A Closer Look at Skip-gram Modelling

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, Yorick Wilks

NLP Research Group, Department of Computer Science, University of Sheffield
Regent court, 211 Portobello Street, Sheffield, S10 4DP
{dguthrie, ben, wei, louise, yorick}@dcs.shef.ac.uk

Abstract

Data sparsity is a large problem in natural language processing that refers to the fact that language is a system of rare events, so varied and complex, that even using an extremely large corpus, we can never accurately model all possible strings of words. This paper examines the use of skip-grams (a technique where by n-grams are still stored to model language, but they allow for tokens to be skipped) to overcome the data sparsity problem. We analyze this by computing all possible skip-grams in a training corpus and measure how many adjacent (standard) n-grams these cover in test documents. We examine skip-gram modelling using one to four skips with various amount of training data and test against similar documents as well as documents generated from a machine translation system. In this paper we also determine the amount of extra training data required to achieve skip-gram coverage using standard adjacent tri-grams.

1. Introduction

Recent corpus based trends in language processing rely on a single premise: that language is its own best model and that sufficient data can be gathered to depict typical (or atypical) language use accurately (Young and Chase, 1998; Church, 1998; Brown, 1990). The chief problem for this central tenet of modern language processing is the data sparsity problem: that language is a system of rare events, so varied and complex, that we can never model all possibilities. Language modelling research uses smoothing techniques to model these unseen sequences of words, yet even with 30 years worth of newswire text, more than one third of all trigrams have not been seen (Allison et al., 2006).

It therefore falls to the linguist to exploit the available data to the maximum extent possible. Various attempts have been made to do this, but they largely consist of defining and manipulating data beyond the words in the text (part-of-speech tags, syntactic categories, etc.) or using some form of smoothing to estimate the probability of unseen text. However, this paper posits another approach to obtaining better model of training data relying only on the words used: the idea of skip-grams.

Skip-grams are a technique largely used in the field of speech processing, whereby n-grams are formed (bi-grams, tri-grams, etc.) but in addition to allowing adjacent sequences of words, we allow tokens to be "skipped". While initially applied to phonemes in human speech, the same technique can be applied to words. For example, the sentence "I hit the tennis ball" has three word level trigrams: "I hit the", "hit the tennis" and "the tennis ball". However, one might argue that an equally important trigram implied by the sentence but not normally captured in that way is "hit the ball". Using skip-grams allows the word "tennis" be skipped, enabling this trigram to be formed. Skip-grams have been used many different ways in language modelling but often in conjunction with other modelling techniques or for the goal of decreasing perplexity (Goodman, 2001; Rosenfeld, 1994; Ney et al., 1994; Siu and Ostendorf, 2000).

The focus of this paper is to quantify the impact skip-gram modelling has on the coverage of trigrams in real text and compare this to coverage obtained by increasing the size of the corpus used to build a traditional language model.

.

2. Defining skip-grams

We define k-skip-n-grams for a sentence $w_1 \ldots w_m$ to be the set

$$\{w_{i_1}, w_{i_2}, \ldots w_{i_n} \mid \sum_{j=1}^{n} i_j - i_{j-1} < k\}$$

Skip-grams reported for a certain skip distance *k* allow a total of *k* or less skips to construct the n-gram. As such, "4-skip-n-gram" results include 4 skips, 3 skips, 2 skips, 1 skip, and 0 skips (typical n-grams formed from adjacent words).

Here is an actual sentence example showing 2-skip-bi-grams and tri-grams compared to standard bi-grams and trigrams consisting of adjacent words for the sentence:

"*Insurgents killed in ongoing fighting.*"

Bi-grams = {insurgents killed, killed in, in ongoing, ongoing fighting}.
2-skip-bi-grams = {insurgents killed, insurgents in, insurgents ongoing, killed in, killed ongoing, killed fighting, in ongoing, in fighting, ongoing fighting}
Tri-grams = {insurgents killed in, killed in ongoing, in ongoing fighting}.
2-skip-tri-grams = {insurgents killed in, insurgents killed ongoing, insurgents killed fighting, insurgents in ongoing, insurgents in fighting, insurgents ongoing fighting, killed in ongoing, killed in fighting, killed ongoing fighting, in ongoing fighting}.

In this example, over three times as many 2-skip-tri-grams were produced than adjacent tri-grams and this trend continues the more skips that are allowed. A typical sentence of ten words, for example, will produce 8 trigrams, but 80 4-skip-tri-grams. Sentences that are 20 words long have 18 tri-grams and 230 4-skip-tri-grams (see Table 1).

Bi-grams

Sentence Length	Bi-grams	1- skip	2-skip	3-skip	4-skip
5	4	7	9	10	10
10	9	17	24	30	35
15	14	29	30	50	60
20	19	37	54	70	85

Tri-grams

Sentence Length	Tri-grams	1- skip	2-skip	3-skip	4-skip
5	3	7	10	10	10
10	8	22	40	60	80
15	13	37	70	110	155
20	18	53	100	160	230

Table 1: Number of n-grams vs. number of k-skip n-grams produced

For an n word sentence, the formula for the number of trigrams with exactly *k* skips is given by:
(*n* - (*k* + 2)) (*k*+ 1), for *n*>*k*+3
But, we use *k*-skip gram to mean *k* skips or less for an *n* word sentence, which can be written as:

$$n\sum_{i=1}^{k+1} i - \sum_{i=1}^{k+1} i(i+1), \quad for\ n > k+2$$
$$= \frac{(k+1)(k+2)}{6}(3n-2k-6)$$

The tables and equations above illustrate that over 12 times as many tri-grams can be generated for large sentences using skip-tri-grams. This is a lot of extra contextual information that could be very beneficial provided that these skip-grams truly expand the representation of context. If a large percentage of these extra tri-grams are meaningless and skew the context model then the cost of producing and storing them could be prohibitive. Later in this paper we attempt to test whether this is the case.

3. Data

3.1. Training data

We constructed a range of language models from each of two different corpora using skip-grams of up to 4 skips:
British National Corpus- The BNC is a 100 million word balanced corpus of British English. It is contains written text and spoken text from a variety of sources and covering many domains.
English Gigaword- The Gigaword English Corpus is a large archive of text data acquired by the Linguistic Data Consortium. The corpus consists of over 1.7 billion words of English newswire from four distinct international sources.

3.2. Testing data

We used several different genres for test data in order to compare skip-gram coverage on documents similar and anomalous to the training.
300,000 words of news feeds- From the Gigaword Corpus
Eight Recent News Documents- From the Daily Telegraph.
Google Translations
Seven different Chinese newspaper articles of approximately 500 words each were chosen and run through the Google automatic translation engine to produce English texts. Web translation engines are know for their inaccuracy and ability to generating extremely odd phrases that are often very different from text written by a native speaker. The intention was to produce highly unusual texts, where meaning is approximately retained but coherence can be minimal. A short sample follows:
BBC Chinese net news: CIA Bureau Chief Gauss told USA the senator, the card you reaches still is attempting to avoid the American information authority, implemented the attack to the American native place goal. Gauss said, the card you will reach or if have the relation other terrorist organizations sooner or later must use the biochemistry or the nuclear weapon attack USA, this possibly only will be the time question. But he said, the card you reach only only are a holy war organization more widespread threat on the one hand.

4. Method

Skip-gram tests were conducted using various numbers of skips, but skips were never allowed to cross sentence boundaries. Training and test corpora were all prepared by removing all non-alphanumeric characters, converting all words to lowercase and replacing all numbers with the <NUM> tag.

We quantify the increase in coverage attained when using skip-grams on both similar and anomalous documents (with respect to the training corpus). To achieve this we compute all possible skip-grams in the training corpus and measure how many adjacent n-grams these cover in test documents. These coverage results are directly comparable with normal n-gram coverage in an unseen text results because we still measure coverage of standard adjacent bi-grams or tri-grams in the test documents and are only collecting skip-grams from the training corpus.

5. Results

5.1. Coverage

Our first experiment (Figure 1, Table 2) illustrates the improvement in coverage achieved when using skip-grams compared to standard bi-grams. We trained on the entire BNC and measured the coverage of k-skip bi-grams

on 300 thousand words of newswire from the Gigaword corpus. The BNC is made up of many different kinds of text other than news, but nonetheless, coverage is still improved. However, in some sense, the results are unsurprising, as there are many more bi-grams observed in training when allowing skips, but it does show that enough of these are legal bi-grams that actually occurred in a test document.

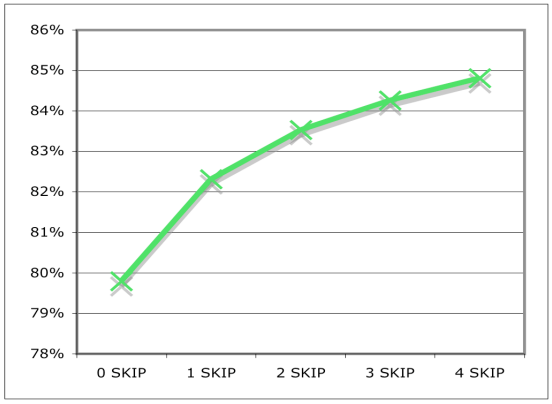

Figure 1: coverage of *k*-skip bi-grams on 300,000 words of news wire

Skips	# of Grams (filesize)	Unfound Bi-Grams	Unique Unfound Bi-grams	Coverage
0	88 M (961M)	58421	49253	79.80%
1	172M (1.9G)	51135	43177	82.32%
2	250M (2.7G)	47635	40350	83.53%
3	323M (3.5G)	45508	38556	84.26%
4	393M (4.2G)	43908	37191	84.82%

Table 2: *k*-skip bi-gram coverage

The next test (Figure 2, Table 3) is the same as the previous, but using tri-grams instead of bi-grams. From these results it seems that skip-grams are not improving tri-gram coverage to a very acceptable level, but as the later results show, this seems to be due to the fact that the BNC is not a specialized corpus of News text. Computing skip-grams on training documents that differ from the domain of the test document seems to add very little to the coverage. This is a promising result, in that it shows that generating random skip-grams from any corpus does not aid in capturing context.

Skips	# of Grams (filesize)	Unfound Tri-Grams	Unique Unfound Tri-Grams	Coverage
0	83 M (1.4G)	153990	138704	45%
1	239M (3.8G)	141618	127536	49.66%
2	457M (7.3G)	134715	121378	52.11%
3	729M (12G)	130521	117543	53.60%
4	1Bill (17G)	127862	115047	54.55%

Table 3: *k*-skip tri-gram coverage

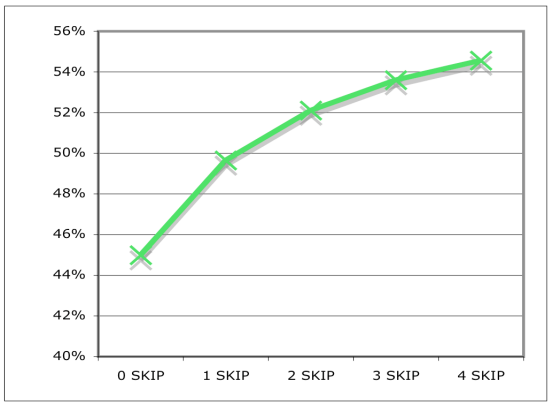

Figure 2: coverage of *k*-skip tri grams on 300,000 words of news wire

5.2. Skip-gram usefulness

Documents about different topics, or from different domains, will have less adjacent n-grams in common than documents from similar topics or domains. It is possible to use this fact to pick documents that are similar to the training corpus based on the percentage of n-grams they share with the training corpus. This is an important feature of n-gram modelling and a good indication the context is being modelled accurately. If all documents, even those on very different topics, had approximately the same percentage of n-grams in common with the training data then we would argue that it is not clear that any context is really being modelled. The use of skip-gram to capture context is dependent upon them increasing the coverage of n-grams in similar documents, while not increasing the n-gram coverage in different (or anomalous) documents to the extent that tri-grams can no longer be used to distinguish documents. We tested this by training on the BNC and testing against British newspaper extracts and texts generated with Google's Chinese to English translation engine (the genre is the same, but the text is generated by an MT system).

The results (Table 4) not only illustrate the difference between machine translated text and standard English, they also show that as skip distance increases coverage increases for all documents, but it does not increase to the extent that one cannot distinguish the Google translations from the News documents. These results demonstrate that skip-grams are accurately modelling context, while not skewing the effects of tri-gram modelling. It seems that most of the skip-grams produced are either useful or they are too random to give false positives.

Subject	0-skip	2-skip	3-skip	4-skip
NEWS 1	47.70%	56.69%	58.66%	62.11%
NEWS 2	55.40%	63.97%	65.67%	66.82%
NEWS 3	56.40%	60.61%	62.29%	65.24%
NEWS 4	52.68%	59.52%	62.04%	66.27%
NEWS 5	58.23%	63.80%	66.60%	71.58%
NEWS 6	54.17%	61.00%	62.95%	65.97%
NEWS 7	54.57%	61.86%	65.81%	70.48%
NEWS 8	56.23%	63.49%	65.75%	71.95%
Average	**54.42%**	**61.37%**	**63.72%**	**67.55%**

Translation 1	37.18%	45.56%	47.93%	50%
Translation 2	15.33%	23.64%	25.45%	22.61%
Translation 3	32.74%	40.22%	42.66%	45.28%
Translation 4	37.01%	33.07%	35.87%	38.05%
Translation 5	33.50%	38.09%	40.70%	42.24%
Translation 6	31.75%	39.20%	41.92%	42.71%
Translation 7	34.26%	38.54%	41.76%	42.52%
Average	**31.68%**	**36.90%**	**39.47%**	**40%**

Table 4: *k*-skip tri-gram coverage on English news and machine translated Chinese news

5.3. Skip-grams or more training data

Often, increasing the size of your training corpus is not an option due to lack of resources. In this section we examine skip-grams as an alternative to increasing the size of training data. The following experiments use different sized portions of the Gigaword corpus as training and a separate randomly chosen 300-thousand word blind section of the Gigaword for testing. We increase the amount of training and compare the results to using skip-grams for coverage. The resulting percentages are very high for trigram coverage, which is not surprising since both training and test documents come from the same domain specific corpus.

Size of Training	base tri-gram coverage	4 skip tri-gram coverage
10 M words	44.36%	53.76%
27.5 M words	53.23%	62.59%
50 M words	60.16%	**69.04%**
100 M words	65.31%	74.18%
200 M words	**69.37%**	79.44%

Table 5: Corpus size vs. skip-gram coverage on a 300,000 word news document.

This experiment demonstrates that skip-grams can be surprisingly helpful when test documents are similar to training documents. Table 5 illustrates that using skip-grams can be more effective than increasing the corpus size! In the case of a 50 million-word corpus, similar results are achieved using skip-grams as by quadrupling the corpus size. This shows an essential use of skip-grams to expand contextual information when training data is limited.

6. Conclusion

We have shown that there is a definite value to using skip-grams to model context. Our results demonstrate that skip-gram modelling can be more effective in covering tri-grams than increasing the size of the training corpus (even quadrupling it), while also keeping misinformation to a minimum. Although skip-grams can generate useless n-grams, these tend not to affect the coverage of n-grams in dissimilar documents. The disadvantage of skip-gram modelling is the sheer size of the training model that can be produced. This can lead to a large increase in processing time that should be leveraged against the time taken to extend the size of the training corpus. In cases where increasing the size of the training data is not an option because of expense or availability, skip-grams significantly lessen the data sparsity problem.

7. References

Ben Allison, David Guthrie, Louise Guthrie, Wei Liu, Yorick Wilks. Quantifying the Likelihood of Unseen Events: A further look at the data Sparsity problem. Awaiting publication, 2005

Peter F. Brown, John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin.1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79--85, June.

Kenneth Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136--143.

Joshua Goodman. 2001. A Bit of Progress in Language Modelling. *Computer Speech and Language*, October 2001, pages 403-434.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer, Speech, and Language*, 8:1-38.

Ronald Rosenfeld. 1994. Adaptive Statistical Language Modelling: A Maximum Entropy Approach. Ph.D. thesis, Carnegie Mellon University, April.

Manhung Siu and Mari Ostendorf. 2000. Variable n-grams and extensions for conversational speech language modelling. *IEEE Transactions on Speech and Audio Processing*, 8:63--75.

S.J. Young and L.L. Chase. 1998. Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes, *Computer Speech and Language*, 12, 263-279.