INTEGRATION OF SUPRA-LEXICAL LINGUISTIC MODELS WITH SPEECH RECOGNITION USING SHALLOW PARSING AND FINITE STATE TRANSDUCERS

Xiaolong Mou, Stephanie Seneff and Victor Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
{mou,seneff,zue}@sls.lcs.mit.edu

ABSTRACT

This paper proposes a layered Finite State Transducer (FST) framework integrating hierarchical supra-lexical linguistic knowledge into speech recognition based on shallow parsing. The shallow parsing grammar is derived directly from the full fledged grammar for natural language understanding, and augmented with top-level *n*-gram probabilities and phrase-level context-dependent probabilities, which is beyond the standard context-free grammar (CFG) formalism. Such a shallow parsing approach can help balance sufficient grammar coverage and tight structure constraints. The context-dependent probabilistic shallow parsing model is represented by layered FSTs, which can be integrated with speech recognition seamlessly to impose early phrase-level structural constraints consistent with natural language understanding. It is shown that in the JUPITER [1] weather information domain, the shallow parsing model achieves lower recognition word error rates, compared to a regular class *n*-gram model with the same order. However, we find that, with a higher order top-level *n*-gram model, pre-composition and optimization of the FSTs are highly restricted by the computational resources available. Given the potential of such models, it may be worth pursing an incremental approximation strategy [2], which includes part of the linguistic model FST in early optimization, while introducing the complete model through dynamic composition.

1. INTRODUCTION

Supra-lexical linguistic modeling in speech recognition refers to formalizing and applying linguistic knowledge above the word level. In highly constrained tasks, it is possible to build a word network that specifies all possible word sequences to be recognized. For speech-based conversational interfaces and more complex domains, however, the most commonly used supra-lexical linguistic modeling approach is the *n*-gram language model. While this approach can successfully model local context-dependencies given sufficient training data, it generally ignores long-distance sentence structure constraints. Furthermore, such unstructured statistical models may not be consistent with the typical rule-based models used in the natural language understanding component of a conversational interface. To address these limitations of *n*-gram models, researchers have explored the use of structured supra-lexical linguistic models [3] with formal grammars, and the integration of semantic constraints into speech recognition [4].

One concern of using formal grammars in speech recognition is to balance sufficient generality and tight constraints. This is particularly important in conversational interfaces where spontaneous speech has to be handled. In many circumstances, such speech inputs may violate the predefined grammar. Another factor concerning supra-lexical linguistic modeling is the integration framework for linguistic knowledge. In a typical speech understanding system, speech recognition and natural language understanding are integrated with an *N*-best list or a word graph, and the language understanding component acts as a post-processor for the recognition hypotheses. It is basically a feed-forward system, with little feedback from natural language understanding to guide the speech recognition search. It can be advantageous to use a unified framework incorporating supra-lexical linguistic knowledge through a tightly coupled interface, which offers early integration of linguistic constraints provided by natural language understanding.

In this work, we propose a two-layer hierarchical linguistic model based on shallow parsing, where meaning-carrying phrases are identified and reduced by a phrase-level shallow parsing grammar. The shallow parsing grammar is derived directly from the full TINA [5] natural language grammar, and augmented with context-dependent probabilities beyond the standard CFG formalism. A layered FST framework is applied to construct the probabilistic shallow parsing model, which has the potential of imposing structural supra-lexical linguistic constraints early during the speech recognition search. FSTs have been used as a flexible framework for integrating different knowledge sources in speech recognition [6], and a similar layered FST approach has been adopted at the sub-lexical level [7] to support generic sub-word structures. In the following sections, we first introduce the phrase-level shallow parsing approach. Then, the FST construction details for the shallow parsing based supra-lexical linguistic model are discussed. Next, FST-based speech recognition experiments are conducted in the JUPITER weather information domain, and the experimental results are shown. Finally, we present our conclusions and discuss the advantages and disadvantages of such an FST-based supra-lexical linguistic modeling approach.

2. INTEGRATION OF LINGUISTIC CONSTRAINTS USING SHALLOW PARSING

Supra-lexical sentence structure is usually described by formal grammars. As was mentioned in section 1, it is important to

This research was supported by a contract from the Industrial Technology Research Institute (ITRI), and by DARPA under contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center.

In *Proceedings of the 7th International Conference on Spoken Language Processing*, Sep. 16-20, 2002, Denver, Colorado, pp. 1289-1292.

balance sufficient generality (i.e., coverage) and tight constraints (i.e., precision) while applying formal grammars. In this work, we model meaningful phrases using a phrase-level shallow parsing grammar. Shallow parsing is a generic term for analyses that are less complete than the output from a conventional natural language parser. A shallow analyzer may identify some phrasal constituents of the sentence, without presenting a complete sentence structure.

Our shallow parsing grammar is derived directly from the full-fledged natural language grammar; therefore, the same phrase structure constraints are applied in speech recognition and natural language understanding. The shallow parse tree is augmented with a probability framework where top-level reduced sentences are supported by an *n*-gram model, and phrases are supported by context-dependent phrase-level probabilities. This is essentially a hybrid approach enforcing longer-distance phrase structure constraints through grammars, as well as retaining the flexibility of allowing arbitrary phrase and filler word sequences with top-level probability constraints.

2.1. TINA natural language grammar

Our study of supra-lexical linguistic modeling is based on TINA, which is a natural language understanding system for spoken language applications introduced in [5]. TINA is designed to perform linguistic analysis for speech recognition hypotheses, and generate a semantic representation encoding the meaning of the utterance.

Like most natural language understanding systems, TINA uses a set of hierarchical CFG rules to describe the sentence structure. The grammars that are designed for our spoken dialogue systems typically incorporate both syntactic and semantic information simultaneously. At the higher levels of the parse tree, major syntactic constituents, such as subject, predicate, object, etc., are explicitly represented through syntax-oriented grammar rules. The syntactic structures tend to be domain-independent, capturing general syntactic constraints of the language. At the lower parse tree levels, major semantic classes, such as "a_location," "a_flight," etc., are constructed according to semantic-oriented grammar rules. The semantic structures tend to be domain-dependent, capturing specific meaning interpretations in a particular application domain. Such a grammar is able to combine syntactic and semantic constraints seamlessly. It also offers an additional convenience that no separate semantic rules are necessary for meaning analysis. The semantic representation can be derived directly from the resulting parse tree.

2.2. Derived phrase-level shallow parsing grammar

The shallow parsing grammar we use is derived directly from the full TINA grammar. It covers the key phrases specified by a set of chosen meaning-carrying TINA categories. This ensures that the phrase-level structural constraints used in speech recognition are consistent with natural language understanding. The derived shallow parsing grammar has a two-layer structure. The top layer allows arbitrary connections between phrases and filler words not covered by phrase-level grammars. The bottom phrase layer represents possible word realizations of phrases. The hierarchical structure within each phrase is not preserved for shallow parsing, because most phrases we model correspond to semantic-oriented categories in the original TINA parse tree, which are usually located at lower levels without deep hierarchy within the phrase. Furthermore, using a two-layer representation simplifies the shallow parsing structure, which facilitates the application of context-dependent probabilities using a layered FST approach. Figure 1 shows an example parse tree according to such a two-layer shallow parsing grammar.

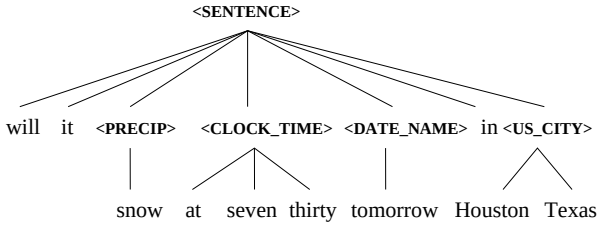

Fig. 1. Example two-layer parse tree according to the shallow parsing grammar.

2.3. Probability framework with shallow parsing

The probability framework associated with the shallow parsing grammar has two major components: the top level *n*-gram probability model, and the phrase level rule-start and rule-internal probability models. The top level *n*-gram probabilities are used to provide local constraints between reduced phrases and filler words. The phrase level probabilities are used to capture the probabilities of word realizations from phrases.

To build the top-level *n*-gram model, the training sentences are first reduced to sequences of phrase tags and filler words according to the shallow parsing grammar. For example, the sentence "will it snow at seven thirty tomorrow in Houston Texas" shown in Figure 1 is reduced to "will it <PRECIP> <CLOCK_TIME> <DATE_NAME> in <US_CITY>." Then, the top-level *n*-gram probability model is trained from the reduced sentences. Since the shallow parsing grammar allows arbitrary phrase and filler word connections at the top level, it is important to have this *n*-gram model to impose additional probability constraints over the the reduced sentences.

The phrase-level rule-start probability is specified for a phrase-start node in the two-layer shallow parse tree. It is conditioned not only on its parent phrase, but also on the left phrase or filler word. Such a *context-dependent* rule-start probability is able to capture context dependency beyond the current phrase boundary. The rule-internal probability is specified for a phrase-internal node, conditioned on the current phrase and the previous phrase-internal node. For example, the probability of the word "tomorrow" in Figure 1 is defined as the conditional probability $P(\text{tomorrow} \mid \text{<DATE\_NAME>}, \text{<CLOCK\_TIME>})$, and the probability of the word "Texas" is defined by the conditional probability $P(\text{Texas} \mid \text{<US\_CITY>}, \text{Houston})$. We also experimented with a *generic* rule-start probability, which is the probability of the phrase-start node conditioned only on the parent phrase. Such generic rule-start probabilities have less detailed conditional context, and do not model context-dependency across phrase rule boundaries.

The rule-start and rule-internal probabilities are trained by parsing a training corpus according to the shallow grammar. The probability of a complete phrase is constructed as a product of the rule-start probability for the phrase-start word and the rule-internal probabilities for other words in the phrase. The complete two-layer parse tree probability is defined as the product of the top-level *n*-gram probabilities and the phrase-level rule-start and rule-internal probabilities.

3. CONTEXT-DEPENDENT PROBABILISTIC SHALLOW PARSING USING LAYERED FSTS

The shallow parsing based supra-lexical linguistic model can be integrated into speech recognition seamlessly within an FST framework. Since the shallow parsing grammar is derived from the full natural language grammar used in natural language understanding, such a tight integration has the potential of providing early feedback consistent with natural language understanding to guide the recognition search. Furthermore, the unified FST framework allows global optimization to be performed on a single composed recognition search space.

The speech recognizer we use is an FST-based system represented by the FST compostion $A \circ L \circ G$, where A is the acoustic model, L encodes the mapping from sub-word acoustic units to words, and G is the supra-lexical linguistic model. Our approach is to substitute the baseline class *n*-gram FST G with the probabilistic shallow parsing model. The shallow parsing grammar can be represented by Recursive Transition Networks (RTNs), which are supported by our FST library. The top-level *n*-gram model probabilities, the *generic* rule-start probabilities and the rule-internal probabilities can all be directly encoded by the transition weights within the sub-networks of the RTNs. It is difficult, however, to incorporate context-dependent probabilities into RTNs directly, because such probabilities are conditioned not only on the parent phrases, but also parse tree nodes beyond the phrase boundaries.

In this work, we decompose G into $R \circ L$, where R is a shallow parsing RTN encoding the top-level *n*-gram probablities, the generic rule-start probablities, and the rule-internal probablities. It takes in a word sequence and outputs a tagged string representing the shallow parse tree. L is the phrase-level rule-start probability FST, which encodes the context-dependent rule-start probabilities. Details of R and L are given below.

3.1. Shallow parsing FST

The shallow parsing FST R is an RTN constructed from the shallow parsing grammar. The sub-networks of R are compiled from the phrase grammar rules. R is configured to output a tagged parse string, which represents the shallow parse tree. Each phrase is enclosed by a phrase open tag and a phrase close tag. For example, the tagged parse string for the shallow parse tree given in Figure 1 is "<SENTENCE> will it <PRECIP> snow </PRECIP> <CLOCK_TIME> at seven thirty </CLOCK_TIME> <DATE_NAME> tomorrow </DATE_NAME> in <US_CITY> Houston Texas </US_CITY> </SENTENCE>." Such a tagged string is used by FST L to apply context-dependent rule-start probabilities.

The supra-lexical parsing RTN incorporates top-level *n*-gram probabilities, as well as phrase-level probabilities, including the *generic* rule-start probabilities and the rule-internal probabilities. The overall structure of R is similar to an FST-based class *n*-gram model, except that the weighted sub-networks representing the word class rules are substituted by the sub-networks representing the shallow phrase rules. R can be used by itself as a supra-lexical linguistic model, without applying context-dependent rule-start probabilities beyond current phrases. This is similar to the hierarchical phrase language model studied by Wang [8].

3.2. Phrase-level rule-start probability FST

The phrase-level rule-start probability FST L is constructed to apply context-dependent rule-start probabilities. The probability of a rule-start node is conditioned on the current parent and its left sibling. The basic context transition of L is designed as follows. It starts with the state representing the current conditional context, i.e., the current parent "P" and its left sibling "a". Then, it applies the context-dependent rule-start probability as it makes the transition accepting the rule-start node "m". Next, it filters out the phrase internal nodes and the filler words between phrases, and finally reaches the state representing the next conditional context, i.e., the next parent "Q" and its left sibling "b". Figure 2 illustrates such a context transition.

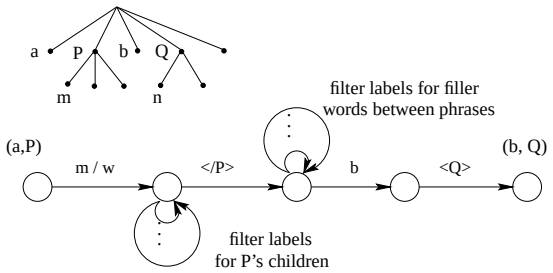

Fig. 2. The context transition diagram in the phrase-level rule-start probability FST, showing the transitions from state (a,P) to state (b,Q), where P, a, Q, and b are the current parent, the current parent's left sibling, the next parent, and the next parent's left sibling, respectively. "w" is the context-dependent rule-start probability: Prob (m | a, P), where "m" is the first child of P.

The context states of L are connected for each trained rule-start probability instance using the basic context transition described above. Since L is composed with R, which already applied generic rule-start probabilities, the context-dependent rule-start probabilities applied in L are normalized by the corresponding generic rule-start probabilities. Note that ill-trained rule-start probabilities without sufficient observations are pruned. In this case, we apply the generic rule-start probability without using context information beyond the current rule. This back-off strategy yields a more robust phrase probability estimation.

3.3. Construction of the complete model

Our approach to constructing the complete shallow parsing based supra-lexical linguistic model FST can be summarized as follows. First, a set of key semantic categories are manually selected in the full TINA natural language grammar. Then, a large training corpus is parsed using the original grammar. The phrases are identified, and the training sentences are reduced to a sequence of phrase tags and filler words between the phrases. The reduced sentences are used to train a top-level bi-gram model, while the identified phrases are used to generate the shallow parsing grammar. Next, the rule-internal probabilities, the generic rule-start probabilities, and the context-dependent rule-start probabilities are trained according to the shallow parsing grammar. Ill-trained context-dependent probability models are pruned. After training the probability models, the top-level *n*-gram probabilities, the generic rule-start probabilities, and the rule-internal probabilities are used to construct the weighted shallow parsing RTN R according to the shallow parsing grammar. The context-dependent rule-start probabilities are used to construct the phrase-level context-dependent rule-start probability FST L. Finally, R and L are composed to obtain the complete shallow parsing based supra-lexical linguistic model. Figure 3 illustrates such an approach.

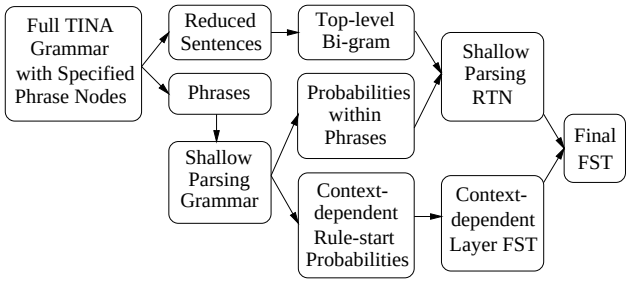

Fig. 3. Construction of FST-based linguistic models derived from TINA language understanding system.

4. EXPERIMENTAL RESULTS

We have experimented with the shallow parsing based supra-lexical linguistic model in the JUPITER weather information domain. Table 1 summarizes the recognition word error rate (WER) results for four different systems: (1) standard class bi-gram, (2) shallow parsing with *generic* rule-start probabilities (using FST R), (3) shallow parsing with *context-dependent* rule-start probabilities (using FST $R \circ L$), and (4) standard class tri-gram. Bi-gram models are used as the top-level probability model in system (2) and (3). The four recognizers are tested on a full test set and an in-vocabulary subset. The ill-trained context-dependent rule-start probability models are pruned, and the pruning threshold is determined through an independent development set.

Supra-lexical Linguistic Model	*WER on Full Test Set (%)*	*WER on In-vocab Test Set (%)*
Class Bi-gram	17.0	12.6
R only	16.8	12.1
$R \circ L$	16.3	11.8
Class Tri-gram	15.3	11.0

Table 1. The recognition word error rate (WER) results in the JUPITER weather information domain.

We can see from the results that, compared to the baseline class bi-gram model, the proposed shallow parsing model with top-level bi-gram and generic-rule start probabilities is able to reduce word error rates on both test sets. The use of context-dependent rule-start probabilities further improves recognition. This suggests that the context-dependent shallow parsing approach with top-level *n*-gram probabilities can offer phrase structure constraints supported by context-dependent phrase probabilities, and may achieve a lower WER compare to a class *n*-gram model with the same order. However, we have found that the FST encoding context-dependent rule-start probabilities has 1100K arcs and 36K states, which is significantly larger than the FST encoding generic rule-start probabilities with 222K arcs and 2K states. The class bi-gram FST consists of only 58K arcs and 1.2K states. If we were to use top-level tri-gram probabilities in the context-dependent shallow parsing model, the recognition FSTs could not be pre-composed and optimized given the computation resources we currently use, though similar recognition improvements are potentially possible. Therefore, the application of the context-dependent shallow parsing model with higher order top-level *n*-gram probabilities can be limited by the complexity of the FSTs.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a shallow parsing approach for supra-lexical linguistic modeling, which helps balance generality and constraints. The proposed context-dependent shallow parsing model is constructed within a layered FST framework, which can be seamlessly integrated with speech recognition. It also has the potential of providing consistent feedback from natural language understanding. Furthermore, with the layered framework, it is convenient to develop the phrase-level context-dependent probability models independently. For example, we can try to use different phrase-level contexts for different domains, which may help produce more effective context-dependent models.

Our speech recognition experiments show that the proposed context-dependent shallow parsing model with top-level *n*-gram probabilities and phrase-level context-dependent probabilities achieves lower recognition word error rates, compared to a regular class *n*-gram model with the same order. However, the final composed FST representing the speech recognition search space can be significantly larger than the regular class *n*-gram model. With a higher order top-level *n*-gram, pre-composition and optimization are restricted by the computational resources available, which can limit the application of this approach. However, given the potential of such models, it may be worth pursing a strategy similar to the FST-based incremental *n*-gram model approach [2], where the complete supra-lexical model is factored into two FSTs. The first one can be statically composed and optimized, and the second one is composed on-the-fly during the recognition search. This approach is able to include part of the supra-lexical linguistic model FST in early pre-composition and optimization, while introducing the complete supra-lexical model through incremental dynamic composition.

6. REFERENCES

[1] V. Zue et al., "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, 2000.

[2] H. Dolfing and I. Hetherington, "Incremental language models for speech recognition using finite-state transducers," in *Proc. IEEE Automatic Speech Recogntion and Understanding Workshop*, Madonna di Campiglio, Italy, 2001.

[3] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 283–332, 2000.

[4] W. Ward and S. Issar, "Integrating semantic constraints into the SPHINX-II recognition search," in *Proc. ICASSP'94*, Adelaide, Australia, 1994, pp. 2017–2019.

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

[6] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducer in speech recognition," in *Proc. ISCA ASR'00*, Paris, France, 2000.

[7] X. Mou, S. Seneff, and V. Zue, "Context-dependent probabilistic hierarchical sub-lexical modeling using finite state transducers," in *Proc. EuroSpeech'01*, Aalborg, Denmark, Sept. 2001, pp. 451–454.

[8] C. Wang et al., "MUXING: A telephone-access mandarin conversational system," in *Proc. ICSLP'00*, Beijing, China, 2000.