Unsupervised Subjectivity-Lexicon Generation Based on Vector Space Model for Multi-Dimensional Opinion Analysis in Blogosphere

Hsieh-Wei Chen[1], Kuan-Rong Lee[2], Hsun-Hui Huang[1], and Yaw-Huang Kuo[1]

[1] Intelligent System/Media Processing (ISMP) Lab, Dept. of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC
[2] Dept. of Information Engineering, Kun Shan University, Yung-Kang, Tainan, Taiwan, ROC
hsiehwei@ismp.csie.ncku.edu.tw, leekr@mail.ksu.edu.tw, {hhhuang,kuoyh}@ismp.csie.ncku.edu.tw

Abstract. This paper presents an unsupervised framework to generate a vector-space-modeled subjectivity-lexicon for multi-dimensional opinion mining and sentiment analysis, such as criticism analysis, for which the traditional polarity analysis alone is not adequate. The framework consists of four major steps: first, creating a dataset by crawling blog posts of fiction reviews; secondly, creating a "subjectivity-term to object" matrix, with each subjectivity-term being modeled as a dimension of a vector space; thirdly, feature-transforming each subjectivity-term into the new feature-space to create the final multi-dimensional subjectivity-lexicon (MDSL); and fourthly, using the generated MDSL for opinion analysis. In the experiments, it shows that the improvement by the feature transform can be up to 31% in terms of the entropy of features. In addition, the subjectivity-terms and objects are also successfully and reasonably clustered in the demonstration of fiction review (literary criticism) analysis.

Keywords: weblogs, information retrieval, opinion mining, sentiment analysis, subjectivity classification, text mining.

1 Introduction

In recent years, more and more internet users own one or more blogs, with which many of the users enjoy sharing opinions and comments. To learn the spectrum of subjective opinions toward certain objects on the ever exploding blogosphere without an efficient opinion retrieval tool will be a difficult task in the study of an emerging area of information retrieval - opinion mining and sentiment analysis [1]-[3]. The process of sentiment analysis and opinion mining can be divided into the following four tasks: (i) topic-relevant document retrieval, (ii) opinionated document retrieval [4]-[8], (iii) opinion orientation classification (mostly polarity, i.e. positive or negative opinion, thumbs up or thumbs down) [9]-[12], and (iv) summarization and visualization [13][14]. Generally speaking, while task 1 can be categorized into the traditional document classification/clustering and retrieval task, tasks 2, 3, and 4 compose the major study area and pose challenges in opinion mining distinct from traditional information retrieval. Specifically, the three distinct major tasks usually involve either

D.-S. Huang et al. (Eds.): ICIC 2010, LNCS 6215, pp. 372–379, 2010.

an internal or external resource to determine the subjectivity and polarity semantic orientation (positive or negative) of a *phrase* and the overall opinion orientation in a document or a set of documents. The resource is commonly a *dictionary*, or a *lexicon* database (e.g. General Inquirer), which consist of entities (word, phrase, or syntactic patterns) tagged with their polarity orientations.

However, for more refined opinion analysis such as criticism analysis (e.g. criticism towards a film, a person, economy, politics, or literature), merely determining the polarity semantic orientation is not adequate Indeed, in literary criticism, a phrase with negative semantic orientation (e.g. poor, pity, and unfortunate) may imply sympathy, disproval, disappointment, tragedy, or sorrow etc. We cannot say that a reviewer has a negative attitude towards a character or a fiction merely because the overall polarity semantic orientation is negative. In political and social criticism, a phrase with positive semantic orientation (e.g. efficient, free, and respectful) may imply social justice, equality, freedom, instrumental rationality, value rationality, idealism etc. A person who is against assembly line system may imply he or she regards human value more important than instrumental rationality. It is difficult to find the author's value behinds an article by merely evaluating its polarity semantic orientation. A policy or a law has a statistically positive semantic orientation may merely due to the effect of advertisement and promotion. It is hard to know whether most people really approve of a government proposal or not. Therefore, the limitations of polarity opinion analysis are obvious and a more sophisticated *subjectivity-lexicon* is needed to achieve a deeper, more thorough, *multi-dimensional* opinion analysis (MDOA) system.

In this paper, a framework is proposed to learn the reviewers' opinions to the characters (objects) of fictions, rather than their recommendations of the fictions as a whole. The task of MDOA is to learn the dimensions of opinions towards objects, and to analyze opinions and objects from the learned dimensions. A *multi-dimensional* subjectivity-lexicon (MDSL) is learned and generated from corpus. A "subjectivity-term to object" matrix is firstly created from modeling fiction reviews in the blogosphere, and then transforming the subjectivity-terms into a feature-space. The transformation is based on measuring the similarity (or redundancy) between subjectivity-terms. The entities in the MDSL are represented by real-value vectors rather than polarity signs (+/-), which are commonly used in previous studies [9]-[12].

The rest of the paper is organized as following: in the next section, the proposed framework is formally described. In section 3, experiment results are shown. Section 4 is conclusions and future work.

2 Multi-dimensional Opinion Analysis

The proposed framework aims to generate a *vector-space-modeled subjectivity-lexicon* for fiction review analysis. The system structure is shown in Figure 1; the framework consists of four major parts, namely: (i) data collecting, (ii) preprocessing, (iii) transformation, and (iv) opinion analysis.

2.1 Data Collecting

The goal of the data collecting phase is to construct a raw dataset of fiction reviews for building the subjectivity-lexicon. In this phase, reviews (blog posts) of fictions are

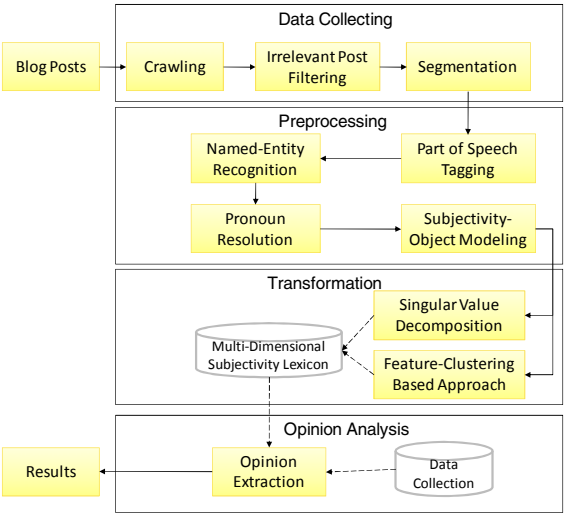

Fig. 1. The overview of multi-dimensional opinion analysis framework

crawled by querying the Google blog search service. Irrelevant posts (not a criticism or irrelevant to the given fictions) are filtered out by a rule-based scheme and human intervention; those remaining posts are considered as relevant reviews, denoted by $R = \{r_i | \ 1 \leq i \leq \text{number of reviews}\}$, and each r_i is segmented into individual sentences, denoted by $\{s_{i,j} | \ 1 \leq j \leq |r_i|\}$, where $|r_i|$=number of sentences in review r_i.

2.2 Preprocessing

The preprocessing phase is to create the "subjectivity-term to object" matrix, the source of the transformation phase, for forming the final vector-space-modeled subjectivity-lexicon. In the current implementation, only adjectives are considered as subjectivity-terms. In this phase, firstly, adjectives (JJs) are identified; secondly, since the objects are referred as person names or person pronouns mentioned in reviews, the named-entity recognition and pronoun resolution are needed; and thirdly, to create the JJ-object matrix, the dependencies between JJs (hereby, denoted as set *JJ*) and objects (hereby, denoted as set *O*) are modeled.

To build the JJ-object matrix, every token in review sentences is tagged by the Stanford POS Tagger and person names are identified by the Stanford Named Entity Recognizer. Also, person pronoun resolution (anaphora resolution) is performed; every third person pronoun is resolved to a person named-entity, which will be resolved to a corresponding fiction character next. Finally, a subjectivity-term to object matrix is constructed by modeling the dependencies of the objects (fiction characters) and their modifiers in reviews. The final "subjectivity-term to object" matrix is denoted as $C_{|JJ|x|O|}=[c_{ij}]$.

2.3 Transformation

The vector model is employed to represent an object $\hat{o}_j$ and a subjectivity-term $\hat{t}_i$. In the transformation phase, the subjectivity-term to object matrix is transformed into a

smaller feature space for three sakes: execution-efficiency in later use, finding semantic relationships between features, and finding subjective relationships between objects. The transformed subjectivity-term vectors are then saved as the multi-dimensional subjectivity-lexicon. Note that, entities in the subjectivity-lexicon are represented as real vectors, and the elements of the vectors are seen as the corresponding degrees between the subjectivity-terms and the un-semantically-tagged attributes, that is, how similar they are from the views of the learned *dimensions*.

The proposed framework employs four different transformation models: Weighting (TF-IDF), Singular Value Decomposition (SVD), Subjectivity-Clustering (SC) (or feature-clustering) based approach [19], and their combination. For the final constructed feature-space, represented by centroids $f_1,f_2,\ldots f_n$, subjectivity-term $\hat{t}_i=(\text{sim}(f_1,t_i), \text{sim}(f_2,t_i),\ldots,\text{sim}(f_n,t_i))$, and object $\hat{o}_j=(\text{P}(f_1|o_j), \text{P}(f_2|o_j),\ldots, \text{P}(f_n|o_j))$ where

$$\text{P}(f_k|o_j) = \log\left(1+\sum_{t_i\in JJ}\text{sim}\left(f_k,t_i\right)c_{ij}\right)$$

2.4 Opinion Analysis

Once the multi-dimensional subjectivity-lexicon (MDSL) is generated, the MDSL can be used for sentiment analysis and mining. A simple clustering is performed on the fiction review collection for evaluation and demonstration, which is given in the experiment section.

3 Experiments

The dataset was created from the blogosphere literature reviews about Gustave Flaubert's *Madame Bovary*, Jane Austen's *Pride and Prejudice*, Fyodor Dostoevsky's *Crime and Punishment*, and Leo Tolstoy's *War and Peace*. The total reviews collected are 191 blog posts, segmented into 8250 sentences, and 73 fiction characters. Taking negation factors into account, the dataset contained 1785 unduplicated subjectivity-terms, as shown in Table 1.

Table 1. Data set

Fiction	Reviews	Sentences	JJs	Characters
Madame Bovary	76	3137	843	13
Pride and Prejudice	61	2561	763	18
Crime and Punishment	23	748	281	19
War and Peace	31	1759	437	23
All Collection	191	8250	1785	73

In the SC transformation and clustering evaluation, the EM algorithm was employed. The generated MDSL was evaluated in terms of the feature entropy of subjectivity-terms and objects. The computation of entropy was specified in [19], in which a lower value indicates a better performance.

The example output of subjectivity-term clustering and object clustering of the proposed MDOA model are shown in Table 2 and 3. The examples show that the

Table 2. Example of subjectivity-term clustering results (MDOA)

Cluster	Words		
Cluster 1	true		
Cluster 2	attractive	decisive	lovely
	clever	smart	quick
	genuine	strong-willed	tolerable
	pretty	female	
Cluster 3	annoyed	generous	ironic
	handsome	ideal	sensible
	inferior	shy	snobbish
	proud	sympathetic	wealthy
Cluster 4	abusive	beloved	dull
	amiable	boring	emotional
	angry	confident	frivolous
	arrogant	crazy	self-centered
Cluster 5	-good	beautiful	desperate
	-happy	young	selfish
	afraid	broken-hearted	naïve
	unhappy	devoted	self-absorbed

Table 3. Example of object-clustering results (MDOA)

Cluster	Characters		
Cluster 1	Lheureux	Homais	Heloise Bovary
	Leon	Justin	Mrs. Bennet
	Berthe Bovary	Lydia Bennet	Miss Bingley
	Mr. Bennet	Charlotte Lucas	Georgiana Darcy
Cluster 2	Emma Bovary	Elizabeth Bennet	Fitzwilliam Darcy
	Charles Bovary		
Cluster 3	Rouault	Katerina Ivanovna	Zossimov
	Hippolyte		
Cluster 4	Rodolphe	Jane Bennet	George Wickham
	Mr. Bingley	Sonya Semyonovna	Rodion
	Mr. Collins	Andrew Bolkonski	Romanovich
	Pierre Bezukhov		

subjectivity-terms and objects clustered together are semantically or subjectively similar to each other. Indeed, in Table 2, the subjectivity-terms in cluster 2 indicate some attractive personalities; cluster 3 consists of some unattractive personalities; cluster 5 implies some unpleasant emotional states. However, cluster 1 consists of only one subjectivity-term, this phenomenon is due to the over-frequently used terms were not filtered out. In Table 3, we can see that characters with common personalities were clustered together; for instance, most characters in cluster 2 are supporting roles with lovely and innocent characteristics (except Lheureux, Homais, and Leon), while in cluster 4, most characters have honest personalities and draw sympathies (except Rodolphe and George Wickham). In cluster 2, those are major characters that encircle love, and are most widely and thoroughly discussed by readers. The result is meaningful and useful for opinion and sentiment analysis. The performance and accuracy may be improved if a larger and more complete dataset is built.

Table 4. Feature evaluation of subjectivity lexicon and objects (with negation and filtering)

Model	Entropy (# of attributes)	
	Subjectivity-Lexicon	Objects
MDOA-weighting	0.966 (73)	0.774 (195)
MDOA-weighting-SVD	0.964 (39)	0.800 (39)
MDOA-weighting-SC	0.921 (4)	**0.644 (4)**
MDOA-weighting-SVD-SC	0.908 (4)	0.805 (4)
MDOA-weighting-SVD-SC (without filtering)	**0.808 (4)**	0.751 (4)
SO	**0.799 (1)**	**0.650 (1)**

To compare with existing polarity opinion analysis approaches, the semantic orientation (hereby, denoted as SO) computed by PMI was implemented [9]. A comparison of the SO model with the proposed MDOA model is shown in Table 4. Here, the MODA model incorporates a cumulative frequency filter and negation factor. The subjectivity-terms (lexicon) and the subjectivity-term to object relationship matrix used in the SO model and the proposed MDOA model are the same; both are constructed by the MDOA model. However, in the SO model, the subjectivity-terms and objects were modeled by calculating their overall SO value, that is object $\hat{o}_j = (\Sigma_i [c_{ij} \mathrm{SO}(t_i)])$. We can see that, both the proposed MDOA model and the SO model had comparable result in terms of the entropy metric. But in the SO model, which was a polarity model, subjectivity-terms and objects were modeled as a value along one dimension.

The example output of object clustering with the SO model used is given in Table 5. We can see that, with the traditional polarity semantic orientation analysis, the clustering could only be done by separating the SO value range. Note that, most characters have negative SO value. Besides, the magnitude of the value could provide little, if any semantic meanings. The magnitude of the value merely implies how often the objects were mentioned in reviews. As mentioned before, a lot of attitudes (such as sympathy, disapproval, disappointment, tragedy, sorrow etc) have negative SO, the SO value alone cannot provide adequate information to distinguish and determine the differences of the subjectivity-terms as well as the objects. Polarity opinion analysis is not adequate for more refined analysis. On the other hand, the proposed MDOA provided a promising result for criticism analysis.

Table 5. Example of object-clustering results (SO)

Cluster	Characters (SO)		
Cluster 1	Leon (4.89)		
Cluster 2	Lheureux (-0.54)	Berthe Bovary (-10.21)	Heloise Bovaryais (0.61)
	Elizabeth Bennet (-5.42)	Homais (-1.89)	Lydia Bennet (-7.545)
	Mr. Bingley (-5.34)	Sofya Semyonovna (0.85)	Miss Bingley (-1.64)
	Mr. Collins (-9.27)	Natasha Rostova (-5.62)	Lizaveta Ivanovna (1.05)
		George Wickham (-4.50)	
Cluster 3	Emma Bovary (-42.1)	Rodion Romanovich (-28.00)	Pierre Bezukhov (-22.23)
	Charles Bovary (-32.56)		
Cluster 4	Rodolphe (-12.37)	Fitzwilliam Darcy (-12.98)	Catherine Bennet (-12.16)
	Jane Bennet (-2.16)	Semyon Zakharovitch (-12.39)	

4 Conclusions and Future Work

In this paper, the process flow of opinion mining and sentiment analysis is introduced, and the need of a more sophisticated subjectivity-lexicon for multi-dimensional opinion analysis (MDOA) is described. Moreover, the proposed multi-dimensional subjectivity-lexicon (MDSL) generation framework engineered for analyzing literary criticism, blog posts of fiction reviews, is also formally described. Finally, the merits of this proposed approach are illustrated in the experimental results: first, a MDSL can be generated by learning the usage of subjectivity-terms; secondly, a performance evaluation in terms of feature entropy achieves an up to 31% improvement by applying the proposed transformations; thirdly, reasonable subjectivity-term and object clustering results were obtained in the demonstration of fiction review analysis, and fourthly, the comparison and limitation of traditional polarity opinion analysis is also demonstrated. The proposed framework can also be adapted to other domains which require MDOA techniques.

In the future, we intend to test the proposed MDOA framework with a larger scale database that includes other literature reviews. The framework is planned to incorporate sentence–level opinion extraction and classification mechanisms. To improve the accuracy, more subjectivity-term selection algorithms and criteria, such as graph-based clustering or one which integrates the existing lexicon (General Inquirer), will also be studied. In addition, more completed experiments, such as precision and recall of retrieval and classification results evaluated by expert judgment, will also be performed.

Acknowledgment

This study was supported by the National Science Council, Taiwan, Grant Nos. NSC 98-2221-E-168-032 and NSC97-2221-E-006-144-MY3.

References

1. Liu, B.: Sentiment Analysis and Subjectivity. In: Handbook of Natural Language Processing, 2nd edn. (2010)
2. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1–135 (2008)
3. Tang, H., Tan, S., Cheng, X.: A Survey on Sentiment Detection of Reviews. Expert Systems with Applications 36, 10760–10773 (2009)
4. Riloff, E., Wiebe, J., Wilson, T.: Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL, vol. 4, pp. 25–32 (2003)
5. Riloff, E., Patwardhan, S., Wiebe, J.: Feature Subsumption for Opinion Analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, pp. 440–448 (2006)
6. Riloff, E., Wiebe, J.: Learning Extraction Patterns for Subjective Expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 105–112 (2003)

7. He, B., Macdonald, C., He, J., Ounis, I.: An Effective Statistical Approach to Blog Post Opinion Retrieval. In: Proceeding of the 17th ACM conference on Information and knowledge management. ACM, Napa Valley (2008)
8. Zhang, W., Yu, C., Meng, W.: Opinion Retrieval from Blogs. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. ACM, Lisbon (2007)
9. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 417–424 (2002)
10. Pang, B.: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the ACL-2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
11. Shandilya, S.K., Jain, S.: Opinion Extraction and Classification of Reviews from Web Documents. In: IEEE International Advance Computing Conference, pp. 924–927 (2009)
12. Mei, Q., Ling, X., Wondra, M., Zhai, C.M.: Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In: Proceedings of the 16th International Conference on World Wide Web. ACM, Banff (2007)
13. Lu, Y., Zhai, C.X., Sundaresan, N.: Rated Aspect Summarization of Short Comments. In: Proceedings of the 18th International Conference on World Wide Web, pp. 131–140. ACM, Madrid (2009)
14. Chang, C.H., Tsai, K.C.: Aspect Summarization from Blogsphere for Social Study. In: The Seventh IEEE International Conference on Data Mining Workshops, pp. 9–14 (2007)
15. Marneffe, M.C., Manning, C.D.: The Stanford Typed Dependencies Representation. In: Workshop on Cross-framework and Cross-domain Parser Evaluation (2008)
16. Ge, N., Hale, J., Charniak, E.: A Statistical Approach to Anaphora Resolution. In: Proceedings of the Sixth Workshop on Very Large Corpora, pp. 161–171 (1998)
17. Brennan, S.E., Friedman, M.W., Pollard, C.J.: A Centering Approach to Pronouns. In: The Proceedings of the 25th Annual Meeting on Association for Computational Linguistics, Stanford, California, pp. 155–162 (1987)
18. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. Comput. Linguist. 20, 535–561 (1994)
19. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 301–312 (2002)