Investigating Variations in Adjective Use across Different Text Categories

Jing Cao[1] and Alex Chengyu Fang[2]

Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR, China
[1]cjing3@student.cityu.edu.hk, [2]acfang@cityu.edu.hk

Abstract. Adjectives are an informative but understudied linguistic entity with good potentials in sentiment analysis, text classification and automatic genre detection. In this article, we report an investigation of the variations in adjective use across different text categories represented in a sizable corpus. In particular, we report the distribution of adjectives across a range of categories grouped together as academic prose in the British National Corpus. We shall measure inter-category similarity in the use of adjectives and demonstrate with empirical data that adjectives are an effective differentia of text categories or domains, at least in terms of arts and sciences as the two major sub-categories within academic prose.

Key Words: corpus, text category, adjective, similarity, BNC

1 Introduction

Adjectives are an informative but understudied linguistic entity [1, 2], drawing more and more attention within the research community. Focus has been mostly on the semantic aspect of adjectives for practical research in sentiment analysis applicable to automatic evaluations of email communication [3], blogs [4] and customer reviews in [5]. Studies in this respect typically focus on evaluative adjectives [6] and size adjectives [7]. In addition to the semantic approach, adjectives are also used for purposes of text categorization and genre detection in [8]. In this respect, [2] and [9] have generally shown with corpus evidence that adjectives occur more often in written texts than in spoken ones, and more frequently in informative writing than in imaginative writing. According to [8], 'the literature suggests that adjectives and adverbs will vary by genre because of their unique patterns of usage in text' (p. 4).

This paper describes one of the recent attempts to study adjectives from the perspectives of text categorization and genre detection. In particular, we investigate the variations of adjective use across various types of academic writing selected from a large-sized corpus. We attempt to ascertain whether adjective-based indices will be able to classify texts in such a way that conforms to manual classification. As we shall show in this article with empirical data, adjectives do differ by text categories and therefore appear to be an important differentia of text categories. More importantly,

Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 207-216

Received 10/11/08
Accepted 15/12/08
Final version 05/02/09

such a difference in adjective use (in terms of token similarity and type similarity between categories) may offer new insights into text categorization and automatic term recognition. Since the texts we used are samples of academic prose grouped according to domain, our study therefore suggests that the grouping criterion offered by adjectives seems to be a semantic one, therefore a useful complement to other studies that have shown adjectives to effectively distinguish between speech and writing in the first place, and formal and informal writings as varying degrees of formality.

The rest of the article will be organized as follows. Section 2 will briefly review three related studies. Section 3 will present a description of the corpus material after a discussion of our methodology. Section 4 will attempt to present the results and demonstrate that similarities in adjective use (in terms of tokens and types) seem to be able to group academic prose according to domains. We shall then draw some initial conclusion in Section 5.

2 Previous Studies

In this section we provide a review of three previous studies on adjectives across text categories. A classic corpus-based study [10] analyzed the distribution of the major word classes across four core fields, such as conversation, fiction, news and academic prose, in the Longman Spoken and Written English Corpus. What concerns us is the use of adjectives across the chosen genres. The results show that adjectives are more common in written texts than in spoken texts. Among written texts, adjectives are most common in academic prose. News has fewer adjectives than academic prose but more adjectives than Fiction. The findings seem to suggest a correlation between adjective use and formality of texts.

Later, [2] studied the 100 most frequent adjectives across genres in three written corpora and also analyzed the syntactic and semantic features of those adjectives. Nevertheless, they also reported and compared the distributional features of adjectives as a whole in Wellington Corpus of Written New Zealand English, Brown Corpus and the LOB corpus. [2] also shows that adjectives are used unevenly across different written texts in all the three corpora. To be more specific, adjectives appear most often in academic prose, reviews and hobbies, while they are less frequent in fiction. The findings echo the results in the written texts in [10].

The two studies touch upon the distribution of adjectives across text categories, whereas [8] not only analyzed adjectives and adverbs across genres and also attempted to examine whether they can discriminate different genres. Rittman [8] employed 44 trait adjectives, 30 speaker-oriented adverbs, and 36 trait adverbs to examine the three chosen genres (i.e. academic, fiction, and news) in the British National Corpus (BNC). First, the investigation was made among the three genres, academic vs. fiction vs. news, or called ‘one-against-one’ classification. Secondly, a one-against-many classification was made when each of the chosen genres used as a host category and the rest of other genres in the BNC as the guest category. For example, the comparison was made between ‘academic’ vs. ‘not-academic’ (fiction, news, non-fiction, other, and spoken). The results show that the one-against-one

classification tends to be more effective that the one-against-many. The study also demonstrates that using adjective and adverb features is generally superior to other models containing features such as nouns, verbs or punctuation. Moreover, among the three features employed, the speaker-oriented adverbs are more effective than the class of trait adjectives and adverbs.

To sum up, previous studies have shown that adjectives can tell speech from writing, and among writing, academic from fiction. Yet, it is still unclear whether adjective use differs across a set of subject domains, which will be the goal of the current study.

3 Methodology and Corpus Description

As mentioned in the last section, previous studies have shown that adjectives can tell speech from writing and also rank texts in a continuum scale of 'formalness'. Nevertheless, it is still unclear whether the variations of adjective use can illustrate domain similarities. When adjective use differs in different text categories from the same genre such as academic writing, the difference is more likely due to the semantic use of adjectives in different domains than due to stylistic difference in the texts. In other words, if the distribution of adjectives differs among various text categories in academic writing, we have reason to conclude that the variations of adjective use can distinguish texts of different domains. To be more specific, if the distribution of adjectives can cluster text categories in academic writing into two broad sub-categories such as arts and sciences, it would be reasonable to say that adjectives can be used as an indicator to distinguish these two different domains. If we can show with empirical data that our assumption is true, variations of adjective use can not only be applied to the ranking of texts according to degrees of formality, but more importantly to the categorization of texts according to different domains.

Given the purpose of our study, the XML Edition of the 100-million-word British National Corpus (BNC) [11] is used as the basis of our experiment. Such a large, balanced and annotated corpus serves effectively the purpose of examining certain word class (in our case, adjectives) across different text categories. According to [12], the texts in the BNC are classified into six genres, namely, academic prose, fiction, newspapers, non-academic prose, other published writing, unpublished writing, conversation, and other spoken. To investigate the variations of adjective use across text categories within a same genre, academic prose (or ACPROSE) is chosen, which has six component text categories: 'humanities and arts', 'medicine', 'natural science', 'politics, law and education', 'social science' and 'technology and engineering'. 500,000 words from each component category were randomly sampled at the text level to compose a 'sub-corpus' as the basis of our experiments. Table 1 summarizes the six categories in terms of tokens, types and type/token ratios.

Table 1. A summary of the six text categories sampled from ACPROSE

Text Category	Text Code	Word Token	Word Type	Type/Token Ratio
humanities and arts	HUM	524224	30780	5.9
Medicine	MEDI	504856	24497	4.9
natural science	NAT	536499	28448	5.3
politics, law and education	POLIT	511935	20137	3.9
social science	SOC	511655	22581	4.4
technology and engineering	TECH	535251	18456	3.4

Since the categories were sampled on a text basis, they do not have exactly the same number of tokens. As is also evident from Table 1, the six categories do not have the same vocabulary size, with 'humanities and arts' (HUM) being the highest in terms of number of types and 'technology and engineering' (TECH) the lowest for that matter.

4 Results and Discussions

On the basis of the sub-corpus created from ACPROSE, the frequencies of adjectives in the six component categories were obtained and summarized in Table 2, which lists the numbers of tokens and types of adjectives in each category as well as type/token ratios for the adjectives. Again, the category HUM has the highest type/token ratio for adjectives and TECH the lowest.

Table 2. Basic data of adjectives in the six text categories

Text Code	ADJ Token	ADJ Type	Type/Token Ratio
HUM	45157	5709	12.6
MEDI	56659	4979	8.8
NAT	54242	6086	11.2
POLIT	43867	3880	8.8
SOC	51602	4240	8.2
TECH	46650	3814	8.2

We next attempt to examine whether adjectives can illustrate the relation between different text categories and to what extent they can achieve that. To be more exact, we aim to measure the similarity or dissimilarity between component categories in terms of adjective use. We therefore define *token similarity*, a measure that reveals the proportion of adjective tokens in common use by any two categories. We also define *type similarity*, which refers to the proportion of types of adjectives that are observed in common use between categories. In this section, we describe our observations made when each text category, serving as the host category, compares with the other five categories (guest categories). We shall first explain the key concepts of *type* and *token similarity*, and then present our data in terms of type similarity and token similarity.

4.1 Key Concepts

In this sub-section, we describe how we calculate *type similarity* and *token similarity*. It takes two steps to determine *type similarity*: First, the number of types of adjectives in common use between a host category and a guest category is calculated, and then the proportions of those shared adjective types in the host category is obtained, which is called *type similarity* of the host category to the guest category, denoted by S_{type}:

$$S_{type} = \frac{\text{Number of shared types by host and guest categories}}{\text{Total number of types in host category}} \times 100\% \quad (1)$$

For example, text A is the host category and text B is the guest category. *Type similarity* of text A to text B is the proportion of shared types of adjectives by texts A and B over the total number of types in text A. The higher the proportion, the higher degree of similarity of text A has towards text B. It is also worth mentioning that *type similarity* is directional, interpreting from the viewpoint of the host category. In other words, the type similarity of text A to text B is not necessarily the same as that of text B to text A because the denominators differ.

Next, based on the shared types by host and guest categories, *token similarity* is then computed. Again the *token similarity* of a host category to a guest category is computed in two steps: Firstly, we count the frequency of shared adjective types in a given host category. Secondly, we calculate the proportions of total number of those shared adjectives in a host category, which is called *token similarity* of the host category to the guest category, denoted by S_{token}:

$$S_{token} = \frac{\text{Frequency of shared types in host category}}{\text{Total number of tokens in host category}} \times 100\% \quad (2)$$

Same as *type similarity*, *token similarity* is also directional, interpreted from the viewpoint of the host category.

4.2 Type Similarity

As for the six chosen text categories, each category is treated as the host category and its *type similarity* to the other five guest categories is calculated according to Equation (1) respectively. Table 3 presents the type similarities between the six chosen text categories, and the similarity scores are interpreted vertically from the viewpoint of host categories.

According to Table 3, with HUM as the host category, it has higher type similarities with POLIT and SOC, both above 35%. On the other hand, HUM has a slightly lower type similarity with the other three guest categories of sciences by a little over 2%. In addition, the guest categories of arts are observed to be grouped together on the top of the similarity scale, while the guest categories of sciences grouped towards the bottom of the scale. When serving as the host category, POLIT has the highest type similarities with HUM and SOC, both above 50%. The other

guest categories are grouped together, all under 46%, a 4% difference between the two groups. Once again, the guest categories are noticed to be grouped neatly into the two broad categories of arts and sciences. When looking into the relation between SOC and its guest categories, we observe the same expected tendency. SOC has a closer similarity to HUM and POLIT, while NAT, MEDI and TECH are grouped together with comparatively lower similarity scores.

Table 3. Type similarity between the six text categories

			S_{type} of Host Category					
			Arts			Sciences		
			HUM	POLIT	SOC	MEDI	NAT	TECH
Guest Category	Arts	HUM	/	57.6	48.5	37.4	31.1	39.8
		POLIT	39.2	/	45.8	35.6	27.7	38.8
		SOC	36.0	50.0	/	36.1	30.6	42.3
	Sciences	MEDI	32.6	45.7	42.4	/	34.2	39.4
		NAT	33.1	43.4	43.9	41.8	/	44.0
		TECH	26.6	38.1	38.0	30.2	27.6	/

On the other hand, the category of MEDI has the highest degree of similarity to NAT, followed by the three guest categories of arts with similar similarity scores. We notice the unexpected behavior of TECH in this group in that it appears at the bottom of the similarity scale. When it comes to NAT as the host category, the three guest categories of arts are grouped together in the similarity scale as we expected. Compared with the sciences category, NAT has the strongest degree of similarity to MEDI, which echoes what can be observed when MEDI is the host category. It can be observed again that TECH is at the bottom of the similarity scale. It is quite within our expectation that TECH is closely related to NAT, and has the weakest relation with POLIT. It is also noted that MEDI, with the second smallest number of adjective types, ranks a little lower than SOC and HUM in the scale of type similarity.

The above observations seem to suggest a division between the two groups (i.e. arts and sciences) in terms of adjective use with a few exceptions. We therefore compute the mean of type similarities between two broad groups of arts and sciences, and the results are presented in Tables 4 and 5.

Table 4. Type similarity from the viewpoint of Arts

		S_{type} of Arts (Host Category)		
			Sub Mean	Mean
Guest Category	Arts	HUM	37.6	46.2
		POLIT	53.8	
		SOC	47.1	
	Sciences	MEDI	30.8	38.2
		NAT	42.4	
		TECH	41.4	
Difference				8.0

Table 5. Type similarity from the viewpoint of Sciences

		S_{type} of Sciences (Host Category)		
			Sub Mean	Mean
Guest Category	Sciences	MEDI	36.0	36.2
		NAT	30.9	
		TECH	41.7	
	Arts	HUM	36.3	35.5
		POLIT	29.8	
		SOC	40.3	
Difference				0.7

According to Table 4, the mean of type similarity between the host and guest categories of arts is 46.2%, and the similarity mean between the host category of arts and the guest categories of sciences is 38.2%. The 8% difference between the two groups apparently suggests a distinction between the arts category and the sciences category. With the sciences category as the host category, Table 5 shows that the mean type similarity between the host and guest sciences categories (36.2%) is slightly higher than the one between the host sciences category and the guest arts category (35.5%). Therefore, the type similarity of adjective use may also be used as an indicator of text categorization in that it has differentiated the arts category from the sciences category.

4.3 Token Similarity

Based on the shared types, the *token similarity* of each host category to the five guest categories is computed according to Equation (2) and the results are presented in Table 6.

Table 6. Token similarity between the six text categories

			S_{token} of Host Category					
			Arts			Sciences		
			HUM	POLIT	SOC	MEDI	NAT	TECH
Guest Category	Arts	HUM	/	77.6	87.9	67.5	73.0	79.8
		POLIT	79.2	/	87.7	68.4	69.8	80.0
		SOC	78.2	89.7	/	69.3	75.3	83.9
	Sciences	MEDI	72.3	85.8	86.9	/	77.6	81.2
		NAT	72.4	80.5	85.2	78.3	/	84.0
		TECH	63.9	80.2	83.8	63.9	73.3	/

When examined across all the guest categories, HUM has the highest token similarities with POLIT and SOC, both above 78%. These two guest categories are often believed to belong to a broader sense of category of 'Arts' as opposed to 'Sciences'. On the other hand, HUM has a comparatively lower token similarity with the other three guest categories of sciences, all under 73%, a 5% difference between the two groups. In other words, sub-categories of the 'Arts' seem to have a closer

relation with each other as opposed to a looser relation with sub-categories of the 'Sciences'. The host category POLIT is closely related to SOC in terms of token similarity and at same time is at a reasonable distance from the sciences category including MEDI, NAT and TECH. It is also worth noticing that although HUM is at the bottom of the guest-category list, the token similarity to the host category is as high as 77.6%. With SOC as the host category, the similarity scores do not show a significant gap between the arts and sciences categories. However, we still observe that the arts are grouped together while the science categories are grouped together at the bottom of the scale.

Next we take a look at the categories of sciences. With MEDI as the host category, the token similarities to the guest categories range from 63.9% to 78.3% according to Table 6. It is significant that the guest category, which belongs to the sciences, demonstrates a greater token similarity with MEDI. The arts categories, in contrast, show a less degree of token similarity, all under 70%. It is interesting, in this regard, to note that TECH has the lowest degree of similarity with MEDI, an unexpected observation that we shall discuss later. As expected, the science categories show a stronger affinity with each other than with the arts categories when NAT has the highest degree of similarity to MEDI and a comparatively lower degree of similarity to SOC, HUM and POLIT. Yet, TECH is observed again to be grouped with the arts categories. The category of TECH is strongly related to NAT by a token similarity of 84.0, which again indicates that they belong to a broad category of the 'Sciences'. The arts categories have a comparatively lower degree of similarity to the host category, with MEDI unexpectedly fall into the same group.

The above observations again seem to suggest a division between the two groups in terms of adjective use. We further examine the mean differences between the arts and sciences categories and the results are summarized in the tables 7 and 8.

4.4 Discussion

Our observations in terms of both *type similarity* and *token similarity* show that text categories can be categorized in a meaningful way according to the proportions of shared adjectives between text categories. A text category of arts often achieves a higher degree of similarity to other text categories of the same broad group but a comparatively lower degree of similarity to text categories of sciences. It is the same case with text categories of sciences. That is, text categories of sciences tend to have a stronger similarity with each other than their similarity to text categories of arts. However, we also observe some unexpected phenomena of text categorization. The text category of TECH is a typical example. TECH is normally to be considered as a sub-category of sciences by intuition. However, the empirical data in our study shows that TECH, as a guest category, is towards the bottom of similarity scale in both token and type similarities when compared with the host categories of MEDI and NAT. In other words, the similarity score between either MEDI and TECH or NAT and TECH is closer to the score of the arts category. There are two possible explanations. One is that the variations of adjective use may not be a perfect differentia to classify text categories although they can distinguish text categories of arts from those of sciences in most cases. The other reason lies in the inconsistency of text categories in the BNC.

Table 7. Token similarity from the viewpoint of Arts

		S_{token} **of Arts (Host Category)**		
			Sub Mean	Mean
Guest Category	Arts	HUM	78.7	83.4
		POLIT	83.7	
		SOC	87.8	
	Sciences	MEDI	69.5	79.0
		NAT	82.1	
		TECH	85.3	
Difference				4.4

Table 8. Token similarity from the viewpoint of Sciences

		S_{token} **of Sciences (Host Category)**		
			Sub Mean	Mean
Guest Category	Sciences	MEDI	71.1	76.4
		NAT	75.5	
		TECH	82.6	
	Arts	HUM	68.4	74.1
		POLIT	72.7	
		SOC	81.3	
Difference				2.3

According to [12], texts on Linguistics in the BNC are found to be classified into both the category of social science and the category of applied science. Therefore, the unexpected results in our investigation could also be caused by such an inconsistency in the pre-defined text categories.

5 Conclusion

In this paper, we described our investigation into the variations of adjective use across different text categories. Our assumption is that when differences in adjective use can be observed across different text categories from the same genre, those differences are more likely pertaining to characteristics of adjectives rather than stylistic features of genres. Six text categories under the same genre 'academic prose' were sampled from the British National Corpus for our investigation. By examining the proportions of adjectives shared between categories, we measure similarity of adjective use in terms of tokens and types, which we define as *token similarity* and *type similarity*. The empirical data show that, when measured in both tokens and types, adjectives in common use do differ across different text categories. Generally speaking, the differences have effectively classified the six text categories into two broad groups of arts and sciences. To put it differently, a text category belonging to arts tends to have a stronger similarity to the other text categories of the arts, but a comparatively weaker similarity to text categories of sciences. On the other hand, a text category of

sciences often achieves a higher degree of similarity to other text category of sciences, and a lower degree of similarity to text categories of arts. Since such categories are constructed according to their domain content, we have found it reasonable to conclude that adjectives demonstrate affinities according to domain and therefore can be used to classify texts according to domain. Our experiment results indicate that the variations of adjective use seem to be a quite reliable indicator to categorize different text categories in a meaningful way.

Acknowledgement. The research reported in this article was supported in part by a research grant (No 7002190) from City University of Hong Kong.

References

1. McNally, L., Kennedy, C.: Adjectives and Adverbs: Syntax, Semantics, and Discourse. Oxford University Press (2008)
2. Yamazaki, S. Distribution of Frequent Adjectives in the Wellington Corpus of Written New Zealand English. In: Saito, T., Nakamura, J., Yamazaki, S. (eds.) English Corpus Linguistics in Japan, pp. 63--75. Rodopi (2002)
3. Oberlander, J. and Gill, A.J. Language with Character: A Stratified Corpus Comparison of Individual Differences in E-mail Communication. Discourse Processes. 42, 239--270 (2006)
4. Chesley, P., Vincent, B., Xu, L., Srihari, R. K.: Using Verbs and Adjectives to Automatically Classify Blog Sentiment. In: AAAI-2006 Spring Symposium on "Computational Approaches to Analyzing Weblogs", pp. 27--33. Stanford, CA. (2006)
5. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168--177. ACM Press (2004)
6. Samson, C.: …*Is Different From*…: A Corpus-Based Study of Evaluation Adjectives in Economics Discourse. IEEE Transaction on Professional Communication. 49 (3), 236--245 (2006)
7. Sharoff, S. How to Handle Lexical Semantics in SFL: A Corpus Study of Purposes for Using Size Adjectives. In: Hunston, S., Thompson, G. (eds.) System and Corpus: Exploring Connections, pp. 184--205. David Brown BK. Co. (2004)
8. Rittman, R.: Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology. VDM Verlag (2008)
9. Rayson, P., Wilson, A., Leech, G.: Grammatical Word Class Variation within the British National Corpus Sampler. Language and Computers. 36, 295--306 (2001)
10. Biber, B., Johansson, S., Leech, G., Conrad, S. and Firegan, E. Longman grammar of spoken and written English. Harlow, England; [New York] : Longman. (1999)
11. The British National Corpus, Version 3 (BNC XML Edition), http://www.natcorp.ox.ac.uk/ (2007)
12. Lee, D.: Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. Language Learning & Technology, 5(3), 37--72 (2001)