INTERSPEECH 2009 BRIGHTON

Exploiting Chinese Character Models to Improve Speech Recognition Performance

J.L. Hieronymus,[1], *X. Liu*[2], *M.J.F. Gales*[2], *P.C. Woodland*[2]

[1]NASA Ames Research Center
Mountain View, CA 94035, USA

[2]Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
[jlh83,xl207,mjfg,pcw]@eng.cam.ac.uk

10.21437/Interspeech.2009-116

Abstract

The Chinese language is based on characters which are syllabic in nature. Since languages have syllabotactic rules which govern the construction of syllables and their allowed sequences, Chinese character sequence models can be used as a first level approximation of allowed syllable sequences. N-gram character sequence models were trained on 4.3 billion characters. Characters are used as a first level recognition unit with multiple pronunciations per character. For comparison the CU-HTK Mandarin word based system was used to recognize words which were then converted to character sequences. The character only system error rates for one best recognition were slightly worse than word based character recognition. However combining the two systems using log-linear combination gives better results than either system separately. An equally weighted combination gave consistent CER gains of 0.1 - 0.2% absolute over the word based standard system.

Index Terms: Mandarin Chinese speech recognition, Mandarin Chinese character modeling, combining Chinese character and word models

1. Introduction

The Chinese language is based on characters which are syllabic in nature and morphological in meaning [1]. There are many characters which have the same pronunciations (homographs). This means that it can be difficult to know what the spoken word is without context or what the character sequence is.

This ambiguity is demonstrated by the practice of signing that Chinese people use to show how to spell their names when they are introduced. By showing how they would write their name in characters on their hand with their fingers, they show how their name is spelled. Otherwise the listener would not know which characters constituted the spelling of the name. Since Chinese characters have a broad meaning in themselves, knowing the character sequence in a name can evoke the underlying meaning.

Written Chinese has no word boundaries marked by spaces or other symbols, so finding the correct word sequence is difficult. The reader must infer the words and word boundaries from the context. In tests of word boundary marking by native speakers, the boundary locations are only agreed on approximately 75% of the time[2, 3].

Since all languages have constrained syllable constructions and syllable sequence rules which enhance intelligibility, it seemed like a good opportunity to use these constraints for Chinese speech recognition explicitly. These constraints are not as restrictive as word sequences, but they provide a different type of information.

In languages like English, the maximum onset principle for spoken syllable construction allows each syllable to have a number of phonetic realizations. Final consonants from the previous syllable can attach to the following syllable, making the onset longer. But these changes must obey English syllable construction rules, so that only limited onset extensions are allowed. When pronouncing the phrase "Up, Up and Away," the second syllable is actually "pup" (/p∧p/ phonetically).

Since Mandarin Chinese has a restricted set of post vocalic consonants (/n/ and /ng/) and no prevocalic consonant clusters, it is reasonable to assume that the maximum onset principle is mostly blocked for Chinese. Thus each syllable should usually have only one segmentation. It is hoped that by combining syllable and word constraints, the resulting speech recognition models will reflect the language better and thus have improved performance.

Because syllable segmented and labeled Chinese speech is generally unavailable in large quantities, character sequences are modeled as an indirect way of modeling syllable sequences. This requires just the usual text and speech data for training.

The character based recognition system has several possible pronunciations per character, expressed as toned phone sequences. The acoustic models are the usual word based phone models used in the HTK Mandarin system. Syllable based acoustic phone models might improve the character recognition performance. Character sequence models are made from 4, 5 and 6-grams based on 4.3 billion characters from a total of 27 text sources. The corpora are listed in Table 1 and are discussed in detail below. These models are used in the character recognition stage, to assure that recognized characters obey the rules of allowed Chinese character (and thus syllable) sequence construction. The result is a character lattice which can be searched for the one best sequence. The lattices can also be examined to test whether the correct characters are present.

A word based system can also be used to construct character sequences, and this is the common method in current Chinese speech recognition systems. Since word sequence models are the most powerful language constraints, word based character sequences recognition has very good performance.

The CU-HTK Mandarin ASR word model based recognition system was trained on the same corpora of 2.8 billion words

which resulted in a 4 gram word sequence model. Once the first best word hypothesis is generated, then words are separated into characters again. The performance of these two systems are compared.

Finally the character and the word based systems are combined using ROVER based hypothesis level, linear or log-linear model level combination to give improved character recognition performance. Two strategies may be considered when applying model level combination schemes. The first starts from a standard word based recognition and lattice generation. This is followed by expanding the resulting word lattices onto subword, i.e. character level prior to a final composition with a character sequence n-gram model to produce the best word sequence. In contrast, the second option starts from character level recognition followed by a transduction of lattices from subword to word level. These lattices are then composed with a word level n-gram sequence model during search. Due to the weaker constraint of character level recognition and the subsequently higher lattice oracle error rate, the first strategy was investigated in this paper. A log-linear model combination was found to yield consistently the lowest character error rate among all the schemes investigated in the paper.

The LIMSI Group (Luo et al)[4] have experimented with a similar character recognition system and found that their word based system gave the best character recognition. Their attempts to perform system combination were not successful in lowering the character error rate.

2. Combining Word and Character Models

When combining a subword based speech recognition system with a word level one, two major categories of techniques may be considered: hypothesis or model level combination methods. Model level combination techniques may be further classified into linear or log-linear combination. In machine learning literature, these are commonly referred to as mixtures of experts (MoE) and products of experts (PoE) [5, 6]. In both cases each expert is a probabilistic distribution. Under the weighted finite state transducers (WFSTs) [8, 9] based framework, where n-gram sequence models are categorized together with many other probabilistic finite state models, MoE and PoE based combination schemes are represented by transducer *union*, and *composition* or *intersection* operations respectively.

Hypothesis level combination: One commonly used form of hypothesis level combination is ROVER [7]. Hypotheses from a total of S component systems are iteratively aligned to create a word transition network first. An interpolated score between voting and confidence measure is then used to find the optimal word sequence within the network. For any set of confusions in the network this is given by,

$$\hat{w} = \arg\max_{w_s} \left\{ \alpha \frac{N_{1:S}(w_s)}{S} + (1-\alpha) c_w^{(s)} \right\} \quad (1)$$

where $N_{1:S}(w_s)$ is number of systems that output word w_s, and $c_w^{(s)}$ the confidence score assigned by the sth system. α is a tunable parameter to balance the contribution between voting and confidence scores. As the outputs from word and character based systems are represented at different linguistic levels, a direct combination between their hypotheses is inappropriate. To handle this issue, the approach adopted in this paper is to perform a character level combination. This requires the outputs from the word based system to mapped to subword, character level. For the Chinese language this process is deterministic. The confidence score of each word is assigned to each character it contains. In general hypothesis level combination method such as ROVER requires the error rate of complimentary components systems have similar error rates in order to be effective in combination.

Linear Model Combination: The linear interpolation based MoE, as a *union* of all the individual experts, tends to to give a broader distribution than individual components alone. Let w_i denote the i^{th} word of a sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$, and $c_{i,j}$ the j^{th} character of word w_i. When combining word and character based language models, the linearly interpolated probability for word w_i is given by

$$\begin{aligned} P_{\mathsf{comb}}(w_i|w_1,...,w_{i-1}) &= \lambda P_{\mathsf{word}}(w_i|w_1,...,w_{i-1}) \quad (2) \\ +(1-\lambda) \prod_{j,c_{i,j}\in w_i} & P_{\mathsf{char}}(c_{i,j}|c_{1,1}, c_{1,2}, ..., c_{i,1}, ..., c_{i,j-1}) \end{aligned}$$

where λ is linear interpolation weight for the word based model. This form of model combination may help overcome the sparsity issue of word based models and thus improve generalization. A linear model level combination may be efficiently implemented using a WFST *union* operation, after the word level n-gram model or lattice is *projected* onto subword level in order to be compatible with the character level transducer symbols.

Log-Linear Model Combination: In contrast, the log-linear interpolation based PoE model provides an *intersection* of individual experts. It typically yields a high likelihood only when all components agree. For the example shown above, the log-linearly interpolated probability for word w_i is

$$\begin{aligned} \ln P_{\mathsf{comb}}(w_i|w_1,...,w_{i-1}) &= \lambda_{\mathsf{w}} \ln P_{\mathsf{word}}(w_i|w_1,...,w_{i-1}) \quad (3) \\ +\lambda_{\mathsf{c}} \sum_{j,c_{i,j}\in w_i} & \ln P_{\mathsf{char}}(c_{i,j}|c_{1,1}, c_{1,2}, ..., c_{i,1}, ..., c_{i,j-1}) \end{aligned}$$

where and λ_{w} and λ_{c} are the log-linear weights for the word and character based models. This form of model combination take the character based model as additional constraints during search. Hypotheses with very different word and character level log-likelihood ranking will be penalized. A log-linear model combination may be efficiently implemented using a WFST *composition* operation between the two transducers that represent word and subword level n-gram sequence models.

The precise nature of the word and subword level models determines which mode of the two model level combination schemes may be appropriate for a combination between the two. If the word based model is too sparse and non robust, a linear model combination may be preferred to improve the combined model's generalization. In contrast, if the word based model lacks of discriminative power, a log-linear combination may be more appropriate as additional subword level linguistic constraints can be introduced in combination. In the following section, both modes of model combination will be investigated.

3. Experiments and Results

The CU-HTK Mandarin ASR system was used to evaluate performance of multi-level language models. The baseline system of word level recognition units was used in an initial lattice generation stage. A 63k word list consists a total of 52k multiple character Chinese words, 6k single character Chinese words and 5k frequent English words. An interpolated 4-gram word based back-off LM and adapted gender dependent cross-word triphone MPE acoustic models were used in decoding. Speech data for acoustic model training consisted of 1673 hours

of broadcast news (BN) and broadcast conversation (BC) data. Acoustic models are trained on word level analyzed transcriptions only. HLDA projected PLP features with CMN normalization and appended pitch parameters were used. A total of 4.3 billion characters from 27 text sources were used in LM training. After a longest first based character to word segmentation, 2.8 billion words of text in total were used train word level n-gram models. Information on corpus size, cut-off settings and smoothing schemes for text sources are given in table 1. The word based model's interpolation weights were perplexity tuned on the combined bn06+bc05+dev07 set. These are shown in the 5th column of table 1. A similar rank ordering of sources weights was also obtained for the character level language model. Due to data sparsity, 5-gram and 6-gram LMs were only built for character level models. Their cut-off settings are shown in brackets. For data sources that are closer in genre to the test data, minimum cut-offs and modified KN smoothing were used. These include two audio transcriptions sources, bcm and bnm, and additional web data from major TV channels such as Phoenix TV and VOA. For the largest corpora of newswire genre and Taiwanese origin, giga-cna, more aggressive cut-offs and Good Turing (GT) discounting were used. Five GALE Mandarin Chinese broadcast speech development sets were used in the experiments: bn06 of 3.4 hours of BN data, bc05 of 2.5 hours of BC data, and three other sets containing data of both genres, 2.6 hour dev07, 1 hour dev08 and 2.6 hour p2ns. Manual audio segmentation was used in decoding.

Comp LM	#Char (M)	#Word (M)	Train Config	Intplt Weight
bcm	14.26	9.21	kn/111(11)	0.260058
bnm	12.29	7.41	kn/111(11)	0.147834
gigaxin	483.65	362.74	kn/112(22)	0.132539
phoenix	144.57	91.38	kn/112(22)	0.107920
gigacna	891.13	604.98	gt/123(33)	0.072665
voarfa	63.54	35.31	kn/112(22)	0.072299
ibmsina2	382.34	253.59	kn/112(22)	0.055601
bbndata	301.39	186.3	kn/112(22)	0.046213
galeweb	556.41	390.8	kn/122(22)	0.045918
agilece	336.78	204.5	kn/112(22)	0.031497
ntdtv	36.44	24.75	kn/112(22)	0.010216
ibmsina1	78.39	51.89	kn/112(22)	0.003814
papersjing	197.75	135.69	kn/112(22)	0.003220
tdt4	2.98	1.76	kn/112(22)	0.003005
tdt2+3	15.87	9.35	kn/112(22)	0.001689
xinhuachina	105.88	76.57	kn/112(22)	0.001587
sriwebconv	163.16	114.6	kn/112(22)	0.001081
gigaafp	40.28	27.24	kn/112(22)	0.000770
cctvcnr	47.31	29.59	kn/112(22)	0.000751
hub4m	0.38	0.22	kn/111(11)	0.000533
chradio	91.55	54.86	kn/112(22)	0.000468
papersyue	52.48	34.14	kn/112(22)	0.000275
gigalhzb	29.16	19.73	kn/112(22)	0.000019
papershu	50.67	34.85	kn/112(22)	0.000012
pdaily	114.51	68.89	kn/112(22)	0.000012
papersning	51.99	33.9	kn/112(22)	0.000006
dongailbo	12.82	8.02	kn/112(22)	0.000000

Table 1: Text source size, 2/3/4-gram cut-off settings, smoothing scheme used in training (5/6-gram cut-offs for character level LMs in brackets), perplexity tuned interpolation weights using the reference of combined bn06+bc05+dev07.

Model sizes of the final interpolated 4-gram word and 6-gram character level LMs are shown in table 2. The total log-probability scores on the combined reference of bn06+bc05+dev07 assigned by these two models are shown in the 6th column of the table. On average the word based system produces approximately 1.5 characters per word. Hence, a 6-gram character based model would be appropriate to compare with a 4-gram word baseline. As expected, with a stronger constraint, the word based model gave a better likelihood than the character based one. The perplexity metric is also commonly used to measure the predictive power of LMs. However, as these two LMs considered here model linguistic units at very different level, a direct comparison between word and subword level perplexity scores is not meaningful. One possible solution is to approximate the subword, or character, level perplexity for the word based model. The number of subword units, instead of the word level sequence length, is used in perplexity computation. This approximated character level perplexity score is in the last column of the first line for the word based system. Consistent with the trend observed on log-likelihood, the word based model also has a lower perplexity by 9% relative.

LM	Model Size(M) 2g	3g	4g	5g	6g	Log Prob	Char PPlex
word	60	228	56	-	-	-511524	25.61
char	10	148	111	130	122	-524473	27.91

Table 2: Model sizes of word and character level LMs, their total log-probability and character level perplexity scores on the combined reference of bn06+bc05+dev07.

A similar error rate performance difference between the 4-gram word baseline and the 6-gram character level model can also be found in the 2nd and 5th line of table 3. The word base system outperformed the character based one by 0.4%-1.5% absolute (6% to 12% rel) across all five test sets. CER performance of various others systems on bn06, bc05, dev07, dev08 and p2ns are shown in table 3. Performance of the 3-gram word based model, the 4-gram and 5-gram character based models are shown in the 1st, 3rd and 4th lines of the table. For the word based model, increasing the n-gram context length from 3-gram to 4-gram gave further CER improvements 0.1%-0.4% absolute (1.0% to 4.0% rel). In contrast, the relative gains on the character based system, for example, between 5-gram to 6-gram models, are only 1.0% relative.

In Table 3 the 1-best error rates of both type of systems are presented. It is also interesting to investigate the lattice oracle error rates for either the word or character level models. These are shown in table 4. Overall and the word based system produced lower oracle lattice error rates, apart from bn06. This is expected as the bn06 set contains more name entity words than other test sets. On this data set, word based LMs trained on texts segmented using a simple left-to-right longest character to word tokenization scheme would be affected more by this issue.

The rest of table 3 shows performance systems using various methods combine information from the word and character based systems. Performance of three ROVER systems between the 4-gram word based system's and various character based ones are shown in the 3rd section of table 3. The best ROVER configuration is between the 4-gram word and 6-gram character based systems. It outperformed the 4-gram word based system by 0.1% on bn06 but also degraded the error rate by 0.1%-0.4% absolute for the other four test sets. As previously dis-

System	CER%				
	bn06	bc05	dev07	dev08	p2ns
w.3g	7.3	16.6	10.0	10.0	9.8
w.4g	7.2	16.4	9.8	9.6	9.6
c.4g	8.0	18.6	11.5	10.4	10.8
c.5g	7.8	18.1	11.1	10.3	10.6
c.6g	7.6	17.9	10.9	10.3	10.5
w.4g $\oplus$ c.4g	7.2	16.8	10.4	9.8	10.0
w.4g $\oplus$ c.5g	7.1	16.5	10.2	9.7	9.9
w.4g $\oplus$ c.6g	7.1	16.5	10.2	9.6	9.8
w.4g $\circ$ c.4g	7.1	16.4	9.7	9.4	9.5
w.4g $\circ$ c.5g	7.1	16.3	9.7	9.4	9.4
w.4g $\circ$ c.6g	7.1	16.3	9.7	9.4	9.4

Table 3: 1-best CER performance of various LMs on bn06, bc05, dev07, dev08 and p2ns. "$\oplus$" denotes hypothesis level ROVER and "$\circ$" finite state grammar composition operations.

cussed, a hypothesis level combination method such as ROVER requires the error rate of complimentary components systems are in close range in order to effective in combination. However, this precondition is not satisfied given the significant performance difference between the word and character based systems of table 3. Furthermore, as there are only two component systems used in ROVER, the combination decision is purely based on confidence scores as voting now has no effect. Poor confidence scores generated by the character based systems can introduce additional errors in combination.

LM	Lattice Oracle CER%				
	bn06	bc05	dev07	dev08	p2ns
word	2.08	5.31	1.71	1.70	1.89
char	2.01	5.72	2.04	1.89	2.02

Table 4: Oracle character level error rate for lattices generated using word or character level LMs on bn06, bc05, dev07, dev08 and p2ns.

Finally it is important to investigate the character recognition performance obtained by using a model level combination between the word and character systems. Given the lattice oracle error rates given in table 4, the lattices produced by the word based LM are used in later rescoring stages using combined multi-level LMs. As discussed previously, a mixture of experts (MoE) model using a linear interpolation, or equivalently a WFST union operation between word and subword n-gram models, tends to broaden the underlying statistical distribution and improve generalization of the combined models. In contrast, a product of experts (PoE) model using a log-linear interpolation, or equivalently an intersection and composition operation, tends to sharpen the underlying distribution and increase its power of discrimination. The precise nature of the subword level model's distribution determines which mode of the two is the appropriate form to use in combination. As character sequence models offer additional subword level constraint in the search, it is expected that a log-linear, rather than linear, interpolation would be more suitable for Chinese speech recognition. This was confirmed by the error rate performance of using a linear model combination, which was consistently outperformed by the standard word based LM. Performing exhaustive tuning of linear interpolation weights on dev08 between the 4-gram word and 6-gram character level models showed the best linear weighting is to use the word based system's probability only. In contrast, an equally weighted log-linear interpolation between the 4-gram word and character based models gave consistent CER gains of 0.1%-0.2% on all test sets over the word based standard system. These are shown in the bottom section of table 3. These results suggest including a character level model in decoding provides additional subword level linguistic constraints and increased discrimination.

4. Conclusion

Character and word level modeling for Chinese produces similar character error rates which differ in the details of the errors produced. Combining these two systems using log-linear combination gives better performance than either of the systems separately. An equally weighted log-linear combination gave consistent CER gains of 0.1 - 0.2% absolute over the standard word based system on a state-of-the-art Chinese broadcast speech recognition task.

5. Acknowledgements

Discussions with Richard Sproat about Chinese character pronunciation and his provision of a character pronunciation dictionary is gratefully acknowledged.

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

6. References

[1] de Francis, J., *The Chinese Language*, University of Hawaii Press, Honolulu, HI., 1994.

[2] Sproat, R., Shih, C., Gale, W. and Chang, N., "A Stocastic Finite-State Word-Segmentation Algorithm for Chinese," Computational Linguistics, 22(3):218-228, 1996.

[3] Wu, D. and Fung, P.,"Improving Chinese tokenization with linguistic filters on statistical lexical acquisition." Proceedings of the Fourth Conference on Applied Natural Language Processing, pp. 180-181, Stuttgart, October, 1994.

[4] Luo, J., Lamel, L., Gauvain, J-L.,"Modeling Characters versus Words for Mandarin Speech Recognition." Proc. ICASSP09, Taipei, Taiwan, 2009.

[5] G. Hinton. Products of Experts, in *Proc. ICANN*, 1999.

[6] G. Hinton. Training Products of Experts by Minimizing Contrastive Divergence, *Neural Computation*, 14:1771–1800, 2002.

[7] J. G. Fiscus (1997). A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proc. IEEE ASRU'97*.

[8] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23:2, 1997.

[9] M. Mohri & M. Riley. Network optimizations for large vocabulary speech recognition. *Speech Communication*, 25:3, 1998.