A compression-based method for ranking n-gram differences between texts

W. J. Teahan

1. Introduction

This paper presents a new method for ranking n-gram differences between two or more texts. The method uses a relative entropy based approach to rank the n-grams (words, bigrams, trigrams) that appear in the texts, with the most 'unusual' being ranked higher in terms of the difference in entropy as measured by the cost of encoding the n-grams with respect to each individual text. The method can be used as the basis for producing tag clouds and is effective at revealing which topics are different between two or more texts. When a common reference corpus (such as the Brown Corpus of American English) is compared against a set of texts taken from a continuous sequence (such as American Inaugural Addresses), the method has also been found effective at revealing trends and emerging topics.

2. The method

The method uses a simple naïve estimate for the probability of each n-gram based on its frequency of use in each text:

$$P_T(g) = C_T(g) / N_T$$

where: $P_T(g)$ is the probability of the n-gram g in the text T; $C_T(g)$ is the frequency of the n-gram, and N_T is the total number of n-grams of the same length (i.e. unigrams, bigrams, trigrams) in the text T. The relative entropy based distance metric used for ranking the 'unusualness' of each n-gram g that appears in both texts T_1 and T_2, is calculated as follows:

$$D_{T_1,T_2}(g) = \left| H_{T_1}(g) - H_{T_2}(g) \right| = \left| -\log_2 P_{T_1}(g) - \log_2 P_{T_2}(g) \right|.$$

From a compression perspective, this measure (which we call '*codelength difference'*) is simply the absolute difference in compression codelengths, the costs of encoding the n-gram using two different naïve models, one trained on the text T_1 and the other trained on the text T_2. The codelength is a measure the "information" (or surprise) for an n-gram compared to the other n-grams.

For example, we can calculate the codelength for encoding the word unigram "Britain" for the balanced Brown Corpus of American English H_{Brown} as follows:

$$H_{Brown}("Britain") = -\log_2 P_{Brown}("Britain") = -\log_2 (61/1014416) = 14.021$$

since the word "Britain" occurs 61 times in 1,014,416 words. In contrast, the word "Britain" occurs 290 times in 1,010,401 words for the Lancaster-Oslo-Bergen (LOB) Corpus of British English:

$$H_{LOB}("Britain") = -\log_2 P_{LOB}("Britain") = -\log_2 (290/1010401) = 11.767.$$

We can use the absolute difference between the two codelength values as a means to measure how unusual the difference in probability is for the two corpora:

$$H_{Brown,LOB}("Britain") = \left| H_{Brown}("Britain") - H_{LOB}("Britain") \right| = \left| 14.021 - 11.767 \right| = 2.254.$$

Table 1 lists the top 20 codelength difference values for words, bigrams, and trigrams that appear in both the Brown corpus and the LOB corpus.

$H_{Brown,LOB}$	**Word**	$H_{Brown,LOB}$	**Bigram**	$H_{Brown,LOB}$	**Trigram**
4.802	Francisco	8.517	– and	5.398	the United Kingdom
4.776	Mercer	7.894	– the	4.761	the centre of
4.762	federal	5.888	the Labour	4.591	the Prime Minister
4.761	geese	5.560	United Kingdom	4.591	that the Government
4.761	Cecil	5.549	toward the	4.518	in favor of
4.749	polynomial	5.498	centre of	4.465	to ensure that
4.749	downtown	5.432	favour of	4.465	in respect of
4.749	Andy	5.215	the centre	4.454	the New York
4.745	toward	4.948	of state	4.398	of State for
4.706	Crown	4.852	the Negro	4.398	in England and
4.695	Franklin	4.734	the Company	4.254	the Earl of
4.695	Alex	4.695	the District	4.254	no need to
4.638	Kansas	4.638	San Francisco	4.254	House of Commons
4.638	Dartmouth	4.620	the Prime	4.176	to the British
4.638	Chandler	4.591	ensure that	4.164	plane of the
4.603	Commonwealth	4.529	respect of	4.164	THE EDITOR OF
4.591	buckling	4.518	in favor	4.164	EDITOR OF THE
4.579	neighboring	4.518	favor of	4.082	the plane of
4.579	dancer	4.498	the Minister	3.913	that at the
4.579	SAM	4.486	the anode	3.901	in New York,

Table 1. Top 20 words, bigrams and trigrams that appear in both the Brown corpus and LOB corpus ranked in descending order according to the codelength difference measure.

For these results, each word has been defined as any consecutive sequence of non white space characters (including punctuation) up until the next white space. The table shows that the method ranks proper nouns such as "*Francisco*" and "*Mercer*" highly – "*Francisco*" because it is contained in the name of the city San Francisco (which understandably appears much less frequently in the LOB corpus of British English), and "*Mercer*" because one of the samples in the Brown Corpus contained a story about Johnny Mercer, the American lyricist, songwriter and singer, where the word "*Mercer*" appeared frequently.

The differences between American English and British English is more revealing when bigrams or trigrams are used in the analysis. Here again, proper name bigram sequences such as "*The Labour*", "*United Kingdom*" "*the Negro*" and "*San Francisco*" are ranked highly, and similarly for trigrams, such as "*the United Kingdom*", "*the Prime Minister*", "*the New York*" and "*in New York*,". However, the well-known differences in spelling between American and British English are also revealed, with bigrams such as "*centre of*", "*favour of*" and "*in favor*", and the trigrams "*the centre of*" and "*in favor of*" appearing.

Although it is often useful to rank all the n-grams together using the absolute codelength difference measure (especially when using the method to reveal trending topics in a stream of texts – see below), it is also sometimes useful to create two separate ranking lists for when the codelengths for the first text is much greater than for the second text, and vice versa. Figure 1 depicts a visualisation of two separate ranking lists of trigram codelength differences produced from the LOB and Brown corpora for the values $H_{LOB} - H_{Brown}$ (shown on the left next to the red circles) and $H_{Brown} - H_{LOB}$ (shown on the right next to the blue circles). The size of the coloured circles reflect the magnitude of the codelength difference value, with the highest (and largest) values appearing at the top of each list.

Figure 1. Trigram tag list produced using the codelength difference measure on the Brown corpus and the LOB corpus. The top ranked trigrams according to the value $H_{LOB} - H_{Brown}$ is shown with the red circles on the left, and according to the value $H_{Brown} - H_{LOB}$ is shown with the blue circles on the right.

Figure 1 clearly shows the different trigram phrases in common use for the American and British dialects, and as a result, is effective at revealing different topics of interest that appear in the two respective texts. Phrases with proper names such as "*the United Kingdom*", "*the Prime Minister*" and "*in England and*" are clearly British; whereas the phrases "*the New York*", "*the American people*" and "*the State Department*" are clearly American. Also, again the difference in spelling appears prominently – the British spelling of "*the centre of*" is ranked second in the left list, and the American spelling of "*in favor of*" appears first in the second list.

3. Producing tag clouds

Tag clouds (also called word clouds) are a common method used to provide a visual representation of textual data. Tags of more weight or 'importance' are depicted more prominently by increasing their font size. One particularly useful application of the method described in section 2 is to use the codelength difference values to calculate the sizes of the tags in a tag cloud of the top-ranked n-grams.

In order to produce a tag cloud from the codelength difference values, the n-grams are randomly placed (as long as there is space left in the visualisation area) in ranking order with the highest first until the minimum codelength difference threshold value has been reached, after which no further n-grams will be included in the visualisation. The font size *f* of the n-gram tag is calculated as follows:

$$f(g) = 2^{\frac{1}{k} \times H_{Brown,LOB}(g)}$$

where k is a divisor constant that can be increased in relation to the ranking order so that greater prominence is given to the highest ranked n-grams. If it is set to 1, it will reflect the raw codelength difference scores. If it is set slightly higher (typically around 1.3 or 1.4), then the size of the lower

ranked n-grams will diminish more quickly.

Figure 2 shows the trigram tag cloud produced using codelength differences for the Brown and LOB corpora. The red tags are for trigrams that appear more prominently in the LOB corpus compared to the Brown corpus, whereas the blue tags are for trigrams that appear more prominently in the Brown corpus rather than the LOB corpus. The intensity of the colours is reduced as the size of the tag is reduced. The figure provides an effective method for visualising the data provided in Table 1 and Figure 1. It clearly reveals the importance of such phrases "*the United Kingdom*", "*the Prime Minister*" and "*the New York*", and also helps reveal how the languages between the two texts differ in a significant way.

Figure 2. Trigram tag cloud produced using the codelength difference measure on the Brown corpus and the LOB corpus. The minimum codelength difference threshold has been set at 3.5 and the initial tag font size divisor *k* at 1.3. The red coloured tags are for trigrams that appear with greater probability in the LOB corpus whereas the blue coloured tags are for those that appear with greater probability in the Brown Corpus.

As another example, the text of the Inaugural Addresses of American Presidents was analysed using this technique. The results are shown in Figures 3 and 4. We can compare language used in the first period of Inaugural Addresses to that used in the more recent speeches. Figure 3 shows the unigram tag cloud produced using the codelength difference method for the texts containing the first ten and the last ten Inaugural Addresses. In this example, the red coloured tags are for unigrams that appear more prominently in the first ten speeches, whereas the blue coloured tags are for unigrams that are more prominently in the last ten speeches. For example, the unigram "*Constitution*" features much more prominently in the last ten speeches compared to the first ten; similarly, the words "*America*" and "*children*" has higher importance in the first ten speeches than in the last ten.

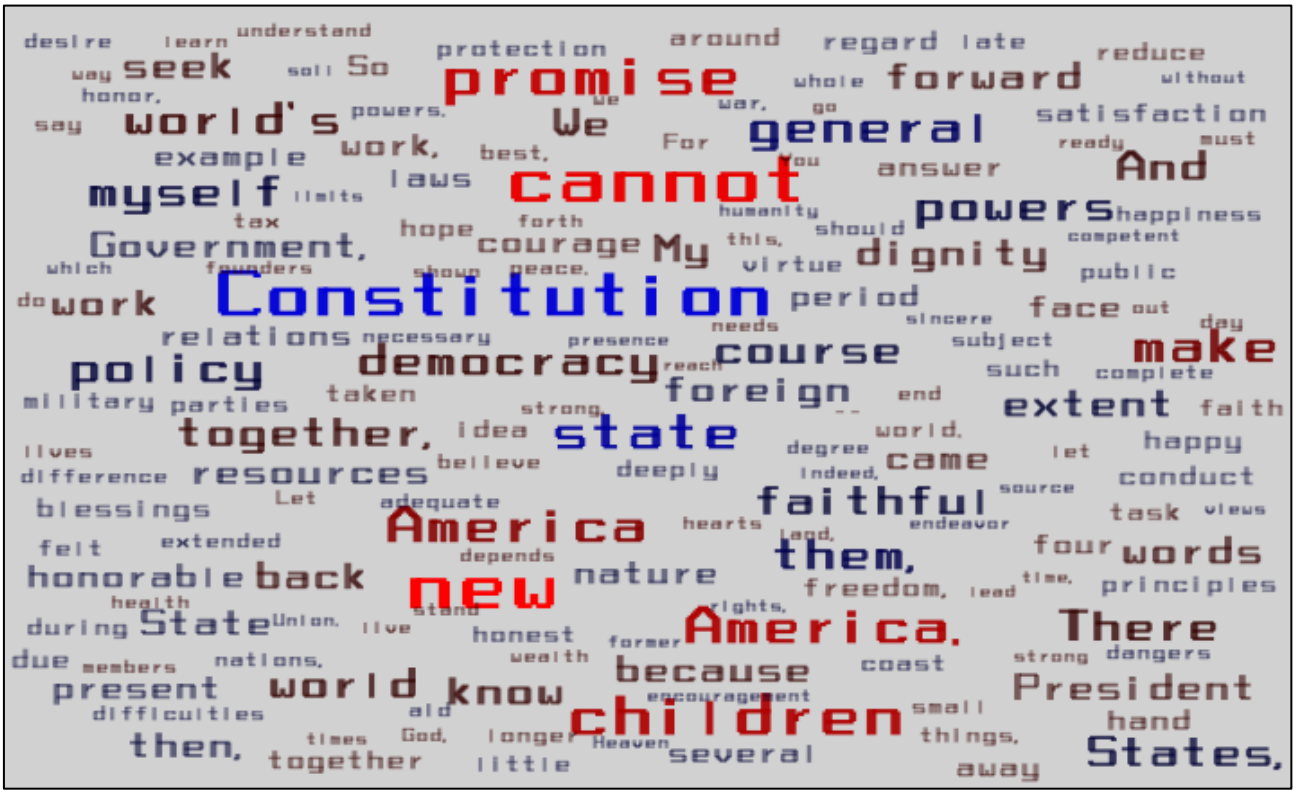

Figure 3. Unigram tag cloud produced using the codelength difference measure on the first and last ten U.S. president inaugural addresses. The minimum codelength difference threshold has been set at 2.2 and the initial tag font size divisor *k* at 1.05. The red coloured tags are for unigrams that appear with greater probability in the first ten speeches whereas the blue coloured tags are for those that appear with greater probability in the last ten.

This method of producing tag clouds is also useful to highlight trends or to reveal emerging topics of interest that appear when a stream of texts is analysed in sequence. One approach is to use a common reference corpus of standard language use as the first text, and then use this to compare against sub-texts while processing sequentially a second stream of text. For example, we can split the American Inaugural Addresses into four equal-sized periods each containing fourteen speeches (since there are $14 \times 4 = 56$ speeches to date). We can use the Brown Corpus of American English as the reference corpus, since it represents a balanced sample of American English text from the 1960s. We can then apply the codelength difference method to rank n-grams for each of the four periods. The resulting unigram tag clouds are shown in Figure 4. The figure depicts the four periods in question –Washington (1789) to Harrison (1841) on the top left; Polk (1845) to McKinley (1897) on the top right; McKinley (1901) to Eisenhower (1953) on the bottom left; and Eisenhower (1957) to Obama (2009) on the bottom right.

Figure 4 provides a useful treasure trove of information that helps to reveal the changes in topics considered important by American Presidents. In the early period of American history, the Presidents considered for example “Government”, “neutrality” and “aboriginal” as important issues, with the later word particularly interesting since its use has largely disappeared in modern usage (for example, the word appears only once in the one million word Brown Corpus of American English in the 1960s). In the middle two periods, words such as “Constitution” and “Democracy” become important, and interestingly the word “negro” appears in both, a word that is non-PC in modern usage. The word “Nation” also appears in the last two periods. In the last period, more words appear which may indicate that the Presidents are focusing on more or different issues.

Note that only a single word – “she” – has been coloured blue, appearing in the top two tag clouds for the first two periods. This indicates that the Brown Corpus provides an effective means for filtering out common American English usage as it provides a balanced sampling of the language. The word “she”

appears prominently as this reflects sexist language use (such as the use of “he” and “him” as impersonal pronouns).

Figure 4. Unigram tag cloud produced using the codelength difference measure on different periods of 14 consecutive U.S. president inaugural addresses. Top left is for the period Washington (1789) to Harrison (1841). Top right is for the period Polk (1845) to McKinley (1897). Bottom left is for the period McKinley (1901) to Eisenhower (1953). Bottom right is for the period Eisenhower (1957) to Obama (2009). The minimum codelength difference threshold has been set at 6.0 and the initial tag font size divisor *k* at 1.10.

3. Conclusions and future work

A new compression-based method for ranking n-gram differences between texts has been proposed. The method can readily be applied to producing n-gram tag clouds and these have been found to be effective at highlighting differences in topics. By using the codelength difference method to compare how a text stream changes over time, the method can be used to reveal trends or emerging topics of interest.

One of the limitations of the method is that it requires n-grams to appear in both texts being compared since the method requires an estimate of the n-gram probabilities to be made. This problem is called the ‘zero frequency problem’, a problem that is well-known in natural language processing. One solution is to use some method of smoothing the probabilities such as back-off estimation or escaping as used in the PPM compression scheme. Another solution is to treat unique n-grams separately since the fact they are unique is an important factor perhaps best dealt with in a different manner. Both solutions are currently being investigated as future research.