Genre-Based In-Document Content Type Classification

Bei Yu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Duane Searsmith
National Center of Supercomputing Applications
University of Illinois at Urbana-Champaign

Abstract

This paper presents an in-document content classification approach that combines genre analysis and shallow natural language processing techniques to do document segment-level content classification. Given a document in a particular genre, we can classify the content of each segment (e.g. a paragraph) based on the recognized content type and typical linguistic features of the genre. The informal evaluative document genre is chosen as the test genre, and the online consumer review is used as the test data set. The classification results support our hypothesis that the content type of segment in a document of a particular genre could be predicted from the linguistic features. This approach may be used as a component in faceted search, multi-document summarization and many other information processing applications.

Introduction

A large amount of research has been done in document-level classification, but not much is done in the content classification within documents. Actually many information processing applications could benefit from being able to classify document segments based on content or style. For example, distinguishing between subjective information and factual information can help multi-document summarization systems to separate opinions from other content (Riloff and Wiebe, 2003). In-document classification can also support faceted search for higher speed and smaller result sets. As an example, research contributions are usually found in the abstract, the introduction or the conclusion section of an article while not possibly in the experiment or the future work segment. (Teufel and Moens, 2002).

This content classification task might be trivial for many formal document with strict and stable structure to identify the basic content units. Resume and Experimental research article are good examples for this type, but this is not the case for the argumentative papers in many social sciences in that they do not share a common subtitling paradigm. It is even worse for the informal text without subtitls to recognize the content unit. Some general-purpose approaches, such as Hearst's TextTiling algorithm (Hearst, 1997), are used to segment a document regarding the topic shift. while our goal is to learn the basic kinds of content within a document given a set of similar document. We define the similarity between the documents as that they belong to the same document genre.

In this paper we use genre analysis and shallow NLP techniques to classify the contents within the documents. After genre analysis we summarized the expected content scope, which were used as the guidelines for content class labels. The linguistic features were extracted using shallow NLP techniques, serving as the classification feature set. We ran two experiments on the online consumer review data chosen as a representative of the informal evaluative document genre. The experiment results shows the content type of segment in a document of a particular genre could be predicted from the linguistic features.

Genre Analysis

Numerous definitions of the concept "genre" have been suggested based on different focuses and scopes. Beghtol (2000) once separated "form" and "content" genre definitions. For example, Orlikowski and Yates (1993) defined "genre" as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form." This definition emphasized the form as an indispensable aspect of genre. Based on this definition we can recognize poems, resumes, homepages and so on as different genres, but it might be hard to discriminate an editorial from a reivew. Some document classification work is based on the "form" genre. Kwasnik et al. (2000) identified different Web document genres in order to improve the search effectiveness. Toms (2000) separated the content version and form version of a document. Because the visual form is defined at the

document level, we can use the "form" based "genre" concept to classify text in the document level, but it will not help the classification in more finer granularity, for example, a paragraph or a section.

On the other hand, Swales (1990) defined another "content" type of "genre" concept, which may help in-document classification. Swales defined "genre" as a class of communicative events with a shared set of communicative purposes and a common expectation of the content scope. For example, research article is a genre in that in the research community the readers have certain expectations for the content of a research article, such as the abstract, introduction, problem statement, methodology, result, discussion, conclusion and future work, etc.

The above two definitions shared a common aspect of genre as "recognized communicative purposes". Considering this aspect alone, we can typically classify the text as the types narrative, descriptive, argumentative, and expository (Kinneavy, Cope and Campbell, 1976). Biber (1993) used factor analysis to investigate the relation between the linguistics style features and the text types (registers). While Biber's study is in document level, it is very common that the narrative, descriptive, argumentative or expository text appear as pieces in one document at the same time. We question if it is still possible to classify smaller text unit by communicative purpose using the linguistic features. Connecting with Swales' genre definition, would we also classify the content parts within the documents of a particular genre?

In order to test these hypotheses, we use the online informal evaluative text as our test genre in that it should include all narrative, descriptive, argumentative and expository types. The subjects vary from consumer product reviews, user feedbacks, newsgroup discussions, etc. The reviews in epinions.com are chosen as the test data due to its good review quality as an independent evaluation website (Kuehl, 1999; Nielsen, 1999).

Applying Swales' definition, we found that readers of the online reviews expect to learn about the reason why they bought this product, the process of decision making, their positive or negative opinions about the aspects and the overall of this product, and the content to support their opinions such as the descriptions of the specific features they like or dislike, or other personal experiences, or the comparison between this brand and others, and finally they make a conclusion or complaint at the end. The content scope of the online reviews exactly match this expectation so that we could fit it into Swales' definition.

Experiment

We did a small-scaled experiment to test our hypotheses that:

1) text types regarding communicative purposes could be predicted at paragraph level from linguistic features;
2) content type within a particular genre could also be predicted at paragraph level from linguistic features.

The first 9 reviews of Dell Inspiron 8200 laptop on epinions.com. They were ordered decreasingly by reviewer' reputation, which assures the quality of the text. The documents are naturally segmented into 112 paragraphs in total. Here we assume paragraph is an appropriate text unit to contain a piece of relatively complete content.

To test hypothesis 1, we manually labeled the paragraphs as two types: expository and narrative. For simplicity reason we just use two types as a simple contrast. The expository text accounts for 66% and the narrative 34%. To test hypothesis 2, we classify the content into 6 classes: evaluation, fact, suggestion, personal experience, subtitle and others. The distributions are: evaluation (41:37%), fact (34:30%), suggestion (5:4%), personal experience (29:26%), subtitle (1: 1%), others (2: 2%).

13 linguistic features were collected in order to predict the paragraph type and content. They are:

1. *pas = VG_PAS / VG;*	2. *pp1 = PP1 / PP;*	3. *n = N / L;*
4. *pre = VG_PRE / VG;*	5. *pp2 = PP2 / PP;*	6. *adj = ADJ / L;*
7. *modal = VG_MODAL / VG;*	8. *pp3 = PP3 / PP;*	9. *adv = ADV / L.*
10. *passive = VG_PASSIVE / VG;*	11. *pp = PP / L;*	12. *vg = VG / L;*
13. *active = VG_ACTIVE /VG;*		

Given:

L: the paragraph length;	*PP*: the number of person pronouns;
VG: the number of predicates;	*PP1*: the number of first person pronouns;
VG_PAS: the number of past tense in predicates;	*PP2*: the number of second person pronouns;
VG_PRE: the number of present tense in predicates;	*PP3*: the number of third person pronouns (exclude "it");

VG_MODAL: the number of modals in predicates;	*N*: the number of nouns;
VG_PASSIVE: the number of passive voice in predicates;	*ADJ*: the number of adjectives;
VG_ACTIVE: the number of active voice in predicates;	*ADV*: the number of adverbs;

GATE, an NLP software, was used to extract the parts-of-speech tags of the words and the verb groups as predicates. The discriminant analysis module in SPSS package were used to classify the text type and the content of each paragraph.

Result

Table 1: The Distribution of Text Types in Documents.

reviews	Expository	Narrative
1	22	2
2	21	4
3	8	6
4	9	7
5	1	3
6	9	1
7	1	0
8	0	15
9	3	0
total	74	38
%	66%	34%

Table 2: Text Type Classification.
Overall Accuracy Rate: 82.1%

		Predicted Group Membership		Total
	Writing Mode	Narrative	Expository	
Original Count	Narrative	29	9	38
	Expository	11	63	74
%	Narrative	76.3	23.7	100.0
	Expository	14.9	85.1	100.0

Table 3: Content Classification. Overall Accuracy Rate: 60.7%

		Predicted Group Membership						Total
	Content	Evaluation	Fact	Suggestion	Personal Experience	Subtitle	Others	
original count	Evaluation	**23**	8	5	5	0	0	41
	Fact	7	**22**	3	1	0	1	34
	Suggestion	1	0	**3**	1	0	0	5
	Personal Experience	5	1	4	**17**	0	2	29
	Subtitle	0	0	0	0	**1**	0	1
	Other	0	0	0	0	0	**2**	2
%	Evaluation	**56.1**	19.5	12.2	12.2	.0	.0	100.0
	Fact	20.6	**64.7**	8.8	2.9	.0	2.9	100.0
	Suggestion	20.0	.0	**60.0**	20.0	.0	.0	100.0
	Personal Experience	17.2	3.4	13.8	**58.6**	.0	6.9	100.0
	Subtitle	.0	.0	.0	.0	**100.0**	.0	100.0
	Other	.0	.0	.0	.0	.0	**100.0**	100.0

Conclusion and Future Work

From the above results we see that the accuracy for text type classification is satisfying (82.1%), which means the chosen linguistic features describe the text type characters in this genre. In the second experiment, the classification accuracy is not so satisfying (60.7%). The lines we draw between evaluation, fact, suggestion and personal experience are not clear enough based on the chosen linguistic features. Various reasons affect the result. It is not easy to verify the validity of human label assignments. Actually the features we use are simple word-level and phrase-level frequencies. We have not have chance to explore many other text style features, such as the dependency between subjects and predicates (Nasukawa and nagano, 2001) and sentence patterns.

The two experiments are just a beginning. We simplified the communicative text types into two contrast ones. The number of the content class labels was also reduced. The data set is also not big enough. To generalize our conclusion, further experiments on larger data sets and more document genres are needed.

Some linguistic resources exist to determine automatically the subjectivity of words and phrases (Levin, 1993). Riloff and Wiebe (2003) recently used information extraction techniques to classify sentences to see whether it presents subjective expression or objective information. The above research would help improve the content classification task.

References:

1. Beghtol, C. (2000). The concept of genre and its characteristics. *ASIST Bulletin* Aug/Sept 2000.
2. Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2).
3. GATE. General Architecture for Text Engineering. http://gate.ac.uk/
4. Hearst, M. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1), 33-64.
5. Johnson, R.A. and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis. (5th edition) Prentice-Hall
6. Kinneavy, J., Cope, J. Q. and Campbell, J.W. (1976). Writing-basic modes of organization. Kendall/Hunt Publishing Company.
7. Kuehl, C. (1999). New world of Web reviews. *Internet World* v5 No 34, p52-4.
8. Kwasnik, B., Crowston, K., Nilan, M. and Roussinov, D. (2000). Identifying document genre to improve web search effectiveness. *ASIST Bulletin* Aug/Sept 2000
9. Levin B. (1993). English verb classes and alternations: a preliminary investigation. University of Chicago Press.
10. Nasukawa, T. and Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal*, 40(4).
11. Nielsen, J. (1999). Reputation managers are happening. *Jakob Nielsen's Alertbox* 09/05/1999. http://www.userit.com/alertbox/990905.html
12. Orlikowski, W.J. and Yates, J. (1994). Genre Repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly*, 33, 541-574.
13. Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*. 2003
14. Swales, J. (1990). Genre analysis: English in academic and research settings. In *Series of Cambridge Applied Linguistics*. Cambridge University Press.
15. Teufel, S. and Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4).
16. Tomes, E.G. (2000). Recognizing digital Genre. *ASIST Bulletin* Aug/Sept 2000

Discussion Points

21. What does the author mean by classification?

22. Does the research involve creation or implementation of a classification scheme?

23. How does the researcher use classification to improve the automated approach?

24. How do these methods compare to current human-generated approaches to classification?

25. How does the reported research expand our understanding of classification?

26. Does the research suggest an improvement over human-generated classification?

27. What do you think are the most important lessons learned in this research?

28. What do you think are the best practices reported in this research?

29. What would you recommend to the researcher as the next step in this approach?

30. Is there other related research that you would recommend the researcher become acquainted with?