The Posit Text Profiling Toolset*

George R. S. Weir

Department of Computer and Information Sciences, University of Strathclyde

george.weir@cis.strath.ac.uk

Abstract

The present paper describes a software toolset developed to assist in making textual analyses and comparisons between and within corpora. This system (Posit) aims to make the textual analysis process as simple as possible by requiring only a single command from the user. The adopted approach seeks to accommodate arbitrarily large corpora. This contrasts with many current tools that are limited in their ability to handle very large file sizes. In the following, we detail the current part-of-speech focus of this toolset and describe developments in progress that will extend its functionality to embrace vocabulary and readability profiling.

Keywords

Text analysis, software, corpora comparisons.

Introduction

With a growing interest in the analysis of text collections, often in order to support language teaching (e.g., Tomlinson, 1998; Granger, et.al, 2002; Aston et al, 2004), has come a desirable increase in available software tools. Tools such as Wordsmith (Scott, 1998) and AntConc (Anthony, 2005) offer approachable means whereby non-computer specialists may analyse their own data collections. More ambitious facilities are freely available in systems such as Nooj (Silberztein, 2005), and GATE (Bontcheva, 2004), which are both under constant development. In addition, NLTK (Bird, 2006) provides a set of programming modules aimed mainly at teaching natural language processing.

When faced with the task of analysing a newspaper text corpus, with a file size exceeding 92Mb (Weir & Anagnostou, 2007), many of the existing tools designed to run under MS-Windows experienced problems in handling a single data file on this scale. In consequence, we sought an approach to analysis that would satisfactorily manage large input files. This led us to adopt a Unix-based scripting approach in which the input text is processed without need for large data structures in memory (the principal obstacle facing the MS-Windows' tools). A convenient advantage of the script approach is that intermediate processing results are held in temporary files, and this further reduces memory overheads. This approach proved both powerful and highly extensible and, as a result, the prospect of a highly functional set of analysis tools became apparent. The Posit Text Profiling Toolset is the term applied to this set of analysis tools, some of which are complete, while others are under development.

1 The Posit Toolset Overview

In its current form, the Posit toolset targets three related aspects of textual analysis, comprising individual software modules whose operation may be combined. The first of these modules concentrates upon parts-of-speech (POS) and performs an analysis of a given text corpus in order to derive statistics on the POS characteristics of that text. This component is known as the POS Profiler.

The second module of the toolset is the Vocabulary Profiler. Based upon the statistical data output by the POS Profiler, the Vocabulary Profiler can determine the relative frequency of occurrence for vocabulary items in the selected corpus. This frequency data may be compared to a reference set of frequency data (derived from the British National Corpus) in order to pinpoint unusual word occurrences or individual terms whose use is likely to prove unfamiliar to English readers.

A third toolset module (presently under development) is the Readability Profiler. This software component will focus on text readability, based upon the statistical analyses from the POS profiler and the frequency data from the vocabulary profiler. In keeping with our research in this area, this module will go beyond current 'simplistic' readability metrics, and apply more sophisticated analyses that include factors such as word commonality (Weir & Ritchie, 2007) and average collocation frequency (Anagnostou & Weir, 2007).

In the following sections, we focus attention on the nature and operation of the POS Profiler component and further outline our plans for the

Vocabulary Profiler and the Readability Profiler components within the Posit Text Profiling Toolset.

2 POS Profiler

The POS profiler supports part-of-speech profiling on any specified text. This command-line facility[1] outputs a detailed account of word occurrences for the selected text corpus. The word occurrence information is provided by raw frequency and by part-of-speech frequency. Totals are given for word tokens, word types, part-of-speech types and part-of-speech tokens. The set of parts-of-speech that can be recorded is a function of the POS tagger used within the POS Profile Tool. While the modular toolset can easily accommodate alternative taggers, we currently use the Lingua::EN::Tagger, which is available as a Perl module from CPAN (http://www.cpan.org). This tagger uses the Penn Treebank tag set (Marcus et al, 1994) so our scripts are currently equipped to collate the occurrence of the constituent tags from the marked-up version of the input corpus. In later versions of the Posit Toolset, users will be able to specify both the requisite tag set and also which range of constituent tags should be recorded when the system is creating the output frequency data.

Table 1: Example summary POS Profile output

Input filename	emma.txt
Total words (tokens)	159826
Total unique words (types)	7364
Type/Token Ratio (TTR)	21.7037
Number of sentences	8585
Average sentence length (ASL)	18.6169
Number of characters	914519
Average word length (AWL)	5.72197
NUMBER OF TOKEN TYPES	
noun_types	4268
verb_types	2603
adjective_types	1346
adverb_types	487
preposition_types	65
personal_pronoun_types	23
determiner_types	18
possessive_pronoun_types	7
interjection_types	5
particle_types	0
NUMBER OF POS TYPES	
nouns	69060
verbs	67678
prepositions	38600
personal pronouns	31192
determiners	26178
adverbs	25432
adjectives	25086
possessive pronouns	9582
interjections	516
particles	0

[1] A version of the Posit Toolset with graphical user interface is also under development.

Table 1, above, shows an example of output from the POS Profiler. This illustrates the summary output for the text of the novel 'Emma' by Jane Austen. In addition to such aggregated summary data, the POS Profiler also details frequency data for specific parts of speech. The module's default POS settings are listed in Table 2, below.

Table 2: POS Profiler default list of parts of speech

adjective_comparatives
adjective_or_numeral_ordinals
adjective_superlatives
adverb_comparative_form
adverb_form
adverb_superlative_form
common_nouns
determiners
interjections
modal_aux
nouns_common_plurals
nouns_proper_plural
particles
prepositions
pronouns_personal
pronouns_possessive
proper_nouns
verbs_base_form
verbs_gerund_form
verbs_past_form
verbs_past_participle_form
verbs_present_3rd_form
verbs_present_not3rd_form
wh_adverbs

A separate output file is created for each of these parts-of-speech and such files list the recorded words of this type. The listed words are ordered by frequency of occurrence and their frequency is included within the data. Table 3 illustrates the most frequently occurring common nouns from 'Emma'. This is extracted from the *common_nouns* output file.

Table 3: Example most frequent common nouns

Frequency	*Common noun*
397	thing
272	time
254	nothing
220	man
206	father
193	body
180	day
177	friend
154	way
132	cannot

In addition to deriving totals for tokens, types, number of sentences and number of characters, the POS Profiler also determines average sentence length and total number of characters. These factors facilitate the calculation of the Flesch Reading Ease

and Flesch-Kincaid Grade Level. An additional function is available for this calculation.

2.1 Process

The POS profiling facility is invoked at the command line on a specified text corpus. The input corpus is processed by the software module in accordance with the following sequence:

1. Create word count and token frequency list
2. Tag the input file using a POS tagger
3. Tokenize the POS tagged file
4. Extract POS counts
5. Analyse results
6. Output results
7. Create results summary

An extensive set of results files is generated by each single analysis run of the POS Profiler. The results summary, also saved as a separate file, is output to the screen as a conclusion to the analysis. As an aid to further comparison across results, files in comma separated value (CSV) format are also produced. Currently, the system converts all input text to lower case, thereby treating all data as case insensitive. (This fact will be apparent from the illustrations of example output.) Of course, this feature is customisable since, for some purposes, the capitalisation within a corpus may be considered significant. In this case, capitalisation can be retained and word counts, as well as other outputs, will recognise distinctive case differences.

3 Vocabulary Profiler

The Posit Vocabulary Profiler uses the analyses produced by the POS Profiler to establish the least common words in any text (with reference to the BNC reference list). This will shortly support the determination of keywords for the specified text, based upon a statistical significance measure of frequency of words in a specified text against frequency of words in a reference frequency list (by applying the log-likelihood measure). In addition, to the analytical value of such insights, this information may provide support or advice to authors wishing feedback on their vocabulary usage.

Vocabulary analysis extends to consider n-gram frequencies within the analysed text. N-gram frequency analysis allows a choice of value for *n* in the n-gram. By default, the system determines frequency lists for bigrams, trigrams and quadgrams. In due course, these may also be compared with reference n-gram frequencies derived from the British National Corpus. The result of quadgram frequency analysis on the text of the novel ‘Emma’ (by Jane Austen) gives the ‘top ten’ results shown in Table 4, below.

Table 4: Example quadgram frequency data

Frequency	*Quadgram*
50	i do not know
26	a great deal of
20	i am sure i
19	it would have been
18	mr and mrs weston
18	it would be a
18	i do not think
16	i have no doubt
16	i am sure you
15	and i am sure

Using the n-gram facility of the Vocabulary Profiler, we can readily contrast the quadgram result with the ‘top ten’ bigram result from the same text (Table 5).

Table 5: Example bigram frequency data

Frequency	*Bigram*
608	to be
566	of the
449	it was
446	in the
395	i am
334	she had
331	she was
308	had been
301	it is
299	mr knightley

4 Readability Profiler

The Readability Profiler component of the Posit Toolset is under development and builds upon the POS Profiler output and output from the Vocabulary Profiler. In turn, this readability module uses collocation analysis in order to establish the contribution made to the readability of the specified text by collocation usage. This analysis relies upon a collocation reference frequency list and applies our Average Collocation Frequency measure as a factor in determining the readability of the complete text. A future extension of this facility will support readability comparisons across texts.

5 Conclusion

The Posit Text Profiling Toolset comprises three software modules that work together to provide a comprehensive textual analysis facility. Built as a set of Unix scripts and Perl programs, the system provides a convenient interface to existing POS taggers and is able to accommodate large text corpora with ease.

In their current form, the POS Profiler and Vocabulary Profiler are being used in a variety of corpus analysis projects. We anticipate that future enhancements will include support for automated corpora comparisons. This will afford POS profile

comparisons, vocabulary comparisons and readability comparisons across sets of texts, in which multiple corpus files may be loaded, analysed and compared with a single user action. Currently, such corpus comparison requires individual processing of each corpus with manual comparison of the resulting data (as in Weir & Ozasa, 2007).

Beyond the future addition of further functional features to the Posit Toolset, a prototype version is under development that includes a graphical user interface. This will provide a more broadly usable system and will seek to simplify the selection of system options. This will greatly enhance the customisability of the toolset and should ensure that it becomes more accessible to a broader base of potential users.

References

Anagnostou, N.K. & Weir, G.R.S. (2007). Average Collocation Frequency as an Indicator of Semantic Complexity. *Proceedings of ICTATLL 2007*, Hiroshima, Japan.

Anthony, L. (2005). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning.

Aston, G., Bernardini, S. & Stewart, D. (2004). Corpora and Language Learners. John Benjamins, Amsterdam.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on interactive Presentation Sessions*. Association for Computational Linguistics, Morristown, NJ, 69-72.

Bontcheva, K., Tablan, V., Maynard, D. & Cunningham, H. (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering*, 10: 349-373.

Granger, S., Hung, J. & Petch-Tyson, S. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* John Benjamins, Amsterdam.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank., 19 (2), 313-330.

Scott, M. (1998). *Wordsmith Tools Version 3*, Oxford University Press, Oxford, UK.

Silberztein, M. (2005). NooJ: A Linguistic Annotation System for Corpus Processing. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, Vancouver, 10-11.

Weir, G.R.S. & Anagnostou, N.K. (2007). Exploring Newspapers: A Case Study in Corpus Analysis. *Proceedings of ICTATLL 2007*, Hiroshima, Japan, 12-19.

Weir, G.R.S. & Ozasa, T., (2007). Estimating Naturalness in Japanese English Textbooks. *Proceedings of PAAL2007*, Pattaya, Thailand.

Weir, G.R.S. & Ritchie, C. (2007). Estimating Readability with the Strathclyde Readability Measure. In *Texts, Textbooks and Readability***.** Edited by G. R. S. Weir and T. Ozasa. University of Strathclyde Publishing. Glasgow, UK, 26-33.

---

* In Proceedings of PAAL2007, Pattaya, Thailand, December 2007.