Mining a corpus of biographical texts using keywords

Mike Conway
National Institute of Informatics, Japan

Abstract

Using statistically derived keywords to characterize texts has become an important research method for digital humanists and corpus linguists in areas such as literary analysis and the exploration of genre difference. Keywords—and the associated concepts of 'keyness' and 'key-keyness'—have inspired conferences and workshops, many and varied research papers, and are central to several modern corpus processing tools. In this article, we present evidence that (at least for the task of biographical sentence classification) frequent words characterize texts better than keywords or key-keywords. Using the naïve Bayes learning algorithm in conjunction with frequency-, keyword-, and key-keyword-based text representation to classify a corpus of biographical sentences, we discovered that the use of frequent words alone provided a classification accuracy better than either the keyword or key-keyword representations at a statistically significant level. This result suggests that (for the biographical sentence classification task at least) frequent words characterize texts better than keywords derived using more computationally intensive methods.

Correspondence:
Mike Conway, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.
E-mail: mike@nii.ac.jp

1 Introduction

Using statistically derived keywords to characterize texts has become an important research method for digital humanists and corpus linguists to explore genre differences (Xiao and McEnery, 2005), analyse character differences in Shakespeare (Culpeper, 2002), and investigate the development of swearing in English (McEnery, 2005), among many other uses. Keywords—and the associated concepts of 'keyness' and 'key-keyness'—have inspired conferences and workshops,[1] many and varied research papers, and are central to several modern corpus processing tools.[2]

In this article, we present evidence that (at least for the task of biographical sentence classification) frequent words characterize texts better than keywords or key-keywords. We used the naïve Bayes learning algorithm in conjunction with frequency-, keyword-, and key-keyword-based text representation to classify a corpus of biographical sentences, and discovered that the frequent word representation provided a classification accuracy better than either the keyword or key-keyword representations at a statistically significant level. This result suggests that (for the biographical sentence classification task) frequent words characterize texts better than keywords derived using more computationally intensive methods.

The article is structured as follows. Section 2 describes the creation of a sentence-level biographical annotation scheme and corpus. Section 3 explores the concept of keyness, key-keyness and introduces the notion of 'naïve' key-keywords. Section 3 also presents our keyword identification software, KWExT a **K**ey**W**ord **Ex**traction **T**ool.

Literary and Linguistic Computing, Vol. 25, No. 1, 2010.
doi:10.1093/llc/fqp035 Advance Access published on 6 October 2009

Finally, Section 4 details our sentence-level text classification experiments based on keyword text representations.

2 Creating a Biographical Corpus

2.1 Developing a biographical annotation scheme

As we are interested in classifying and identifying biographical sentences, our first step was to develop a sentence-level biographical annotation scheme. While there are several such schemes in existence (for example, the Text Encoding Initiative biography module,[3] *Oxford Dictionary of Biography* guidelines (OUP, 2003) and a biographical scheme developed at the University of Southern California (Zhou *et al.*, 2004)), none of these fit well with the needs of the current work, which requires 'sentence'-level classification using a well-defined decision procedure.

We therefore developed a sentence-level biographical scheme, where sentences are tagged with zero or more of six biographical tags. If the sentence contains no biographical information, then the sentence remains untagged. The six biographical tags used are:

- <key>: key information about a person's life course:
  - Information about date of birth, date of death, or age at death.
  - Names and alternative names (for example, nicknames).
  - Place of birth: 'Orr was born in Ann Arbor, Michigan but was raised in Evansville, Indiana'.
  - Place of death: 'He died of a heart attack while holidaying in the resort town of Sochi on the Black Sea coast'.
  - Nationality: 'He became a naturalized citizen of the United States in 1941'.
  - Cause of death: 'He died of a heart attack in Bandra, Mumbai'.
  - Longstanding illnesses or medical conditions: 'He stepped down from the position on grounds of poor health in February 2004'.
  - Place of residence: 'Sontag lived in Sarajevo for many months of the Sarajevo siege'.
  - Physical appearance: 'With his movie star good looks he was a crowd favourite'.
  - Major threats to health and wellbeing (for example, assassination attempts, car crashes).
- <fame>: what a person is famous for. This kind of information can be broadly positive (for example, rewards, prizes, honours, and so on) or negative (for example, scandal, jail terms, and so on). Examples of <fame> tags include:
  - 'His study of Dalton won him the Whitbread prize'.
  - 'In 1976 heroin landed him in Los Angeles County Jail, where he spent two months for possession of narcotics'.
- <character>: attitudes, qualities, character traits, and political or religious views. For example:
  - 'He was raised Catholic, the faith of his mother'.
  - 'Jones is recalled as a gentle and unassuming man'.
- <relationships>: information concerning relationships with intimate partners and sexual orientation. Relationship with parents, siblings, children, and friends.
  - 'Her mother died when she was eleven'.
- <education>: institutions attended, dates, educational choices, and qualifications awarded (with dates if available). General comments on educational experiences. For example:
  - 'Corman studied for his master's degree at the University of Michigan, but dropped out when two credits short of completion'.
- <work>: references to positions, job titles, affiliations (for example, employers), lists of publications, films, or other work orientated achievements. General areas of interest (for example, industries, sectors, and geographical regions).
  - 'He returned to England in 1967 to work for the offshore pirate radio station Wonderful Radio, London'.

Table 1 Descriptive statistics for biographical corpora

	Guar News	BBC Obits	Guar Obits	Stop
No. of Docs in Corpus	37	11	17	15
Avg. length of Docs (in words)	824	643	778	2257
Total No. of Bio Tags	194	173	327	107
Avg. No. of Bio Tags per Doc	6.5	15.7	19.7	7.1
Total No. of Bio Sent.	170	150	247	90
Avg. No. of Bio Sent per Doc	4.6	13.6	14.5	6.0

The experimental work reported in this article (Section 4) uses a binary scheme (that is, the six biographical tags described above are subsumed into a single biographical category). We developed the six tag scheme for two reasons:

- In preliminary annotation scheme development, the use of six tags helped us to understand the thought processes of annotators, and thus facilitated iterative annotation scheme refinement.
- We wished to create a resource that could be used for further, finer grained work. The use of a single biographical tag would mean the loss of information acquired during the annotation process.

In order to test how consistently the scheme could be applied, we conducted an agreement study with 25 participants annotating 100 sentences as biographical or non-biographical. We discovered that the scheme could be applied with a high level of consistency between annotators (Conway, 2007).

2.2 Constructing the biographical corpus

The next step required the selection of a range of texts, and then the annotation of these texts using our biographical scheme. The corpus consists of 84,305 word tokens from 80 different documents. As our goal is the creation of an automatic biographical sentence classifier, we selected texts that contain biographical sentences (according to our scheme) but which are not necessarily explicitly biographical in intention.

Four text sources were used: news text from *The Guardian*[4] newspaper, text from BBC Obituaries, obituaries from *The Guardian* newspaper, and finally literary texts selected from the multi-genre Stop Corpus.[5]

Texts were sampled from *The Guardian* newspaper online edition on 3 days [(11 August 2006 (13 documents), 12 September 2006 (12 documents), and 24 September 2006 (12 documents)]. News items only were chosen, though theme or subject was not restricted.

Seventeen obituaries were sampled from *The Guardian* newspaper from the first half of 2006. The obituaries include those of prominent lawyers, civil servants, diplomats, and journalists.

The 11 BBC Obituaries used in the corpus were downloaded from the BBC web site in July 2006.[6] They include writers, actors, politicians, and princes.

Fifteen texts were included from the Stop corpus. Although the Stop corpus includes texts from newspaper sources, only texts from the (auto)biography and literary categories were included. Each text is of around 2000 words in length.

Descriptive statistics for all the text sources that constitute the biographical corpus are presented in Tables 1[7] and 2.

3 Keyword Generation Methods

In this section, we describe our keyword extraction methodology, introduce the 'naïve' key-keyword extraction method, and describe Scott's well-known keyword extraction method (Scott, 2008). We also briefly introduce the **K**ey**W**ord **Ex**traction **T**ool (KWExT), which we used to derive all keywords in the current study.

3.1 Keywords and key-keywords

Keyword extraction is designed to capture salient words or concepts from texts using an algorithm—normally chi-square (χ^2) (Oakes *et al.*,

Table 2 Average number of biographical tag types per text

Source Type	Key	Fame	Char	Relation	Edu	Work
Guardian News (11 August 2006)	0.38	0.61	0.15	0	0	3.38
Guardian News (12 September 2006)	3.01	0.50	0.08	0.75	0	3.33
Guardian News (24 September 2006)	0.41	0.25	1.16	0.33	0.25	3.41
BBC Obituaries	3.82	3.54	2.36	2.18	0.64	5.45
Guardian Obituaries	6.35	1.05	2.82	4.23	1.29	8.94
Stop Corpus	1.47	0.27	1.67	1.80	0.27	3.00

2001), or log-likelihood (Dunning, 1993)—that compares the frequency of each word type in the corpus of interest (COI), to the frequency of that word type in a 'reference corpus' (that is, a corpus of general text). The algorithm ascribes a 'surprise' score to each word in the COI according to how much it deviates from the expected frequency (as determined from the reference corpus). Keyword extraction has been used in a variety of research contexts (for example, genre analysis (Xiao and McEnery, 2005), analysis of political bias in election manifestos (Rayson, 2008), bioinformatics (Kim and Tsujii, 2006), and so on). The process of keyword extraction is particularly associated with the `WordSmith` software tool (Scott, 2008).

In order to improve the 'keyness' of the keywords, 'key-keywords' are used (Scott and Tribble, 2006). These key-keywords are words that are keywords in 'more than one' text in the COI. That is, those words that are only key in one document from the COI are not key-keywords. The central idea here is that by eliminating those words that are only key in one text, we are left with key-keywords that better reflect the 'essence' of a given corpus, rather than the specific topicality of individual texts.

In order to produce key-keywords two corpora were constructed, a biographical corpus (consisting of short biographical documents from *wikipedia* and *Chambers Dictionary of Biography* (Chambers, 2004)—this is our COI) and a 'reference corpus' (the Flob corpus). The COI consisted of 47,967 words taken from 383 documents. These documents were randomly selected from *wikipedia* Biographies (194 documents used) and *Chambers* Biographies (189 documents used). Note that these summary biographies consist almost entirely of biographical text (see Fig. 1, for examples). It is important that attempts are made to make the reference corpus 'balanced' (that is, containing text from various different sources), hence the use of the Flob corpus, which, despite its—by modern standards—relatively small size (approximately one million word tokens) does cover a large number of text types (including 'general fiction' and 'reportage').[8] Note that Tribble (1998) found that the 'size' of the reference corpus used is not of vital importance, a result also gained by Xiao and McEnery (2005), who discovered that the Flob corpus and the 100 million word British National Corpus,[9] yielded a similar keyword list. Berber-Sardinha (2000) suggests that a reference corpus five times larger than the COI is sufficient. These results indicate that the Flob corpus is a suitable choice for the task in terms of its size and balance. Additionally, the Flob corpus—like the *Chambers* biographical entries and most of the *wikipedia* biographies—is written in British English, thus minimizing problems associated with British and American spelling variations.

Two related methods for extracting key-keywords were used in this work. First, the '*naïve*' key-keywords method and second, the `WordSmith` key-keywords method. The naïve key-keywords method is essentially a simplified version of the established `WordSmith` key-keywords technique developed for the current study.

3.2 Naïve key-keywords method

The process of identifying 'naïve' key-keywords can usefully be divided into two stages:

(1) The most discriminating keywords were identified by comparing the COI with a reference corpus (the Flob corpus) using the log-likelihood feature selection method.

Chambers
Babbage, Charles Born in Teignmouth, Devon, and educated at Trinity and Peterhouse colleges, Cambridge, he spent most of his life attempting to build two calculating machines. The first, the difference engine, was designed to calculate tables of logarithms and similar functions by repeated addition performed by trains of gear wheels. A small prototype model described to the Astronomical Society in1822 won the Society' sfirst gold medal, and Babbage received government funding to build a full-sized machine. However, by 1842 he had spent large amounts of money without any substantial result, and government support was withdrawn...

Wikipedia
Charles Babbage (26 December 1791–18 October 1871) was an English mathematician, analytical philosopher, mechanical engineer and (proto-) computer scientist who originated the idea of a programmable computer. Parts of his uncompleted mechanisms are on display in the London Science Museum. In 1991, working from Babbage's original plans, a difference engine was completed, and functioned perfectly. It was built to tolerances achievable in the 19th century, indicating that Babbage's machine would have worked. Nine years later, the Science Museum completed the printer Babbage had designed for the difference engine; it featured astonishing complexity for a 19th century device...

Fig. 1 Excerpts from *Chambers* and *wikipedia* biographies (Charles Babbage)

Table 3 Unigrams in the biographical corpus ranked by frequency (with additional information about the number of biographical documents in which the unigram occurs)

Rank	Unigram	Percentage of Bio Docs	No. of Bio Docs
1	the	97	370
2	in	94	359
3	of	87	334
4	and	89	342
5	he	78	300
6	a	81	310
7	was	83	319
8	to	79	268
9	his	63	242
10	as	53	202

(2) Selected keywords as identified by the log-likelihood method were reranked according to the number of biographical documents in which they occur, remembering that there are 383 biographical documents in total. The resulting ranking is the *naïve key-keyword*[10] ranking. For example, if the unigram 'born' occurs in 220 biographical documents, and the unigram 'became' occurs in 111 biographical documents, then the unigram 'born' will be ranked above the unigram 'became' in the key-keywords list. That is, the unigram 'born' will have a higher key-keyword ranking than 'became'. The intuition here is that while a high-ranked 'keyword' may occur in only one or two biographical document, a high-ranked naïve *key-keyword* is likely to appear in *many* biographical documents.

Table 3 presents the 10 most frequent unigrams in the biographical corpus, together with information about the number of biographical documents (that is, *Chambers* or *wikipedia* biographies) in which the unigram occurs. Table 4 shows the 10 unigrams with the highest naïve key-keyword value (that is, of the most discriminating keywords identified by the log-likelihood algorithm, those 20 that appear in the most biographical documents). Note that column 3 of Tables 3 and 4 refers to the proportion of biographical texts in which the keyword occurs, and column 4 gives the number of texts in which the keyword occurs (of which there were 383 in total). Note also that ordinary function words appear high on both lists (for example, 'in', and 'and'). The word 'in' is used disproportionately frequently in the biographical texts to indicate the time of a biographically significant

Table 4 Unigrams in the biographical corpus ranked by naïve key-keyness (with additional information about the number of biographical documents in which the unigrams occur)

Rank	Unigram	Percentage of Bio Docs	No. of Bio Docs
1	in	97	370
2	and	89	342
3	was	83	319
4	he	78	300
5	his	63	242
6	born	57	220
7	as	53	202
8	at	50	191
9	an	40	153
10	became	29	111

Table 5 Unigrams in the biographical corpus ranked by WordSmith key-keyness (with additional information about the number of biographical documents in which the unigrams occur)

Rank	Unigram	Percentage of Bio Docs	No. of Bio Docs
1	2004	20	76
2	he	78	300
3	she	18	67
4	his	63	242
5	in	94	359
6	2001	4	17
7	stub	3	10
8	college	7	27
9	1998	3	12
10	university	13	49

event (for example, 'He died ''in'' *1964*') or the location of an event ('He was born ''in London''').
Of the 20,714 instances of '[iI]n' (that is 'in' or 'In') in the FLOB corpus, only 800 (4%) were followed by a four digit date. When the biographical texts were analysed, the proportion of instances of 'in' followed by four digits was 26% (504 out of 1983 instances).[11] Additionally, '[Ii]n' occurs more than twice as often in the biographical texts as in the reference (FLOB) corpus (4.24% and 2.02%, respectively). It is possible that the large discrepancy in the frequency of 'in' is likely to arise—at least partially—from the increased use of the word 'in' to associate an event with a year in biographical text.

3.3 WordSmith key-keywords method

The process for identifying WordSmith key-keywords falls into two stages:

(1) For each of the 383 biographical documents a keyword list was produced (using the log-likelihood algorithm and the FLOB corpus as a reference corpus.
(2) A key-keyword list was then generated by identifying those words that appeared as keywords in the greatest number of biographical documents 'A ''key key-word'' is one which is ''key'' in more than one of a number of related texts. The more texts it is ''key'' in, the more ''key-key'' it is.'[12] This method can be contrasted with the naïve key-keywords method. Instead of reranking the keywords according to the number of biographical documents in which they occur, the WordSmith method simply ranks words according to the number of documents in which they are key. For example, if the unigram 'he' is a keyword in 300 biographical documents and the unigram 'college' is a keyword in 27 biographical documents, then the keyword 'he' will have a higher WordSmith key-keyword ranking than 'college'.

Table 5 shows 10 unigrams from the biographical corpus ranked by WordSmith key-keyness. It is noticeable that the unigrams identified using the WordSmith key-keyword method differ from those identified by the 'naïve' key-keyword method. The relative lack of function words among the highest ranking WordSmith key-keywords is noticeable, as is the appearance of unigrams that are perhaps tied to the particular biographical corpora used, rather than biographical texts in general. For example, the unigrams '2004', '2001', and '1998' appear very high in the WordSmith key-keyword list because these represent the death dates for the *wikipedia* biographies. Similarly the unigram 'stub'—a word used by *wikipedia* to indicate that an entry is a short summary—appears as a keyword in 3% of biographical

Table 6 Unigrams in the biographical corpus, ranked by frequency, keyness, naïve key-keyness, and WordSmith key-keyness

Rank	Frequency	Keyness	Naïve KKW	WS KKW
1	the	he	in	2004
2	in	born	and	he
3	of	in	was	she
4	and	2004	he	his
5	he	was	his	in
6	a	his	born	2001
7	was	became	as	stub
8	to	died	at	college
9	his	university	an	1998
10	as	peel	became	university
11	for	studied	after	her
12	at	career	also	won
13	on	poe	first	worked
14	with	won	died	1999
15	by	served	2004	series

KKW, key-keywords.

documents. Place names ('edinburgh', 'istanbul', and 'pennsylvania') also appear on the list, as well as a single personal name ('john'), whereas the paradigmatically biographical word 'born' does not occur. For ease of comparison, Table 6 shows the 15 highest ranked unigrams produced by each method.

The important difference between the *naïve* and WordSmith key-keyword methods is that the naïve method ranks keywords according to the number of biographical documents in which the keyword occurs, where as the WordSmith method ranks keywords according to the number of biographical documents in which the word is key. For each key-keyword identification method, the 250 top key-keywords (ranked by key-keyness) are retained.

3.4 KWExT—a keyword extraction tool

To perform keyword extraction in this work, we developed the KWExT Graphical User Interface-based software tool (Conway, 2009). KWExT performs keyword and key-keyword extraction (naïve and WordSmith style) using the log-likelihood and χ^2 methods (see Fig. 2 for a screenshot). Additionally, KWExT is capable of extracting key *n*-grams and key-key *n*-grams. The tool is freely available and runs on Mac OS and Linux operating systems.[13]

4 Classification Experiments

In this section, we report the results of a series of machine learning experiments testing the utility of various keyword representations for the task of classifying our corpus of biographical sentences.

4.1 Experimental design

Feature selection is a commonly used technique in machine learning (Witten and Frank, 2005), and it has been shown that aggressive feature selection increases classification accuracy for some kinds of text classification tasks (Yang and Pedersen,1997). It is hypothesized that key-keywords-based methods will provide features more able to discriminate between biographical and non-biographical sentences than either frequent unigrams, or log-likelihood-derived keywords, alone. Note that feature selection was not performed on our annotated biographical corpus. Rather, features were identified using the COI (that is, a corpus constructed from *wikipedia* and *Chambers* data), in order that unigram features characteristic of biographical text in general could be identified.

The feature sets are fully described in Section 3 but summarized here:

- The 250 most frequent unigrams from the biographical corpus.
- The 250 most discriminating keywords identified by the log-likelihood algorithm.
- The 250 most discriminating key-keywords identified using the *naïve* key-keywords method.
- The 250 most discriminating key-keywords identified using the WordSmith key-keywords method.

The 250 most frequent unigrams were included as a baseline against which to test the performance of the keywords and key-keywords representations.

For all these experiments we use three machine learning (classification) algorithms: naïve Bayes, C4.5 (decision tree) and the support vector machine (SVM) algorithm (Mitchell, 1997).[14] The naïve Bayes and SVM algorithms are commonly used in text classification work (Sebastiani, 2002). To quantify classifier performance, we use 'accuracy' (that is, the percentage of correctly assigned sentences) in conjunction with stratified 10-fold cross-validation.

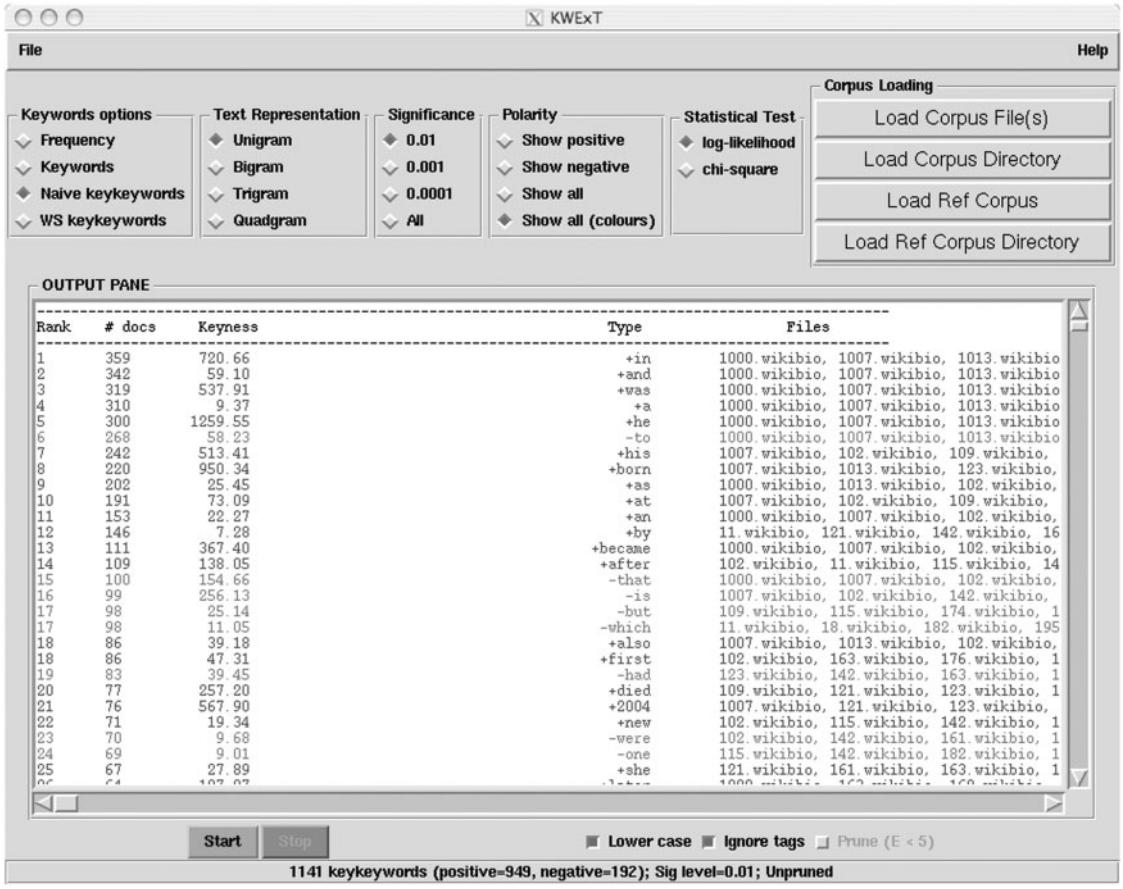

Fig. 2 KWExT tool

Where statistical tests are described, we use the corrected resampled *t*-test and 10×10-fold cross-validation ($\alpha = 0.05$) (Boukaert and Frank, 2004). We used a Boolean feature representation for two reasons. First, as we are performing *sentence* classification, the term weighting often used in document classification is likely to be inappropriate. Second, Boolean features have been shown to be useful for text classification tasks that are not focused on raw topicality (for example, Yu's (2008) work on the classification of erotic prose).

The data for these experiments was constructed from the annotated biographical corpus using the following method:

(1) We tokenized the tagged biographical corpus at the sentence level.
(2) From those sentences that contained biographical tags (for example, <key>, <fame>), we sampled 235 sentences.
(3) From those sentences that did not contain biographical tags, we sampled 265 sentences.

This process resulted in a test/training corpus of 500 sentences (235 biographical plus 265 non-biographical). As we are assessing classification performance using 10×10-fold cross-validation, manually separating test and training data was not required.

4.2 Results

Table 7 and Fig. 3 show that (using the naïve Bayes algorithm) the 250 most frequent unigrams feature

Table 7 Performance of keyword and key-keyword features relative to a baseline

Feature set	Naïve Bayes (%)	SVM (%)
250 Frequent unigrams	**81.24**	74.65
250 Keywords	76.65	77.05
250 Naïve key-keywords	78.84	78.24
250 `WordSmith` key-keywords	75.65	72.85

Bold value indicates best result.

set performed at 81.24%. The keyword feature set achieved 76.65%. The `WordSmith` key-keywords and naïve key-keywords achieved 75.65% and 78.84%, respectively. The difference between the 250 frequent unigram and the 250 'naïve' key-keywords feature sets was not statistically significant. The difference between the 250 frequent unigrams and the `WordSmith` method was significant, however, with the `WordSmith` key-keywords feature set performance significantly worse than the frequent unigram feature set. This was a surprising result, as it was expected that the log-likelihood feature selection (that is, the keyword feature set) and both the key-keyword feature sets would achieve better results than the simple frequent unigram-based representation. Indeed, the frequent unigram representation outperforms both the log-likelihood (keyword) feature set and the `WordSmith` key-keywords feature set at a statistically significant level.[15]

Table 7 also shows that the naïve Bayes algorithm outperforms the SVM algorithm for all but the keywords representation, and provides the best overall accuracy (81.24%) by a substantial measure. We have therefore focused our discussion on the results obtained by the naïve Bayes algorithm.

4.3 Discussion

These results show that, for the biographical sentence categorization task at least, the use of key-keywords reduces classification accuracy. Note that feature selection was performed using external data (*wikipedia* and *Chambers*—the COI—as a biographical corpus, and the FLOB corpus as a reference corpus), in order to avoid artificially inflating classification accuracy.

In order to gain insight into the differing performance of the four feature sets, and the surprising

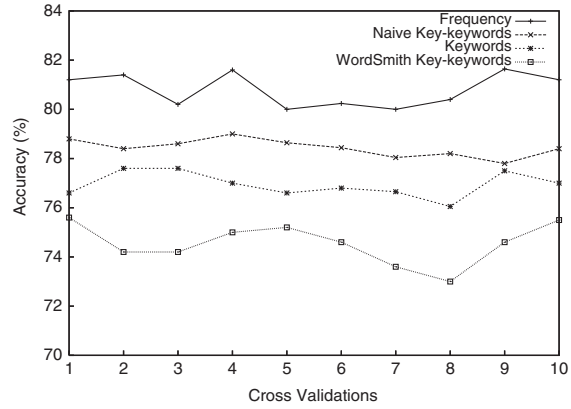

Fig. 3 Comparison of the performance of keywords, key-keywords, and frequencies using the Naïve Bayes algorithm

success of the frequency feature set, the `C4.5` decision tree algorithm was used as a tool for data exploration. Fig. 4 shows that—for the top levels and with the exception of the `WordSmith` key-keywords representation—the trees are similar, with the major difference between the top performing frequency feature set, and the keywords and naïve key-keyword feature sets, being that 'best' is used as the second node of the frequency tree, and does not occur in the top part of the other trees. The `WordSmith` key-keyword tree is very different from the other three trees as there is little overlap between the features selected by the `WordSmith` method, and those selected by the alternatives.

It is notable that there is a substantial difference between the frequency-based and naïve key-keywords feature sets. One hundred and thirty-nine words appear in the frequency list that do not appear in the naïve key-keywords list. While some biographically important function words appear in the naïve key-keywords list—for example, the preposition 'in', the connective 'and' and the pronoun 'he'—many are absent. For example, 'the', 'of', and 'to' appear in the frequency list but not in the naïve key-keywords list. Similarly, words that we would intuitively regard as biographical appear in the frequency list—words like 'home' and 'father'—but do not appear in the naïve key-keywords list.

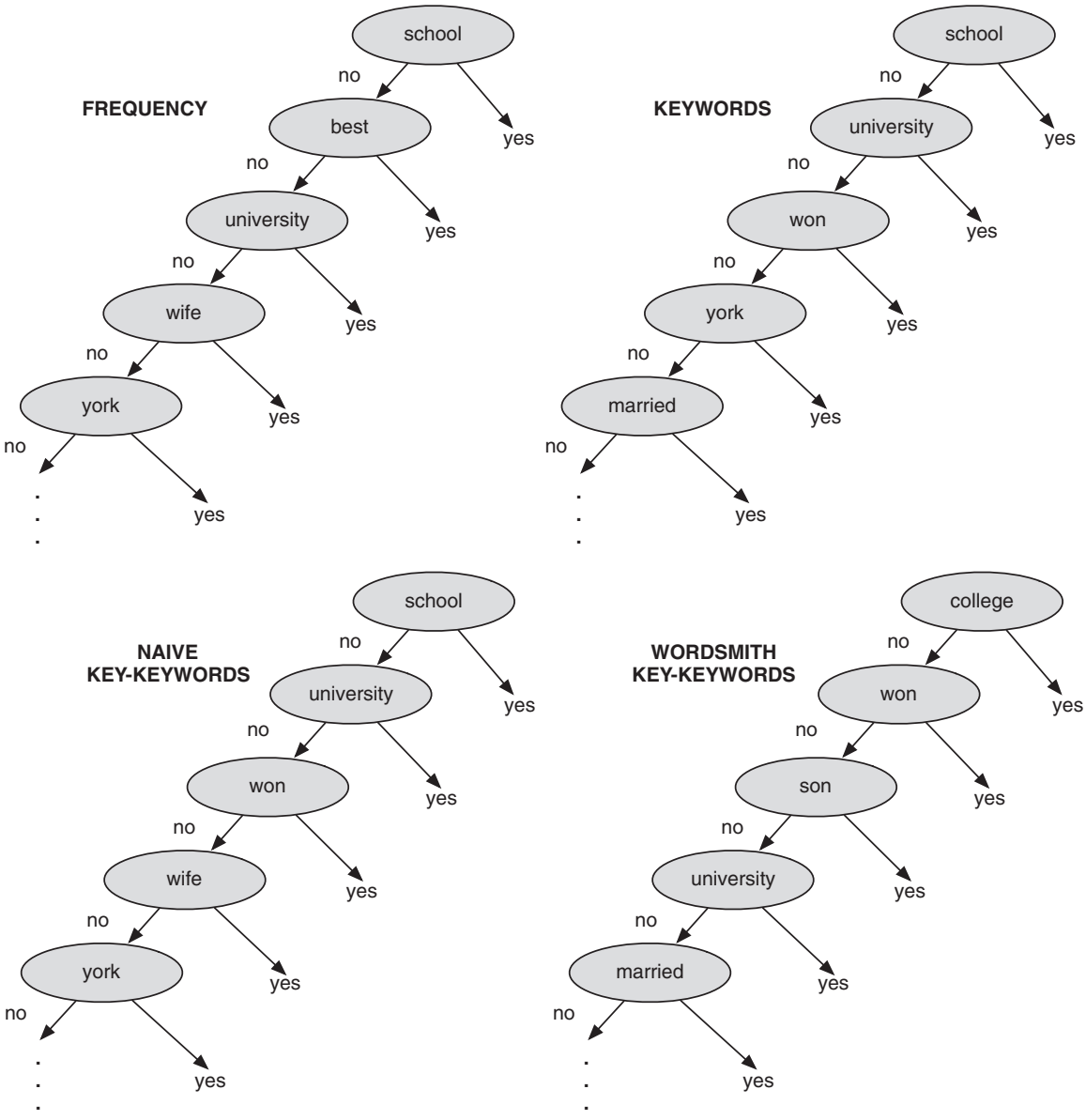

Fig. 4 Comparison of partial decision trees for each feature set

The difference between the frequency-based feature set and the `WordSmith` key-keyword feature set is even more marked than the difference between the frequency-based feature set and the 'naïve' key-keywords feature set. One hundred and eighty-two words occur in the frequency-based feature set that do not occur in the `WordSmith` key-keywords feature set. Biographically relevant function words (like 'the' and 'of') are missing from the `WordSmith` feature set, as are more obviously biographical words like 'home' and 'father'.

There are a number of possible reasons why both the key-keyword feature sets failed to provide a better (in terms of classification accuracy) representation than the simple frequency list:

- The biographical corpus, consisting of the *wikipedia* and *Chambers* data, while large enough to provide a 'biographical' frequency list, was not large or varied enough to counter the inclusion of ostensibly non-biographical unigrams. For example, 'detroit' and 'constantinople' occur in the `WordSmith` key-keyword list, whereas 'philadelphia' and 'york' occur in the naïve key-keyword list.
- It is possible that the number of features used was too low for the benefits of the key-keyword approaches to be clear. Perhaps if more features were used in each case, key-keywords may outperform the simple frequency list approach. On the other hand, the purported benefit of the `WordSmith` key-keyword method is that genre or topic salient words are pushed to the top of the list.
- The frequency and keyword lists were derived from biographical *documents* rather than biographical *sentences*, whereas the classification task involved the classification of biographical *sentences*. The non-biographical sentences in the biographical documents counted equally with the biographical sentences in the frequency calculations. It is possible that the key-keywords method discarded many features that are characteristic of biographical 'sentences'. This possibility is weakened, however, if we consider the high level of biographical sentences in *wikipedia* and *Chambers* (>85%, based on a sample of 10 biographies).
- It is possible that the key-keyword methods are capturing 'corpus'-specific features, rather than 'genre'-specific features, and that frequency lists derived from corpora of a given genre of interest provide a better insight into that genre. In other words, a frequency list derived from a corpus of a given genre may reflect that 'genre's' characteristics, better than key-keywords, which are too specific to the particular topical idiosyncracies of the corpus.
- Another interpretation of the results is that the nature of the current task—sentence classification rather than document classification—may not be suitable for a key-keyword approach as genre analysis techniques are only appropriate at the document level. Indeed, the systemic functional linguistics tradition holds that single sentences cannot be described as having a genre (Halliday and Matthiessen, 2004). Rather (according to Systemic Functional Linguistics theory), genre is a phenomenon of the discourse level.

In conclusion, on the basis of the work presented in this section, 'key-keyword' methodologies are not suitable techniques for the identification of unigram features for biographical sentence classification using the annotation scheme and corpus developed in this work, as simple frequency counts provide better performance. This result is surprising, however, as the opposite result—that key-keywords would prove to be a 'better' feature set than frequent unigrams—was expected. Whether this result is generalizable to other corpora (of different types, size, and genre) is a question that requires further investigation.

References

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith tools: how large must the reference corpus be? In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL 2000): Workshop on Comparing Corpora*, Morristown, NJ: ACL, pp. 7–13.

Bouckaert, R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In Dai, H., Srikant, R., and Zhang, C. (eds), *Advances in Knowledge Discovery and Data Mining*. Berlin: Springer, pp. 3–12.

Chambers (2004). *Chambers Biographical Dictionary*. Edinburgh: Chambers-Harrap.

Conway, M. (2007). *Approaches to Automatic Biographical Sentence Classification: An Empirical Study*. Ph.D. thesis, University of Sheffield.

Conway, M. (2009). `KWExT` – A Prototype Keyword Extraction Tool. In *Corpus Linguistics 2009*, Liverpool p. 294.

Culpeper, J. (2002). Computers, language and characterisation: an analysis of six characters in Romeo and Juliet. In *Conversations in Life and in Literature: Papers from the ASLA Symposium* Uppsala: ASLA, pp. 11–30.

Dunning, T. (1993). Accurate methods for the statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1): 61–74.

Halliday, M. and Matthiessen, C. (2004). *An Introduction to Functional Grammar, 3rd edn*. London: Hodder Arnold.

Kim, J. and Tsujii, J. (2006). Corpora and their annotation. In Ananiadou, S. and McNaught, J. (eds), *Text Mining for Biology and Biomedicine*. London: Artech House, pp. 179–211.

McEnery, A. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Mitchell, T. (1997). *Machine Learning*. Singapore: McGraw-Hill International.

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, A., Wan V. and Beaulieu, M. (2001). A method based on the Chi-Square test for document classification. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 21)*. Morristown, NJ: ACM, pp. 440–1.

OUP (2003). *New Dictionary of National Biography: Notes for Contributors*. Oxford: Oxford University Press.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, **13**(4): 519–49.

Scott, M. (2008). *WordSmith Tools*. Liverpool: Lexical Analysis Software.

Scott, M. and Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins Publishing Company.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1): 1–47.

Semino, E. and Short, M. (2004). *Corpus Stylistics*. London: Routledge.

Tribble, C. (1998). *Writing Difficult Texts*. Ph.D. thesis, University of Lancaster.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn*. San Francisco: Morgan-Kaufmann.

Xiao, R. and McEnery, A. (2005). Two approaches to genre analysis: three genres in modern American English. *Journal of English Linguistics*, **33**: 62–82.

Yang, Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, San Francisco, CA: Morgan Kauemann, pp. 412–20.

Yu, B. (2008). An evaluation of text classification methods for literary studies. *Literary and Linguistic Computing*, **28**: 327–43.

Zhou, L., Ticrea, M. and Hovy, E. (2004). Multi-document biography summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*. Morristown, NJ: ACL, pp. 434–41.

Notes

1 Conferences and Workshops include *Keyness in Text* held at Sienna in 2007 (http://www.disas.unisi.it/keyness/index.php), and the *Word Frequency and Keyword Extraction Seminar* held at Lancaster in 2005 (http://www.methodsnetwork.ac.uk/activities/es01mainpage.html) (accessed 2 January 2007).

2 For example, `WordSmith` (www.lexically.net/wordsmith/), AntConc (www.antlab.sci.waseda.ac.jp), and `WMatrix` (http://ucrel.lancs.ac.uk/wmatrix) (accessed 2 January 2007).

3 *Report on* XML *Markup of Biographical and Prosopographical data* (http://www.tei-c.org, accessed 1 August 2006). Prosopography is a research method in history which examines the relationships between historical figures in order to identify common experiences (among other things).

4 http://www.guardian.co.uk (accessed 2 January 2007).

5 *The Lancaster Speech, Thought and Writing Presentation Corpus*, described in Semino and Short (2004) and available from the Oxford Text Archive at http://ota.ox.ac.uk (accessed 2 January 2007). Note that we did not use the speech and thought presentation annotation encoded in the STOP corpus.

6 http://news.bbc.co.uk/obituaries (accessed 8 February 2007).

7 Note that as single sentences can have multiple tags, there are fewer biographical sentences than biographical tags for each data source (see Table 1, rows 4 and 6).

8 Freiburg-LOB corpus (http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM, accessed 2 January 2007).
9 British National Corpus (http://www.natcorp.ox.ac.uk) (accessed 2 January 2007).
10 We have named this method the 'naïve' method as it is less computationally intensive than the `WordSmith` method.
11 The regular expressions used to identify 'in', and 'in' followed by a four digit year, were '\s[Ii]n(\s|,|.)' and '\s[Ii]n\s\d\d\d\d(\s|,|.)', respectively.
12 `WordSmith` documentation (http://www.lexically.net/downloads/version4/, (accessed 1 May 2007).
13 Binaries for `KWExT` are available at Google Code (http://code.google.com/p/kwext/) (accessed 2 January 2007).
14 The `weka` implementation of these algorithms was used (http://www.cs.waikato.ac.nz/ml/weka/) (accessed 2 January 2007).
15 Note that we performed the same series of experiments using χ^2 and achieved similar results.