氏　　名	AYE AYE MAR
学位の種類	博士（情報科学）
学位記番号	博情第 526 号
学位授与年月日	令和 6 年 3 月 22 日
論文題目	Implicit Aspect Classification of Online Reviews by Clustering-based Weak Supervision
論文審査委員	白井　清昭　北陸先端科学技術大学院大学 准教授
	Nguyen Minh Le　同　教授
	井之上　直也　同　准教授
	岡田　将吾　同　准教授
	新納　浩幸　茨城大学　教授

論文の内容の要旨

Aspect Term Extraction (ATE) which is a process of extracting an aspect (also known as opinion target) from a customer review sentence plays a vital role in Aspect-Based Sentiment Analysis (ABSA). Many previous work of ATE focused on explicit aspect but only a few work considered to extract implicit aspects. However, customer reviews containing implicit aspects are widespread on the Web (such as Amazon.com) and these sentences are also important to fully understand the opinions and sentiments of the customers.

One of the bottleneck problem of implicit aspect extraction is lack of a large dataset of reviews annotated with implicit aspects. Although the corpus annotated with implicit aspects is required for every domain due to different types of aspects in different domains, constructing a corpus is labour intensive and time consuming. Therefore, a system to automatically construct a dataset annotated with implicit aspects is required. This study proposed a novel approach that automatically constructs a dataset annotated with implicit aspects using unlabelled Amazon reviews to address the challenge of implicit aspect extraction. To the best of our knowledge, no prior work has been performed on the automatic construction of such a dataset.

The goal of this study is to develop a system of ATE for implicit aspects. A dataset labeled with implicit aspects is automatically constructed by guessing implicit aspects in unlabeled review sentences. The proposed method involves clustering review sentences labeled with explicit aspects (which were extracted by CRF model trained on golden explicit review sentences) and unlabeled review sentences. In this study, using a K-means clustering approach with a relatively large number of clusters (10% of total review sentences) aims to generate many small but accurate clusters.

Cluster labels, considered as implicit aspects, are automatically assigned based on the assumption that sentences with similar context share a common aspect. When selecting the most relevant cluster label among the explicit aspects in the cluster, the frequency of the aspect in the list of aspects extracted by CRF and its occurrence in the review sentences within the cluster are considered to determine the relevance of the chosen cluster label. When there is more than one aspect that can be the cluster label, we did not consider such kind of cluster since the cluster label is not unique. Moreover, the reliability of the cluster label to be chosen was determined by the threshold value (T_r). Unlabeled sentences in clusters matching pre-defined implicit aspect categories are then obtained as implicit-aspect-labeled sentences. To increase the number of clusters related with the implicit

aspects, the aspect synonym list was identified.

The accuracy of the constructed corpus was evaluated by a human annotator by checking manually on 50 random sentences for each implicit aspect. The results showed that accuracy of the sentences in the constructed corpus was reasonably high, i.e., from 0.58 to 0.82. The study presents findings and observations regarding with constructing the corpus annotated with implicit aspects.

In this study, implicit aspect extraction problem is formulated as classification problem. Then, BERT model is fine-tuned for implicit aspect classification using the constructed dataset by investigating the best values of hyper-parameters. Experiments results of implicit aspect classification show that our method achieves 82% and 84% accuracy for the mobile phone and PC reviews respectively, which are 20 and 21 percentage points higher than the baseline.

Furthermore, the study explores the impact of explicit review sentences for implicit aspect classification by combining the explicit sentences and implicit sentences and then by training classification model on the combined dataset. The experimental results showed that it further boosts the performance of implicit aspect classification in both phone and PC domain.

Keywords: Aspect-based Sentiment Analysis, Aspect Extraction, Implicit Aspect, Weakly-supervised Learning, Online Review

論文審査の結果の要旨

本論文はオンライン上に書かれた製品に関するユーザレビューから暗黙的属性を抽出する新しい手法を提案している。属性とは、スマートフォンにおける「デザイン」「価格」など、レビューで評価対象となっている製品の機能や評価項目を指す。多くの従来研究が単語によって表された明示的な属性を抽出していたのに対し、本研究ではレビューで暗に言及されている暗黙的な属性を抽出することに取り組む。暗黙的属性を分類するモデルを教師あり学習するには、レビューに対して正解の暗黙的属性が付与されたラベル付きデータセットが必要であるが、大規模なデータセットは現時点では公開されていない。本論文は暗黙的属性のラベル付きデータセットを以下の手順により自動的に構築する。(1)既存の明示的な属性がラベル付けされたデータセットから、明示的な属性を抽出するモデルを学習する。(2)大量のラベルなしレビュー文の集合に対し、学習したモデルを用いて明示的な属性を抽出する。このとき、明示的な属性が抽出されなかった文には暗黙的属性が含まれる可能性がある。(3)レビュー文のクラスタリングを行う。それぞれの文に対し、SCDV と呼ばれる手法を用いて、その文の意味を表すベクトル表現を得る。次に、得られたベクトル表現間の距離を文間距離とし、k-means 法を用いてクラスタリングを行う。ここでは明示的属性、暗黙的属性に関わらず、同じ属性に言及したレビュー文がひとつのクラスタにまとめられる。(4)同一クラスタ内に明示的属性を含む文と含まない文が混在するとき、明示的属性を含まない文には暗黙的属性が含まれているとして、これを抽出する。さらに、抽出した文に対し、同一クラスタに存在する(明示的)属性を暗黙的属性としてラベル付けする。最後に、構築したデータセットを用いて事前学習済み言語モデル BERT をファインチューニングし、暗黙的属性を分類するモデルを学習する。提案手法を評価するために、スマートフォンとパソコンに関するレビュー文を対象に、暗黙的属性を分類する実験を行った。実験の結果、提案手法によって構築された暗黙的属性のラベル付きデータセットは、十分に高い品質を持つことを確認した。また、構築したデータセットを用いて学習したモデルは、既存

の明示的属性がラベル付けされたデータセットから学習されたモデルと比べて、暗黙的属性の分類の正解率や F 値が明確に高いことを確認した。さらに、上記 2 つのデータセットを組み合わせてモデルを学習することで、分類の性能がさらに改善した。以上、本論文は、レビューにおける暗黙的属性を分類するための新しい弱教師あり機械学習手法を示したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。