MULTILINGUAL REDUCED *N*-GRAM MODELS

Tran Thi Thu Van and Le Quan Ha

Abstract*: Statistical language models should improve as the size of the n-grams increases from 3 to 5 or higher. However, 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 [1993] can be applied. A reduced n-gram language model can store an entire corpus's phrase-history length within feasible storage limits. Another theoretical advantage of reduced n-grams is that they are closer to being semantically complete than traditional models, which include all n-grams. In our experiments, the reduced n-gram Zipf curves are first presented, and compared with conventional n-grams for all Irish, Chinese and English. The reduced n-gram model is then applied for large Irish, Chinese and English corpora. For Irish, we can reduce the model size, compared to the 7-gram traditional model size, with a factor of 15.1 for a 7-million-word Irish corpus while obtaining 41.63% improvement in perplexities; for English, we reduce the model sizes with factors of 14.6 for a 40-million-word corpus and 11.0 for a 500-million-word corpus while obtaining 5.8% and 4.2% perplexity improvements; and for Chinese, we gain a 16.9% perplexity reductions and we reduce the model size by a factor larger than 11.2. This paper is a step towards the modeling of Irish, Chinese and English using semantically complete phrases in an n-gram model.*

Keywords*: Reduced n-grams, Overlapping n-grams, Weighted average (WA) model, Katz back-off, Zipf's law.*

ACM Classification Keywords*: I. Computing Methodologies - I.2 ARTIFICIAL INTELLIGENCE - I.2.7 Natural Language Processing - Speech recognition and synthesis*

Introduction

Shortly after this laboratory first published a variable *n*-gram algorithm by [Smith and O'Boyle, 1992], [O'Boyle, 1993] proposed a statistical method to improve language models based on the removal of overlapping phrases.

A distortion in the use of phrase frequencies had been observed in the small railway timetable Vodis Corpus when the bigram "RAIL ENQUIRIES" and its super-phrase "BRITISH RAIL ENQUIRIES" were examined. 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 bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it. At the time of their research, O'Boyle reduced the frequencies of "RAIL ENQUIRIES" by subtracting the frequency of the larger trigram, which gave a probability of zero for "ENQUIRIES" following "RAIL" if it was not preceded by "BRITISH." The phrase with a new reduced frequency is called a reduced phrase.

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 [O'Boyle, 1992], O'Boyle was able to use it to analyse the Brown corpus of American English [Francis and Kucera, 1964] (of one million word tokens, whose longest phrase-length is 30), which was a considerable improvement at the time. The results were used in an *n*-gram language model by O'Boyle, but with poor results, due to lack of statistics from such a small corpus. We have developed here a modification of his method, and we discuss its usefulness for reducing *n*-gram perplexity.

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. For the LOB corpus, the varigram model obtained 11.3% higher perplexity than the word-trigram model [Niesler and Woodland, 1996.]

[Kneser, 1996] built up variable-context length language models based on the North American Business News (NAB-240 million words) 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 a 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 depended 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.

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**	**Perplexity**	**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		

[Siu and 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. They obtained word error reductions of 0.1 and 0.5% (absolute) in development and evaluation test sets, respectively. 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 model size by more than 50% compared to a regular trigram model, and improved both the test-set perplexity and recognition performance. They also reduced the model size by an additional 8%.

Other related work are those of [Seymore and Rosenfeld, 1996]; [Hu, Turin and Brown, 1997]; [Blasig, 1999]; and [Goodman and Gao, 2000.]

In order to obtain an overview of variable *n*-grams, Table 1 combines all of their results.

Reduced N-Gram Algorithm

The main goal of this algorithm [Ha, Seymour, Hanna and Smith, 2005] 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.
- The file that contains any overlapping *n*-grams that occur at least *m* times in the *SUB* file is called the *LOS* file.

The final list of reduced phrases is called the *FIN* file, where

$$FIN := PHR + LOS - SUB \tag{1}$$

Before O'Boyle's work, a student Craig [O'Boyle, 1993] in an unpublished project used a loop algorithm that was equivalent to *FIN*:=*PHR*–*SUB*. This yields negative frequencies for some resulting *n*-grams with overlapping, hence the need for the *LOS* file.

There are 2 additional files

- 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.)
- To create the *LOS* file, O'Boyle 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's reduced *n*-gram algorithm is limited to small corpora, such as the Brown corpus (American English) of 1 million words [Smith and O'Boyle, 1992], in which the longest phrase has 30

words. Now their algorithm, re-checked by us, still works for medium size and large corpora. In order to work well for very large corpora, it has been implemented by file distribution and sort processes.

By re-applying O'Boyle and Smith's algorithm, Ha et al. [2005] investigated a reduced *n*-gram model for the Chinese TREC corpus of the Linguistic Data Consortium (LDC) (www.ldc.upenn.edu), catalog no. LDC2000T52. Later on [Ha, Hanna, Stewart and Smith, 2006] obtained reduced *n*-grams from two English large corpora and a Chinese large corpus. The two English corpora used in their experiments were the full text of articles appearing in the Wall Street Journal (WSJ) [Paul and Baker, 1992] of 40 million tokens respectively; and the North American News Text (NANT) corpus from the LDC (catalog no. LDC95T21 and LDC98T30) sizing 500 million tokens. Their employed Chinese corpus was the compound word version in [Ha, Sicilia-Garcia, Ming and Smith, 2003] of the Mandarin News corpus with 50,000 word types, originally from the LDC, catalog no. LDC95T13 of over 250 million syllables.

Reduced N-Grams and Zipf's Law

By re-applying O'Boyle and Smith's algorithm, we obtained the Zipf curves [Zipf, 1949] for the English, Chinese and Irish reduced *n*-grams.

The Irish is a highly-inflected Indo-European Celtic language. Both the beginning and end of words are regularly inflected. The Irish corpus in our experiments is taken from a corpus of 17th and 18th century Irish from the Royal Irish Academy (www.ria.ie) with sizes 7,122,537 tokens with 449,968 types [Harvey, Devine and Smith, 1994.]

We next present the Zipf curves for the English, Chinese and Irish reduced *n*-grams: All of our reduced *n*-grams were created on a Pentium II 586 of 512MByte RAM.

Wall Street Journal corpus (English)

The WSJ reduced *n*-grams can be created by the original O'Boyle-Smith algorithm for over 40 hours, the disk storage requirement being only 5GBytes.

The Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 1 showing all the curves have slopes within [-0.6, -0.5]. The WSJ reduced bigram, trigram, 4-gram and 5-gram curves become almost parallel and straight, with a small observed noise between the reduced 4-gram and 5-gram curves when they cut each other at the beginning. Note that information theory tells us that an ideal information channel would be made of symbols with the same probability. So having a slope of –0.5 is closer than –1 to this ideal.

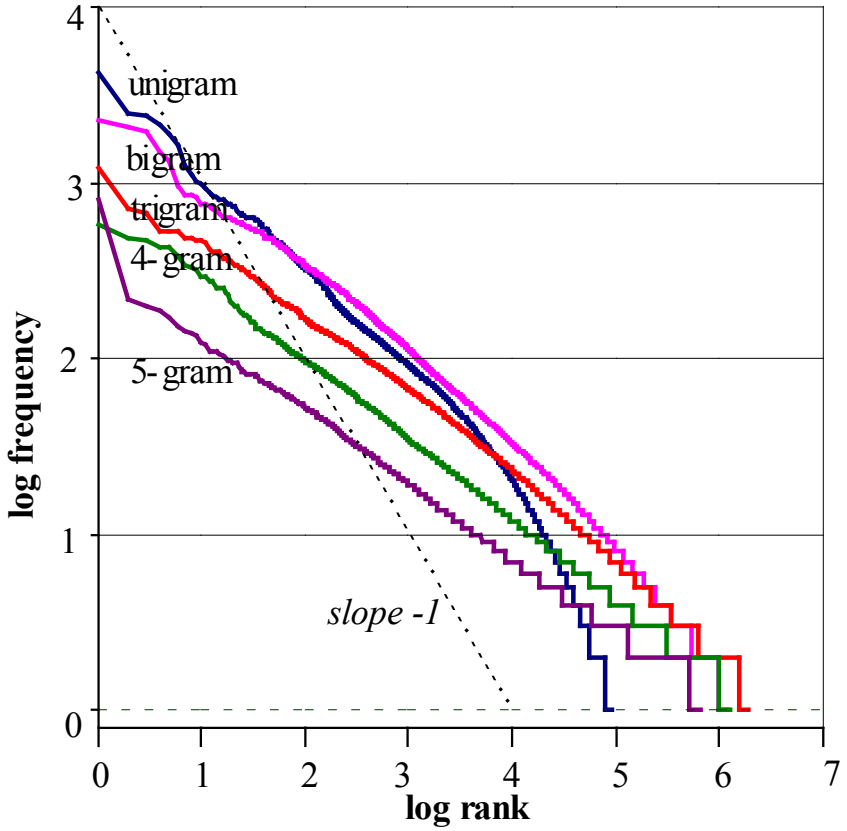

Figure 1. The WSJ reduced *n*-gram Zipf curves.

Table 2. Most common WSJ reduced *n*-grams.

Rank	Unigrams		Bigrams		Trigrams	
	Freq	Token	Freq	Token	Freq	Token
1	4,273	Mr.	2,268	he said	1,231	terms weren't disclosed
2	2,469	but	2,052	he says	709	the company said
3	2,422	and	1,945	but the	664	as previously reported
4	2,144	the	1,503	but Mr.	538	he said the
5	1,918	says	1,332	and the	524	a spokesman for
6	1,660	or	950	says Mr.	523	the spokesman said
7	1,249	said	856	in addition	488	as a result
8	1,101	however	855	and Mr.	484	earlier this year
9	1,007	while	832	last year	469	in addition to
10	997	meanwhile	754	for example	466	according to Mr.

The conventional 10-highest frequency WSJ words have been published by [Ha, Sicilia-Garcia, Ming and Smith, 2002] and the most common WSJ reduced unigrams, bigrams and trigrams are shown in Table 2. It illustrates that the most common reduced word is not THE; even OF is not in the top ten. These words are now mainly part of longer *n*-grams with large *n*.

North American News Text corpus (English)

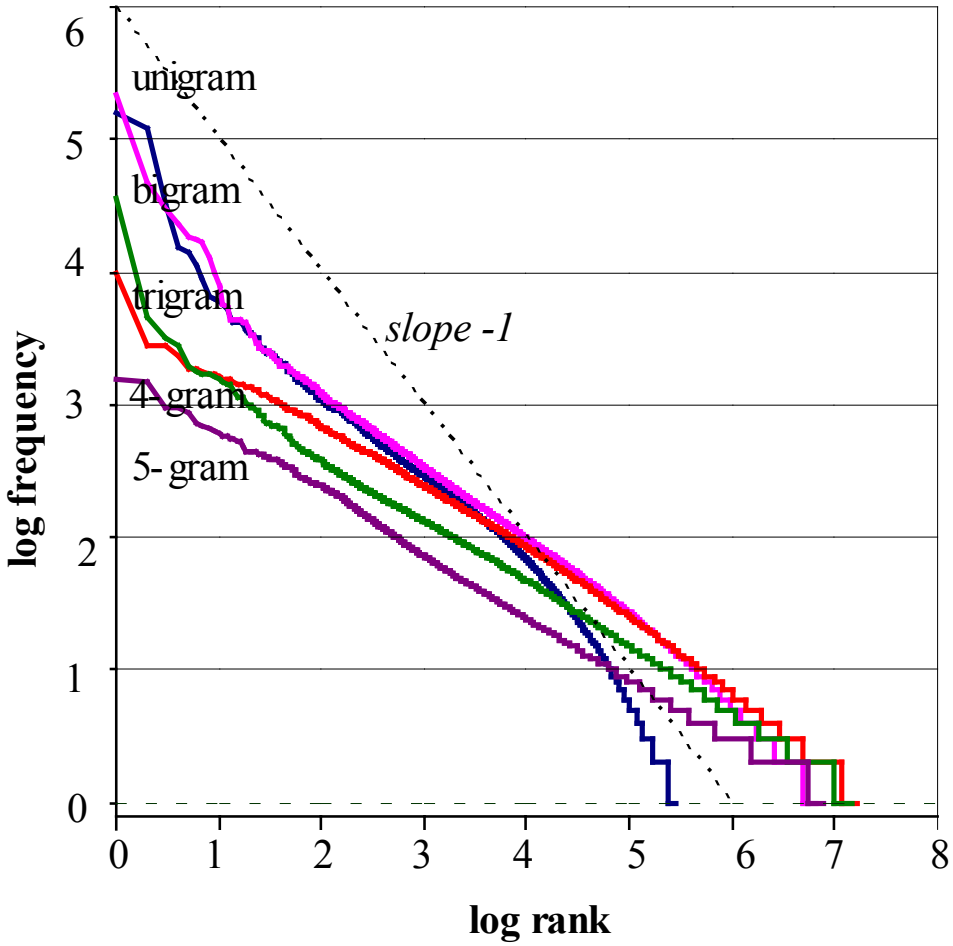

Figure 2. The NANT reduced *n*-gram Zipf curves.

The NANT reduced *n*-grams are created by the improved algorithm after over 300 hours processing, needing a storage requirement of 100GBytes.

Their Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 2 showing all the curves are just sloped around [-0.54, -0.5]. The reduced unigrams of NANT still show the 2-slope behavior when it starts with slope

–0.54 and then drop with slope nearly –2 at the end of the curve. We have found that the traditional *n*-grams also show this behaviour, with an initial slope of –1 changing to –2 for large ranks [Ha and Smith, 2004; Ferrer and Solé, 2002.]

Mandarin News compound words

The Mandarin News reduced word *n*-grams were created in 120 hours, using 20GB of disk space. The Zipf curves are plotted in Figure 3 showing that the unigram curve now has a larger slope than –1, it is around –1.2. All the *n*-gram curves are now straighter and more parallel than the traditional *n*-gram curves, have slopes within [-0.67, -0.5]. Usually, Zipf's rank-frequency law with a slope –1 is confirmed by empirical data, but the reduced *n*-grams for English and Chinese shown in Figures 1, 2 and 3 do not confirm it. In fact, various more sophisticated models for frequency distributions have been proposed by [Baayen, 2001] and [Evert, 2004.]

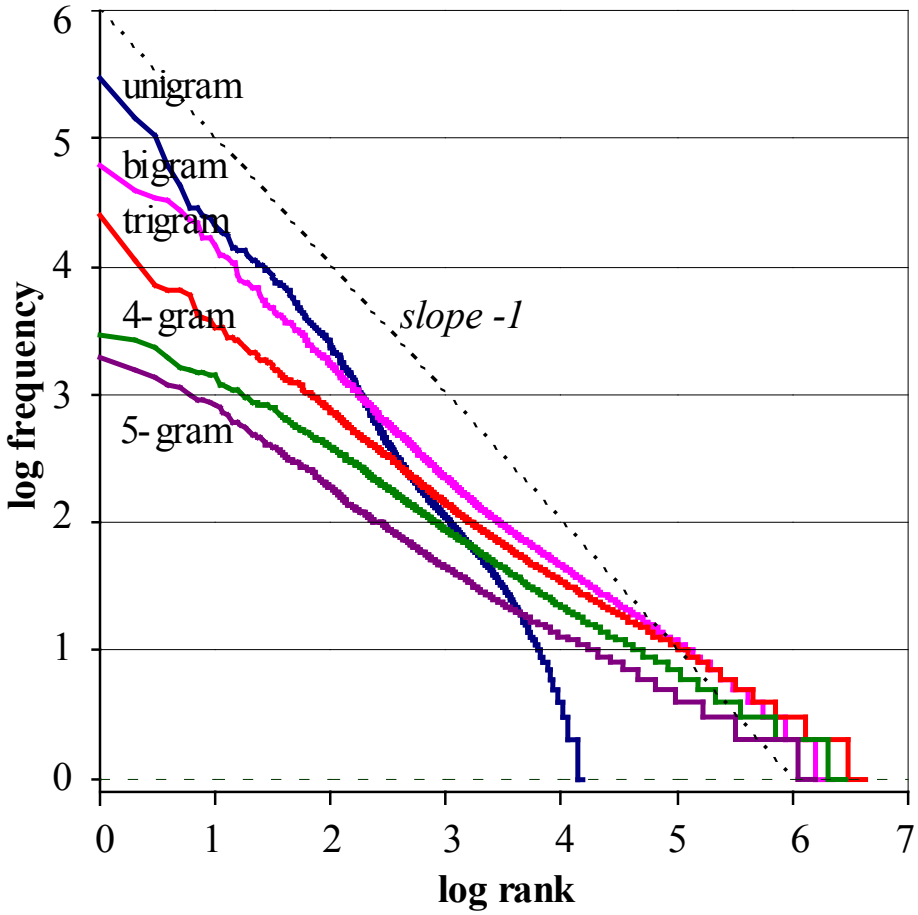

Figure 3. The Mandarin News reduced *n*-grams Zipf curves.

The Irish RIA corpus

The Irish reduced *n*-grams can be created by the original O'Boyle-Smith algorithm in 15 hours, the disk storage requirement being only 600 MBytes.

The most common Irish reduced and traditional unigrams are shown in Table 3 but nine words in the top ten are the same ones. This is very different from English and Chinese that only one or two most common words re-appeared within the top ten reduced unigrams. This fact is coming from the numerous word inflections in Irish language so that there are less overlapping Irish words and phrases.

The Irish Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 4. Because of the word inflections, the Irish reduced unigram curve has a slope of -0.92, closer to the original Zipf's law than the previous English and Chinese reduced unigrams while the Irish reduced bigram, trigram, 4-gram and 5-gram Zipf curves have slopes within [-0.67, -0.45].

Table 3. The ten most common Irish traditional and reduced unigrams.

Rank	Traditional unigrams			Reduced unigrams		
	Freq	Token	Meaning	Freq	Token	Meaning
1	265,288	a	[relative particle]	74,505	a	[relative particle]
2	260,938	agus	and	71,145	an	the
3	259,093	do	to your	68,675	do	to your
4	253,247	an	the	45,063	na	the (pl.)
5	148,703	na	the (pl.)	37,252	agus	and
6	125,486	ar	on	34,006	go	that
7	111,170	go	that	31,398	ar	on
8	92,142	is	is	19,252	i	in
9	66,809	i	in	16,753	is	is
10	57,160	sin	that	15,991	#NO	[number]

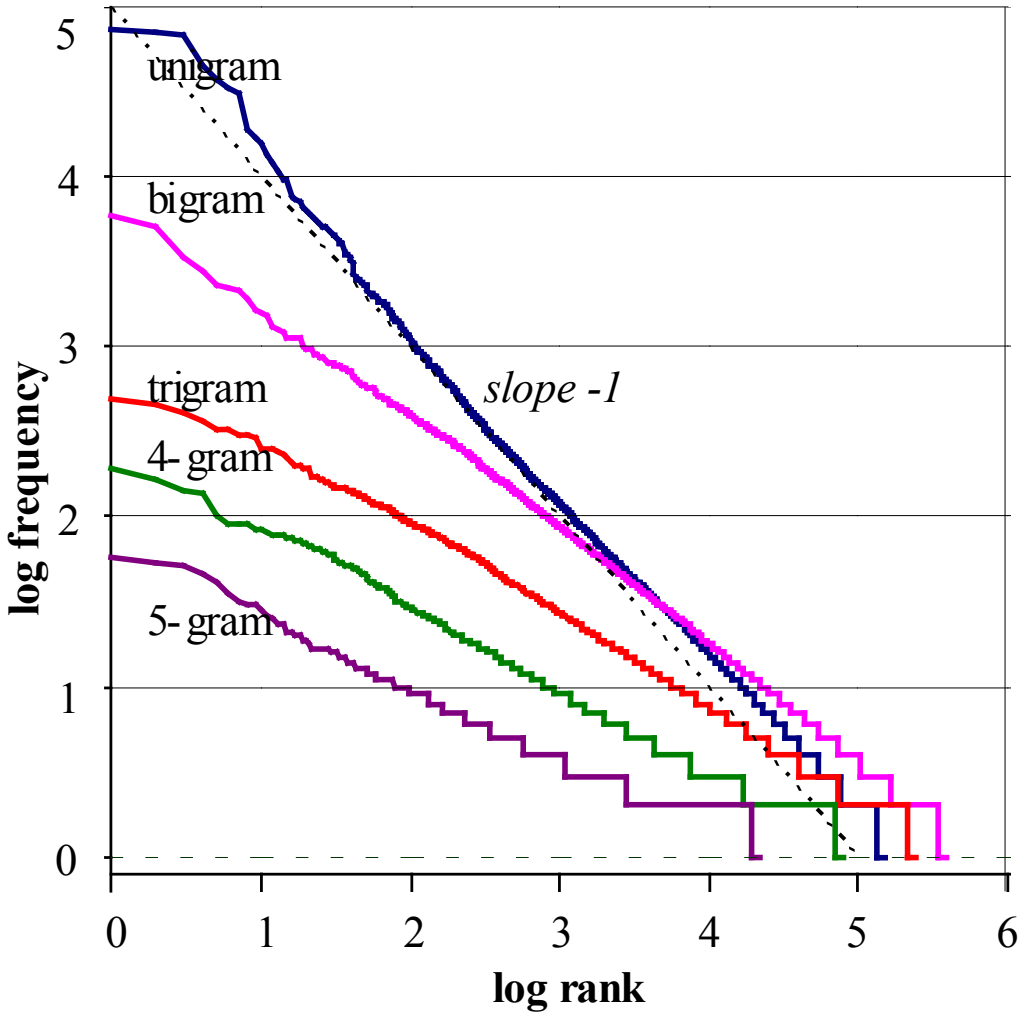

Figure 4. The Irish reduced *n*-gram Zipf curves.

Methods of Testing

The reduced *n*-gram approach was used to build a statistical language model based on the weighted average model of [O'Boyle, Owens and Smith, 1994.] We rewrite this model in formulae (2) and (3)

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

$$P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = \frac{wgt(w_i) \times P(w_i) + \sum_{l=1}^{N-1} wgt\left(w_{i-l}^i\right) \times P\left(w_i \middle| w_{i-l}^{i-1}\right)}{\sum_{l=0}^{N-1} wgt\left(w_{i-l}^i\right)} \tag{3}$$

This averages the probabilities of a word w_i following the previous one word, two words, three words, etc. (i.e. making the last word of an *n*-gram.) The averaging uses weights that increase slowly with their frequency and rapidly with the length of *n*-gram.

The probabilities of all of the sentences w_1^m in a test text are then calculated by the weighted average (WA) model

$$P\left(w_1^m\right) = P_{WA}\left(w_1\right)P_{WA}\left(w_2\middle|w_1\right)...P_{WA}\left(w_m\middle|w_1^{m-1}\right) \tag{4}$$

and an average perplexity of each sentence is evaluated using Equation (5)

$$PP\left(w_1^m\right) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} Ln\left(P_{WA}\left(w_i\middle|w_1w_2...w_{i-1}\right)\right)\right) \tag{5}$$

This weighted average model is a variable length model that gives results comparable to the Katz back-off method [Katz, 1987], but is much quicker to use.

Perplexity for Reduced N-Grams

[Ha et al., 2005, 2006] already investigated and analysed the main difficulties arising from perplexity calculations for our reduced model: a statistical model problem, an unseen word problem and an unknown word problem. Their solutions are applied in this paper also. Similar problems have been found by other authors, e.g. [Martin, Liermann and Ney, 1997]; [Kneser and Ney, 1995.]

The perplexity calculations of reduced *n*-grams includes statistics on phrase lengths starting with unigrams, bigrams, trigrams, etc. and on up to the longest phrase which occur in the reduced model.

The nature of the reduced model makes the reporting of results for limited sizes of *n*-grams to be inappropriate, although these are valid for a traditional *n*-gram model. Therefore we show results for several *n*-gram sizes for the traditional model, but only one perplexity for the reduced model. The perplexities of the WSJ reduced model by the weighted average model and the Katz Back-off method are shown in Table 4, North American News Text corpus in Table 5, Mandarin News words in Table 6 and the Irish RIA corpus in Table 7.

Table 4. Reduced perplexities for English WSJ.

Unknowns	**Tokens**	0	
	Types	0	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	762.69	762.69
	Bigrams	144.33	108.04
	Trigrams	75.36	59.71
	4-grams	60.73	53.16
	5-grams	56.85	52.84
	6-grams	55.66	51.61
	7-grams	55.29	51.10
Reduced Model by WA model		70.98	
%Improvement of Reduced Model on baseline WA trigrams		5.81%	
Model size reduction		14.56	

Table 5. Reduced perplexities for English NANT.

Unknowns	**Tokens**	24	
	Types	23	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	1,442.99	1,442.99
	Bigrams	399.61	339.26
	Trigrams	240.52	217.22
	4-grams	202.59	189.24
	5-grams	194.06	181.55
	6-grams	191.91	179.09
	7-grams	191.23	178.97
Reduced Model by WA model		230.46	
%Improvement of Reduced Model on baseline WA trigrams		4.18%	
Model size reduction		11.01	

Table 6. Reduced perplexities for Mandarin News words.

Unknowns	**Tokens**	84	
	Types	26	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	1,620.56	1,620.56
	Bigrams	377.43	328.32
	Trigrams	179.07	158.24
	4-grams	135.69	116.27
	5-grams	121.53	105.61
	6-grams	114.96	102.69
	7-grams	111.69	102.17
Reduced Model by WA model		148.71	
%Improvement of Reduced Model on baseline WA trigrams		16.95%	
Model size reduction		11.28	

Table 7. Reduced perplexities for Irish RIA corpus.

Unknowns	**Tokens**	36	
	Types	36	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	412.80	412.80
	Bigrams	162.99	144.81
	Trigrams	134.93	134.56
	4-grams	133.47	130.81
	5-grams	133.25	126.97
	6-grams	133.19	126.11
	7-grams	133.18	125.89
Reduced Model by WA model		78.75	
%Improvement of Reduced Model on baseline WA trigrams		41.63%	
Model size reduction		15.1	

In all cases of Irish, Chinese and English, their reduced models produced various perplexity improvements over the traditional 3-gram model (41.63% for Irish, 16.95% for Chinese and 4.18% for English). However, [Ha et al., 2006] obtained a significant reduction in model size, from a factor of 11.2 to almost 15.1 compared to the traditional Irish, Chinese and English model sizes. The Irish reduced model produces a perplexity improvement much better than previous results in English and Chinese languages and the reason is that the Irish language has numerous word inflections and inflected words' meanings are much related together.

For further work, we also need missing word tests.

Conclusion

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, 1993] and created a reduced *n*-gram model for Irish language. Our aim was to try to create an *n*-gram model that used semantically more complete *n*-grams than traditional *n*-grams in the expectation that this might lead to an improvement in language modeling. The good improvements in perplexity and model size reduction are better for Irish than for similar work by [Ha et al., 2005, 2006] in English and Chinese because Irish is a highly inflected language. It has numerous Irish word inflections while the inflected words' meanings are related together. So this represents an encouraging step forward, although still very far from the final step in language modelling.

Acknowledgements

The authors would like to thank the Royal Irish Academy for their Irish corpus. Our special acknowledgement is heart-fully sent to Prof Vladimir Polyakov and Prof Valery Solovyev.

Bibliography

[Baayen, 2001] H.R. Baayen. Word Frequency Distributions. Kluwer Academic Publishers, 2001.

[Blasig, 1999] R. Blasig. Combination of Words and Word Categories in Varigram Histories. In: ICASSP'99, Vol. 1, 529-532. 1999.

[Evert, 2004] S. Evert. A Simple LNRE Model for Random Character Sequences. In: Proc. of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles, 411-422. 2004.

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

[Francis and Kucera, 1964] N. Francis 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.

[Goodman and Gao, 2000] J. Goodman and J. Gao. Language Model Size Reduction by Pruning and Clustering. In: ICSLP'00. Beijing, China, 2000.

[Ha and Smith, 2004] L.Q. Ha and F.J. Smith. Zipf and Type-Token rules for the English and Irish languages. In: MIDL workshop. Paris, 2004.

[Ha, Hanna, Stewart and Smith, 2006] L.Q. Ha, P. Hanna, D.W. Stewart and F.J. Smith. Reduced n-gram Models for English and Chinese Corpora. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, 309-315. Sydney, Australia, 2006.

[Ha, Seymour, Hanna and Smith, 2005] L.Q. Ha, R. Seymour, P. Hanna and F.J. Smith. Reduced N-Grams for Chinese Evaluation. In: CLCLP, 10(1):19-34. 2005.

[Ha, Sicilia-Garcia, Ming and Smith, 2002] L.Q. Ha, E.I. Sicilia-Garcia, J. Ming and F.J. Smith. Extension of Zipf's Law to Words and Phrases. In: COLING'02, Vol. 1, 315-320. 2002.

[Ha, Sicilia-Garcia, Ming and Smith, 2003] L.Q. Ha, 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. In: CLCLP, 8(1):77-102. 2003.

[Harvey, Devine and Smith, 1994] A. Harvey, K. Devine and F.J. Smith. Archive of Celtic-Latin Literature ACLL-1 Royal Irish Academy. Dictionary of Medieval Latin from Celtic sources. Brespols, 1994.

[Hu, Turin and Brown, 1997] J. Hu, W. Turin and M.K. Brown. Language Modeling using Stochastic Automata with Variable Length Contexts. In: Computer Speech and Language, Vol. 11, 1-16. 1997.

[Katz, 1987] S.M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. In: IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-35, 400-401. 1987.

[Kneser and Ney, 1995] R. Kneser and H. Ney. Improved Backing-off for M-Gram Language Modeling. In: ICASSP'95, Vol. 1, 181-184. Detroit, 1995.

[Kneser, 1996] R. Kneser. Statistical Language Modeling Using a Variable Context Length. In: ICSLP'96, Vol. 1, 494-497. 1996.

[Martin, Liermann and Ney, 1997] S.C. Martin, J. Liermann and H. Ney. Adaptive Topic-Dependent Language Modelling Using Word-Based Varigrams. In: EuroSpeech'97, Vol. 3, 1447-1450. Rhodes, 1997.

[Niesler and Woodland, 1996] T.R. Niesler and P.C. Woodland. A Variable-Length Category-Based N-Gram Language Model. In: ICASSP'96, Vol. 1, 164-167. 1996.

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

[O'Boyle, Owens and Smith, 1994] P. O'Boyle, M. Owens and F.J. Smith. A weighted average N-Gram model of natural language. In: Computer Speech and Language, Vol. 8, 337-349. 1994.

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

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

[Paul and Baker, 1992] D.B. Paul and J.B. Baker. The Design for the Wall Street Journal based CSR Corpus. In: Proc. of the DARPA SLS Workshop, 357-361. 1992.

[Seymore and Rosenfeld, 1996] K. Seymore and R. Rosenfeld. Scalable Backoff Language Models. In: ICSLP'96, 232-235. 1996.

[Siu and Ostendorf, 2000] M. Siu and M. Ostendorf. Integrating a Context-Dependent Phrase Grammar in the Variable N-Gram framework. In: ICASSP'00, Vol. 3, 1643-1646. 2000.

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

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

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

Authors' Information

Tran Thi Thu Van *– Lecturer, Hochiminh City University of Technology HUTECH, 41/32 Le Duc Tho, Ward 16, Go Vap District, Hochiminh City, Vietnam; e-mail: Nlp.Sr@Shaw.ca*

Major Fields of Scientific Research: Natural language processing, Speech recognition

Le Quan Ha *– PhD Research Assistant, PhD, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom; Vice-Head of Computer Science, Faculty of IT, Hochiminh City University of Industry, Ministry of Industry and Trade, Vietnam; contact: 8 Bermondsey Court N.W., Calgary, Alberta T3K 1V7, Canada; e-mail: lequanha@hui.edu.vn, lequanha@fit-hui.edu.vn*

Major Fields of Scientific Research: Natural language processing, Speech recognition