Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 19-34

Reduced *N*-Grams for Chinese Evaluation

Le Quan Ha[*], R. Seymour[+], P. Hanna[*] and F. J. Smith[*]

Abstract

Theoretically, an improvement in a language model occurs as the size of the *n*-grams increases from 3 to 5 or higher. As the *n*-gram size increases, the number of parameters and calculations, and the storage requirement increase very rapidly if we attempt to store all possible combinations of *n*-grams. To avoid these problems, the reduced *n*-grams' approach previously developed by O' Boyle and Smith [1993] can be applied. A reduced *n*-gram language model, called a reduced model, can efficiently store an entire corpus's phrase-history length within feasible storage limits. Another advantage of reduced *n*-grams is that they usually are semantically complete. In our experiments, the reduced *n*-gram creation method or the O' Boyle-Smith reduced *n*-gram algorithm was applied to a large Chinese corpus. The Chinese reduced *n*-gram Zipf curves are presented here and compared with previously obtained conventional Chinese *n*-grams. The Chinese reduced model reduced perplexity by 8.74% and the language model size by a factor of 11.49. This paper is the first attempt to model Chinese reduced *n*-grams, and may provide important insights for Chinese linguistic research.

Keywords: Reduced *n*-grams, reduced *n*-gram algorithm / identification, reduced model, Chinese reduced *n*-grams, Chinese reduced model

1. Introduction to the Reduced *N*-Gram Approach

P O' Boyle and F J Smith [1992, 1993] proposed a statistical method to improve language models based on the removal of overlapping phrases.

The distortion of phrase frequencies were first observed in the Vodis Corpus when the bigram "RAIL ENQUIRIES" and its super-phrase "BRITISH RAIL ENQUIRIES" were examined and reported by O' Boyle. Both occur 73 times, which is a large number for such a small corpus. "ENQUIRIES" follows "RAIL" with a very high probability when it is preceded by "BRITISH." However, when "RAIL" is preceded by words other than "BRITISH," "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the

[*] Computer Science School, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK.
Email: {q.le, p.hanna, fj.smith}@qub.ac.uk
[+] Email: rowan@rowan.ws

bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it. At the time of their research, Smith and O' Boyle reduced the frequencies of "RAIL ENQUIRIES" by using the frequency of the larger trigram, which gave a probability of zero for "ENQUIRIES" following "RAIL" if it was not preceded by "BRITISH." This problem happens not only with word-token corpora but also corpora in which all the compounds are tagged as a unit since overlapping *n*-grams still appear.

Therefore, a phrase can occur in a corpus as a reduced *n*-gram in some places and as part of a larger reduced *n*-gram in other places. In a reduced model, the occurrence of an *n*-gram is not counted when it is a part of a larger reduced *n*-gram. One algorithm to detect/identify/extract reduced *n*-grams from a corpus is the so-called reduced *n*-gram algorithm. In 1992, P O' Boyle and F J Smith were able to store the entire content of the Brown corpus of American English [Francis and Kucera 1964] (of one million word tokens, whose longest phrase-length is 22), which was a considerable improvement at the time. There was no additional way for O' Boyle to evaluate the reduced *n*-grams, so his work was incomplete. We have developed and present here our perplexity method, and we discuss its usefulness for reducing *n*-gram perplexity.

2. Similar Approaches and Capability

Recent progress in variable *n*-gram language modeling has provided an efficient representation of *n*-gram models and made the training of higher order *n*-grams possible. Compared to variable *n*-grams, class-based language models are more often used to reduce the size of a language model, but this typically leads to recognition performance degradation. Classes can alternatively be used to smooth a language model or provide back-off estimates, which have led to small performance gains but also an increase in language model size.

For the LOB corpus, the varigram model obtained 11.3% higher perplexity in comparison with the word-trigram model [Niesler and Woodland 1996], but it also obtained a 22-fold complexity decrease.

Reinhard Kneser [1996] built up variable-context length language models based on North American Business News (NAB - 240 million words of newspaper data) and the German Verbmobil (300,000 words with a vocabulary of 5,000 types). His results show that the variable-length model outperforms conventional models of the same size, and if a moderate loss in performance is acceptable, that the size of a language model can be reduced drastically by using his pruning algorithm. Kneser's results improve with longer contexts and the same number of parameters. For example, reducing the size of the standard NAB trigram model by a factor of 3 results in a loss of only 7% in perplexity and 3% in the word error rate.

The improvement obtained by Kneser's method depends on the length of the fixed context and on the amount of available training data. In the case of the NAB corpus, the improvement was 10% in perplexity.

M. Siu and M. Ostendorf [2000] developed Kneser's basic ideas further and applied the variable 4-gram, thus improving the perplexity and word error rate results compared to a fixed trigram model. The obtained word error reductions of 0.1 and 0.5% (absolute) in development and evaluation test sets, respectively, were not statistically significant. However, the number of parameters was reduced by 60%. By using the variable 4-gram, they were able to model a longer history while reducing the size of the model by more than 50%, compared to a regular trigram model, and at the same time improve both the test-set perplexity and recognition performance. They also reduced the size of the model by an additional 8%.

Another related work was that of Hu, Turin, Brown [1997].

2.1 The first algorithm [R Kneser 1996]

Variable-length models are determined by the set S of word sequences. If T is the set of all word sequences in the training data with a maximal length of M, then variable-length models can be created by finding a suitable subset S of the set T of all the M-gram sequences in the training data with a given maximal context length M. The distance measure between model P_S and model P_T is as follows:

$$D_2(P_T \parallel P_S) := \sum_{k=0}^{M-1} \sum_{(h_k, w) \in T \setminus S} d_1(h_k, w), \tag{1}$$

where the terms of the sum are defined by the average Kullback Leiber distance

$$d_1(h_k, w) := P_T(h_k, w) \log \frac{P_T(w \mid h_k)}{\gamma_T(h_k) P_T(w \mid h_{k-1})}, \tag{2}$$

where h_k is a phrase history of word w and γ is the normalisation factor.

In the implementation, they store the word sequences of S in a tree structure. Each node of the tree corresponds to a word sequence, and each arc is labeled with a word identity. For each node $W = (h_k, w) \in S$, $Succ(W)$ is the set of all longer word sequences starting with the same words as W. If a node W is removed, then all $Succ(W)$ will be removed.

Therefore, the average contribution to the sum d_2 is

$$d_2(W) = \frac{d_1(w) + \sum_{V \in Succ(W)} d_1(V)}{1 + |Succ(W)|}. \tag{3}$$

The pruning algorithm is as follows:

Start with $S = T$

While $(|S| > K)$

For all nodes in S calculate d_2

Remove node with lowest d_2

2.2 The second algorithm [T R Niesler and P C Woodland 1996]

1. *Initialisation:* $L = -1$
2. $L = L + 1$
3. *Grow:* Add level #L to level #$(L-1)$ by adding all the $(L+1)$-Grams occurring in the training set for which the L-Grams already exist in the tree.
4. *Prune:* For every (newly created) leaf in level #L, apply a quality criterion and discard the leaf if it fails.
5. *Termination:* If there is a nonzero number of leaves remaining in level #L, goto step 2.

The quality criterion checks for improvement in the leaving-one-out cross-validation training set likelihood achieved by the addition of each leaf.

2.3 Combination of variable *n*-grams and other language model types

Using the first algorithm, M Siu and M Ostendorf [2000] combined their variable *n*-gram method with the skipping distance method and class-based method in a study on the Switchboard corpus, consisting of 2 million words. In 1996, using the second algorithm, T R Niesler and P C Woodland developed the variable *n*-gram based category in a study on LOB, consisting of 1 million English words. In order to obtain an overview of variable *n*-grams, we combine all of these authors' results in Table 1.

3. O' Boyle and Smith's Reduced *N*-Gram Algorithm and Application Scope

The main goal of this algorithm is to produce three main files from the training text:

- The file that contains all the complete *n*-grams appearing at least *m* times is called the PHR file ($m \geq 2$).
- The file that contains all the *n*-grams appearing as sub-phrases, following the removal of the first word from any other complete *n*-gram in the PHR file, is called the SUB file.

Table 1. Comparison of combinations of variable n-grams and other Language Models

COMBINATION OF LANGUAGE MODEL TYPES								
Basic *n*-gram	**Variable *n*-grams**	**Category**	**Skipping distance**	**Classes**	**#params**	**Perpl-exity**	**Size**	**Source**
Trigram √					987k	474	1M	LOB
		Bigram √			-	603.2		
		Trigram √			-	544.1		
	√	√			-	534.1		
Trigram √					743k	81.5	2M	Switch board Corpus
	Trigram √				379k	78.1		
	Trigram √		√		363k	78.0		
	Trigram √		√	√	338k	77.7		
	4-gram √				580k	108		
	4-gram √		√		577k	108		
	4-gram √		√	√	536k	107		
	5-gram √				383k	77.5		
	5-gram √		√		381k	77.4		
	5-gram √		√	√	359k	77.2		

- The file that contains any overlapping *n*-grams that occur at least *m* times in the SUB file is called the LOS file.

Therefore, the final result is the FIN file of all reduced *n*-grams, where

$$\textbf{FIN} := \textbf{PHR} + \textbf{LOS} - \textbf{SUB}. \tag{4}$$

Before O' Boyle and Smith's work, Craig used a loop algorithm that was equivalent to FIN := PHR – SUB. This yields negative frequencies for resulting *n*-grams with overlapping, hence the need for the LOS file.

There are 2 additional files:

1. To create the PHR file, a SOR file is needed that contains all the complete *n*-grams regardless of *m* (the SOR file is the PHR file in the special case where $m = 1$). To create the PHR file, words are removed from the right-hand side of each SOR phrase in the SOR file until the resultant phrase appears at least *m* times (if the phrase already occurs more than *m* times, no words will be removed).

2. To create the LOS file, O' Boyle and Smith applied a POS file: for any SUB phrase, if one word can be added back on the right-hand side (previously removed when the PHR file was created from the SOR file), then one POS phrase will exist as the added phrase. Thus, if any POS phrase appears at least *m* times, its original SUB phrase will be an overlapping *n*-gram in the LOS file.

The application scope of O' Boyle and Smith 's reduced *n*-gram algorithm is limited to small corpora, such as the Brown corpus (American English) of 1 million words [1992], in which the longest phrase has 22 words. Now their algorithm, re-checked by us, still works for medium size and large corpora with training sizes of 100 million word tokens.

4. Reduced *N*-Grams and Zipf's Law

By re-applying O'Boyle and Smith's algorithm, we obtained reduced *n*-grams from the Chinese TREC corpus of the Linguistic Data Consortium[1], catalog no. LDC2000T52. TREC was collected from full articles in the People's Daily Newspaper from 01/1991 to 12/1993 and from Xinhua News Agency articles from 04/1994 to 09/1995. Originally, TREC had 19,546,872 syllable tokens but only 6,300 syllable types. Ha, Sicilia-Garcia, Ming and Smith [2002] proposed an extension of Zipf 's law and applied it to the TREC syllable corpus. Then in 2003, they produced a compound word version of TREC with 50,000 types, this version was employed in our study for reduced *n*-gram creation.

We will next present the Zipf curves for Chinese reduced *n*-grams, starting with syllables.

4.1 Chinese syllables

The TREC syllable reduced *n*-grams were created in 28 hours on a Pentium II with 512 MB of RAM and 2 GB of free hard-drive space.

The most common TREC syllable reduced unigrams, bigrams, trigrams, 4-grams and 5-grams are shown in Table 3. It can be seen that much noise existed in the unigram frequency observations when only one syllable "年 YEAR" re-appeared in the top ten syllable unigrams [Ha, Sicilia-Garcia, Ming and Smith 2002], listed in Table 2.

[1] http://www.ldc.upenn.edu/

Table 2. The 10-highest frequency unigrams in the conventional Chinese TREC syllable corpus [Ha, Sicilia-Garcia, Ming and Smith 2002]

Rank	Unigrams		
	Freq	**Token**	**Meaning**
1	620,619	的	Of
2	308,826	国	State
3	219,543	一	One
4	209,497	中	Centre / Middle
5	176,905	在	In / At
6	159,861	和	And
7	143,359	人	Human
8	139,713	了	Perfective Marker
9	133,696	会	Get Together / Meeting / Association
10	128,805	年	Year

The Zipf [1949] curves are plotted for TREC syllable reduced unigrams and *n*-grams in Figure 1. It can be seen that none of the syllable unigram, bigram, trigram, 4-gram and 5-gram curves are straight. The unigram curve has an average slope of –1, while the bigram, trigram, 4-gram and 5-gram curves have slopes of around –0.5. At the beginning, they are very turbulent, crossing each other due to much observed noise at high frequencies.

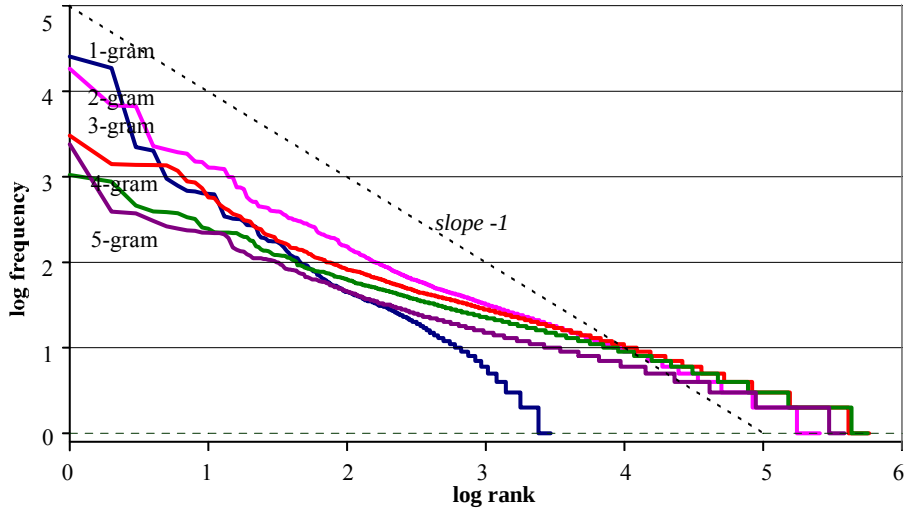

Figure 1. TREC syllable reduced n-gram Zipf curves

Table 3. Most common TREC syllable reduced n-grams

Rank	Unigrams			Bigrams			Trigrams			4-grams			5-grams		
	Freq	**Tokens**	**Meaning**	**Freq**	**Tokens**	**Meaning**	**Freq**	**Tokens**	**Meaning**	**Freq**	**Tokens**	**Meaning**	**Freq**	**Tokens**	**Meaning**
1	25,434	月	Month	18,344	日期	Date	3,034	小标题	Subtitle	1,047	与此同时	At the same time	2,402	新华社北京	Xinhua News Agency Beijing
2	18,633	完	Complete/ Finish	6,828	日电	News on	1,406	据介绍	As is reported	876	本报北京	The Beijing Agency of the Paper	390	新华社东京	Xinhua News Agency Tokyo
3	2,211	年	Year	6,698	他说	He said	1,377	亿美元	Hundred million dollars	459	据新华社	According to Xinhua News	371	到目前为止	Up to date
4	2,019	秒	Second	2,270	目前	At the present	1,366	近年来	In recent years	394	今年以来	Since this year	303	新华社巴黎	Xinhua News Agency Paris
5	949	个	Individual	2,080	因此	Therefore	1,360	据了解	According (to) investigation	385	近几年来	In recent years	263	新华社伦敦	Xinhua News Agency London
6	786	到	Arrive/ Reach	1,929	但是	But	1,168	据统计	According (to) statistics	374	容加金旁	Contain plus gold beside	247	新华社广岛	Xinhua News Agency Guang Dao
7	683	倍	Double	1,854	此外	In addition	881	他指出	He indicates	337	钱其琛说	Qian Qi Chen said	237	今年上半年	In the first half of the year
8	671	米	Meter/ Rice	1,506	据悉	As is known	859	据报道	It is reported	321	江泽民说	Jiang Zemin said	233	新华社天津	Xinhua News Agency Tian Jin
9	641	二	Two	1,482	同时	At the same time	735	他认为	He thinks	255	另一方面	On the other hand	223	他表示相信	He said he believed
10	630	元	Dollar	1,278	今年	This year	573	李鹏说	Li Peng said	246	今天上午	This morning	221	新华社波恩	Xinhua News Agency Bonn

4.2 Chinese Compound Words

The TREC compound word reduced *n*-grams obtained using O' Boyle and Smith 's algorithm were created in 20 hours (we executed the algorithm non-stop for less than one day on a Pentium II with 512 MB of RAM) with a storage requirement of only 1 GB.

The most common TREC word reduced unigrams, bigrams, trigrams, 4-grams and 5-grams are shown in Table 5. One can observe noises in the unigram frequency observations when words with more than 1 syllable appeared in the top ten ("日期 Date," "目前 Currently," "小标题 Subtitle," "因此 Therefore," and "同时 Simultaneously"), but the 2-syllable word "中国 China" disappeared, as shown in Table 4, which lists the most common traditional TREC word unigrams [Ha *et al*. 2003].

Our observations of reduced *n*-grams show that they increase the semantic completeness of longer *n*-grams with large *n* in comparison with conventional Chinese word *n*-grams [Ha *et al*. 2003].

Table 4. The 10-highest frequency unigrams in the conventional Chinese TREC word corpus [Ha et al. 2003]

Rank	Unigrams		
	Freq	**Token**	**Meaning**
1	609,395	的	Of
2	154,827	在	In / At
3	144,524	和	And
4	126,134	了	Perfective Marker
5	99,747	是	Be
6	86,928	一	One
7	77,037	中国	China
8	69,253	中	Centre / Middle
9	60,230	日	Sun
10	57,045	为	For

Table 5. Most common TREC compound word reduced n-grams

Rank	Unigrams			Bigrams			Trigrams			4-grams			5-grams		
	Freq	Tokens	Meaning	Freq	Tokens	Meaning	Freq	Tokens	Meaning	Freq	Tokens	Meaning	Freq	Tokens	Meaning
1	26,831	月	Month	7,253	日 电	Daily News	2,849	星期 星期二 版次	Week Tuesday order	642	现在 广播 完 了	Modern broadcast finishes	405	版名 政治 法律 社会 标题	Political social law page title
2	19,325	完	Complete/ Finish	6,849	他 说	He says	2,703	星期 星期一 版次	Week Monday order	374	容 加 金 旁	Contain plus gold beside[2]	366	一 九 九 四 年	Year nineteen ninety four
3	18,406	日期	Date	2,404	新华社 北京	Xinhua news agency Beijing	2,686	星期 星期六 版次	Week Saturday order	273	人民币 市场 汇 价	Renminbi market exchange price	340	版名 教育 科技 文化 标题	Educational cultural & technology page title
4	8,423	年	Year	1,530	万 元	Ten-thousand yen	2,629	星期 星期四 版次	Week Thursday order	263	据 不 完全 统计	According (to) incomplete statistics	271	外 币 名称 中间 价	Foreign money Ming-Chien transferring price
5	4,188	日	Sun	1,516	亿 美元	Ten-million US Dollars	2,585	星期 星期五 版次	Week Friday order	239	中国 经济 简 讯	China economic brief news	270	一 九 九 三 年	Year nineteen ninety three
6	3,998	目前	Currently	1,450	亿 元	Ten-million yen	2,491	星期 星期日 版次	Week Sunday order	221	十 万 日 元	One hundred thousand Japanese yen	185	个 国家 和 地区 的	Individual international and region of
7	3,292	分	Unit	1,407	据 介绍	According (to) introduction	2,408	星期 星期三 版次	Week Wednesday order	220	一百 欧洲 货币 单位	A hundred European currency units	160	日 电 据 外电 报道	Daily news according to foreign newspaper report
8	3,034	小标题	Subtitle	1,360	据 了解	According (to) understanding	876	本 报 北京	Our newspaper Beijing	212	上 接 第一 版	Continue from No.1 edition	139	一 九 九 二 年	Year nineteen ninety two
9	2,348	因此	Therefore	1,019	日 至	Date to	788	版名 要闻 标题	Abstract page title	134	中国 文化 简 讯	China cultural brief news	96	版名 要 闻 正 文	Brief news major content page
10	2,299	同时	Simultaneously	891	他 指出	He indicates	572	版名 经济 标题	Economic page title	127	日 电 综 述	Daily News generally speaking	91	一 九 九 一 年	Year nineteen ninety one

[2]It is remarkable that many of the *n*-grams with larger *n* values contain content markup information.

Zipf curves for TREC word reduced unigrams and *n*-grams are plotted in Figure 2. It can be observed that the unigram curve not straight, but rather exhibits a two-slope behaviour, beginning with a slope of –0.67 and then falling-off with a slope of approximately -2 at the end. All the bigram, trigram, 4-gram, and 5-gram curves have slopes in the range [-0.6, -0.5] and have become more parallel and straighter. Noise is visible among the TREC word reduced bigrams, trigrams, 4-grams and 5-grams where they turbulently cross each other at the beginning.

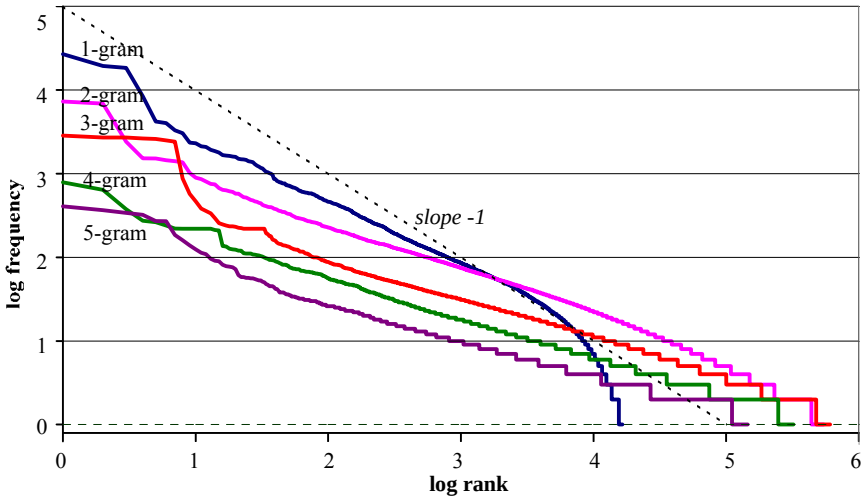

Figure 2. TREC word reduced n-gram Zipf curves

Usually, Zipf 's rank-frequency law is contradicted by empirical data, and the syllable and compound-word reduced *n*-grams from Chinese shown in Figure 1 and Figure 2 also contradict it. In fact, various more sophisticated models for frequency distributions have been proposed by Baayen [2001] and Evert [2004].

5. Perplexity for Chinese Reduced *N*-Grams

The reduced *n*-gram approach was also checked by means of Chinese compound-word perplexity calculations based on the Weighted Average Model of O' Boyle and Smith [1993, 1994, 1995, 1997], which was further developed by Sicilia-Garcia, Ming, Smith and Hanna [1999, 2000, 2001]. We rewrote this famous model in formulae (5) and (6):

$$wgt\left(w_j^i\right) = \log\left(f\left(w_j^{i-1}\right)\right) \times 2^{i-j+1}, \tag{5}$$

$$P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = \frac{wgt\left(w_i\right) \times P\left(w_i\right) + \sum_{l=1}^{N-1} wgt\left(w_{i-l}^i\right) \times P\left(w_i \mid w_{i-l}^{i-1}\right)}{\sum_{l=0}^{N-1} wgt\left(w_{i-l}^i\right)}. \tag{6}$$

Next, we will analyse the main difficulties arising from perplexity calculations for our reduced model: the statistical model problem, unseen word problem and unknown word problem.

5.1 Statistical model problem

In a reduced model, the following rules apply:

- If $f\left(w_{i-l}^i\right) > 0$, but $f\left(w_{i-l}^{i-1}\right) = 0$, then the maximum likelihood $P\left(w_i \middle| w_{i-l}^{i-1}\right) = \frac{f\left(w_{i-l}^i\right)}{f\left(w_{i-l}^{i-1}\right)}$ and the weight $wgt\left(w_{i-l}^i\right) = \log\left(f\left(w_{i-l}^{i-1}\right)\right) \times 2^{l+1}$ will be undefined.
- Once the weight calculation has been performed, if $P\left(w_i \middle| w_{i-l-L_0}^{i-1}\right) > 0$ and the previous L_0 phrases $P\left(w_i \middle| w_{i-l-L_0+1}^{i-1}\right)$, $P\left(w_i \middle| w_{i-l-L_0+2}^{i-1}\right)$, …, $P\left(w_i \middle| w_{i-l}^{i-1}\right)$ are all 0, then we should include the L_0 phrases' weights of zero probability $wgt\left(w_{i-l-L_0+1}^i\right)$, $wgt\left(w_{i-l-L_0+2}^i\right)$, …, $wgt\left(w_{i-l}^i\right)$ into the sum of weights in the denominator of formula (6).

5.2 Unseen word problem

If $P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = 0$ but w_i occurs in other reduced n-grams, then how can we calculate the probability ?

5.3 Unknown word problem

If $P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = 0$ and w_i does not occur in any other reduced n-grams, then w_i will be totally unknown and we will not be able to apply the Turing-Good probability. This is because an unusual phenomenon will occur with the hapax legomena n_1 and dis legomena n_2 when $n_1 < n_2$, and because the Turing-Good probabilities will become too high if we use T_{reduced}, as shown in Table 6.

Table 6. Unusual Turing-Good observations with respect to reduced models

	n_1	N_2	T_{reduced}	Turing-Good reduced probability
TREC reduced words	1,620	2,171	17,515	0.0001530258

The Turing-Good probability for the conventional TREC word corpus is 7.082435E-08. For the reduced model shown in Table 6, the Turing-Good value is 2,161 times higher, which is unusual.

5.4 Solutions for reduced perplexities

- If $f\left(w_{i-l}^{i}\right) > 0$ but $f\left(w_{i-l}^{i-1}\right) = 0$ and the reduced training size is R, then the degraded weight $wgt\left(w_{i-l}^{i}\right) = \ln(R)$ and the maximum likelihood $P\left(w_i \middle| w_{i-l}^{i-1}\right) = \dfrac{f\left(w_{i-l}^{i}\right)}{R}$ are defined in the case of an isolated unigram.
- If $P\left(w_i \middle| w_{i-l-L_0}^{i-1}\right) > 0$ but all the previous L_0 phrases $P\left(w_i \middle| w_{i-l-L_0+1}^{i-1}\right)$, $P\left(w_i \middle| w_{i-l-L_0+2}^{i-1}\right)$, …, $P\left(w_i \middle| w_{i-l}^{i-1}\right)$ are 0, then we will include all the weights of zero probability, i.e., $wgt\left(w_{i-l-L_0+1}^{i}\right)$, $wgt\left(w_{i-l-L_0+2}^{i}\right)$, …, $wgt\left(w_{i-l}^{i}\right)$, into the sum of the weight denominator in formula (6). This should reduce the weighted average probability in comparison with the probabilities in other cases where all the previous L_0 phrases exist.
- Unseen word problem: If $P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = 0$ but w_i occurs in other reduced n-grams, then we degrade w_i when words have been eliminated from the reduced model because they appear less than m times. Therefore, w_i will have an estimated probability of $\dfrac{m-1}{T}$, where T is the overall conventional training size.
- Unknown word problem: If $P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = 0$ and w_i does not occur in any other reduced n-grams, then w_i will be assigned the Turing-Good probability.

5.5 Results and Discussion

The perplexities for the Chinese TREC compound-word corpus were calculated. We obtained very poor and confusing perplexity results when we investigated short contexts of reduced n-grams, but coped well with long contexts since the purpose and the strength of reduced

models are their ability to store phrase histories that are as long as possible as well as an entire large corpus in a compact database. The test text file had 27,485 words in 3,093 sentences and 927 paragraphs (along with 7 unknown words of 6 types and 109 unseen words of 48 unseen types) for the TREC conventional *n*-gram model and also for the conventional and reduced models.

Our Chinese TREC word reduced model was stored in 50 MB of memory, and the perplexity investigation started with phrase-lengths of 10 words and more and increased until all the phrases had been analysed. The perplexity results obtained using the TREC word reduced model are shown in Table 7.

Table 7. Reduced perplexities for Chinese TREC words obtained using the weighted average model

	Traditional *n*-grams		**Reduced *n*-grams**		**The cost of reduced *n*-grams on baseline trigrams**	**Factor of reduced model size**
Phrase Length	**Unigram**	1,515.03	**10-grams**	128.35	**-19.25%**	
	Bigram	293.96	**11-grams**	131.95	**-16.99%**	
	Trigram	**158.95**	**12-grams**	134.77	**-15.21%**	
	4-gram	140.81	**13-grams**	136.98	**-13.82%**	
	5-gram	137.61	**14-grams**	138.68	**-12.75%**	
	6-gram	137.31	**15-grams**	140.01	**-11.91%**	
	7-gram	137.25	**Complete contexts**	**145.06**	**-8.74%**	**11.49**

Surprisingly for TREC word reduced *n*-grams, we achieved an 8.74% perplexity reduction, and the model size was reduced by a factor of 11.49. Thus, in our study on Chinese TREC words, we achieved improvement in both perplexity and model size.

6. Conclusions

The conventional *n*-gram language model is limited in terms of its ability to represent extended phrase histories because of the exponential growth in the number of parameters. To overcome this limitation, we have re-investigated the approach of O' Boyle and Smith [1992, 1993] and created a Chinese reduced *n*-gram model. The main advantage of Chinese reduced *n*-grams is that they have quite complete semantic meanings thanks to their creation process, starting from execution of whole sentence contexts.

Chinese reduced word and character Zipf curves and perplexity calculations along with the model size for TREC, a large Chinese corpus, have been presented. The reduced Chinese

syllable unigram Zipf curve has a slope of –1, which satisfies Zipf's law, and the reduced TREC word unigram Zipf curve shows a two-slope behaviour, similar to the curves reported by Ferrer and Solé [2002]. The difficulties with reduced model perplexity calculations due to statistical, unseen and unknown problems have been solved using the Weighted Average Model, a back-off probability model developed by O' Boyle and Smith [1993, 1994, 1995, 1997]. By extending TREC word reduced *n*-grams, we achieved an 8.74% perplexity reduction, and we were able to reduce the model size by a factor of 11.49. This remarkable improvement in the Chinese TREC reduced *n*-gram distribution may be smaller that that possible with the English language, in which the meaning of a word is clearer. This confirms Siu and Ostendorf 's [2000] conclusions concerning the potential application of their variable *n*-grams to Chinese (and Japanese) and other languages besides English.

Acknowledgements

The authors would like to thank Maria Husin and Dr Sicilia-Garcia for valuable support, the reviewers for their valuable comments and David Ludwig for his revision.

References

Baayen, H., "Word Frequency Distributions," Kluwer Academic Publishers, 2001.

Evert, S., "A Simple LNRE Model for Random Character Sequences," *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, 2004, pp. 411-422.

Ferrer I Cancho, R. and R. V. Solé, "Two Regimes in the Frequency of Words and the Origin of Complex Lexicons," *Journal of Quantitative Linguistics*, 8(3) 2002, pp. 165-173.

Francis, W. N. and H. Kucera, "Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers," Department of Linguistics, Brown University, Providence, Rhode Island, 1964.

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming and F. J. Smith, "Extension of Zipf 's Law to Word and Character *N*-Grams for English and Chinese," *Journal of Computational Linguistics and Chinese Language Processing*, 8(1) 2003, pp. 77-102.

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming and F. J. Smith, "Extension of Zipf 's Law to Words and Phrases," *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, 2002, pp. 315-320.

Hu, J., W. Turin and M. K. Brown, "Language Modeling using Stochastic Automata with Variable Length Contexts," *Computer Speech and Language*, vol. 11, 1997, pp. 1-16.

Kneser, R., "Statistical Language Modeling Using a Variable Context Length," *ICSLP*, vol. 1, 1996, pp. 494-497.

Niesler, T. R., "Category-based statistical language models," St. John's College, University of Cambridge, 1997.

Niesler, T. R. and P. C. Woodland, "A Variable-Length Category-Based *N*-Gram Language Model," *IEEE ICASSP*, vol. 1, 1996, pp. 164-167.

O' Boyle, P., J. McMahon and F. J. Smith, "Combining a Multi-Level Class Hierarchy with Weighted-Average Function-Based Smoothing," *IEEE Automatic Speech Recognition Workshop*, Snowbird, Utah, 1995.

O' Boyle, P. and M. Owens, "A Comparison of human performance with corpus-based language model performance on a task with limited context information," *CSNLP*, Dublin City University, 1994.

O' Boyle, P., M. Owens and F. J. Smith, "A weighted average *N*-Gram model of natural language," *Computer Speech and Language*, vol. 8, 1994, pp. 337-349.

O' Boyle, P. L., J. Ming, M. Owens and F. J. Smith, "Adaptive Parameter Training in an Interpolated *N*-Gram language model," *QUALICO*, Helsinki, Finland, 1997.

O' Boyle, P. L., "A study of an *N*-Gram Language Model for Speech Recognition," PhD thesis, Queen's University Belfast, 1993.

Sicilia-Garcia, E. I., "A Study in Dynamic Language Modelling," PhD thesis, Queen's University Belfast, 2001.

Sicilia-Garcia, E. I., J. Ming and F. J. Smith, "A Dynamic Language Model based on Individual Word Domains," *Proceedings of the 18th International Conference on Computational Linguistics* COLING 2000, Saarbrucken, Germany, vol. 2, 2000, pp. 789-794.

Sicilia-Garcia, E. I., J. Ming and F. J. Smith, "Triggering Individual Word Domains in *N*-Gram Language Model," *Proceedings of the European Conference on Speech Communication and Technology* (EuropeSpeech), vol. 1, 2001, pp. 701-704.

Sicilia-Garcia, E. I., F. J. Smith and P. Hanna, "A Dynamic Language Model based on Individual Word Models," *Pre-proceedings of the 10th Irish Conference on Artificial Intelligence and Cognitive Science* AICS 99, Cork Ireland, 1999, pp. 222-229.

Siu, M. and M. Ostendorf, "Integrating a Context-Dependent Phrase Grammar in the Variable *N*-Gram framework," *IEEE ICASSP*, vol. 3, 2000, pp. 1643-1646.

Siu, M. and M. Ostendorf, "Variable *N*-Grams and Extensions for Conversational Speech Language Modelling," *IEEE Transactions on Speech and Audio Processing,* 8(1) 2000, pp. 63-75.

Smith, F. J. and P. O' Boyle, "The *N*-Gram Language Model," *The Cognitive Science of Natural Language Processing Workshop*, Dublin City University, 1992, pp. 51-58.

Zipf, G. K., "Human Behaviour and the Principle of Least Effort," Reading, MA: Addison-Wesley Publishing Co., 1949.