Integrating Multi-level Linguistic Knowledge with a Unified Framework for Mandarin Speech Recognition

Xinhao Wang, Jiazhong Nie, Dingsheng Luo, Xihong Wu[*]
Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education),
School of Electronics Engineering and Computer Science,
Peking University, Beijing, 100871, China
{wangxh,niejz,wxh,dsluo}@cis.pku.edu.cn

Abstract

To improve the Mandarin large vocabulary continuous speech recognition (LVCSR), a unified framework based approach is introduced to exploit multi-level linguistic knowledge. In this framework, each knowledge source is represented by a Weighted Finite State Transducer (WFST), and then they are combined to obtain a so-called analyzer for integrating multi-level knowledge sources. Due to the uniform transducer representation, any knowledge source can be easily integrated into the analyzer, as long as it can be encoded into WFSTs. Moreover, as the knowledge in each level is modeled independently and the combination is processed in the model level, the information inherently in each knowledge source has a chance to be thoroughly exploited. By simulations, the effectiveness of the analyzer is investigated, and then a LVCSR system embedding the presented analyzer is evaluated. Experimental results reveal that this unified framework is an effective approach which significantly improves the performance of speech recognition with a 9.9% relative reduction of character error rate on the HUB-4 test set, a widely used Mandarin speech recognition task.

1 Introduction

Language modeling is essential for large vocabulary continuous speech recognition (LVCSR), which aims to determine the prior probability of a supposed word string W, $p(W)$. Although the word-based n-gram language model remains the mainstream for most speech recognition systems, the utilization of linguistic knowledge is too limited in this model. Consequently, many researchers have focused on introducing more linguistic knowledge in language modeling, such as lexical knowledge , syntax and semantics of language (Wang and Vergyri, 2006; Wang et al., 2004; Charniak, 2001; Roark, 2001; Chelba, 2000; Heeman, 1998; Chelba et al., 1997).

Recently, structured language models have been introduced to make use of syntactic hierarchical characteristics (Roark, 2001; Charniak, 2001; Chelba, 2000). Nevertheless, the computational complexity of decoding will be heavily increased, as they are parser-based models. In contrast, the class-based language model groups the words that have similar functions of syntax or semantics into meaningful classes. As a result, it handles the questions of data sparsity and generalization of unseen event. In practice, the part-of-speech (POS) information, capturing the syntactic role of words, has been widely used in clustering words (Wang and Vergyri, 2006; Maltese et al., 2001; Samuelsson and Reichl, 1999). In Heeman's POS language model (Heeman, 1998), the joint probability of word sequence and associated POS sequence was estimated directly, which has been demonstrated to be superior to the conditional probability previously used in the class-based models (Johnson, 2001). Moreover, a SuperARV language model was presented (Wang and Harper, 2002), in which lexical features and syntactic constraints were tightly integrated into a linguistic structure of SuperARV serving as a class in the model. Thus, these knowledge was integrated in the representation level, and then the joint probabilities

[*] Corresponding author: Xihong Wu

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 820–828,
Honolulu, October 2008.

of words and corresponding SuperARVs were estimated. However, in the class-based language models, words are taken as the model units, while other units smaller or larger than words are unfeasible for modeling simultaneously, such as the Chinese characters for Chinese names.

Usually, speech recognition systems can only recognize the words within a predefined dictionary. With the increase of unknown words, i.e., out-of-vocabulary (OOV) words, the performance will degrade dramatically. This is because not only those unknown words cannot be recognized correctly, but the words surrounding them will be affected. Thus, many efforts have been made to deal with the issue of OOV words (Martins et al., 2006; Galescu, 2003; Bazzi and Glass, 2001), and various model units smaller than words have been examined to recognize OOVs from speech, such as phonemes (Bazzi and Glass, 2000a), variable-length phoneme sequence (Bazzi and Glass, 2001), syllable (Bazzi and Glass, 2000b) and sub-word (Galescu, 2003). Since the proper name is a typical category of OOV words and usually takes a very large proportion among all kinds of OOV words, it has been specially addressed in (Hu et al., 2006; Tanigaki et al., 2000).

All those attempts mentioned above succeed in utilizing linguistic knowledge in language modeling in some degree respectively. In this study, a unified framework based approach, which aims to exploit information from multi-level linguistic knowledge, is presented. Here, the Weighted Finite State Transducer (WFST) turns to be an ideal choice for our purpose. WFSTs were formerly introduced to simplify the integration of models in speech recognition, including acoustic models, phonetic models and word n-gram (Mohri, 1997; Mohri et al., 2002). In recent years, the WFST has been successfully applied in several state-of-the-art speech recognition systems, such as systems developed by the AMI project (Hain et al., 2006), IBM (Saon et al., 2003) and AT&T (Mohri et al., 1996), and in various fields of natural language processing, such as smoothed n-gram model, partial parsing (Abney, 1996), named entities recognition (Friburger and Maurel, 2004), semantic interpretation (Raymond et al., 2006) and machine translation (Tsukada and Nagata, 2004). In (Takaaki Hori and Minami, 2003), the WFST has been further used for language model adaptation, where language models of different vocabularies that represented different styles were integrated through the framework of speech translation. In WFST-based systems, all of the models are represented uniformly by WFSTs, and the general composition algorithm (Mohri et al., 2000) combines these representations flexibly and efficiently. Thereby, rather than integrating the models step by step in decoding stage, a complete search network is constructed in advance. The combined WFST will be more efficient by optimizing with determinization, minimization and pushing algorithms of WFSTs (Mohri, 1997). Besides, the researches on optimizing the search space and improving WFST-based speech recognition has been carried out, especially on how to perform on-the-fly WFSTs composition more efficiently (Hori et al., 2007; Diamantino Caseiro, 2002).

In this study, we extend the linguistic knowledge used in speech recognition. As WFSTs provide a common and natural representation for lexical constraints, n-gram language model, Hidden Markov Model models and context-dependency, multi-level knowledge sources can be encoded into WFSTs under the uniform transducer representation. Then this group of WFSTs is flexibly combined together to obtain an analyzer representing knowledge of person and location names as well as POS information. Afterwards, the presented analyzer is incorporated into LVCSR to evaluate the linguistic correctness of recognition candidates by an n-best rescoring.

Unlike other methods, this approach holds two distinct features. Firstly, as all multi-level knowledge sources are modeled independently, the model units such as character, words, phrase, etc., can be chosen freely. Meanwhile, the integration of these information sources is conducted in the model level rather than the representation level. This setup will help to model each knowledge source sufficiently and may promote the accuracy of speech recognition. Secondly, under this unified framework, it is easy to combine additional knowledge source into the framework with the only requirement that the new knowledge source can be represented by WFSTs. Moreover, since all knowledge sources are finally represented by a single WFST, additional efforts are not required for decoding the new knowledge source.

The remainder of this paper is structured as follows. In section 2, we introduce our analyzer in detail, and incorporate it into a Mandarin speech recognition system. In section 3, the simulations are performed to evaluate the analyzer and test its effectiveness when being applied to LVCSR. The conclusion appears in section 4.

2 Incorporation of Multi-level linguistic knowledge in LVCSR

In this section, we start by giving a brief description on WFSTs. Then some special characteristics of Chinese are investigated, and the model units are fixed. Afterwards, each knowledge source is represented with WFSTs, and then they are combined into a final WFST, so-called analyzer. At last, this analyzer is incorporated into Mandarin LVCSR.

2.1 Weighted Finite State Transducers

The Weighted Finite State Transducer (WFST) is the generalization of the finite state automata, in which, besides of an input label, an output label and a weight are also placed on each transition. With these labels, a WFST is capable of realizing a weighted relation between strings. In our system, log probabilities are adopted as transition weights and the relation between two strings is associated with a weight indicating the probability of the mapping between them.

Given a group of WFSTs, each of which models a stage of a mapping cascade, the composition operation provides an efficient approach to combine them into a single one (Mohri et al., 2002; Mohri et al., 1996). In particular, for two WFSTs R and S, the composition $T = R^{o}S$ represents the composition of relations realized by R and S. The combination is performed strictly on R's output and S's input. It means for each path in T, mapping string r to string s, there must exist a path mapping r to some string t in R and a path mapping t to s in S. Decoding on the combined WFST enables to find the joint optimal results for multi-level weighted relations.

2.2 Model Unit Selection

This study primarily takes the person and location names as well as the POS information into account. To deal with Chinese OOV words, different from the western language in which the phoneme, syllable or sub-word are used as the model units (Bazzi and Glass, 2000a; Bazzi and Glass, 2000b; Galescu, 2003), Chinese characters are taken as the basic units. In general, a person name of Han nationality consists of a surname and a given name usually with one or two characters. Surnames commonly come from a fixed set that has been historically used. According to a recent investigation on surnames involving 296 million people, 4100 surnames are found, and 129 most used surnames account for 87% (conducted by the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences). In contrast, the characters used in given names can be selected freely, and in many situations, some commonly used words may also appear in names, such as ”胜利” (victory) and ”长江” (the Changjiang River). Therefore, both Chinese characters and words are considered as model units in this study, and a word re-segmentation process on recognition hypotheses is necessary, where an n-gram language model based on word classes is adopted.

2.3 Representation and Integration of Multi-level Knowledge

In this work, we ignore the word boundaries of n-best hypotheses and perform a word re-segmentation for names recognition. Given an input Chinese character, it is encoded by a finite state acceptor FSA_{input}. For example, the input ”合成分子时” (while synthesizing molecule) is represented as in Figure 1(a). Then a dictionary is represented by a

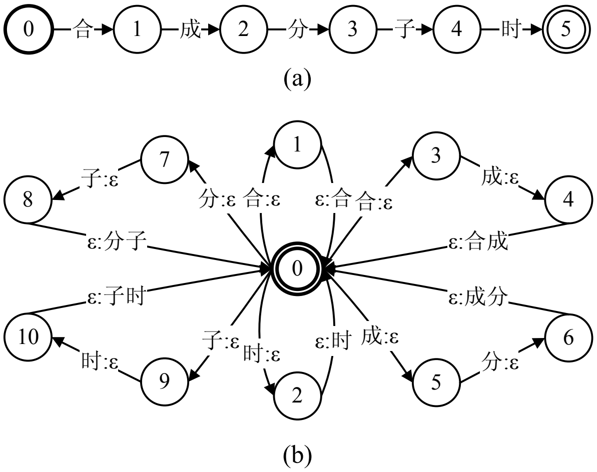

Figure 1: (a) is an example of the FSA representing a given input; (b) is the FST representing a toy dictionary.

transducer with empty weights, denoted as FST_{dict}. Figure 1(b) illustrates a toy dictionary listed in Table 1, in which a successful path encodes a mapping from a Chinese character sequence to some word in the dictionary. In practice, all Chinese characters should appear in the dictionary for further incorporating models of names. Then the combination of FSA_{input} and FST_{dict}, $FST_{seg} = FSA_{input} \circ FST_{dict}$, will result in a WFST embracing all the possible candidate segmentations. Afterwards an n-gram language model based on word classes is used to weight the candidate segmentations. As in Figure 2, a toy bigram with three words is depicted by $WFST_{n-gram}$, and the word classes are defined in Table 2. Here, both in the training and test stages, the strings of numbers or letters in sentences are extracted according to the rules, and then substituted with the class tags, ”NUM” and ”LETTER” respectively. At the same time, the words, such as ”三月” and ”A型”, are replaced with ”NUM月” and ”LETTER型” in the dictionary. In addition, name classes, including ”CNAME”, ”TNAME” and ”LOC”, will be set according to names recognition.

Chinese Words	English Words
合成	synthesize
成分	element
分子	molecule
子时	the period of the day from 11 p.m.to l a.m.
合	together
时	present

Table 1: The Toy dictionary

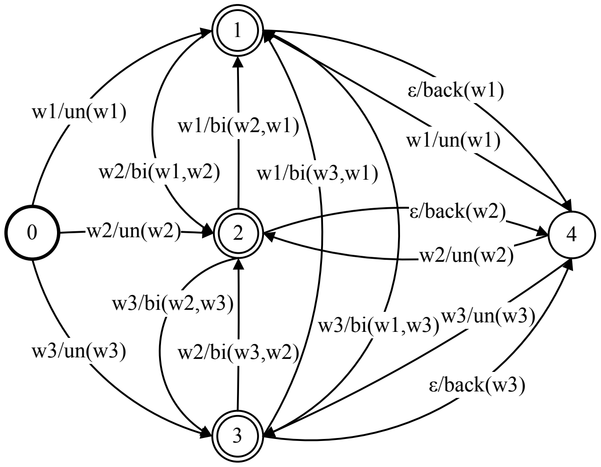

Figure 2: The WFST representing a toy bigram language model, in which $un(w1)$ denotes the unigram of $w1$; $bi(w1, w2)$ and $back(w1)$ respectively denotes the bigram of $w2$ and the backoff weight given the word history $w1$.

Classes	Description
w_i	Each word w_i listed in the dictionary
CNAME	Person names of Han nationality
TNAME	Translated person names
LOC	Location names
NUM	Number expressions
LETTER	Letter strings
NON	Other non Chinese character strings
BEGIN	Beginning of sentence
END	End of sentence

Table 2: The Definition of word classes

Hidden Markov Models (HMMs) are adopted both for names recognition and POS tagging. Here, each HMM is represented with two WFSTs. Taking the POS tagging as an example, the toy POS WFSTs with 3 different tags are illustrated in Figure 3. The emission probability of a word by a POS, $(P(word/pos))$, is represented as in Figure 3(a), and the bigram transition probabilities between POS tags are represented as in Figure 3(b), similar to the word n-gram. In terms of names recognition, the HMM states correspond to 30 role tags of names, some for model units of Chinese characters, such as surname, the first or second character of a given person name with two characters, the first or last character of a location name and so on, but others for model units of words, such as the word before or after a name, the words in a name and so on. When recognizing the person names, since there is a big difference between the translated names and the names of Han nationality, two types of person names are modeled separately, and substituted with two different class tags in the segmentation language model, as ”TNAME” and ”CNAME”. Some rules, which can be encoded into WFSTs, are responsible for the transformation from a role sequence to corresponding name class (for example, a role sequence might consist of the surname, the first character of the

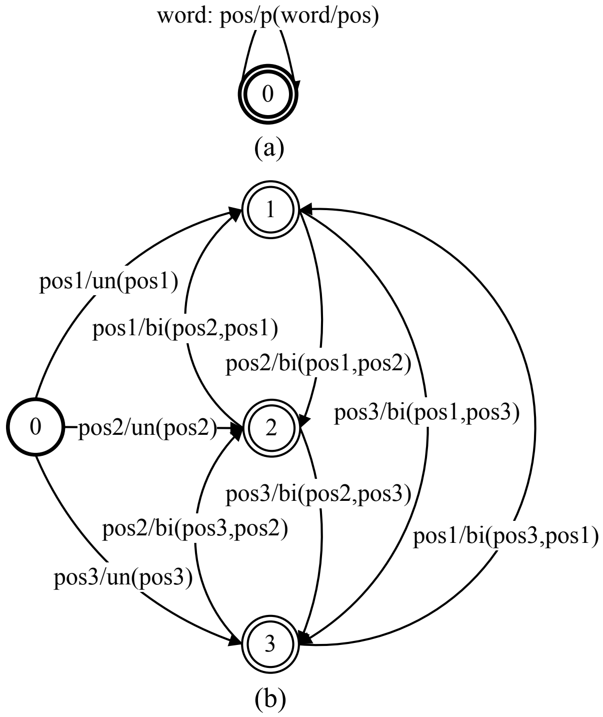

Figure 3: The toy POS WFSTs. (a) is the WFST representing the relationship between the word and the pos; (b) is the WFSA representing the bigram transition probabilities between POS tags

given name, and the second character of the given name, which will be transformed to "CNAME" in FST_{seg}). Hence, taking names recognition into account, a WFST, including all possible segmentations as well as recognized candidates of names, can be obtained as below, denoted as $WFST_{words}$:

$$FSA_{input} \circ FST_{dict} \circ WFST_{ne} \circ WFSA_{n-gram} \tag{1}$$

POS information is integrated as follows.

$$(\alpha * WFST_{words}) \circ WFST_{POS} \tag{2}$$

Consequently, the desired analyzer, a combined WFST that represents multi-level linguistic knowledge sources, has been obtained.

2.4 Incorporation in LVCSR

The presented analyzer models linguistic knowledge at different levels, which will be useful to find an optimal words sequence among a large number of speech recognition hypotheses. Thus in this research, the analyzer is incorporated after the first pass recognition, and the n-best hypotheses are reranked according to the total path scores adjusted with the analyzer scores as follows.

$$\hat{W} = \arg\max_{W} \left(\begin{array}{l} \log\left(P_{AM}\left(O|W\right)\right) \\ +\beta * \log\left(P_{LM}\left(W\right)\right) \\ +\gamma * \log\left(P_{Analyzer}\left(W\right)\right) \end{array} \right) \tag{3}$$

where $P_{AM}\left(O|W\right)$ and $P_{LM}\left(W\right)$ are the acoustic and language scores produced in first pass decoding, and $P_{Analyzer}\left(W\right)$ reflects the linguistic correctness of one hypothesis scored by the analyzer. Through the reranking paradigm, a new best sentence hypothesis is obtained.

3 Simulation

Under the unified framework, multi-level linguistic knowledge is represented by the analyzer as mentioned above. To guarantee the effectiveness of the introduced framework in integrating knowledge sources, the analyzer is evaluated in this section. Then the experiments using an LVCSR system in which the analyzer is embedded are performed.

3.1 Analyzer Evaluation

Considering the function of the analyzer, cascaded subtasks of word segmentation, names recognition and POS tagging can be processed jointly, while they are traditionally handled in a pipeline manner. Hence, a comparison between the analyzer and the pipeline system can be used to evaluate the effectiveness of the introduced framework for knowledge integration. As illustrated in Figure 4, two systems based on the presented analyzer and the pipeline manner are constructed respectively.

The evaluation data came from the People's Daily of China in 1998 from January to June (annotated by the Institute of Computational Linguistics of Peking University[1]), among which the January to May data was taken as the training set, and the June data was taken as the test set (consisted of 21,143 sentences and about 1.2 million words). The first two thousand sentences from the June data were extracted as the development set, used to fix the composition weight α in equation 2. A dictionary including about 113,000 words was extracted from the training data,

[1] http://icl.pku.edu.cn/icl_res/

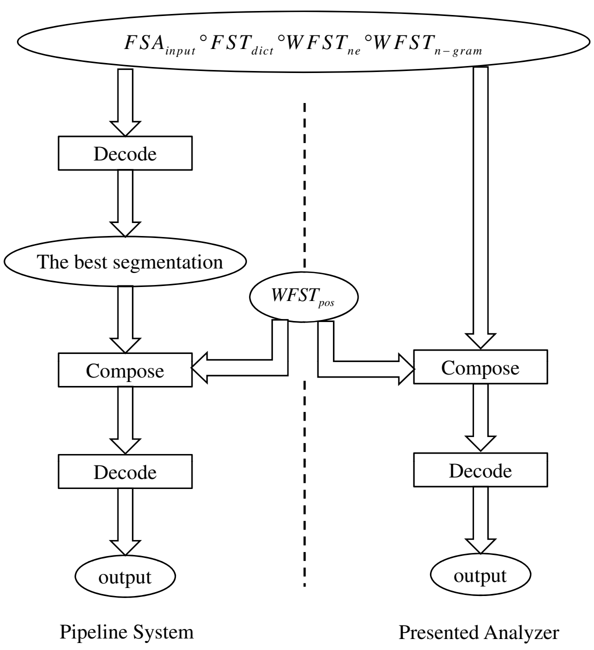

Figure 4: The pipeline system vs The analyzer

in which a person or location name was accounted as a word in vocabulary, only when the number of its appearances was no less than three.

In Figure 5, the analyzer is compared with the pipeline system, where the analyzer outperforms the pipeline manner on all the subtasks in terms of $F1$-score metric. Furthermore to detect the differences, the statistical significance test using approximate randomization approach (Yeh, 2000) is done on the word segmentation results. Since there are more than 21,000 sentences in the test set, which is not appropriate for approximate randomization test, ten sets (500 sentences for each) are randomly selected from the test corpus. For each set, we run 1048576 shuffles twice and calculate the significance level p-value according to the shuffled results. It has been shown that all p-value are less than 0.001 on the ten sets. Accordingly the improvement is statistically significant. Actually, this significant improvement is reasonable, since the joint processing avoids error propagation and provides the opportunity of sharing information between different level knowledge sources. The superiority of this analyzer also shows that the integration of multi-level linguistic knowledge under the unified framework is effective, which may lead to improved LVCSR.

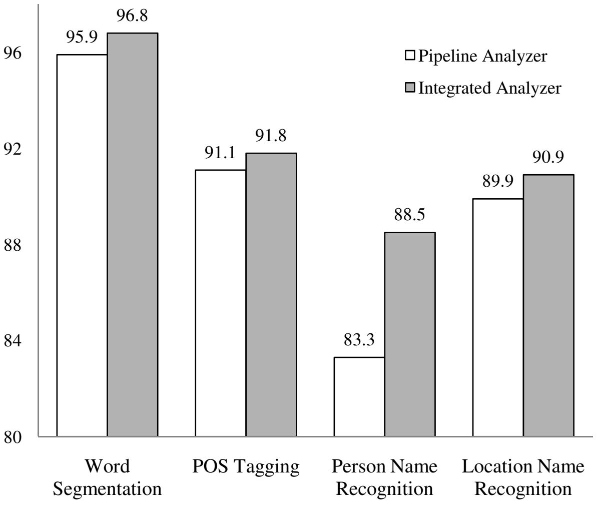

Figure 5: The Performance comparison between the pipeline system and the analyzer. The system performances are measured with the F1-score in the tasks of word segmentation, POS tagging, the person names recognition and the location names recognition.

3.2 Experimental Setup for Mandarin Speech Recognition

In the baseline speech recognition system, the acoustic models consisted of context-dependent Initial-Final models, in which the left-to-right model topology was used to represent each unit. According to the phonetic structures, the number of states in each model was set to 2 or 3 for initials, and 4 or 5 for tonal finals. Each state was trained to have 32 Gaussian mixtures. The used 39-dimension feature vector comprised 12 MFCC coefficients, energy, and their first-order and second-order deltas. Since in this work we focused on modeling knowledge of language in Mandarin LVCSR, only clean male acoustic models were trained with a speech database that contained about 360 hours speech of over 750 male speakers. This training data was picked up from three continuous Mandarin speech corpora: the 863-I, 863-II and Intel corpora. The brief information about these three speech corpora was listed in Table 3. As in this work, the evaluation data was the 1997 HUB-4 Mandarin broadcast news evaluation data (HUB-4 test set), to better fit this task, the acoustic models were adapted by the approach of maximum a posterior (MAP) adaptation. The adaption data was drawn from the HUB4 training set, excluding the HUB-4 develop-

Corpus	Speakers	Amount of Speech (hours)
863-I (male)	83	56.67
863-II(male)	120	78.08
Intel (male)	556	227.30
total	759	362.05

Table 3: The information of the speech training data

System	Err.	Sub.	Del.	Ins.
Baseline	14.85	13.02	0.76	1.07
Analyzer incorporation	13.38	11.78	1.00	0.60

Table 4: The Speech recognition results

ing set, where only the cleaned male speech data (data under condition f0 defined as (Doddington, 1996)) was used. The partition for the clean data was done with the acoustic segmentation software CMUseg_0.5[2] (Siegler et al., 1997), and finally 8.6 hours adaptation data was obtained.

The language model was a word-based trigram built on 60,000 words entries and trained with a corpus about 1.5 billion characters. The training set consisted of broadcast news data from the Xinhua News Agency released by LDC (Xinhua part of Chinese Gigaword), seven years data of People's Daily of China from 1995 to 2002 released by People's Daily Online[3], and some other data from news websites, such as yahoo, sina and so on.

In addition, the analyzer incorporated in speech recognition was trained with a larger corpus from People's Daily of China, including the data in 1998 from January to June and the data in 2000 from January to November (annotated by the Institute of Computational Linguistics of Peking University). The December data in 2000 was taken as the development set used to fix the composition weight α in equation 2.

3.3 Experimental Results

In our experiments, the clean male speech data from the Hub-4 test set was used, and 238 sentences were finally extracted for testing. The weight of the analyzer was empirically derived from the development set, including 649 clean male sentences from the devSet of HUB-4 Evaluation. The recognition results are shown in Table 4. The baseline system has a character error rate (CER) of 14.85%. When the analyzer is incorporated, a 9.9% relative reduction is achieved. Furthermore, we ran the statistical significance test to detect the performance improvement, in which the approximate randomization approach (Yeh, 2000) was modified to output the significance level, p-value, for the CER metric. The p-levels produced through two rounds of 1048576 shuffles are 0.0058 and 0.0057 respectively, both less than 0.01. Thus the performance improvement imposed by the utilization of the analyzer is statistically significant.

4 Conclusion

Addressing the challenges of Mandarin large vocabulary continuous speech recognition task, within the unified framework of WFSTs, this study presents an analyzer integrating multi-level linguistic knowledge. Unlike other methods, model units, such as characters and words, can be chosen freely in this approach since multi-level knowledge sources are modeled independently. As a consequence, the final analyzer can be derived from the combination of better optimized models based on proper model units. Along with two level knowledge sources, i.e., the person and location names as well as the part-of-speech information, the analyzer is built and evaluated by a comparative simulation. Further evaluation is also conducted on an LVCSR system in which the analyzer is embedded. Experimental results consistently reveal that the approach is effective, and successfully improves the performance of speech recognition by a 9.9% relative reduction of character error rate on the HUB-4 test set. Also, the unified framework based approach provides a property of integrating additional linguistic knowledge flexibly, such as organization name and syntactic structure. Furthermore, the presented approach has a benefit of efficiency that additional efforts are not required for decoding as new knowledge comes, since all knowledge sources are finally encoded into a single WFST.

[2] Acoustic segmentation software downloaded from http://www.nist.gov/speech/tools/CMUseg_05targz.htm.

[3] http://www.people.com.cn

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the National Key Basic Research Program of China (2004CB318005), and the New-Century Training Program Foundation for the Talents by the Ministry of Education of China.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4):337–344.

Issam Bazzi and James R. Glass. 2000a. Modeling out-of-vocabulary words for robust speech recognition. In *Proc. of 6th International Conference on Spoken Language Processing*, pages 401–404, Beijing, China, October.

Issam Bazzi and James Glass. 2000b. Heterogeneous lexical units for automatic speech recognition: preliminary investigations. In *Proc. of ICASSP*, pages 1257–1260, Istanbul, Turkey, June.

Issam Bazzi and James Glass. 2001. Learning units for domain-independent out-of-vocabulary word modelling. In *Proc. of EUROSPEECH*, pages 61–64, Aalborg, Denmark, September.

Eugene Charniak. 2001. Immediate-head parsing for language models. In *Proc. of ACL*, pages 116–123, Toulouse, France, July.

Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenez, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Ristad, Ronald Rosenfeld, Andreas Stolcke, and Dekai Wu. 1997. Structure and performance of a dependency language model. In *Proc. of EUROSPEECH*, pages 2775–2778, Rhodes, Greece.

Ciprian Chelba. 2000. *Exploiting Syntactic Structure for Natural Language Modeling*. Ph.D. thesis, Johns Hopkins University.

Isabel Trancoso Diamantino Caseiro. 2002. Using dynamic WFST composition for recognizing broadcast news. In *Proc. of ICSLP*, pages 1301–1304, Denver, Colorado, USA, September.

George Doddington. 1996. The 1996 hub-4 annotation specification for evaluation of speech recognition on broadcast news. In *ftp://jaguar.ncsl.nist.gov/csr96/h4/h4annot.ps*.

N. Friburger and D. Maurel. 2004. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104.

Lucian Galescu. 2003. Recognition of out-of-vocabulary words with sub-lexical language models. In *Proc. of EUROSPEECH*, pages 249–252, Geneva, Switzerland, September.

Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Jithendra Vepa, and Vincent Wan. 2006. The AMI meeting transcription system: Progress and performance. In *Proc. of Rich Transcription 2006 Spring Meeting Recognition Evaluation*.

Peter A. Heeman. 1998. Pos tagging versus classes in language modeling. In *Proc. of the 6th Workshop on very large corpora*, pages 179–187, Montreal, Canada.

Takaaki Hori, Chiori Hori, Yasuhiro Minami, and Atsushi Nakamura. 2007. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Transactions on audio, speech, and language processing*, 15(4):1352–1365.

Xinhui Hu, Hirofumi Yamamoto, Genichiro Kikui, and Yoshinori Sagisaka. 2006. Language modeling of chinese personal names based on character units for continuous chinese speech recognition. In *Proc. of INTERSPEECH*, pages 249–252, Pittsburgh, USA, September.

Mark Johnson. 2001. Joint and conditional estimation of tagging and parsing models. In *Proc. of ACL*, pages 322 – 329, Toulouse, France.

G. Maltese, P. Bravetti, H. Crépy, B. J. Grainger, M. Herzog, and F. Palou. 2001. Combining word- and class-based language models: A comparative study in several languages using automatic and manual word-clustering techniques. In *Proc. of EUROSPEECH*, pages 21–24, Aalborg, Denmark, September.

Ciro Martins, Antonio Texeira, and Joao Neto. 2006. Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system. In *Proc. of Spoken Language Technology Workshop*, pages 146–149, December.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 1996. Weighted automata in text and speech processing. In *ECAI-96 Workshop*.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88.

Mehrya Mohri. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311.

Christan Raymond, Fre de ric Be chet, Renato D. Mori, and Ge raldine Damnati. 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48(3-4):288–304.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Christer Samuelsson and Wolfgang Reichl. 1999. A class-based language model for large-vocabulary speechrecognition extracted from part-of-speech statistics. In *Proc. of ICASSP*, pages 537–540, Phoenix, Arizona, USA, March.

George Saon, Geoffrey Zweig, Brain KingsBury, Lidia Mangu, and Upendra Canudhari. 2003. An architecture for rapid decoding of large vocabulary conversational speech. In *Proc. of Eurospeech*, pages 1977–1980, Geneva, Switzerland, September.

Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. 1997. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. of DARPA Speech Recognition Workshop*, pages 97–99, Chantilly, Virginia, February.

Daniel Willett Takaaki Hori and Yasuhiro Minami. 2003. Language model adaptation using WFST-based speaking-style translation. In *Proc. of ICASSP*, pages I.228–I.231, Hong Kong, April.

Koichi Tanigaki, Hirofumi Yamamoto, and Yoshinori Sagisaka. 2000. A hierarchical language model incorporating class-dependent word models for oov words recognition. In *Proc. of 6th International Conference on Spoken Language Processing*, pages 123–126, Beijing, China, October.

Hajime Tsukada and Masaaki Nagata. 2004. Efficient decoding for statistical machine translation with a fully expanded WFST model. In *Proc. of EMNLP*, pages 427–433, Barcelona, Spain, July.

Wen Wang and Mary P. Harper. 2002. The superarv language model: investigating the effectiveness of tightly integrating multiple knowledge sources. In *Proc. of EMNLP*, pages 238–247, Philadelphia, USA, July.

Wen Wang and Dimitra Vergyri. 2006. The use of word n-grams and parts of speech for hierarchical cluster language modeling. In *Proc. of ICASSP*, pages 1057–1060, Toulouse, France, May.

Wen Wang, Andreas Stolcke, and Mary P. Harper. 2004. The use of a linguistically motivated language model in conversational speech recognition. In *Proc. of ICASSP*, pages 261–264, Montreal, Canada, May.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proc. of COLING*, pages 947–953, Saarbrücken, August.