $n_S(p, X, q)$ containing chunk <X> as its second element, would be updated by the following two conditions: (1) length-one extension, as in Equation 4: replacement by single words x_i for those with $n_X(b, x_i, e)$ observed, where *b* denoting phrase beginning; (2) length-two extension, as in Equation 5: replacement by word pairs (x_i, x_j) for those with $n_X(b, x_i, x_j)$ observed.

$$n(p, x_i, q) = n_S(p, X, q) \cdot n_X(b, x_i, e) / n_X \tag{4}$$

$$n(p, x_i, x_j) = n_S(p, X, q) \cdot n_X(b, x_i, x_j) / n_X \tag{5}$$

Finally, as <*X*> is the third element in $n_S(p, q, X)$, length-one extension could be applied with $n_X(b, x_i)$ observed.

$$n(p, q, x_i) = n_S(p, q, X) \cdot n_X(b, x_i) / n_X \tag{6}$$

Komatani explored similar combination of language models but limited on bigram format [17]. The underlining assumption is that every sentence with a form like *"I would like <route> <time>"* should share its observations with all grammatical sentences in the same form. In our study, we found phrases with length larger than two words would perform better in its sharing of probabilities. Therefore, the merged N-gram model was trained conceptually based on more grammatical sentences of larger coverage and could be better in dealing with data sparseness. It is different from the phrase-based language modeling approach [18][19]. The later extends the length of the context information to further reduce the perplexity of the language model, while the former enhance the data sparseness over probability estimation.

Since the merged model is in the format of N-gram model, it could be easily adopted by many speech recognition systems without modification and remains its robustness to spontaneous speech, especially as comparing to the Finite-State-Transducer decoder. On the contrary, earlier researches such as the two-layer bigram model [11], the unified language model of N-grams and Stochastic Finite State Automata [12], and the unified language model of N-grams and SCFG [13][14] would need a specialized recognizer.

2.2. Three-pass understanding process

In the Mandarin Mercury system, a set of rules was written for complete full-sentence parsing. In our initial experiment on the proposed N-gram modeling, we use a newly designed three-pass parsing approach but with the existing rules.

Speech recognition generates a sequence of words with additional concept information about chunk phrase categories and lexical attributes, such as "*May you please en give me a ticket ah <route>from Taipei to Boston</route> ah <time>on March seven</time>?"* The complete parsing of the full sentences acts as the first pass. Once it fails, the second pass is to parse the joint phrase of <route> and <time> chunks like "from Taipei to Boston on March seven" with the same parser and grammar, as illustrated in Figure 3. Once it fails again, the lexical attributes would be detected and check their consistency for the final possible concept understanding.

```
Recog:     May you give me a ticket of Thai_'airline'
           <route> from Taipei_'city-departure' to
           Boston_'city-arrival' </route> <time> on
           March_'month' seven_'day' </time>?
Sentence:  May you give me a ticket of Thai from Taipei
           to Boston on March seven?          ➔ 1-pass
<route>+<time> joint phrase:
           from Taipei_'city-departure' to Boston
           _'city-arrival' on March_'month'
           seven_'day'                        ➔ 2-pass
Lexical attributes:
           Thai_'airline' & Taipei_'city-departure' &
           Boston_'city-arrival' & March_'month' &
           seven_'day'                        ➔ 3-pass
```

Figure 3. Three-pass understanding process

Unavoidable recognition errors and incompleteness of grammar might result in understanding errors. Our proposed approach provides a similar phrase-spotting ability as partial parsing does. The partial parsing approach is constrained to less range of syntactic information and can work efficiently and reliably as the complete parsing fails. However, it sacrifices completeness and depth of analysis [4][5]. On the contrary, the merged N-gram models in our approach possess the continuity of word sequences, instead of phrase sequences, over the whole sentence range in computing probabilities.

An apparent difference exists between our approach and partial parsing: the former provide in the front-end speech recognizer both the chunk and word probabilities seamlessly in the merged N-gram model, while the later should perform in the back-end language understanding processing with fixed recognizer output word graphs or n-best sequences. Furthermore, sophisticated utilization of the recognized sentence pattern in our understanding process could be developed to enhance the sentence level analysis, which was disregarded in our preliminary study.

3. Experiments

Our system is composed of a segment-based speech recognizer SUMMIT [15] and a natural-language understanding system TINA [9] for spoken languages. The acoustic model for SUMMIT is trained using Mandarin telephony speech corpora MAT-2000 [16], while language modeling of N-grams for SUMMIT and of SCFG for TINA using 2,928 utterances collected through the Mandarin Mercury system. Vocabulary of the recognizer has 2,424 base words in 186 base categories.

Our language model training set was annotated with the mentioned concept information manually in the experiments. Statistics of the data is summarized in Table 1 and 2. The training set was divided into two sub-sets: Set A is composed of sentences with <route> and/or <time> chunk phrases, while Set B of neither. The average length of Set A is 5.59 words per sentence, which is 1.6 times of that of Set B. After decomposing target phrases, the average length of sentences decreases to 3.17. The average lengths of <time> and <route> phrases are 3.15 and 2.85, respectively. It shows that decomposition of the utterances into phrases might largely shorten them and probably provide an easier framework in developing of the grammar and collecting the corpora.

Table 1. Statistics of the training utterances

	Set A	Set B
#Utterance	1,980	948
Word / sent.	5.59	3.53

Table 2. Statistics of Set A

	<route>	<time>	Abstr-form
#Utterance	1,358	1,059	1,980
Ave. #Word	2.85	3.15	3.17

A set of 3,389 utterances is used as the test set. The baseline system uses conventional category-based trigram modeling and complete parsing strategy (called 'one-pass' in the paper). Its performance is listed in Table 3: toneless syllable error rate (SER) 11.92%, concept error rate (CER) 21.00% and parsing failure rate (PFR) 10.24%.

Table 3. Conventional trigram modeling (baseline)

(%)	SER	PFR	CER	Rel. Impr.
1-pass	11.92	10.24	21.00	-

3.1. Experiments using conventional trigram modeling

A conventional trigram modeling is experimented here based on the annotated training sentences with additional chunk and attribute information. The results were used for a comparison to the proposed two-layer organization of corpora for trigram modeling, which is shown in the next subsection.

Table 4. Conventional trigram modeling, attributed words

(%)	SER	PFR	CER	Rel. Impr.
1-pass	12.11	10.30	21.06	-0.3
2-pass	12.11	4.90	18.66	+11.1
3-pass	12.11	2.54	18.21	+13.3

The SER in Table 4 is larger than that of the baseline in Table 3, probably due to the enlarged vocabulary and data sparseness. The second-pass parsing dropped PFR from 10.39% to 4.90%. The huge drop of PFR might contribute to a relative 11.1% CER reduction. The chunk information generated from the speech recognizer seemed helpful. The joint phrases, by excluding the words outside the <route> and <time> chunks, were largely accepted by the same grammar and parser. The third pass of understanding process did help a big drop of PFR but not in CER. Its concepts understood showed much less accuracy contribution than that of the second pass. Probably, the lexical attribute attachment had limited adjacent contextual dependency accounts for its less confidence in understanding.

3.2. Experiments using two-layer trigram modeling

Here are the major experiments of the proposed approach.

Table 5. Two-layer trigram modeling, attributed words

(%)	SER	PFR	CER	Rel. Impr.
1-pass	11.30	5.02	16.18	+23.0
2-pass	11.30	2.63	14.75	+29.8
3-pass	11.30	0.77	14.12	+32.8

The proposed two-layer organization of corpora contributed to a better N-gram model for the speech recognizer. Both the SER and CER of the one-pass parsing showed significant improvements of relative 5.2% and 23.0% error rate reduction, respectively. The second pass parsing accounted for a significant PFR reduction, similar to the result in previous subsection, and accounted for relatively 29.8% error rate reduction. Comparing to the two-pass performance in the Table 4, the two-layer trigram modeling gained more improvement than the conventional trigram modeling did. That encourages the collection of separated chunk corpora in composing of spoken dialogue systems.

The following two reasons might explain the significant improvement by using the proposed two-layer trigram modeling. Firstly, the trigram model might be well trained within the scope of the phrase layer by the constraint of a smaller vocabulary size and in a shorter context range. In addition, the trained trigrams of word sequences across chunk phrases were influenced by the trigrams in the sentence-pattern layer, which possessed longer distance dependency.

Importantly, the use of chunk structure probably provides an efficient way in developing spoken dialogue applications. The reusability of the knowledge in the grammars and the information in the corpora is one of the major concerns. The proposed approach provides a way of flexibly reusing them because of its shorter context range. Like the time chunk, both the time corpus and the time grammar would be well adopted by another applications concerning about time information. Besides, the proposed trigram modeling is still mainly based on data-driven approach, so as to rely less on the experts for grammars. Collection of the phrase sets across applications may make it more mature and enhance the efficiency of system development.

4. Conclusions

In the paper, we experimented the integration of multiple layers of concept information, including both chunk phrases and lexical attributes, into trigram modeling for spoken language understanding. It outperformed the conventional way by more than 30% concept error rate reduction in our Mandarin Mercury system. Firstly, it provides a robust way to spoken language understanding by parsing phrases instead of full sentences. Like partial parsing, it outperforms the complete parsing and salvages lots of parsing failures. Secondly, it improves the recognition performance by smoothing away the data sparseness problem via the construction of two-layer organization of corpora for N-gram modeling. Finally, via the more reusable chunk phrase grammars and corpora the effectiveness and efficiency of system development could be enhanced.

5. Acknowledgements

The author would like to thank all colleagues in Delta Electronics speech group for a lot of valuable discussions and cooperation. We would like to thank the MIT-SLS group for the long-term support on spoken language technologies. Finally, we would like to express our gratitude to Alan Lee and Ching-Ho Tsai for their endeavor to developing the Mandarin Mercury system.

6. References

[1] Zue, V. W. and Glass, J. R., “Conversational Interface: Advances and Challenges”, *Proc. IEEE, Special Issue on Spoken Language Processing,* Vol. 88, August 2000.

[2] Seneff, S., "Response Planning and Generation in the MERCURY Flight Reservation System", *Computer Speech and Language,* Vol. 16, pp. 283-312, 2002.

[3] Giachin, E. and McGlashan, S., “Spoken Language Dialogue Systems”, chapter three in *Corpus-Based Methods in Language and Speech Processing,* edited by Yong, S. and Bloothooft, G., published by Kluwer Academic, 1997.

[4] Abney, S., “Part-of-Speech Tagging and Partial Parsing”, chapter four in *Corpus-Based Methods in Language and Speech Processing,* see above.

[5] Kellner, A., Rueber, B., Seide, F. and Tran, B.-H., “PADIS – an Automatic Telephone Switchboard and Directory Information System”, *Speech Communication,* 1996.

[6] Abney, S., “Parsing by chunks”, In Berwick, R., Abney, S., and Tenny, C., editors, *Principle-Based Parsing,* Kluwer Academic Publishers, 1991.

[7] Wang, N., Shen, J. L., and Tsai, C. H., “Integrating Layer Concept Information into N-gram Modeling for Spoken Language Understanding”, to appear in *Proc. ICSLP, 2004*.

[8] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L., “Class-based n-gram Models of Natural Language”, *Computational Linguistics* 18(4) pp. 467-479.

[9] Seneff, S., "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1., pp. 61-86, March 1992.

[10] Pla, F., Molina, A., Sanchis, E., Segarra, E. and Garcia, F., “Language Understanding Using Two-Level Stochastic Models with POS and Semantic Units”, *Text Speech and Dialogue*, pp. 403-409, 2001.

[11] Goblirsch, D. M., "Viterbi Beam Search with Layered Bigrams", *Proc. ICSLP,* 1996.

[12] Nasar A, et al, "A Language Model Combining N-grams and stochastic Finite State Automata", *Proc. Eurospeech 1999*.

[13] Wang, Y. Y., Mahajan, M. and Huang, X. "A Unified Context-Free Grammar and N-gram model for Spoken Language Processing", *Proc. ICASSP 2000*.

[14] Wang, K. “Semantics Synchronous Understanding for Robust Spoken Language Applications”, *Proc. ASRU, 2003*.

[15] Glass, J., Hazen, T. J. and Hetherington, L., "Real-time telephone-based speech recognition in the JUPITER domain", *ICASSP,* 1999.

[16] Wang, H. C., Seide, F., Tseng, C. Y. and Lee, L.-S., “MAT-2000 – Design, Collection, and Validation of a Mandarin 2000-Speaker Telephony Speech Database”, *Proc. ICSLP,* 2000.

[17] Komatani, K., Tanaka, K., Kashima, H and Kawahara, T., “Domain-independent spoken dialogue platform using key-phrase spotting based on combined language model.”, *Proc. Eurospeech,* 2001.

[18] Heeman, P. A. and Damnati, G., “Deriving Phrase-based Language Models”, *Proc. ASRU,* 1997.

[19] Kuo, H.-K. J. and Reichl, W., “Phrase-based Language Models for Speech Recognition”, *Proc. Eurospeech,* 1999.