CAN STATISTICAL LANGUAGE MODELS BE USED TO DISTINGUISH BETWEEN DIFFERENT GENRES OF NEWS ?

Sabine Dreibe* and Gordon Hunter
Faculty of Science, Engineering and Computing, Kingston University, KT1 2EE, U.K.

* Now at King's College London, Strand, London WC2, U.K.

1 INTRODUCTION

Statistical Language Models (SLMs) have found widespread applications in many fields, including Automatic Speech Recognition systems, Automated Translation systems, and Cryptographic Analysis. It was been previously observed that lexical unigram, bigram and trigram distributions, which form the foundations of such SLMs, heavily depend on the type of data from which they were acquired – popular or serious literature, news, non-fiction text, formal speeches and structured or spontaneous dialogue. It has also been proposed that the lexical distributions also heavily depend on the theme or topic within each of the above styles of language.

In this paper, we investigate the extent to which such distributions vary between two different types of news – business and sports – within a dataset compiled by the BBC. We discuss our findings, particularly focusing on whether such models could form the basis of an automated genre or topic detector or classifier for news text or broadcasts.

2 STATISTICAL LANGUAGE MODELS

2.1 History and Essentials

It was established hundreds of years ago that natural language is not "uniform". Some words are much more common than others – for example, the verb form "is" is much more common than the verb form "establishes" – and some written letter characters are more common than others, for instance the characters "e","t" and "s" are much more common in written English than are "q", "x" or "z". The latter property was used as early as 1586 by Queen Elizabeth I's spymaster Sir Francis Walsingham (or rather, his cryptography advisor, Thomas Phelippes) to uncover the Babington Plot to overthrow Queen Elizabeth and replace her by her cousin, Mary Stuart, Queen of Scots (Dooley, 2013). This could perhaps be regarded as one of the first practical applications of statistical language modelling. In the 20th Century, linguists started to compare "stochastic" (statistical) models of language structure – in which the probabilities of each theoretically possible "next word" in a sequence such as a sentence can be estimated based on a model and statistical evidence from previous "experience" – with the "phrase structure grammar" of Chomsky (1957, 1965) and his followers, amongst other types of model. Whilst theoretical linguistics – particularly syntacticans and semanticians – have criticized purely statistical models due to the facts that they do not impose grammatical rules or distinguish between meaningful and nonsensical sentences, language engineers disagree (e.g. Young 1996, 2000) and statistical models have proved highly valuable at the core of applications such as Automatic Speech Recognition, Text Prediction and Machine Translation.

The basis of most types of SLM is the N-gram model (Jurafsky and Martin 2019) – based on occurrence statistics over previously observed data from appropriate "training" sources of sequences of N consecutive words. Individual words are known as unigrams, ordered pairs of consecutive words are called bigrams, whilst ordered triplets of consecutive words are called trigrams. Having compiled occurrence statistics of N-grams over a training "corpus" of data, these can be used to estimate probabilities of particular given words occurring at some specific point in a new document or

utterance. In the absence of any context information, we would make our “best guesses” for a word at a particular position in a new text or utterance purely based on unigram statistics – the most common words in the language currently under consideration would be the most likely to be present in the new document, whilst the rarest words would be the least likely. For example, the word “is” is very likely to occur several times in most documents, whilst “econometric” would be expected to be absent from most documents, apart from some which related to subjects related to economics. Longer N-grams allow the additional use of context information, using the Markov assumption (Jurafsky and Martin 2019) and Bayes’ Theorem – we can use bigram and unigram statistics to estimate the probably the probability of the second word in a pair, given that the first word has appeared. For example, the probability of the next word being “fast” would be higher if we knew that the previous word was “ran” than it would have been in ordinary circumstances without that contextual information. Similarly, trigram and bigram statistics can be used together to predict the third word in a sequence given the first two words. For example, if the previous two words were “the cat”, likely third words would include “sat”, “slept”, “ate”, “flap”, “burglar” and so on, whilst options such as “green”, “rat” and “dog” – which might have been quite probable in other contexts - would be unlikely candidates for the third word.

It is possible to consider N-grams for values of $N > 3$, but this is rarely done in practice. For a language with a vocabulary of V words, the number of theoretically possible N-grams is of the order of V^N, so for large values of V this will grow very rapidly with N, but most of those distinct N-grams will occur very rarely (if at all) for $N > 3$, making their occurrence probabilities very difficult to estimate accurately (Hunter 2004).

2.2 Sensitivity of Language Models to Genre and Topic

A number of previous authors (Rosenfeld 1996, 2000, Young 2000, Hunter 2004, Hunter & Huckvale 2006) have noted that both the statistics and performance of SLMs are highly sensitive to the nature of the material use to compile the model, and the material to which the model is applied. For example, suppose a model developed on specialist medical texts – say in relation to the study of cancers - were applied to a scenario whether the task was to recognize words uttered during the commentary of a football (soccer) match, the performance of the model would not be expected to be very good. The two domains have very different specialized vocabulary, and words such as “shot”, kick, pass, corner, keeper would be expected to be very common in a dataset relating to football, but not at all common in a dataset relating to the study of cancers. Similar features would be expected to be true for the lexical content of different types of news. These observations have been proposed as the basis of “topic classifiers”, for detecting the topic of a conversation, speech or text, for example for Finnish language material (Lagus & Kuusisto, 2002). Also, models developed from material relating to different topics or genres have been used to construct “adaptive” SLMs, where the model is “fine tuned” over time, to allow for the topic of current interest altering as time progresses – for example over the course of a lengthy conversation between two friends (Rosenfeld 1996).

In the remainder of this paper, we compare N-gram statistics for two genres of news material – sports news and business news – in a dataset available in the public domain. This was done with a view to using these statistics in a classifier to distinguish between these two topic types.

3 DATA AND MODELS USED IN THIS STUDY

The data used in this study came from a set of BBC news reports from 2005, now available in the public domain (Greene and Cunningham, 2006) and freely available from University College Dublin. We decided to focus on two distinct topic areas – namely Sports News and Business News. Both datasets contained reports, with the Business News dataset totaling 168 569 words, whilst the Sports News dataset contained 169 818 words in all. The sizes of the two datasets therefore only differed by 1249 words, or about 0.75% of the size of either dataset. Unigram (individual occurrences of particular words) bigram (pairs of successive words) and trigram (triplets of consecutive words) statistics were compiled for each dataset. The process of compiling these was aided by the use of

the tool *WordCounter* (Databasic.io, 2020). Two sets of N-gram statistics (for N = 1, 2, 3) were compiled for each dataset – the first including all distinct words observed, the second with the 80 most common "grammatical function" words (e.g. "a", "the", "this", "that", "and", "but", "is") – sometimes called "stop words" excluded, since these are very common in most situations and are not generally considered useful in identification of the genre or topic of a text or conversation. The unigram statistics for each dataset, and the rankings of the most common words, were first compared with a widely-accepted ranked list of the most common words in written English (Empire Skola, no date) – see Figure 1.

1. the	35. were	69. has
2. of	36. we	70. look
3. and	37. when	71. two
4. a	38. your	72. more
5. to	39. can	73. write
6. in	40. said	74. go
7. is	41. there	75. see
8. you	42. use	76. number
9. that	43. an	77. no
10. it	44. each	78. way
11. he	45. which	79. could
12. was	46. she	80. people
13. for	47. do	81. my
14. on	48. how	82. than
15. are	49. their	83. first
16. as	50. if	84. water
17. with	51. will	85. been
18. his	52. up	86. call
19. they	53. other	87. who
20. I	54. about	88. oil
21. at	55. out	89. its
22. be	56. many	90. now
23. this	57. then	91. find
24. have	58. them	92. long
25. from	59. these	93. down
26. or	60. so	94. day
27. one	61. some	95. did
28. had	62. her	96. get
29. by	63. would	97. come
30. word	64. make	98. made
31. but	65. like	99. may
32. not	66. him	100. part
33. what	67. into	
34. all	68. time	

Figure 1 : Ranked list of the most common 100 words in written English (from Empire Skola https://www.empire-skola.sk/data/USR_042_IMAGES/The_100_Most_Common_Written_Words_in_English.pdf). Although this list is not identical to others available through other sources, the "Top 20" are essentially the same in all cases, and variations between the available lists are quite slight.

The results are presented and discussed in the following section.

4 RESULTS AND DISCUSSION

4.1 Unigram Statistics

We firstly present the most common words in each dataset, including the "stop words". Table 1 gives the statistics – frequency of occurrence, fraction (percentage) of occurrence within that dataset (relative to the total number of words in that dataset), the ranking of that word within the dataset (Rank 1 being the most common), and the ratio of occurrences, relative to the most common word in that dataset.

Unigrams- Business and Sport News with stop words included									
Business News					*Sport News*				
word	**frequency**	**percentage**	**rank**	**ratio**	**word**	**frequency**	**percentage**	**rank**	**ratio**
the	10810	6.43460%	Rank 1	1	the	9628	5.70285%	Rank 1	1
to	5087	3.02801%	Rank 2	0.47058	to	4687	2.77620%	Rank 2	0.48681
of	4356	2.59289%	Rank 3	0.40296	a	3850	2.28043%	Rank3	0.39988
in	4311	2.56610%	Rank 4	0.39880	and	3678	2.17855%	Rank 4	0.38201
a	3423	2.03752%	Rank 5	0.31665	in	3656	2.16552%	Rank 5	0.37973
and	3212	1.91193%	Rank 6	0.29713	of	2807	1.66264%	Rank 6	0.29155
said	1676	0.99763%	Rank 7	0.15504	for	1744	1.03300%	Rank 7	0.18114
is	1625	0.96727%	Rank 8	0.15032	he	1614	0.95600%	Rank 8	0.16764
for	1620	0.96430%	Rank 9	0.14986	I	1596	0.94534%	Rank 9	0.16577
that	1575	0.93751%	Rank 10	0.14570	on	1506	0.89203%	Rank 10	0.15642
it	1417	0.84346%	Rank 11	0.13108	is	1490	0.88256%	Rank 11	0.15476
on	1384	0.82382%	Rank 12	0.12803	but	1443	0.85472%	Rank 12	0.14988
has	1256	0.74763%	Rank 13	0.11619	was	1419	0.84050%	Rank 13	0.14738
its	1112	0.66191%	Rank 14	0.10287	that	1208	0.71552%	Rank 14	0.12547
by	1091	0.64941%	Rank 15	0.10093	with	1200	0.71078%	Rank 15	0.12464
at	944	0.56191%	Rank 16	0.08733	it	1193	0.70664%	Rank 16	0.12391
as	923	0.54941%	Rank 17	0.08538	at	1170	0.69301%	Rank 17	0.12152
was	922	0.54882%	Rank 18	0.08529	his	1142	0.67643%	Rank 18	0.11861
with	921	0.54822%	Rank 19	0.08520	have	1142	0.67643%	Rank 19	0.11861
from	861	0.51251%	Rank 20	0.07965	has	965	0.57159%	Rank 20	0.10023

Table 1 : The most common words (Unigrams) in each dataset : Sports News and Business News, including the most common “function” or “stop” words.

It can be observed from Table 1 that, as far as “function” or “stop” words are concerned, the two datasets are fairly similar. The rankings and proportions of particular words are not identical in both lists, but most entries in those “top 20s” appear in both lists. However, there are exceptions : “said” was ranked 7th in the business news, accounting for almost 1% of the total words, but did not appear in the top 20 words in the sports news (it was actually ranked 22, accounting for about 0.5% of the word count). Conversely, “I” was ranked 9th in the sports news vocabulary (again accounting for just under 1% of the total for that set), but was only ranked 184 (accounting for less than 0.07%) of the business news text. Even at the level of very common words, there are some differences between the two datasets.

Unigram Statistics of Business News and Sport News without stop words										
Business News						***Sport News***				
Word	**Frequency**	**Percentage**	**Rank**	**Ratio**		**Word**	**Frequency**	**Percentage**	**Rank**	**Ratio**
said	1676	0.994%	Rank 1	1.000		said	932	0.55%	Rank 1	1.000
us	813	0.482%	Rank 2	0.485		first	481	0.28%	Rank 2	0.516
year	684	0.406%	Rank 3	0.408		game	478	0.28%	Rank 3	0.513
mr	592	0.351%	Rank 4	0.353		year	449	0.26%	Rank 4	0.482
would	465	0.276%	Rank 5	0.277		time	419	0.25%	Rank 5	0.450
also	442	0.262%	Rank 6	0.264		win	410	0.24%	Rank 6	0.440
1	435	0.258%	Rank 7	0.260		England	396	0.23%	Rank 7	0.425
market	434	0.257%	Rank 8	0.259		would	392	0.23%	Rank 8	0.421
new	411	0.244%	Rank 9	0.245		two	392	0.23%	Rank 9	0.421
growth	395	0.234%	Rank 10	0.236		last	384	0.23%	Rank 10	0.412
last	374	0.222%	Rank 11	0.223		world	382	0.22%	Rank 11	0.410
company	369	0.219%	Rank 12	0.220		6	382	0.22%	Rank 12	0.410
economy	345	0.205%	Rank 13	0.206		one	381	0.22%	Rank 13	0.409
firm	327	0.194%	Rank 14	0.195		back	375	0.22%	Rank 14	0.402
sales	320	0.190%	Rank 15	0.191		also	329	0.19%	Rank 15	0.353
economic	313	0.186%	Rank 16	0.187		players	301	0.18%	Rank 16	0.323
2004	313	0.186%	Rank 17	0.187		team	294	0.17%	Rank 17	0.315
bank	310	0.184%	Rank 18	0.185		cup	292	0.17%	Rank 18	0.313
could	306	0.182%	Rank 19	0.183		play	292	0.17%	Rank 19	0.313
oil	302	0.179%	Rank 20	0.180		new	289	0.17%	Rank 20	0.310

Table 2 : Unigram Statistics for the most common words in each of the two datasets – Business News and Sports News – with the 80 most common “function” or “stop” words excluded. Major contrasts can now be seen between the two datasets.

Once the 80 most common “stop words” have been excluded, it is clear that the lexical content differs quite considerably between the two datasets (see Table 2). Whilst “said” is now the most common word in both datasets, it is considerably more prevalent in the Business News dataset (0.994% of total words) than in the Sports News dataset (0.55% of total words). The rankings and prevalences of many other words differ radically between the two datasets, with “year” (ranked 3 for Business News and 4 for Sports News) and “would” – which possibly should have been considered to be a “stop word” – (ranked 5 in Business and 8 in Sport) being notable exceptions. “Us” was ranked second in Business News, but only 39th in Sports News, whilst “first” ranked second in Sports News, but was only 58th most common in the Business News data. Both these genres of News text appear to be somewhat atypical of written English – “said” is only ranked 40th in Figure 1, whilst “first” (the second most common non-stop word in the Sports News data) in only 83rd in Figure 1. Not surprisingly, the Sports News data has the words “game”, “win”, “players”, “team”, “cup”, “play” and (since it is BBC news) “England” highly ranked, whilst “market”, “growth”, “company”, “economy”, “sales”, “economic” and “bank” all feature prominently in the Business News dataset.

4.2 Bigram Statistics

Once again, we compiled ranked lists of bigrams, ordered by frequency of occurrence for each dataset. The most common of these are shown in Table 3.

Bigrams - Business News and Sport News with stop words included					
Business News			***Sports News***		
bigram phrase	**frequency**	**percentage**	**bigram phrase**	**frequency**	**percentage**
in the	1016	1.25088%	in the	1290	1.5008%
of the	996	1.22625%	of the	804	0.93538%
for the	400	0.49247%	for the	464	0.5398%
to the	390	0.48016%	at the	458	0.53284%
the US	381	0.46908%	on the	366	0.4258%
on the	330	0.40629%	to the	360	0.41883%
that the	306	0.37674%	to be	293	0.3409%
said the	270	0.33242%	will be	250	0.29085%
and the	251	0.30903%	it was	236	0.2746%
to be	246	0.30287%	with a	231	0.26875%
in a	234	0.28810%	the first	230	0.2676%
said it	214	0.26347%	he said	210	0.24432%
at the	211	0.25978%	has been	207	0.2408%
the company	211	0.25978%	with the	203	0.23617%
of a	206	0.25362%	in a	199	0.2315%
it is	202	0.24870%	it is	198	0.23036%
by the	195	0.24008%	and the	194	0.2257%
more than	195	0.24008%	year old	187	0.21756%
from the	192	0.23639%	from the	181	0.2106%
with the	182	0.22407%	and I	180	0.20941%

Table 3 : The most common bigram phrases in the Business News and Sports News datasets when the common “function” or “stop” words are included.

From Table 3, we can see that most of the bigrams in both datasets involve rather common words, and many of these are common to both lists, if not in exactly the same orders. Only a few of the “Top 20” bigrams for the Business News data seem particularly noteworthy : “the US” and “the company”. Further down the rankings, “the firm” (27th), “chief executive” (31st), “the government” (35th), “the economy” (48th) and “the country’s” (50th) are bigrams relevant to the nature of the dataset, whilst other such examples, including “the market”, “stock market”, “economic growth” and “interest rates” appear further down the rankings. Of the “Top 20” bigrams for Sports News, only “the first” could really be considered as particularly appropriate for this dataset, but “six nations” (for Rugby – 21st), “the game” (22nd), “the second” (38th), “to play” (42nd), “the club” (50th), “the final” (52nd), “the ball” (54th) and “the match” (65th) are relevant to the dataset. For only a few of these examples are both words of the bigram particularly related to the topic in question, suggesting that bigram statistics might give little additional advantage over a unigram-based model when trying to distinguish between, or classify, example documents from these datasets.

4.3 Trigram Statistics

Statistics of trigrams occurring in each dataset were compiled in a similar way to the bigram statistics. The "Top 20" results obtained for the two datasets are shown in Table 4.

Trigram Statistics for Business News and Sport News							
Business News				***Sport News***			
Trigram phrase	**frequency**	**percentage**	**TTC**	**Trigram phrase**	**frequency**	**percentage**	**TTC**
in the US	92	0.27002%	34071	a lot of	100	0.28076%	35618
one of the	53	0.15556%		in the first	71	0.19934%	
the end of	51	0.14969%		the end of	69	0.19934%	
according to the	51	0.14969%		in the second	65	0.19372%	
as well as	48	0.14088%		out of the	64	0.18249%	
in a statement	48	0.14088%		the six nations	63	0.17968%	
is expected to	48	0.14088%		one of the	62	0.17688%	
said it was	44	0.12914%		it was a	60	0.17407%	
said in a	42	0.12327%		in the world	57	0.16845%	
the bank of	39	0.11447%		to win the	46	0.16003%	
said that the	39	0.11447%		the Champions League	46	0.12915%	
as part of	38	0.11153%		told BBC sport	46	0.12915%	
in the UK	36	0.10566%		the Australian Open	44	0.12915%	
a number of	35	0.10273%		for the first	41	0.12353%	
of the US	34	0.09979%		in the final	39	0.11511%	
as a result	34	0.09979%		the first time	38	0.10950%	
has said it	33	0.09686%		end of the	37	0.10669%	
Bank of England	32	0.09392%		the second half	36	0.10388%	
said it would	31	0.09099%		of the season	35	0.10107%	
the world's biggest	31	0.09099%		the first half	34	0.09826%	

Table 4 : Trigram statistics for the Business News and Sports News datasets. TTC denotes "Total Trigram Count" for that dataset.

Although the frequency counts for even the most common trigrams are quite low – with no count exceeding 100 over either dataset, quite a number of the "Top 20" of each dataset do appear to be somewhat distinctive for that dataset – for example "Bank of England" for the Business News dataset, and "the Champions League" or "the Australian Open" for the Sports News dataset. This suggests that trigram statistics could prove useful for distinguishing between these genres.

5 CONCLUSIONS AND FUTURE WORK

We have compiled unigram, bigram and trigram statistics for two different themes of British English news text which are in the public domain – namely Business News and Sports News from the BBC. Whilst unigram (including "stop word") and bigram statistical distributions were found to be rather similar for both datasets, the statistics of unigrams (excluding "stop words") and trigrams from those same datasets were notably different. These could be used to form the basis of a "clustering" or "classification" system, making use of a Bayesian classifier (e.g. Peng et al 2004) or one of the lexically-based or entropy-based approaches to clustering described by Hunter & Huckvale (2006).

6 REFERENCES

1. J.F. Dooley, "A Brief History of Cryptology and Cryptographic Algorithms", Springer, p 21 https://doi.org/10.1007/978-3-319-01628-3 , ISBN 978-3-319-01628-3, 2013
2. N. Chomsky, "Syntactic Structures", Mouton, The Hague, Netherlands, 1957, ISBN 9027933855,
3. N. Chomsky, "Aspects of the Theory of Syntax", MIT Press, Cambridge, MA, USA,1965, ISBN 9780262260503
4. S. Young, "Large Vocabulary Continuous Speech Recognition: A Review", IEEE Signal Processing Magazine, 13 (5), 1996, pp 45 – 57, 1996
5. S. Young "Probabilistic Methods in Spoken Dialogue Systems", Philosophical Transactions of the Royal Society (London), Series A, Vol. 358, No. 1769, pp 1389-1402, 2000
6. D. Jurafsky & J.H Martin, "Speech and Language Processing", Chapter 3, 2019 (Advance draft of 3rd Edition ahead of publication, available from https://web.stanford.edu/~jurafsky/slp3/ Last accessed 11 September 2020)

6. D. Greene & P. Cunningham, Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 2006,
   BBC datasets available at http://mlg.ucd.ie/datasets/bbc.html
7. R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modelling", Computer Speech & Language, Vol. 10, pp187 – 228, 1996
8. R. Rosenfeld, "Two Decades of Statistical Language Modelling : Where do we go from here ?", Proceedings of the IEEE, Vol. 88 (8) pp 1270 - 1278, 2000
9. G.J.A. Hunter, "Statistical Language Modelling of Dialogue Material in the British National Corpus", PhD Thesis, UCL, University of London, 2004
10. G. Hunter & M. Huckvale, "Is it Appropriate to Model Dialogue in the Same Way as Text?", Proceedings of the European Modelling Symposium, London, U.K., September 2006, pp 199 – 203, 2006 ISBN 0951650939
11. K. Lagus & J. Kuusisto, J. (2002) "Topic Identification in Natural Language Dialogues Using Neural Networks", Proceedings of the 3rd SIGDial Workshop on Discourse & Dialogue (ACL-02), Philadelphia, USA, July 2002, pp 95-102
12. Databasic.io (2020) "Wordcounter", available at https://www.databasic.io/en/wordcounter/
13. Empire Skola "The 100 Most Common Words" in written English", https://www.empire-skola.sk/data/USR_042_IMAGES/The_100_Most_Common_Written_Words_in_English.pdf (Last accessed 11 September 2020)
14. F. Peng, D. Schuurmans & S. Wang, "Augmenting Naive Bayes Classifiers with Statistical Language Models", Information Retrieval, 7, pp 317 – 345, 2004. https://doi.org/10.1023/B:INRT.0000011209.19643.e2
15. G. Hunter & M. Huckvale, "Cluster-Based Approaches to the Statistical Modelling of Dialogue Data in the British National Corpus", Proceedings of 2nd IET International Conference on Intelligent Environments, Athens, Greece, 2006, https://doi.org/10.1049/cp:20060647