TOWARD CONTEXTUAL VALENCE SHIFTERS IN VIETNAMESE REVIEWS

Thien Khai Tran[1,2], Tuoi Thi Phan[1]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology
VNU-HCM, Ho Chi Minh City, Vietnam

[2] Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and
Information Technology, Ho Chi Minh City, Vietnam

[1,2]thientk@cse.hcmut.edu.vn, [1]tuoi@cse.hcmut.edu.vn

Abstract. Valence shifters are complex linguistic structures that can modify the sentiment orientations of texts. In this paper, the authors concentrate on the study of shifters in Vietnamese texts and a discussion on the distribution of different types of shifters in the hotel reviews is presented. Finally, an approach for extracting the contextual valance shifters is proposed.

Keywords: valance shifters, polarity shifters, sentiment shifter, Vietnamese shifters, sentiment analysis for Vietnamese, valance shifters of Vietnamese

1. Introduction

Sentiment analysis assumes the task of identifying positive, negative, and neutral thoughts as well as emotional and subjective attitudes of the holder to the target audience, such as products, persons, or topics. Sentiment analysis is emerging as a research field attracting the scientific community. This includes the construction of emotional lexicons as a basis for other sentiment analysis problems at the document level, sentence level, and aspect level. This work can be done in many ways, with the simplest approach being to manually decide the polarities of sentiment words, and then to have a way of identifying the sentiment for each sentence or for the whole document based on the sentiment values of the words. However, this approach is not appropriate for subjectivity analysis because the sentiment values may be changed in context by the so-called valence shifters [1]. Valence shifters (or "polarity shifters", "sentiment shifter") are words (or phrases) that can change the sentiment orientations of texts. They are complex linguistic structures that may include explicit negations, contrasts, intensifiers, and diminishers, etc. [2].

In this paper, the authors present some approaches for contextual valence shifting detection of a Vietnamese sentiment analysis problem. We focus on rule-based methods that may be suitable for the complexity of the linguistic characteristics of Vietnamese.

With this paper, the authors describe the research contributions:

- Identifying many situations that cause valence shifters in Vietnamese texts.

- Proposing an approach to deal with the problem of valence shifters of Vietnamese texts.

To the best of our knowledge, this is the first work to investigate valence shifters in Vietnamese texts.

The authors have organized the rest of this paper as follows: in Section 2 the related work is presented. In Section 3 the valence shifters for Vietnamese are introduced. In Section 4 the authors conclude the paper and discuss possibilities for future work.

2. Related Work

The rise of the "contextual valence shifters" phenomenon has made traditional methods become ineffective when used to extract individual terms that indicate prior positive or negative polarity and build a set of emotional words. The fact that the valence of a word/phrase may be modified by one or more words founds the basis of the so-called "contextual valence shifters". These shifters were categorized into several types by the authors in [1], some of them are Negators, Intensifiers, Modals and conditional words, Presuppositional items, and Connectors.

Rule-based methods

SO-CAL [2], an early publication, deals with valence shifters by pattern rules. The authors created a set of emotional words annotated with their semantic orientation and then used the Mechanical Turk to check the consistence and reliability of the method.

The authors in [3,4] adopted dependency grammar to develop some syntastic rules for determining the scope of each negator as well as other shifters.

Machine learning, data mining, and deep learning approaches

Early sentiment classification work did not pay attention enough to the effect of negators and other shifters, as the authors only used a bag-of-words and n-grams. This meant that two reviews such as "I like this hotel" and "I don't like this hotel" would be classify to the same emotional category since both contain one

sentiment word "like", although the first one shows a positive sentiment while the second shows a negative sentiment.

Recently, some remarkable works have adopted machine learning, data mining, or deep learning to successfully consider these problems. In [5], a semi-automatic approach based on sequence mining was proposed to extract valence shifter patterns that inverted, attenuated, or canceled polarity. This approach covered many valence shifter patterns and reduced the cost of human annotating.

The authors in [6] used a hybrid approach to deal with the valence shifting problem. At first, a rule-based method was designed to detect shifters. These shifters were then used to train a component classifier of an ensemble method. Along with this, another component classifier was trained on the processed reviews, where the negators were removed and an antonym dictionary (which was built by adopting a weighted log-likelihood ratio algorithm took place of the negators.

3. VALENCE SHIFTERS OF VIETNAMESE TEXTS

Tien et al. [7] developed VietSentiWordNet, which contains approximately 1,000 lexicons. Hong et al. [8] built Vietnamese sentiment lexicons for product domains. Son et al. [9] built a Vietnamese emotional dictionary with five sub-dictionaries (noun, verb, adjective, adverb, and proposed features). In those, each lexicon is inherently carrying a sentiment polarity that is positive, negative, or neutral. As mentioned in the above section, these polarities may be shifted by the context of the texts. In this section, the authors identify many valence shifter situations in Vietnamese texts and propose some approaches to settle these problems.

Based on Vietnamese language characteristics, the authors realized five kinds of shifters, and these were Modifier (or Negator), Intensifier, Booster, Diminisher, and Minimizer [10,11].

For the corpus, 14,460 hotel reviews were extracted from mytour.vn. There were 3,829,253 words. The "AntConc" software [12] was destined to perform the corpus linguistics research and produced a model of the distribution of different types of valance shifters.

3.1 Valence shifting situations in Vietnamese texts

Modifier

A Modifier (or Negator) is the most common kind of valence shifter. For example, in the sentence "Cô ấy không thích cái laptop này" *("She doesn't like this laptop.")*, the negator "không" doesn't reverses the valance

of the sentiment word "thích" ₗᵢₖₑ. We can list some other forms of Modifiers, such as "không" ₍ₐₒₙ'ₜ₎, "chả" ₍ₙₒₜ₎, "chẳng" ₍ₙₒ₎, and "chẳng bao giờ" ₍ₙₑᵥₑᵣ₎, etc. Table I presents the Modifiers that occurred in the corpus very often, the maximum number of occurrences was for the word "không" with 9,778 from a total of 3,829,253 words in the corpus.
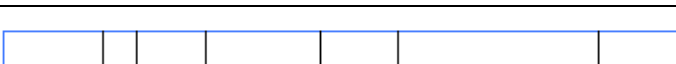
TABLE I. STATISTICS OVER THE MODIFIER FREQUENCY IN THE HOTEL REVIEWS CORPUS.

Shifters	Occurrences in the corpus	
không don't/doesn't		No. of Hits = 9778 File Length (in chars) = 3829253
chẳng no		No. of Hits = 260 File Length (in chars) = 3829253
chả not		No. of Hits = 7 File Length (in chars) = 3829253

Diminisher and Minimizer

The sentiment words (or phrases) will decrease their sentiment strength when occurring with a Diminisher or Minimizer. For example, the valence of the sentence "Cô ấy học khá chăm chỉ" *("She studies rather hard.")* is lesser than the valence of the sentence "Cô ấy học chăm chỉ" *("She studies hard.")*. Some Diminisher/Minimizer words are "khá" ᵣₐₜₕₑᵣ, "hơi" ₍qᵤᵢₜₑ₎, and "phần nào" ₛₒₘₑwₕₐₜ, etc. Table II presents the Diminishers or Minimizers that occurred in the corpus, the maximum is reached by the word "khá" ᵣₐₜₕₑᵣ with 5,977 occurrences from a total of 3,829,253 words in the corpus.
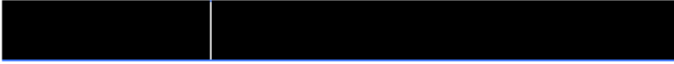
TABLE II. STATISTICS OVER THE DIMINISHER OR MINIMIZER FREQUENCY IN THE HOTEL REVIEWS CORPUS.

Shifters	Occurrences in the corpus	
khá ᵣₐₜₕₑᵣ		No. of Hits = 5977 File Length (in chars) = 3829253
hơi ₍qᵤᵢₜₑ₎		No. of Hits = 2464 File Length (in chars) = 3829253
phần nào somewhat		No. of Hits = 6 File Length (in chars) = 3829253

Intensifier and Booster

The sentiment words (or phrases) will increase their sentiment strength when they occur with an Intensifier or Booster. For example, the valence of the sentence "Cô ấy học rất chăm chỉ" *("She studies very hard.")* is greater than valence of the sentence "Cô ấy học chăm chỉ" *("She studies hard.")*. Some Intensifiers/ Boosters are "rất" very, "cực kỳ" extremely, and "vô cùng" exceedingly, etc. Table III presents the Intensifiers or Boosters that occur in the corpus, the maximum is reached by the word "rất" very with 8,373 occurrences from a total of 3,829,253 words in the corpus.
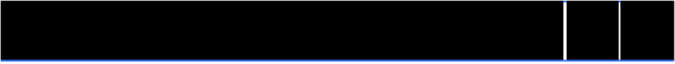
TABLE III. STATISTICS OVER THE INTENSIFIER AND BOOSTER FREQUENCY IN THE HOTEL REVIEWS CORPUS.

Shifters	Occurrences in the corpus	
rất very		No. of Hits = 8373 File Length (in chars) = 3829253
cực kỳ extremely		No. of Hits = 40 File Length (in chars) = 3829253
vô cùng exceedingly		No. of Hits = 145 File Length (in chars) = 3829253

Connectors

Connectors, such as "mặc dù" although, "tuy nhiên" however, and "nhưng" but, etc., can both modify information and work on information elsewhere in the sentence to decrease the force of that information. For example, the valence of the phrase "Cô ta thì xinh nhưng không tốt" *("She is pretty but is not kind.")* is equal to the valence of the phrase "nhưng không tốt" *("is not kind")*. Table IV presents the Connectors that occur in the corpus, the maximum is reached by the word "nhưng" but with 3,728 occurrences from a total of 3,829,253 words in the corpus.

TABLE IV. STATISTICS OVER THE CONNECTOR FREQUENCY IN THE HOTEL REVIEWS CORPUS.

Shifters	Occurrences in the corpus	
mặc dù although		No. of Hits = 193 File Length (in chars) = 3829253
tuy nhiên however		No. of Hits = 1450 File Length (in chars) = 3829253
nhưng but		No. of Hits = 3728 File Length (in chars) = 3829253

Cause-Effect sentences

The valance of a cause-effect sentence is the valance of the effect clause. For example, in the sentence "Vì cô ấy chăm chỉ nên cô ấy thi đậu." *("She passed the exam because she studied hard."),* the valence of this sentence is equal the valence of the clause *"she studied hard"*. Some cause-effect words are "vì...nên" because … that, "vì…mà" because … that, and "bởi vì…mà" because … that, etc.

Conditional sentences

Conditional sentences often do not contain emotion. We cannot determine the valance of these sentences because of the "if-then" statement. For example, in the sentence "Nếu laptop đó rẻ thì tôi sẽ mua một cái" (*"If that laptop is cheap then I will buy it."*), we do not know whether the laptop is *"cheap"* or not. Some conditional words are "nếu … thì" if…then, "hễ … thì" if…then, and "giả sử … thì" if…then, etc.

Questions

Questions also do not contain emotion because we cannot determine the polarities of the words in the texts. For example, in the question "Laptop đó có ok không?" *("Is that laptop ok?"),* we do not know whether the laptop is *"ok"* or not.

3.2 Approach to deal with contextual valance shifters in Vietnamese texts

We can capture the above situations by using some rule-based methods built on the presence of many predefined patterns. In [13], we proposed an approach for mining features and opinion words based on an upgraded double propagation algorithm [14], some regular expression rules and ontologies. However, as mentioned in [10,11], there are several exceptions based on the linguistics characteristics of Vietnamese. For example, the kind of adjective that goes along with negators or the position between the shifters and adjectives (or verbs), etc.

Moreover, we also pay attention to some special words/phases that can modify the valence of Vietnamese texts, as follows:

- Example 1: "Khách sạn ấy <u>đã từng</u> được ưa chuộng" *("This hotel <u>used to be</u> popular.")*. In this review, the word "ưa chuộng" popular shows a positive emotion, but the whole sentence shows a negative emotion because of the word "đã từng" used to be.

- Example 2: "Hotel ấy <u>mới nhìn</u> có vẻ tốt" *("<u>At first glance</u>, this hotel seems to be good.")*. In this example, the word "tốt" good shows a positive emotion but the whole sentence does not show a positive emotion because of the phrase "mới nhìn" at first glance.

After capturing the words/phrases that are the valence shifters in texts, we can identify their sentiment scores in the same way as the authors in [10,11] have done.

4. Conclusions

This paper presented valence shifters in the Vietnamese language and proposed some approaches to deal with this problem. In the paper, a discussion of the distribution of different types of shifters was conducted using hotel reviews. Based on this, and via investigating the linguistic characteristics of Vietnamese, the authors intend to build effective rules for extracting reliable shifters. In future work, the proposed rules will be actualized and machine learning or deep learning methods will be adopted. This will help the system become more flexible and robust.

5. References

1. L. Polanyi and A. Zaenen, 2006. Contextual valence shifters. Computing attitude and affect in text: Theory and applications, pages 1–10, 2006.

2. M. Taboada, J. Brooke, M. Tofiloski, K.D. Voll, M. Stede, 2011. Lexicon-based methods for Sentiment Analysis. Computational Linguistics, 37 (2011), pp. 267-307.

3. J. Carrillo de Albornoz, L. Plaza, and P. Gervas. A hybrid approach to emotional sentence polarity and intensity classification. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL'10, pages 153–161, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

4. L. Jia, C. Yu, and W. Meng. The effect of negation on sentiment analysis and retrieval effectiveness. In Proceeding of the 18th ACM conference on Information and knowledge management, CIKM'09, pages 1827–1830, New York, NY, USA, 2009. ACM.

5. Xu, G. & Huang, CR. Extracting Chinese polarity shifting patterns from massive text corpora. Lingua Sinica (2016) 2: 5. https://doi.org/10.1186/s40655-016-0014-z

6. Xia, R.; Xu, F.; Yu, J.; Qi, Y. Erik Cambria: Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Inf. Process. Manag. 2016, 52, 36–45.

7. Tien-Thanh Vu, Huyen-Trang Pham, Cong-To Luu, Quang-Thuy Ha.: A feature-based opinion mining model on product reviews in Vietnamese. In: Semantic Methods for Knowledge Discovery and Communication, Polish-Taiwanese Workshop, Springer Berlin Heidelberg (2011) 22–23.

8. Hong Nam Nguyen, Thanh Van Le, Hai Son Le, Tran Vu Pham, Domain Specific Sentiment Dictionary for Opinion Mining of Vietnamese Text. The 8th Multi-Disciplinary International Workshop on Artificial Intelligence (MIWAI 2014): 136-148.

9. Son Trinh, Luu Nguyen, Minh Vo, Phuc Do, (2016) "Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language," Recent Developments in Intelligent Information and Database Systems Volume 642 of the series Studies in Computational Intelligence. pp 263-276

10. Thien Khai Tran, Tuoi Thi Phan: Computing Sentiment Scores of Verb Phrases for Vietnamese. ROCLING 2016.

11. Thien Khai Tran, Tuoi Thi Phan: Computing Sentiment Scores of Adjective Phrases for Vietnamese. MIWAI 2016: 288-296.

12. Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

13. Tran, T.K. & Phan, T.T. Int. j. inf. tecnol. (2017) 9: 239. https://doi.org/10.1007/s41870-017-0032-9.

14. Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1): 9–27.