Last but *Definitely* not Least:

On the Role of the Last Sentence in Automatic Polarity-Classification

Israela Becker and Vered Aharonson
AFEKA – Tel-Aviv Academic College of Engineering
218 Bney-Efraim Rd.
Tel-Aviv 69107, Israel
{IsraelaB,Vered}@afeka.ac.il

Abstract

Two psycholinguistic and psychophysical experiments show that in order to efficiently extract polarity of written texts such as customer-reviews on the Internet, one should concentrate computational efforts on messages in the final position of the text.

1 Introduction

The ever-growing field of polarity-classification of written texts may benefit greatly from linguistic insights and tools that will allow to efficiently (and thus economically) extract the polarity of written texts, in particular, online customer reviews.

Many researchers interpret "efficiently" as using better computational methods to resolve the polarity of written texts. We suggest that text units should be handled with tools of discourse linguistics too in order to reveal where, within texts, their polarity is best manifested. Specifically, we propose to focus on the last sentence of the given text in order to efficiently extract the polarity of the whole text. This will reduce computational costs, as well as improve the quality of polarity detection and classification when large databases of text units are involved.

This paper aims to provide psycholinguistic support to the hypothesis (which psycholinguistic literature lacks) that the last sentence of a customer review is a better predictor for the polarity of the whole review than other sentences in the review, in order to be later used for automatic polarity-classification. Therefore, we first briefly review the well-established structure of text units while comparing notions of topic-extraction vs. *our* notion of polarity-classification. We then report the psycholinguistic experiments that we ran in order to support our prediction as to the role of the last sentence in polarity manifestation. Finally, we discuss the experimental results.

2 Topic-extraction

One of the basic features required to perform automatic topic-extraction is sentence position. The importance of sentence position for computational purposes was first indicated by Baxendale in the late 1950s (Baxendale, 1958): Baxendale hypothesized that the first and the last sentence of a given text are the potential topic-containing sentences. He tested this hypothesis on a corpus of 200 paragraphs extracted out of 6 technical articles. He found that in 85% of the documents, the first sentence was the topic sentence, whereas in only 7% of the documents, it was the last sentence. A large scale study supporting Baxendale's hypothesis was conducted by Lin and Hovy (Lin and Hovy, 1997) who examined 13,000 documents of the Ziff-Davis newswire corpus of articles reviewing computer hardware and software. In this corpus, each document was accompanied by a set of topic keywords and a small abstract of six sentences. Lin and Hovy measured the yield of each sentence against the topic keywords and ranked the sentences by their average yield. They concluded that in ~2/3 of the documents, the topic keywords are indeed mentioned in the title and first five sentences of the document.

Baxendale's theory gained further psycholinguistic support by the experimental results of Kieras (Kieras, 1978, Kieras, 1980) who showed that subjects re-constructed the content

Proceedings of the ACL 2010 Conference Short Papers, pages 331–335,
Uppsala, Sweden, 11-16 July 2010.

of paragraphs they were asked to read by relying on sentences in initial positions. These finding subsequently gained extensive theoretical and experimental support by Giora (Giora, 1983, Giora, 1985) who correlated the position of a sentence within a text with its degree of informativeness.

Giora (Giora, 1985, Giora, 1988) defined a discourse topic (DT) as the least informative (*most uninformative)* yet *dominant* proposition of a text. The DT best represents the redundancy structure of the text. As such, this proposition functions as a reference point for processing the rest of the propositions. The text position which best benefits such processing is text initial; it facilitates processing of oncoming propositions (with respect to the DT) relative to when the DT is placed in text final position.

Furthermore, Giora and Lee showed (Giora and Lee, 1996) that when the DT appears *also* at the end of a text it is somewhat informationally redundant. However, functionally, it plays a role in wrapping the text up and marking its boundary. Authors often make reference to the DT at the end of a text in order to summarize and deliberately recapitulate what has been written up to that point while also signaling the end of discourse topic segment.

3 Polarity-classification *vs.* Topic-extraction

When dealing with polarity-classification (as with topic-extraction), one should again identify the *most uninformative* yet *dominant* proposition of the text. However, given the cognitive prominence of discourse final position in terms of memorability, known as "*recency effect*" (see below and see also (Giora, 1988)), we predict that when it comes to polarity-classification, the *last* proposition of a given text should be of greater importance than the *first* one (contrary to topic-extraction).

Based on preliminary investigations, we suggest that the DT of any customer review is the customer's evaluation, whether negative or positive, of a product that s/he has purchased or a service s/he has used, rather than the details of the specific product or service. The message that customer reviews try to get across is, therefore, of evaluative nature. To best communicate this affect, the DT should appear at the end of the review (instead of the beginning of the review) as a means of recapitulating the point of the message, thereby guaranteeing that it is fully understood by the readership.

Indeed, the cognitive prominence of information in final position - the *recency-effect* - has been well established in numerous psychological experiments (see, for example, (Murdock, 1962)). Thus, the most *frequent* evaluation of the product (which is the *most uninformative* one) also should surface at the end of the text due to the ease of its retrieval, which is presumably what product review readers would refer to as "the bottom line".

To the best of our knowledge, this psycholinguistic prediction has not been supported by psycholinguistic evidence to date. However, it has been somewhat supported by the computational results of Yang, Lin and Chen (Yang et al., 2007a, Yang et al., 2007b) who classified emotions of posts in blog corpora. Yang, Lin & Chen realized that bloggers tend to emphasize their feelings by using emoticons (such as: ☺,☹ and 👍) and that these emoticons frequently appear in final sentences. Thus, they first focused on the last sentence of posts as representing the polarity of the entire posts. Then, they divided the positive category into 2 sub-categories - happy and joy, and the negative category - into angry and sad. They showed that extracting polarity and consequently sentiments from last sentences outperforms all other computational strategies.

4 Method

We aim to show that the last sentence of a customer review is a better predictor for the polarity of the whole review than any other sentence (assuming that the first sentence is devoted to presenting the product or service). To test our prediction, we ran two experiments and compared their results. In the first experiment we examined the readers' rating of the polarity of reviews in their entirety, while in the second experiment we examined the readers' rating of the same reviews based on reading single sentences extracted from these reviews: the last sentence *or* the second one. The second sentence could have been replaced by any other sentence, *but* the first one, as our preliminary investigations clearly show that the first sentence is in many cases devoted to presenting the product or service discussed and does not contain any polarity content. For example: "I read Isaac's storm, by Erik Larson, around 1998. Recently I had occasion to thumb through it again which has prompted this review…..All in all a most interesting and rewarding book, one that I would recommend highly." (Gerald T. Westbrook, "GTW")

4.1 Materials

Sixteen customer-reviews were extracted from Blitzer, Dredze, and Pereira's sentiment database (Blitzer et al., 2007). This database contains product-reviews taken from Amazon [1] where each review is rated by its author on a 1-5 star scale. The database covers 4 product types (domains): Kitchen, Books, DVDs, and Electronics. Four reviews were selected from each domain. Of the 16 extracted reviews, 8 were positive (4-5 star rating) and the other 8 – negative (1-2 star rating).

Given that in this experiment we examine the polarity of the last sentence relative to that of the whole review or to a few other sentences, we focused on the first reviews (as listed in the aforementioned database) of at least 5 sentences or longer, rather than on too-short reviews. By "too-short" we refer to reviews in which such comparison would be meaningless; for example, ones that range between 1-3 sentences will not allow to compare the last sentence with any of the others.

4.2 Participants

Thirty-five subjects participated in the first experiment: 14 women and 21 men, ranging in age from 22 to 73. Thirty-six subjects participated in the second experiment: 23 women and 13 men ranging in age from 20 to 59. All participants were native speakers of English, had an academic education, and had normal or corrected-to-normal eye-vision.

4.3 Procedure

In the first experiment, subjects were asked to read 16 reviews; in the second experiment subjects were asked to read 32 single sentences extracted from the same 16 reviews: the last sentence and the second sentence of each review. The last and the second sentence of each review *were not* presented together but *individually*.

In both experiments subjects were asked to guess the ratings of the texts which were given by the authors on a 1-5 star scale, by clicking on a radio-button: "In each of the following screens you will be asked to read a customer review (or a sentence extracted out of a customer review). All the reviews were extracted from the www.amazon.com customer review section. Each review (or sentence) describes a different product. At the end of each review (or sentence) you will be asked to decide whether the reviewer who wrote the review recommended or did not recommend the reviewed product on a 1-5 scale: Number 5 indicates that the reviewer highly recommended the product, while number 1 indicates that the reviewer was unsatisfied with the product and did not recommend it."

In the second experiment, in addition to the psychological experiment, the latencies following reading of the texts up until the clicking of the mouse, as well as the biometric measurements of the mouse's trajectories, were recorded.

In both experiments each subject was run in an individual session and had an unlimited time to reflect and decide on the polarity of each text. Five seconds after a decision was made (as to whether the reviewer was in favor of the product or not), the subject was presented with the next text. The texts were presented in random order so as to prevent possible interactions between them.

In the initial design phase of the experiment we discussed the idea of adding an "irrelevant" option in addition to the 5-star scale of polarity. This option was meant to be used for sentences that carry no evaluation at all. Such an addition would have necessitated locating the extra-choice radio button at a separated remote place from the 5-star scale radio buttons, since conceptually it cannot be located on a nearby position. From the user interaction point of view, the mouse movement to that location would have been either considerably shorter or longer (depending on its distance from the initial location of the mouse curser at the beginning of each trial), and the mouse trajectory and click time would have been, thus, very different and difficult to analyze.

Although the reviews were randomly selected, 32 sentences extracted out of 16 reviews might seem like a small sample. However, the upper time limit for *reliable* psycholinguistic experiments is 20-25 minute. Although tempted to extend the experiments in order to acquire more data, longer times result in subject impatience, which shows on lower scoring rates. Therefore, we chose to trade sample size for accuracy. Experimental times in both experiments ranged between 15-35 minutes.

5 Results

Results of the distribution of differences between the authors' and the readers' ratings of the texts are presented in Figure 1: The distribution of differences for whole reviews is (unsurprisingly) the narrowest (Figure 1a). The dis-

[1] http://www.amazon.com

tribution of differences for last sentences (Figure 1b) is somewhat wider than (but still quite similar to) the distribution of differences for whole reviews. The distribution of differences for second sentences is the widest of the three (Figure 1c).

Pearson correlation coefficient calculations (Table 1) show that both the correlation between authors' ratings and readers' rating for whole reviews and the correlation between authors' rating and readers' rating upon reading the last sentence are similar, while the correlation between authors' rating and readers' rating when presented with the second sentence of each review is significantly lower. Moreover, when correlating readers' rating of whole reviews with readers' rating of single sentences, the correlation coefficient for last sentences is significantly higher than for second sentences.

As for the biometric measurements performed in the second experiment, since all subjects were computer-skilled, hesitation revealed through mouse-movements was assumed to be attributed to difficulty of decision-making rather than to problems in operating the mouse. As previously stated, we recorded mouse latency times following the reading of the texts up until clicking the mouse. Mouse latency times were not normalized for each subject due to the limited number of results. However, the average latency time is shorter for last sentences (19.61±12.23s) than for second sentences (22.06±14.39s). Indeed, the difference between latency times is not significant, as a paired t-test could not reject the null hypothesis that those distributions have equal means, but might show some tendency.

We also used the WizWhy software (Meidan, 2005) to perform combined analyses of readers' rating and response times. The analyses showed that when the difference between authors' and readers' ratings was $\leq |1|$ and the response time *much* shorter than average (<14.1 sec), then 96% of the sentences were *last* sentences. Due to the small sample size, we cautiously infer that last sentences express polarity better than second sentences, bearing in mind that the second sentence in our experiment represents any other sentence in the text except for the first one.

We also predicted that hesitation in making a decision would effect not only latency times but also mouse trajectories. Namely, hesitation will be accompanied by moving the mouse here and there, while decisiveness will show a firm movement. However, no such difference between the responses to last sentences or to second sentences appeared in our analysis; most subjects laid their hand still while reading the texts and while reflecting upon their answers. They moved the mouse only to rate the texts.

6 Conclusions and Future Work

In 2 psycholinguistic and psychophysical experiments, we showed that rating whole customer-reviews as compared to rating final sentences of these reviews showed an (expected) insignificant difference. In contrast, rating whole customer-reviews as compared to rating second sentences of these reviews, showed a considerable difference. Thus, instead of focusing on whole texts, computational linguists should focus on the last sentences for efficient and accurate automatic polarity-classification. Indeed, last but definitely not least!

We are currently running experiments that

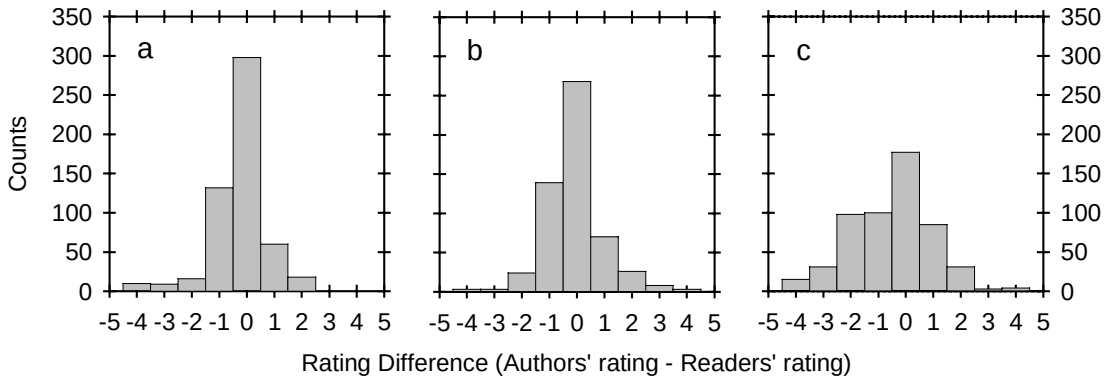

Figure 1. Histograms of the rating differences between the authors of reviews and their readers: for whole reviews (a), for last sentence only (b), and for second sentence only (c).

Readers' star rating of:	Correlated with:	Pearson Correlation Coefficient (P<0.0001)
Whole reviews	Authors' star rating of whole reviews	0.7891
Last sentences		0.7616
Second sentences		0.4705
Last sentences	Readers' star rating of whole reviews	0.8463
Second sentences		0.6563

Table 1. Pearson Correlation Coefficients

include hundreds of subjects in order to draw a profile of polarity evolvement throughout customer reviews. Specifically, we present our subjects with sentences in various locations in customer reviews asking them to rate them. As the expanded experiment is *not* psychophysical, we added an additional *remote* radio button named "irrelevant" where subjects can judge a given text as lacking any evident polarity. Based on the rating results we will draw polarity profiles in order to see where, within customer reviews, polarity is best manifested and whether there are other "candidates" sentences that would serve as useful polarity indicators. The profiles will be used as a feature in our computational analysis.

Acknowledgments

We thank Prof. Rachel Giora and Prof. Ido Dagan for most valuable discussions, the 2 anonymous reviewers – for their excellent suggestions, and Thea Pagelson and Jason S. Henry - for their help with programming and running the psychophysical experiment.

References

Baxendale, P. B. 1958. Machine-Made Index for Technical Literature - An Experiment. *IBM journal of research development* 2:263-311.

Blitzer, John, Dredze, Mark, and Pereira, Fernando. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. Paper presented at *Association of Computational Linguistics (ACL)*.

Giora, Rachel. 1983. Segmentation and Segment Cohesion: On the Thematic Organization of the Text. *Text* 3:155-182.

Giora, Rachel. 1985. A Text-based Analysis of Non-narrative Texts. *Theoretical Linguistics* 12:115-135.

Giora, Rachel. 1988. On the Informativeness Requirement. *Journal of Pragmatics* 12:547-565.

Giora, Rachel, and Lee, Cher-Leng. 1996. Written Discourse Segmentation: The Function of Unstressed Pronouns in Mandarin Chinese. In *Reference and Reference Accessibility* ed. J. Gundel and T. Fretheim, 113-140. Amsterdam: Benjamins.

Kieras, David E. 1978. Good and Bad Structure in Simple Paragraphs: Effects on Apparent Theme, Reading Time, and Recall. *Journal of Verbal Learning and Verbal Behavior* 17:13-28.

Kieras, David E. 1980. Initial Mention as a Cue to the Main Idea and the Main Item of a Technical Passage. *Memory and Cognition* 8:345-353.

Lin, Chen-Yew, and Hovy, Edward. 1997. Identifying Topic by Position. Paper presented at *Proceeding of the Fifth Conference on Applied Natural Language Processing*, San Francisco.

Meidan, Abraham. 2005. Wizsoft's WizWhy. In *The Data Mining and Knowledge Discovery Handbook*, eds. Oded Maimon and Lior Rokach, 1365-1369: Springer.

Murdock, B. B. Jr. 1962. The Serial Position Effect of Free Recall. *Journal of Experimental Psychology* 62:618-625.

Yang, Changua, Lin, Kevin Hsin-Yih, and Chen, Hsin-Hsi. 2007a. Emotion Classification Using Web Blog Corpora. In *IEEE/WIC/ACM/ International Conference on Web Intelligence*. Silicon Valley, San Francisco.

Yang, Changua, Lin, Kevin Hsin-Yih, and Chen, Hsin-Hsin. 2007b. Building Emotion Lexicon from Weblog Corpora. Paper presented at *Proceeding of the ACL 2007 Demo and Poster Session*, Prague.