Summery

A Closer Look at Skip-gram Modelling

1 Introduction

Data sparsity is a large problem in NLP. A language is a system of rare events, which are complex and vary a lot. The n-gram model becomes more sparse with increasing n. Skip-grams try to address this problem by allowing to skip tokens.

2 Defining skip-grams

The k-skip-n-grams for a sentence $w_1 \dots w_m$ is defined as the set

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

Example: *"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 }

For a sentence with n words the number of k-skip grams of k skips or less and $n < k + 2$ is given by,

$$\frac{(k+1)(k+2)}{6}(3n-2k-6)$$

If skip-grams are a good representation of context, then they are very beneficial otherwise they might skew the context model.

3 Data

3.1 Training data

British National Corpus

- 100 million words
- written and spoken text
- various sources
- many domains

English Gigaword

- 1.7 billion words
- news texts

3.2 Testing data

300,000 words of news feeds Gigaword corpus

Eight News Documents Daily Telegraph

Google Translate Seven Chinese newspaper articles translated.

4 Method

Skips do not expand over to the next sentence. All data was preprocessed by removing all non-alphanumeric characters, lower case and replace numbers with <Num>.

All skip-grams are computed from the training corpus and all adjacent n-grams from the test document. The coverage of the n-grams over skip-grams is measured.

5 Results

5.1 Coverage

Trainging on BNC. Measured the coverage of k-skip bi-grams on 300k words from Gigaword.

Skips	Coverage
0	$\sim 79\%$
1	$\sim 82\%$
2	$\sim 83\%$
3	$\sim 84\%$
4	$\sim 84\%$

k-skip tri-grams

Skips	Coverage
0	$\sim 45\%$
1	$\sim 49\%$
2	$\sim 52\%$
3	$\sim 53\%$
4	$\sim 54\%$

This result is interpreted to be caused by the fact that training was not performed on a specialized news corpus. Therefore, context might be captured if random skip-grams do not perform well.

5.2 Skip-gram usefulness

The distribution of adjacent n-grams is similar for documents with the same topic, otherwise hardly any context would have been modeled. The use of skip-grams 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 documents to the extent that tri-grams can no longer be used to distinguish documents. Use of Chinese text and Google machine translation.

5.3 Skip-grams or more training data

Another experiment to make the point of the conclusion.

6 Conslusion

Skip-grams are more efficient in covering tri-grams than increasing the size of the training corpus (even quadrupling it).