Actor Identification and Relevance Filtering in Movie Reviews

Julia Romberg
Heinrich-Heine-Universität Düsseldorf
Institut für Informatik
Universitätsstraße 1
40225 Düsseldorf, Deutschland
Julia.Romberg@uni-duesseldorf.de

ABSTRACT

With a large amount of data it is not always useful to run analyses on the entire corpus. Sometimes, it is helpful to previously preprocess data by filtering relevant information in order to form a fitting basis for the examination of particular aspects such as sentiment analysis. As a result, the amount of data that needs to be explored is reduced and concentrated, and thus the performance is enhanced. For example, a correct recognition of the rating of acting performances in movie reviews assumes that only judgements on the movie's actors are used as a basis.

In this paper, we discuss different approaches for a rule-based selection of sentences from movie reviews. Our aim is the filtering of sentences in order to facilitate analyses about single actors. Thereby actor identification is used to preselect a set of sentences that mention a specific actor. This is done individually for every actor involved in the movie. Furthermore, filtering is used to identify sentences that not only mention an actor but also state facts about him. To evaluate the developed methods, a test corpus consisting of ten movies with 30 reviews each, taken from the online movie platform IMDb, was built. Based on this data and the presented feature selection rules, an average F_1 score of 77.9% is achieved as best result.

Categories and Subject Descriptors

I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing—*language parsing and understanding, text analysis*; D.2.8 [**DOCUMENT AND TEXT PROCESSING**]: Document Capture—*document analysis*; H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*data mining*

Keywords

Natural Language Processing, Text Mining, Text Analysis, Coreference Resolution, Movie Reviews, IMDb

28th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 24.05.2016 - 27.05.2016, Nörten-Hardenberg, Germany.

1. INTRODUCTION

The internet is a highly frequented medium for the description and rating of a wide variety of matters. Products, such as furniture, books, and movies, can be reviewed by users to reflect their experiences and to help other users to decide whether they should buy a product or not. In addition, review platforms provide the possibility of reading different opinions and comparing them with the own one. Especially in the field of entertainment, there is a a need of exchanging opinions and notes after watching a movie or a series, reading a book or listening to a music album and the resulting reviews offer great potential for analyses.

In this work we focus on movie reviews. Movie reviews mostly consist of a summary of the movie plot and an overall assessment of the movie. In addition, many reviewers expand on the cast and, hence, on the acting performance the actors and actresses are showing in this particular movie. On the basis of reviews, various aspects can be analysed. One way to support a user's decision for or against watching a movie could be an automatic text analysis that gives a general overview of the popularity of the movie instead of letting him read several reviews.

At least of equal interest is the individual performance of the participating actors. By examining the reviewers' impressions, the quality of acting can be investigated and, as a result, recommendations can be given matching the user's preferences. Furthermore general advices can be given such as film rankings for single actors and overall rankings for outstanding performances in movies. In order to facilitate these examinations and to increase the efficiency, it is necessary to extract sentences that are of interest for the later analysis.

Mainly, we want to identify actors in movie reviews. This is the basis for analysis on the actor's performances and acting skills. Therefore we use name recognition and coreference resolution approaches. To select informative sentences, filtering techniques are then applied.

2. RELATED WORK

In the field of movie review mining different aspects have been examined. Especially opinion mining and sentiment analysis are important issues. Overall sentiment classification of movie review documents into negative and positive opinions [10, 12, 5] or into deeper levels of granularity [6, 13] has been investigated intensely.

Closely related to opinion mining is informative review summarisation. This has been analysed for product and customer reviews in general [1, 3] and for movie reviews [9, 14] in particular. In [3] product features in customer reviews are mined, positive and negative phrases are identified and then the discovered information is used to summarize the reviews. Thereby product features are product characteristics such as picture quality for a digital camera. As the structure of movie reviews differs from that one of product and customer reviews, extra research was done. In [9] overall sentiment analysis on movie reviews is done but on subjective parts of the documents. To extract those parts, techniques for finding minimum cuts in graphs are used. Subsequently, reviews are summarized for a cleaner representation of the polarity. These concepts are not applicable for our purpose as they focus on overall sentiment analysis.

Concerning the approaches evaluated in our paper, [14] is of interest: In there, movie reviews are mined in order to determine whether opinions are negative or positive and summarized. Thereby features on which the reviewers express their opinions are extracted. Features are split in different classes such as "*screenplay*" and "*music and sound effects*". The class "*actor and actress*" is treated explicitly. To ease the actor identification, movie cast lists are used. Then first name only, last name only, full name and abbreviations are used to identify a feature. The obtained features are used to mine feature-opinion pairs on which the sentiment analysis is based. In contrast to our work, the focus in [14] is not on actors in particular. Instead, a number of classes of features is used to examine opinions for which reason more general approaches were chosen. Actor identification through name recognition was used. In [4] the opinion mining approach presented in [14] is extended with anaphora resolution. However, name recognition by including spelling mistakes is not further examined. We also add coreference resolution to improve the identification, but instead of searching directly for opinion target-opinion words pairs, we take reference to a list of words that may indicate irrelevant identifications for eventual sentiment analysis.

3. ACTOR IDENTIFICATION ON SENTENCE LEVEL

Aiming for an extraction of sentences that refer to specific actors, we first have to define what we consider as an extraction. Under the assumption that we have at least one review, a set of sentences exists. Not every sentence contains information related to an actor. Thus, a filtering of relevant sentences is needed. As the focus is on individual actors, the selection also has to be done distinctly:

Definition 1. *Let S be the set of all sentences of a movie review. For any actor a_j exists a subset $S_{a_j} \subseteq S$ of relevant sentences.*

Below, after the definition of relevance, approaches for actor identification and filtering of irrelevant sentences are presented.

3.1 Definition of relevant sentences

First, the relevance of a sentence according to a certain actor has to be clarified. Not every mention of an actor really refers to him. For the purpose of this study, only sentences that not only take reference to but also contain information about an actor are interesting.

Definition 2. *Let $s_i \in S$ be any sentence of a movie review and let a_j be an actor. Then $s_i \in S_{a_j}$ if and only if s_i contains information about the actor himself.*

Thereby, the knowledge about the played role is excluded since this is assumed as previously known and gives no further description about the actor and his play. In addition, only cases in which it is clear that the actor is meant are regarded.

To clarify the difference between relevant and irrelevant information concerning the played role consider the following sentences:

- Leonardo DiCaprio plays Hugh Glass.
- Leonardo DiCaprio plays Hugh Glass very well.

The first sentence only states the relation of DiCaprio and his role in the movie "*The Revenant*" without any further information regarding the actor, whereas the latter one additionally describes the quality of DiCaprio's play. Therefore, only "*Leonardo DiCaprio plays Hugh Glass very well.*" is relevant for us.

Equally, sentences like "*Hugh Glass is played by Leonardo DiCaprio*" and "*Hugh Glass (Leonardo DiCaprio) is out for vengeance.*" are not relevant for Leonardo DiCaprio according to Definition 2.

3.2 Approaches

Several approaches for the extraction of relevant sentences are developed. The first approach focuses on the explicit use of names, in the next one coreferences are considered and the last approach is about filtering irrelevant sentences.

3.2.1 Names

The most naive approach for finding an actor in a sentence is the search for his name. The full name of an actor can be considered as well as only parts of the actor's name. In general, a person's full name is used in the beginning of a text passage for the purpose of introduction. Then, the first or the last name can serve as representatives.

In many cases movie reviews contain spelling mistakes. To take that into account, the Levenshtein distance [7] is used for allowing deviations: The Levenshtein distance of two words is the minimum number of edit operations that are necessary to convert one word into the other. Permitted edit operations are the insertion of characters, the replacement of characters and the deletion of characters.

3.2.2 Coreferences

On the one hand, names and parts of names are important clues for the reference to an actor. On the other hand, personal pronouns are used as well to substitute full names. According to A. Radford [11], two expressions are coreferential if they refer to the same entity. In the present case, entities are actors and all the expressions building a coreference towards a specific actor within a movie review are to be found.

For example, given the following sentences: "***DiCaprio*** *plays Jack in James Cameron's Titanic.*" and "*This year* ***he*** *finally won the Academy Award for Best Actor.*". Both sentences are about Leonardo DiCaprio as referred entity and the coreferential expressions are "*DiCaprio*" in the first one and "*he*" in the second one.

Through intense examination of the test corpus (see section

4.1), "*he*" and "*she*" were identified as the most frequently used single personal pronouns. For this reason, we focus on "*he*" as male coreferential expression and "*she*" as female coreferential expression.

Please note that also descriptive nouns like "*the actor*" provide an additional way of referring to an entity. They are not explicitly examined in this paper as there are many different ways of describing someone. Additionally, they depend on the appearance and characteristics of the referenced person. This explicit knowledge is not given through the resources we use here. The special case of "*the actor*" is supposed to be used independently for every actor and thus more frequently in general but after examination of the test corpus (see section 4.1) this is not the case. However, a complete coreference resolution system is run for comparison with the here developed techniques in section 4.2.

3.2.3 *Filtering*

To take into account the relevance of information as mentioned previously, the feature set for an actor obtained by the name and the coreference approach has to be filtered. The filtering is motivated by the three kinds of irrelevant sentences presented in the section 3.1.

A sentence like "*Leonardo DiCaprio plays Hugh Glass.*" gives no information about the actor and hence is irrelevant. However, sentences of this kind are included in the feature set at this point because of the explicit mention of the actor's name. To correct this, the actor's name followed by "*plays*" is not treated as a mark for relevance.

The phrase "*played by*" as in "*Hugh Glass is played by Leonardo DiCaprio*" is also an indicator for irrelevance and is similarly taken out of the sentence set of the involved actor.

The last case is a note about the actor in brackets. An example is "*Hugh Glass (Leonardo DiCaprio) is out for vengeance.*". To solve this, actor names in brackets are filtered out.

It should be noted that these filtering rules do not exclude every sentence that includes one of the cases from the set of relevant sentences of a specific actor. "*Leonardo DiCaprio plays Hugh Glass and DiCaprio is great.*" is still correctly detected as relevant for DiCaprio.

4. APPLICATION

For the evaluation, the approaches explained in section 3.2 are now implemented. First, an overview of the used tools and the database on which we evaluate is given. Then the results are discussed.

4.1 Database and pipeline

Internet platform IMDb[1] is chosen as a freely available database. Ten films with 30 reviews each from different genres are selected arbitrarily to build a test corpus for the evaluation of the approaches described above. The selected films are "*Blue Valentine*", "*Cruel Intentions*", "*Fast & Furious 7*", "*Philadelphia*", "*Pretty Woman*", "*Sex and the City*", "*The Lord of the Rings: The Fellowship of the Ring*" and "*Walk the Line*". To provide a good text quality, the first 30 reviews according to the IMDb filter "*Best*" are extracted. For every movie the cast list and for every actor the gender are crawled. Every review is processed with the Stanford CoreNLP[8]. Stanford CoreNLP was chosen for several reasons. On the one hand, basis features, such as tokenization and part of speech-tagging, seem to perform well. On the other hand, they provide a coreference resolution system we work with for comparison (see section 4.2). Subsequently, based on the cast list, each sentence is manually annotated with actor names. At times, poor text quality makes it difficult to understand references. Furthermore, the distinction between roles and actors is not consistent in some reviews. Instead of using the role name when talking about a character, the actor's name is used. Biopics are hard to handle as well. For example the movie "*Walk the Line*" is about the life of Johnny Cash but there is a distinction between the role Johnny Cash and the real Cash in the reviews.

The evaluation is based on each actor of the ten movies in the test corpus to which, at least one time, reference has been made by some reviewer. For every movie's actor, recall and precision are built. The F_1-measure was chosen to include both measures, recall and precision, for comparison of the approaches. Then, a F_1 score for each movie is calculated by averaging the F_1 scores of the involved actors. In order to compare approaches, an average F_1 score is built over all movies.

4.2 Evaluation

The presented approaches are evaluated sequentially. Actor identification by the full name serves as a baseline as this is the most intuitive way of finding sentences that could be relevant. The average F_1 score for this baseline is 61.7%.

4.2.1 *Actor identification through First and Last Names*

Firstly, we evaluate if the use of parts of the actor's name can improve the baseline. Therefore, it is checked if the test sentences contain the first or the last name. Due to spelling mistakes, we experiment with a constant Levenshtein distance as a threshold and with a threshold that is relative to the word length. For a constant threshold, the values $0, 1, 2$ and 3 are tested and for a threshold in respect of the worth length, the values $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}$ and $\frac{1}{16}$ are tested.

The inclusion of selection through first names yields an average score of 64.7%. The application of the Levenshtein distance does not lead to better results. For short words already one edit operation can change their meaning. First names tend to be short and, therefore, respond strongly to allowed deviations. For instance the name "*Ron*" transforms into "*on*" with just one delete operation.

By contrast the reached score regarding last names can be maximized to 73.5% by using a Levenshtein distance of $\frac{1}{3}$ of the worth length. This behavior is as expected: Because of the fact, that last names are usually longer, the idea of detecting spelling mistakes like "*Gossling*" (correctly "*Gosling*") with the Levenshtein distance works.

After recognizing that both approaches individually enhance the results, we take a closer look on combinations of them. Only the search for the first and the last name without consideration of spelling mistakes leads to a barely noticeable improvement (73.6%). Overall, by combining first and last name approaches after running the baseline, the recall increases but the precision worsens.

By comparison, last names are used more frequently to

[1] `http://www.imdb.com`

Table 1: Comparison of the different coreference approaches.

	(0)	(1)	(2)	(3)	(4)
b+first&last	0.736	0.735	0.733	0.736	0.659
b+last($\frac{1}{3}$)	0.735	0.739	0.734	0.736	0.658

talk about actors in the test corpus whereas that does not hold true for roles. The former may originate in the missing personal connection towards the actor. However, as characters in movies are merely speaking to each other using full names, in the test corpus, first names are more frequently used for descriptions of roles.

These results are achieved with lowercased words. Case sensitiveness is also investigated with the same parameters. In the test corpus it is observed that reviewers do not always use capitalization correctly. Some tend to use no capitals at all while others vary or write names completely in upper case. As expected, the resulting scores are slightly worse in general and thus we decide to proceed without spending attention to the case of capitalization.

4.2.2 Actor identification through Coreferences

After name based strategies, approaches that use coreferences are evaluated. Based on the name recognition, the personal pronouns "he" and "she" are assigned. For every actor and already selected sentence, the next sentence is observed. Because of the known gender of each actor, a distinction in female and male is possible. If the actor is a woman, the preceding sentence will be searched for "she". Equally, for a male actor the word "he" is required.

As one single personal pronoun only refers to one person, an assignment is only done in a clear case. This means that only one actor of the same gender is found in a sentence and the next sentence contains the fitting personal pronoun. For comparison we also analyse an assignment of all male/female actors of a sentence. Table 1 shows the results. The first row displays the F_1 scores that are reached by combining name recognition through first and last name without allowed deviation (**b+first&last**) with the below explained coreference approaches. The second row refers to name recognition by only the last name with threshold $\frac{1}{3}$ of the length (**b+last($\frac{1}{3}$)**) as starting point for the coreference resolution. The columns show the coreference techniques that are combined with **b+first&last** and **b+last($\frac{1}{3}$)**. The first column (0) shows the score only achieved by **b+first&last** and **b+last($\frac{1}{3}$)**. (1) stands for the first coreference resolution approach, that only assigns "he" and "she" in a clear case (as described above), whilst (2) represents the approach, in which "he" and "she" are also assigned in ambiguous cases. The more restrictive version achieves better results whereas the other version even leads to decreases in the performance. However the improvement in regard to the names-only approach is poor. The best F_1 score 73.9% is reached by **b+last($\frac{1}{3}$)** and the restrictive pronoun resolution (1).

Stanford's CoreNLP includes a coreference resolution system. Taking this into account, two further techniques are developed and the results can be seen in Table 1 as well. Similar to (1), in (3) for every assigned actor it is checked if a coreference chain to the next sentence exists. If so the occurrence of "he" or "she" depending on the actor's gender is the critical factor for a relevant coreference. Since Stanford's coreference solution in general reveals coreferences of various types, in (4) every chain to the next sentence is comprehended as relevant and as an indicator for an assignment. Although an improvement is expected, (3) does not significantly change the F_1 score. By incorporating more aspects of coreference, the results are even worse.

For completeness, each of the name recognition approaches mentioned in section 4.2.1 has been combined with (1), (2), (3), and (4). The two that perform best were listed in Table 1.

4.2.3 Improvement through Filtering

Finally, the approaches for filtering are evaluated. The phrases "*actor/coreference* plays", "played by *actor/coreference*" and "(*actor*)" are used for this purpose. They are tested with all possible combinations for names and pronoun resolution as described in section 4.2.1 and 4.2.2. Using "played by *actor/coreference*" and "(*actor*)" as filter after the baseline and assignment by the last name under Levenshtein distance ($\frac{1}{3}$ of word length) with subsequent use of (1) results in a F_1 score of 77.9%.

The individual F_1 scores according to this approach are shown in Figure 1. Each bar represents one of the ten movies that are considered in the test corpus. The y-axis shows the reached F_1 scores. Depending on the movie, the F_1 scores vary widely. For "*Pretty Woman*" only a score of 55.8% is reached and also the score of 65.7% for "*Sex and the City*" is comparatively poor. In contrast, for "*Sicario*" a score of 89.8% is achieved. Closer inspection of the test corpus reveals that role and actor names are mixed more in the included reviews for "*Pretty Woman*" and "*Sex and the City*" than in the other movies' 30 reviews.

5. CONCLUSIONS AND FUTURE WORK

In this work different approaches for the assignment of actors to sentences in movie reviews were discussed. An average F_1 score of 77.9% was achieved by the presented approaches. Thereby the performance varies strongly depending on the movie. For most of the tested movies, good scores between 75% and 90% could be reached. For one movie, only a F_1 score slightly above 55% was reached, which must be rated as rather poor.

Names offer good reference points for the treatment of an actor in a sentence. In order to exclude irrelevant sentences, in effect sentences that only mention the actor without giving any information, filtering techniques are useful. Especially the elimination of phrases like "*role(actor)*" enhanced the performance. To improve the results, further filtering approaches are to be developed. Likewise the two other filtering approaches presented offer a potential for further research on descriptive sentence structures used with "*plays*" and "*played by*". Coreference resolutions do not lead to a significant improvement in our experiments. In fact, some of them even lead to a decrease of the F_1 score. A closer examination of co-referential expressions for actors is planned. The recognition of persons by paraphrases has not further been discussed in the course of this paper. Nevertheless this aspect should not be omitted and needs intense research. Other rule-based techniques that might achieve better results can be developed. Besides different strategies such as machine learning concepts may be useful for the invented

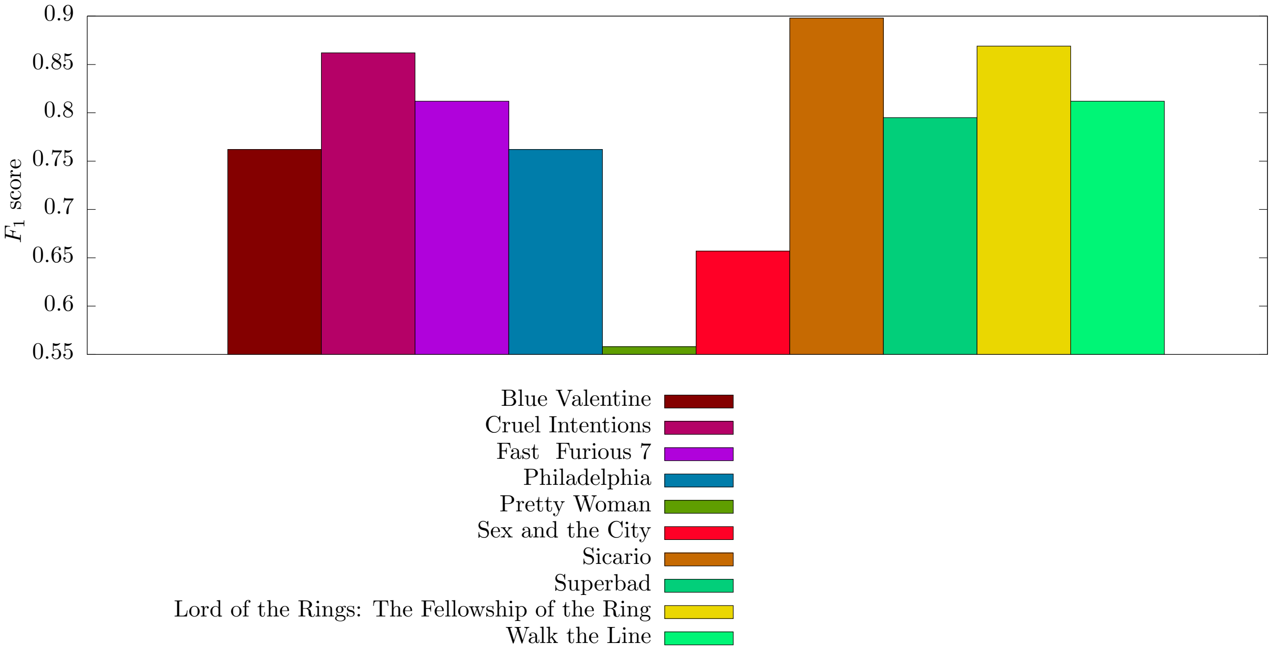

Figure 1: F_1 scores for the movies and reviews that are contained in the test corpus when using "*played by actor/coreference*" and "*(actor)*" as filter after the baseline and assignment by the last name under Levenshtein distance ($\frac{1}{3}$ of word length) with subsequent use of (1).

problem. In general, exclamations like "*Great performance!*" that state facts about actors and their play seem to be hard to solve. These sentences are not correctly assigned with our approaches as neither a name is mentioned nor an explicit coreference is used. Likewise the mix-up of role and actor names can not be handled without contextual knowledge. Based on the assignment of sentences, the reviewer's view towards an actor can be analysed. In future work, we want to examine the polarity of things being said about an actor in movie reviews. Therefore, a classification in {positive, neutral, negative} is done as an initial approach. A code book will be used for a manual classification of the test corpus. As an instrument, SentiWordNet [2] or Stanford's sentiment analysis tool could be applied.

6. REFERENCES

[1] M. Abulaish, Jahiruddin, M. N. Doja, and T. Ahmad. Feature and Opinion Mining for Customer Review Summarization. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, 2009.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), may 2010.

[3] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[4] N. Jakob and I. Gurevych. Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 263–268. Association for Computational Linguistics, 2010.

[5] A. Kennedy and D. Inkpen. Sentiment Classification of Movie Reviews using Contextual Valence Shifters. *Computational intelligence*, 22(2):110–125, 2006.

[6] A. Koumpouri, I. Mporas, and V. Megalooikonomou. Evaluation of Four Approaches for "Sentiment Analysis on Movie Reviews": The Kaggle Competition. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, EANN '15, pages 23:1–23:5. ACM, 2015.

[7] V. I. Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*. Elsevier Science & Technology, 1965.

[8] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[9] B. Pang and L. Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04. Association for Computational Linguistics, 2004.

[10] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs Up?:

Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. Association for Computational Linguistics, 2002.

[11] A. Radford. *English Syntax: An Introduction.* Cambridge, UK: Cambridge University Press, 2004.

[12] K. Tsutsumi, K. Shimada, and T. Endo. Movie Review Classification Based on a Multiple Classifier. In *Proceedings of the annual meetings of the Pacific Asia conference on language, information and computation (PACLIC)*, pages 481–488, 2007.

[13] C. Whitelaw, N. Garg, and S. Argamon. Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 625–631. ACM, 2005.

[14] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie Review Mining and Summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.