MODELING CHARACTERS VERSUS WORDS FOR MANDARIN SPEECH RECOGNITION

*Jun Luo, Lori Lamel, Jean-Luc Gauvain**

Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
{luo,lamel,gauvain}@limsi.fr http://www.limsi.fr/tlp

ABSTRACT

Word based models are widely used in speech recognition since they typically perform well. However, the question of whether it is better to use a word-based or a character-based model warrants being for the Mandarin Chinese language. Since Chinese is written without any spaces or word delimiters, a word segmentation algorithm is applied in a pre-processing step prior to training a word-based language model. Chinese characters carry meaning and speakers are free to combine characters to construct new words. This suggests that character information can also be useful in communication. This paper explores both word-based and character-based models, and their complementarity. Although word-based modeling is found to outperform character-based modeling, increasing the vocabulary size from 56k to 160k words did not lead to a gain in performance. Results are reported for the Gale Mandarin speech-to-text task.

Index Terms: Speech recognition, language modeling, Mandarin Chinese, speech-to-text transcription

1. INTRODUCTION

Language models are an important component of state-of-the-art speech recognition systems. Almost all recognition systems make use of word-based N-gram language models, which have been shown to be effective for a variety of languages [1]. One of the challenges in speech recognition is going beyond simple n-gram language models to better suit the characteristics of a specific language. The question of whether it is better to use a word-based or a character-based language model warrants being asked when addressing the Mandarin Chinese language. There are two reasons for this. First, there is no standard definition of a word in Chinese and there are no specific word separators, i.e. words are not delimited by blank spaces. A word segmentation algorithm is therefore required, and is applied in pre-processing step before training a word-based language model. The quality of word language models for Chinese is therefore highly dependent on this text pre-processing procedure which is used both for word selection and to segment the character sequence into words. According to Sproat et al. [2] and Wu and Fung [3], there is only about a 75% agreement between native speakers as to what is the "correct" segmentation. This lack of agreement between humans makes segmentation a difficult and ill-defined task.

Second, each Chinese character represents a syllable and has a corresponding meaning. It is possible to construct new words by combining multiple characters. Native Chinese speakers typically do not have any difficulty in understanding the new words, and can recognize them in texts. Such new words cause serious out-of-vocabulary (OOV) problems for dictionary-based segmentation methods and can dramatically affect their accuracy. Character based modeling for Mandarin speech recognition is interesting to explore at both the language and acoustic model levels. At the language model level, it is not very clear if there exists a "best" definition of words in Chinese, thus the use of different units, i.e, word and character is potentially interesting to examine. At the acoustic level, each character represents a syllable which can provide a natural acoustic context.

Compared with word based language modeling, modeling characters has the following benefits:

- It eliminates the pre-processing procedure needed to select words and to segment the text into word sequences;
- The vocabulary is much smaller than a word vocabulary, which makes it possible to train higher order models;
- It does not suffer from an out-of-vocabulary (OOV) problem since all characters are known.

It should be emphasized that these benefits do not imply that character based language models would be a better choice than word based models for speech recognition. In fact, most word based language model also include all Chinese characters thus eliminating the OOV problem. In [4] different size word and word/character language models were compared on a Mandarin broadcast news transcription

*This work was in part supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, and in part under the research programme Quaero, funded by OSEO, the French State agency for innovation.

Group	*Source*	*# Words (M)*	*# Chars (M)*
Audio Transcripts	BC	4.8	7.4
	BN	3.8	6.3
	Hub4 man	0.2	0.4
	TDT2+3	8.0	13.5
	TDT4	1.8	3.0
LDC	China Radio	54.0	90.6
	Giga XinHua	245.4	416.6
	People's Daily	68.7	114.4
	Giga CNA	405.3	652.6
	Giga ZaoBao	16.7	26.3
CU Webtexts	CCTV CNR	26.6	43.7
	VOA RFA	31.5	57.0
	NTDTV	11.6	19.4
	XinHua China	33.0	53.3
	Papers Jing	83.2	133.0
	Papers Ning	28.8	45.7
	Papers Hu	21.1	33.6
	Papers Yue	27.2	43.7
	DongA I1 Bo	7.9	12.7
Phoenix TV	Phoenix	76.5	121.5

Table 1: *Summary of the various Mandarin LM text sources.*

task, demonstrating that including only the 17K most frequent words in the word list and the splitting the remaining into characters resulted in recognition performance close to that of the best system.

Most word-based systems make use of word position-dependent (PD) acoustic models. In a character based system, the position relates to not the word but to where in the character the phone is located. These result in different acoustic units from word position-dependent models.

In this paper both character-based and word-based models are explored, as well as their complementarity. The comparison is performed at two levels: at the language model level, since on average there are 1.5 characters per word, a 4-gram word based language model is compared with higher order character language model (6-gram); at the acoustic model, position dependent triphone models are built both for character and word with the different definition of modeled units. Combining word-based and character-based systems is explored using the ROVER technique [5]. Since word selection is an important aspect of the segmentation procedure, increasing the recognition vocabulary was also explored. Experimental results are reported on the DARPA GALE Mandarin broadcast data.

2. USING CHARACTER VS. WORD UNITS FOR LANGUAGE MODELING

The text resources for language model training come from the following 4 sources:

- Transcriptions of audio training data in the GALE Y1, P2R[1-3] releases and prior LDC releases;
- LDC distributed Chinese Gigaword corpus;
- Cambridge University downloaded web texts;
- SRI Phoenix web downloads.

	Perplexity	$-\log Likelihood$
Word LM (4-gram)	207.997	149352
Char. LM (6-gram)	28.611	152198

Table 2: *Perplexity of the* $bcmdev05 + bnmdev06$ *devdata with the 4-gram word LM and the 6-gram character LM.*

The longest match algorithm [6] was used to segment the characters into Chinese words. This approach is a dictionary based method, which segments sentences by matching the longest entries in the dictionary. Since the focus is on developing Mandarin Chinese language models, any non-Chinese characters remaining after normalization are removed from each source. A 54K vocabulary is used for word language model and 6K character list is used for character model. It should be noted that the 6K characters are also included in the 54K word list.

The distribution of text sources and the number of words and characters for each individual source is given in Table 1. Individual LMs are first generated for each text corpus, and then interpolated to generate the final language model. The interpolation weights were optimized on a development data set (bcmdev05+bnmdev06) defined by Cambridge University, The modified Kneser-Ney [7] discounting is used for language model parameter estimation. Both word based and character based language models are created in this way.

The perplexities and log likelihoods of the development data with the word 4-gram and character 6-gram language models are reported in Table 2. According to the log Likelihood, the word 4-gram is only slightly better than character 6-gram, which does not account for the difference in character error rate (CER) based on these two LMs (see Section 5).

3. CHARACTER VS. WORD POSITION DEPENDENT ACOUSTIC MODELING

The phone context is important when building acoustic models, each phone-in-context having corresponding acoustic model. When cross-word triphone models are used, the phone context also specifies the word position of the phone. The phone position can be word internal, word initial, word final or monophone. If characters are used in language modeling, it can make sense to also develop acoustic models that make use of character (rather than word) position. A character in Mandarin corresponds to one syllable and is more or less independent from the surrounding characters. The number of modeled contexts is greatly reduced when using character position dependent models rather than word position dependent ones (to 35868 from. 46912). Both model sets give similar log-likelihoods in training.

4. SELECTING A LARGE VOCABULARY

The word selection is an important process in any word based recognizer. Generally speaking, when large text corpora are available for language model estimation, as is the

case for Mandarin, it has been found that increasing the recognition vocabulary, increases performance by reducing errors engendered by out-of-vocabulary words. Mandarin Chinese is a bit of special case, in that OOV words are not very problematic since any character sequence not represented as a word is simply split into constituent characters. At the same time since the word list is used for segmentation, the segmentation algorithm and word list have been shown to affect recognition accuracy.

We therefore posed the question of if it is of interest to include the frequent words in the recognition word list, increasing the vocabulary size. Additional vocabulary items were selected from a 200K list of words appearing frequently on the Internet. The training texts were segmented using this 200K word list, and any words not seen in the training data were discarded. This resulted in a 160K word list. The corresponding pronunciations were automatically generated by concatenating all possible character pronunciations, and the words with too many possible pronunciations (> 40) were manually corrected.

5. EXPERIMENTAL RESULTS

Modeling Words vs. Characters

The first experiments are carried out on the Broadcast News (BN) portion of the 2007 development data from the DARPA GALE program. The development set, bnmdev07, contains 1.1 hours of speech extracted from 40 shows. Up-to-date results are also given on the GALE dev07 data containing 2.4 hours of broadcast news and conversation data.

All acoustic models are sets of 3-state left-to-right hidden Markov models with Gaussian mixture. For the word models, 46.9k phone contexts are covered and for the character models 35.9k phone contexts are covered. Both model sets contain 11.5k tied states, which 32 Gaussians per state with 2048 Gaussians for silence. Gender-dependent triphones are estimated using MAP adaptation of gender independent seed models, which are trained on 891 hours (468h female, 423h male) of data.

The first model set (AM1) is context-dependent, position-independent non-pitch model that can thus be used both with Character and Word language models[1]. Since Mandarin is a tonal language and it is helpful to incorporate pitch feature into the front-end for recognition, the second model set (AM2) uses the same configuration as AM1 except pitch features are included to improve the performance. The final two sets of acoustic models are either word or character position dependent (AM3).

A first experiment compares the difference in language models. The position-independent acoustic model is used to makes it possible to use exactly the same acoustic model with both the word language model and character language model. The difference only lies in the language model level. A three-pass decoding is used in this experiment. For the first pass decoding adaptation is performed using a speech GMM. There is a 2-class MLLR adaptation in the second pass, whereas the third pass uses decision tree based MLLR adaptation and generates the final outputs. In each pass a word/character lattice is first generated with a 2-gram LM, and then rescored with a word 4-gram or a character 6-gram to find the best hypothesis respectively.

AM Set	*Word LM*	*Char. LM*	
		app. 1	app. 2
AM1	7.83%	9.77%	8.13%
+pitch,AM2	5.09%	6.86%	5.34%

Table 3: CER using word/character position-independent acoustic models with different language models.

There are two possible approaches to using the character language model. The first approach (app.1) is to use character language model both for the generation and re-scoring of lattices. In this case the character lattice is generated by character 2-gram and then pruned using 3-gram/4-gram/5-gram in turn. Finally a 6-gram character language model is used to generate the best hypothesis. As can be seen in Table 3, the difference of CERs is quite large, with the character based system having a CER about 20% worse (relative) than the word based one.

Concerned that the context in a character 2-gram is too limited and generating lattices with it might remove too many useful hypotheses, the second approach (app.2) uses a word bigram model to generate the word lattices, as for the word based approach, but preserving character boundaries. The word lattice is then transformed to a character lattice and pruned/re-scored via character LMs. The character language model is only used in re-scoring stage to find the best hypothesis. The character system is seen to improve but still performs worse than word system.

The quality of the lattices generated by different order character language models as well as word 2-gram is compared in Table 4. The 1-best CER and the lattice CER, are given as performance measures. When using a 4-gram character LM, the 1-best CER can be seen to approach that of using 2-gram word LM, however, the lattice CER is much worse with character LM, which could explain why the word language model always does better.

The difference at acoustic level was explored using the third set of acoustic models, which are word or character position dependent. Results are given in Table 5 for each model set alone, and for the result of applying confidence based ROVER to the hypotheses. The character based model performs substantially less well than the word based one, and due to the large difference in CERs between these systems, the ROVER result is also less good than word based system.

[1]The transcripts for the training data are word based, which means during training the silence could only be inserted between words. But silence could be inserted between each character if we are using character based language model for recognition.

Config.	ID	*LM*	*1-best CER*	*lat CER*	size
un-pruned	1	c2g	13.33%	2.67%	27.6M
	2	c3g	7.35%	2.88%	11.7M
	3	c4g	6.70%	3.02%	10.6M
	4	w2g	6.39&	1.69%	8.8M
pruned	1	c3g	7.57%	2.96%	13.8M
	2	c3g	7.35%	2.96%	9.5M
	3	c4g	6.70%	3.05%	9.0M
	4	w2g	6.39%	1.76%	6.2M

Table 4: Measure of lattice quality. "LM" is the language model used to generate the lattice; c2g/c3g/c4g corresponds to character 2-gram/3-gram/4-gram and w2g to the word 2-gram; "un-pruned": results with the un-pruned lattice, and in the "pruned" case the "LM" is used to prune the lattice obtained with original LM.

AM Set	*Word*	*Char.*	*ROVER*
AM3	5.07%	8.99%	5.36%

Table 5: CER with word/character position dependent AMs and using ROVER to combine system outputs.

The word position context is seen to have essentially no effect on the performance for word based system (5.09% vs 5.07%). This is not the case for character based system, where using the character position context trained on character-based transcripts is quite a bit worse than the position independent models that were trained on word transcripts. An experiment was carried out to test the hypothesis that the increase in CER is due to silence being optionally inserted between characters during training. Character position dependent AMs were trained using segmentations converted from word transcripts so that intraword silences were not allowed. These models also performed less well than character position independent models, so it appears that the character position information does not help.

Increasing the Vocabulary Size and Other Experiments

Several experiments were carried out on the DARPA GALE dev07 data set. The acoustic models are word position dependent models trained on 1000 hours of data (11.5k tied states). A 56K word list (2K words were added to the previous 54K list), was compared to a 160K word list selected as described above. The same text sources [2] are used to train the language models. As can be seen in Table 6, with a single decoding pass, the 56K LM is seen to outperform the larger 160K LM which may be due to a variety of factors: the text segmentation during acoustic model training which also means that there are no pronunciation probabilities for these words, or to poor estimates of the extra words included in the word list. Therefore the remaining experiments were carried out with the 56k LM. Although not described in this paper, acoustic models were trained on discriminative features obtained from a multi layer perceptron [8], result in a CER of 11.2%, and rescoring this output by a neural network LM [9] interpolated with a 4-gram backoff LM reduces the CER to 10.8%.

[2] But different character to word segmentations since this depends on the word list.

Word List	*CER 1-pass*	*2-pass*
160K	14.6%	–
56K	14.0%	12.3%
56K + MLP	–	11.2%
56K + MLP + NNLM	–	10.8%

Table 6: CER on dev07 data set for alternate models.

6. ACKNOWLEDGMENTS

The authors thank Petr Fousek for training the MLP network.

7. CONCLUSIONS

This paper has compared character based and word based modeling for Mandarin Chinese speech recognition at the acoustic, lexical and language model level. Explicit modeling of words, as is the common convention even though results are measured at the character level, was found to significantly outperform modeling characters, even when higher order character n-grams are applied. Different acoustic model configurations with comparable sizes were considered, taking into account or not the phone position in the word or character. Although our original hypothesis was that words and characters can carry complementary information, no gain was observed while combining the outputs probably due to the large difference of CERs. Despite the better performance of the word-based system over the character-based system, increasing the vocabulary size led to a degradation of performance. This indicates that the interaction between word selection and segmentation is still an important problem for Mandarin speech recognition.

REFERENCES

[1] T. Schultz, K. Kirchhoff, *Multilingual Speech Processing*, Academic Press, 2006.

[2] R. Sproat, S. Chilin, W. Galw, N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," *Computational Linguistics*, **22**(3):218–228, 1996.

[3] D. Wu, P. Fung, "Improving Chinese tokenization with linguistic filters on statistical lexical acquisition," *ANLP-94*, 180–181.

[4] L. Chen, L. Lamel, G. Adda, J.L. Gauvain, "Broadcast news transcription in Mandarin," in *ICSLP'2000*, **II**:1015–1018.

[5] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," 1997.

[6] K.S. Cheng, G.H. Young, K.F. Wong, "Study on word-based and integral-bit Chinese text compression algorithms," *J. Am. Soc. Information Science*, **50**:218–228, 1999.

[7] S.F. Chen, J. Goodman, "An empirical study of smoothing techniques for language modeling," *34th Mtg ACL*, 310–318, 1996.

[8] F. Grezl, P. Fousek, " Optimizing Bottle-Neck Features for LVCSR," *ICASSP'08*, Las Vegas, 2008.

[9] Holger Schwenk, "Continuous space language models," *Computer Speech and Language*, **21**:492–518, 2007.