Aspect clustering methods for sentiment analysis

Francielle Alves Vargas and Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos, Brazil
franciellealvargas@gmail.com, taspardo@icmc.usp.br

Abstract. Automatic aspect identification and clustering are critical tasks for opinion mining/sentiment analysis, as users employ varied terms (explicitly or not) to evaluate objects of interest and their characteristics. In this paper, we focus on aspect clustering methods and present a new approach to group implicit and explicit aspects from online reviews. We evaluate four linguistic methods inspired in the literature and one statistical method (using word embeddings), and also propose a new one, based on varied linguistic knowledge. We test the methods in three commonly used domains and show that the method that we propose significantly outperforms the other methods by a large margin.

Keywords: Clustering, aspect-based sentiment analysis, opinion mining

1 Introduction

The expansion of the social networks and e-commerce services resulted in the growth of online reviews in the web. Websites as Amazon and Buscape encourage users to write reviews for products, where users may do objective or subjective descriptions for a product and its aspects or properties. Subjective descriptions are characterized by a personal language, with opinions, sentiments, emotions and judgments. The research area in charge of identifying, extracting and summarizing subjective information in texts is called opinion mining or sentiment analysis [17] [22] [21]. According to [27], this area is different from the traditional text mining area, which is mostly based on objective topics rather than on subjective perceptions.

According to [24], due to the huge number of reviews in unstructured and varied formats, it is impractical for interested users to fully read and understand what other users comment. Therefore, the semantic organization of such information is mandatory, and the area of sentiment analysis consists in a first step towards this, enabling "mining" and synthesizing the relevant information, which may be employed by final users and companies for supporting their decision making process. Although its usefulness, according to [13], sentiment analysis represents a "delicious challenge", as natural languages are very rich and allow to express subjectivity in different ways.

In this paper, our particular interest lies on how users refer to the aspects of the products that they evaluate in their reviews, as the area has struggled with the way users employ several different terms to refer to the same aspects. For example, in the review passage "she considered the camera price very expensive", the consumer employed the term "price" to evaluate an aspect of the camera; however, consumers might also use the terms "cost", "value", "investment", etc. In addition, consumers may use implicit or explicit aspects to refer to the same aspect, e.g., the sentences "she got calls at the São Francisco river" and "working anywhere" have been employed in actual reviews to evaluate the (implicit) "signal" aspect of a smartphone. It is also interesting to notice that, in some domains, proper names may be employed to refer to the aspects. For instance, the proper names "Sony" and "Nikon" may be used to evaluate the "product brand" aspect of digital cameras.

As there are some previous work on aspect identification for the Portuguese language (see, e.g., [3]), we focus our efforts on the next step of aspect clustering, which aims at automatically grouping aspect terms that refer to the same thing. Such process is a core step for several sentiment analysis tasks, as aspect-based polarity classification and opinion summarization.

In this paper, we investigate six aspect clustering methods for both explicit and implicit aspects in product reviews. We test four linguistic-based methods inspired in the literature, a statistical method (based in word embeddings), and a new (linguistic) method that we propose, which was motivated by an empirical study of the relevant linguistic phenomena in Portuguese reviews. We compared the six methods on three different domains - smartphones, digital cameras, and books - and demonstrate that our method significantly outperforms the others.

The rest of this paper is organized as follows. Section 2 introduces the essential related work. Section 3 presents the clustering methods, while Section 4 describes their evaluation. Some final remarks are made in Section 5.

2 Related Work

According to [25], two kinds of similarity measures are usually applied in the aspect clustering task: (i) those relying on knowledge resources (e.g., thesauri and semantic networks) [2] [10] [23], and (ii) those relying on distributional properties of the words in corpora [4] [19] [26].

In the knowledge source approach, ontologies are frequently explored, using WordNet lexical relations [15] or the categories of Wikipedia[1], as in [8] and [18]. In [18], for instance, the hypernymy/is-a relation is the basis for clustering aspects. However, such approach did not perform very well. The authors reported that specific domain aspects were not found on ontologies and lexicons, e.g., in smartphone domain, relevant aspects as "gps", "3G", "wap" and "hit" are too specific to happen in the above ontologies.

The distributional similarity methods generally employ measures such as Cosine, Jaccard, Dice and PMI (Pointwise Mutual Information) [12], as in [4], [1], [26] and [28]. In this line, [28] also makes use of the widely adopted word

[1] http://wiki.dbpedia.org/

embeddings produced by *word2vec* algorithm [14] to find aspect categories (which refer to the type of entity being evaluated, and not the groups that we look for). Some good results were produced, but difficulties to group domain specific aspects were also reported.

Overall, we could not find proposals for clustering implicit aspects. Only explicit aspects are tackled.

In this paper, we have developed and tested aspect clustering methods that were inspired in such previous attempts. We also try to overcome some of their limitations, as we describe in what follows.

3 Clustering Methods

We compared six aspect clustering methods: 4 linguistic methods inspired in the literature, 1 statistical method and 1 new method that we propose. All the methods receive as input a list of implicit and explicit aspects in their user reviews and produce as output clusters of similar aspects, i.e., aspects that refer to the same property or feature of an object. As we commented before, we consider that the task of aspect identification in the reviews was already performed (e.g., by one of the methods of [3]). We only focus on the aspect clustering task.

The implicit aspects are represented by their indicative terms in the reviews, for instance, the "working anywhere" n-gram to refer to the signal aspect of smartphones.

The general overview of the 4 linguistic methods inspired in the literature is shown in Figure 1. The methods were incrementally implemented in order to evaluate the results obtained at each level of increment. For example, the first implemented method creates clusters of aspects using only synonymy relations (i.e., aspects that happen to be in a synonymy relation are clustered together). The second method creates clusters using synonymy and is-a relations. The third method uses synonymy, is-a, and part-of relations. The fourth method uses synonymy, is-a, part-of, and coreference relations. As we work for the Portuguese language, we have used the Onto-PT lexical ontology [16] for extracting synonymy, is-a and part-of relations. To find coreference relations, we have employed the CORP coreference resolution system for Portuguese [6] [7]. These resources are widely used for this language.

For our statistical method, we adopted the ready-to-use trained word embedding models proposed by [9] and available in the NILC word embeddings repository[2]. These models have been widely used for Portuguese. We used the word2vec version [14] with 300 dimensions. The idea is that aspects with "similar" corresponding vectors should be clustered together.

The last method is the one that we propose, which we refer by OpCluster-PT. The OpCluster-PT algorithm was motivated by a linguistic empirical study, in which we studied the aspect-related linguistic phenomena that happen in product reviews. This study showed that, besides the traditional lexical relations

[2] http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

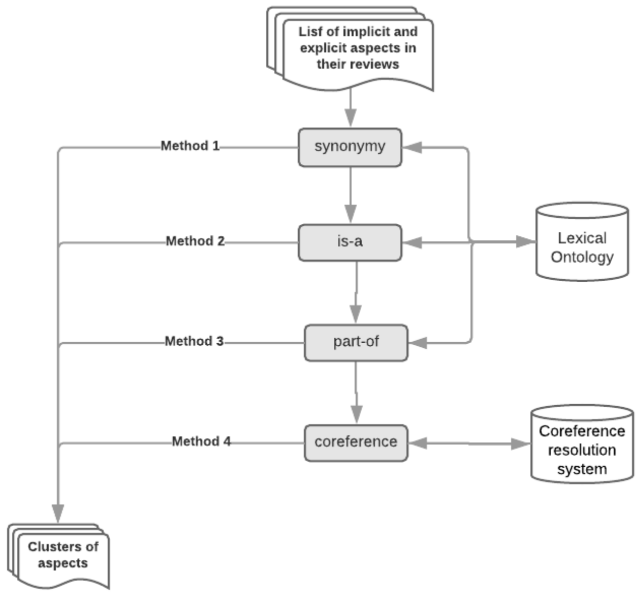

Fig. 1. Aspect clustering methods

(such as synonymy, hypernymy/is-a and meronymy/part-of), it is necessary to use causative, deverbal, diminutive (or augmentative), foreignism, and substring relations to find aspect clusters. To find these new relations, we have used Onto-PT [16], CORP [6] [7] and the Portuguese foreignism and deverbal iLteC lexicons [11] [5]. A list of diminutive/augmentative words was compiled for this proposal, as we could not find an available one. It is important to say that, as we use CORP, we also take advantage of this tool to (indirectly) find the hypernymy relations in our new method, as CORP proved do be better than Onto-PT in our domains for identifying aspects grouped by such relations.

The relevance of the new relations may be easily exemplified. Causative relations help finding that "to finish" may indicate the implicit aspect "end of story" (of a book); deverbal relations may indicate that "to write" (in a specific manner) refers to the implicit aspect "writing style" (in book domain); the diminutive "little book" (*livrinho*, in Portuguese) refers to the aspect "book"; the aspect "display" is a foreignism in Portuguese language that is similar to the aspect "screen" (*tela*, for some electronic product); and substring relation allows to detect that "image" and "image quality" may refer to the same aspect "image" (for some electronic product again).

The proposed method works as the previous linguistic methods. It incrementally forms groups with the aspects that show some of the predicted relations. The full method is shown in Algorithm 1. It receives as input a set of reviews R and the list of implicit and explicit aspects A that have occurred in the reviews. It then performs three main steps. In the first one (starting in the first *repeat*), for each aspect a, it looks for other aspects that show some relation with the previous one. In the second step (starting in the second *repeat* inside the first one), the method looks for other aspects that show some specific relations[3] with the ones that were clustered together with a. This second step is, in fact, a recursive step, which looks for any other related aspect that was left outside the cluster. In the last step (in the third *repeat*), we look for any remaining unitary clusters that might be joined with other clusters. This happens if the unitary aspect is in a substring relation with some aspect of other cluster. As output, the algorithm provides a list of clusters G.

In what follows, we report the evaluation of the methods.

4 Evaluation

To evaluate the aspect clustering methods, we have manually annotated a corpus of smartphone, digital camera, and book reviews. We selected three commonly used domains to test how robust and generalizable the methods are. Each domain counted with 60 reviews.

In each review, we marked and clustered the explicit and implicit aspects. The implicit aspects were indicated by the clue terms that signaled them. Such data consisted in the reference annotation to which the automatic output of the clustering methods was compared to.

Table 1 shows the relevant numbers of the reference annotation. One may see that there is a significant difference among the domains. In the smartphone and digital camera domains, we have identified more domain specific aspects, maybe due to the popularity of such devices, which allows users to comment about their technical details. Books, otherwise, are usually not evaluated on their technical details (as the type of paper and weight), but on more prototypical aspects in this domain (as characters and story).

Table 1. Reference annotation

Domain	**Book**	**Camera**	**Smartphone**
total number of aspects	103	132	180
number of explicit aspects	91	109	142
number of implicit aspects	12	23	38
number of clusters	21	36	48

For evaluating the methods, we have computed the traditional clustering evaluation measures of Precision, Recall, F-measure and Global F-measure (as defined in [20]) over the reference clusters. Precision indicates the proportion of aspects of each automatic cluster that is correctly clustered (according to

[3] We only look for coreference, foreignism and diminutive-augmentative relations, because we empirically observed that they were the most accurate ones in this step.

Algorithm 1 OpCluster-PT

Input: List of aspects $A = \{a_1, a_2, ..., a_n\}$ sorted by frequency (in decreasing order) and their corresponding reviews $R = \{r_1, r_2, ..., r_m\}$ preprocessed by CORP
Output: Clusters of aspects $G = \{g_1, g_2, ..., g_p\}$, where each g_i contains a subset of aspects of A

Let **B** = $\{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim\text{-}augm}, b_{coref}, b_{subs}\}$, where each $b_{relation}$ in **B** contains the result of searching for aspects in synonymy, part-of, causative, deverbal, foreignism, diminutive-augmentative, coreference, and substring relations (for example, b_{syn} contains the synonymous aspects)
Let **j** and **k** be integers initiliazed with zero

repeat
  Consider a_i as the most frequent aspect in **A**
  if a_i in **A** contains *synonymous* words in Onto.PT **then**
    Add such words to $\mathbf{b_{syn}}$
  end if
  if a_i in **A** contains *part-of* related words in Onto.PT **then**
    Add such words to $\mathbf{b_{part}}$
  end if
  if a_i in **A** contains *causative* related words in Onto.PT **then**
    Add such words to $\mathbf{b_{caus}}$
  end if
  if a_i in **A** contains *deverbal construction* related words in iLteC lexicon **then**
    Add such words to $\mathbf{b_{devb}}$
  end if
  if a_i of **A** contains *foreignism* related words in iLteC lexicon **then**
    Add such words to $\mathbf{b_{fore}}$
  end if
  if a_i in **A** contains *diminutive or augmentative* related words in our compiled list **then**
    Add such words to $\mathbf{b_{dim\text{-}augm}}$
  end if
  if a_i in **A** has related *coreference* terms in the corresponding reviews, as indicated by CORP **then**
    Add such terms to $\mathbf{b_{coref}}$
  end if
  if a_i in **A** contains *substring* relations with other aspects in **A** **then**
    Add such aspects to $\mathbf{b_{subs}}$
  end if
  Remove duplicate items from **B** = $\{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim\text{-}augm}, b_{coref}, b_{subs}\}$
  Increment **j**
  Create cluster $\mathbf{g_j}$ and add to it the aspects of intersection**(A,B)**
  Remove from **A** the aspects of intersection**(A,B)**
  Empty **B**
  repeat
    Consider a_k as each aspect in $\mathbf{g_j}$, ignoring a_i, which was already processed
    if $\mathbf{a_k}$ in $\mathbf{g_j}$ has related *coreference* terms in the corresponding reviews, as indicated by CORP **then**
      Add such terms to $\mathbf{b_{coref}}$
    end if
    if $\mathbf{a_k}$ in $\mathbf{g_j}$ contains *foreignism* related words in iLteC lexicon **then**
      Add such words to $\mathbf{b_{fore}}$
    end if
    if $\mathbf{a_k}$ in $\mathbf{g_j}$ contains *diminutive or augmentative* related words in our compiled list **then**
      Add such words to $\mathbf{b_{dim\text{-}augm}}$
    end if
    Remove duplicate items from **B** = $\{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim\text{-}augm}, b_{coref}, b_{subs}\}$;
    Add to $\mathbf{g_j}$ the aspects of intersection**(A,B)**
    Remove from **A** the aspects of intersection**(A,B)**
    Empty **B**
  until every $\mathbf{a_k}$ in $\mathbf{g_j}$ is tested
until **A** is empty
repeat
  Consider each unitary cluster $\mathbf{g_j}$ in **G**
  if the aspect in this unitary cluster $\mathbf{g_j}$ has a substring relation with some aspect in other cluster $\mathbf{g_k}$ in **G** **then**
    Add this aspect in $\mathbf{g_j}$ to the cluster $\mathbf{g_k}$
    Eliminate $\mathbf{g_j}$
  end if
until every unitary cluster $\mathbf{g_j}$ in **G** is tested

the reference clusters). Recall indicates the proportion of aspects of the reference clusters that was covered by the automatically generated clusters. As these measures are complementary, we also compute the F-measure score, which represents the harmonic average between precision and recall. The global F-measure of each automatically generated cluster, relative to the entire set of clusters, is based on the cluster that best describes each reference cluster. The achieved results are shown in details in Tables 2 to 5.

One may see that the synonymy-based method (the simplest one, under the linguistic point of view) presented the best Precision results (in Table 2) for the task. However, we observed a very high number of unitary clusters (in relation to the reference annotation), which increases Precision (as the Precision of each unitary cluster is equal to 100%), but seriously harms Recall, which is confirmed in Table 3. The method that we propose here produced significantly higher recall numbers for all the domains. Overall, looking at F-measure values, the method we propose was the best one, outperforming all the others.

It is also interesting to notice that our method achieved the best results in book domain, probably because there are more prototypical aspects and, therefore, less domain specific aspects, which are more difficult to find in the linguistic repositories that we adopt.

Surprisingly, the word embeddings performed very poorly. We believe that this happened because we have used a widely used model trained on general corpora, and not corpora of reviews. However, one might argue that most of the aspects are general enough to be used in general language corpora too. This remains as an open question to investigate in the future.

Table 2. Precision results

	Methods	**Book**	**Camera**	**Smartphone**
1	synonymy	**0.974**	**0.987**	**0.973**
2	synonymy + is-a	0.916	0.967	0.940
3	synonymy + is-a + part-of	0.916	0.967	0.943
4	synonymy + is-a + part-of + coreference	0.945	0.963	0.953
5	word embeddings	0.953	0.962	0.956
6	opcluster-pt	0.925	0.933	0.947

Table 3. Recall results

	Methods	**Book**	**Camera**	**Smartphone**
1	synonymy	0.231	0.281	0.296
2	synonymy + is-a	0.242	0.287	0.314
3	synonymy + is-a + part-of	0.242	0.287	0.310
4	synonymy + is-a + part-of + coreference	0.321	0.307	0.364
5	word embeddings	0.231	0.292	0.300
6	opcluster-pt	**0.748**	**0.687**	**0.550**

Table 4. F-measure results

	Methods	Book	Camera	Smartphone
1	synonymy	0.374	0.438	0.454
2	synonymy + is-a	0.383	0.442	0.471
3	synonymy + is-a + part-of	0.383	0.442	0.466
4	synonymy + is-a + part-of + coreference	0.480	0.466	0.527
5	word embeddings	0.372	0.448	0.457
6	opcluster-pt	**0.827**	**0.792**	**0.702**

Table 5. Global F-measure results

	Methods	Book	Camera	Smartphone
1	synonymy	0.300	0.351	0.347
2	synonymy + is-a	0.249	0.319	0.333
3	synonymy + is-a + part-of	0.244	0.319	0.333
4	synonymy + is-a + part-of + coreference	0.399	0.409	0.508
5	word embeddings	0.280	0.336	0.350
6	opcluster-pt	**0.711**	**0.605**	**0.583**

As illustration, our method automatically generated a cluster composed by the aspects “cost benefit”, “price”, “value”, “investment” and “cheap” (which is an implicit aspect), which is very good. An example of problematic cluster is the one composed by “enterprise”, “lg”, “nokia”, “sony”, “sony_ericson”, “program”, “design”, “system” and “model’, in which the 4 last words are clearly misplaced in this cluster.

We have checked that, for correcting the remaining errors and improving the results, we might also incorporate knowledge about proper names and slangs, using, e.g., Wikipedia data and specialized lexicons. This remains for future work.

5 Final Remarks

According to [17], the aspect-based sentiment analysis task requires deep understanding of natural language characteristics and textual context. Therefore, in this paper, we present the OpCluster-PT algorithm, designed to cluster explicit and implicit aspects in product reviews. We achieved the best results when comparing to other four linguistic-based methods and one statistical method.

As a side effect of this work, a reference dataset was produced, with indicated explicit and implicit aspects, as well as manually produced aspect clusters. Additionally, we have also produced aspect ontologies for the investigated domains.

More information about this work and the related tools and resources may be found at the OPINANDO project website[4].

[4] https://sites.google.com/icmc.usp.br/opinando/

Acknowledgments

The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

1. Abu-Jbara, A., King, B., Diab, M.T., Radev, D.R.: Identifying opinion subgroups in arabic online discussions. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. pp. 829–835. Sofia, Bulgaria (2013)
2. Alvarez, M., Lim, S.: A graph modeling of semantic similarity between words. In: Proceedings of the Conference on Semantic Computing. pp. 355–362. Irvine, United States (2007)
3. Balage Filho, P.P.: Aspect extraction in sentiment analysis for Portuguese. Ph.D. thesis, University of São Paulo, São Carlos, Brazil (2017)
4. Chen, Y., Zhao, Y. and, Q.B., Liu, T.: Product aspect clustering by incorporating background knowledge for opinion mining. PLOS ONE **11**(8), 1–16 (2016)
5. Ferreira, J.P., Janssen, M.: Dicionário de Formas Não Adaptadas. Instituto de Linguística Teórica e Computacional, 1ª edn. (2017)
6. Fonseca, E., Sesti, V., Antonitsch, A., Vanin, A., Vieira, R.: Corp: Uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. Linguamática **9**(1), 3–18 (2017). https://doi.org/10.21814/lm.9.1.241, http://linguamatica.com/index.php/linguamatica/article/view/v9n1p1
7. Fonseca, E.B., Vieira, R., Vanin, A.A.: Corp: Coreference resolution for portuguese. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese. pp. 9–11. Tomar, Portugal (2016)
8. García, A., Cuadros, M., Rigau, G., Gaines, S.: V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation). pp. 833–837. Dublin, Ireland (2014)
9. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: Proceedings of the Symposium in Information and Human Language Technology. pp. 122–131. Uberlandia, Brazil (2017)
10. Hughes, T., Ramage, D.: Lexical Semantic Relatedness with Random Graph Walks. Computational Linguistics **7**(1), 581–589 (2007)
11. Janssen, M., Ferreira, J.P.: Dicionário de nomes deverbais. Intituto de Linguística Teórica e Computacional, 1ª edn. (2007)
12. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 25–32. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999). https://doi.org/10.3115/1034678.1034693, https://doi.org/10.3115/1034678.1034693
13. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 1ª edn. (2012)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computing Research Repository **1301.3781**(1) (2013)
15. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Wordnet: An online lexical database. International Journal of Lexicography **3**, 235–244 (1990)

16. Oliveira, H.G.: Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT. In: Proceedings of the Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. pp. 64–68. São Carlos, Brazil (2014)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 79–86. Stroudsburg, United States (2002)
18. Patra, B.G., Mandal, S., Das, D., Bandyopadhyay, S.: Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 370–374. Dublin, Ireland (2014)
19. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of english words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. pp. 183–190. Stroudsburg, United States (1993)
20. Seno, E.R.M.: Um método para fusão automática de sentenças similares em português. Ph.D. thesis, University of São Paulo, São Carlos, Brazil (2010)
21. Taboada, M.: Sentiment analysis: An overview from linguistics **2**(1), 325–347 (2016), http://www.annualreviews.org/doi/full/10.1146/annurev-linguistics-011415-040518
22. Wu, C.W., Liu, C.L.: Ontology-based text summarization for business news articles. In: Proceedings of the 3th International Symposium on Computer Architecture. pp. 389–392. Honolulu, United States (2003)
23. Yang, H., Callan, J.: A metric-based framework for automatic taxonomy induction. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. pp. 271–279. Suntec, Singapore (2009)
24. Yu, J., Zha, Z., Wang, M., Wang, K., Chua, T.: Domain-assisted product aspect hierarchy generation: Towards hierarchical organization of unstructured consumer reviews. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 140–150. Edinburgh, United Kingdom (2011)
25. Zhai, Z., Liu, B., Xu, H., Jia, P.: Clustering product features for opinion mining. In: Proceedings of the 4th International Conference on Web Search and Data Mining. pp. 347–354. New York, United States (2011)
26. Zhang, S., Jia, W., Xia, Y., Meng, Y., Yu, H..: Product features extraction and categorization in chinese reviews. In: Proceedings of the 6th International Multi-Conference on Computing in the Global Information Technology. pp. 38–42. Nice, France (2011)
27. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: Proceedings of the 3th International Conference on Knowledge Science, Engineering and Management. pp. 204–214. Berlin, Germany (2009)
28. Zhou, X., Wan, X., Xiao, J.: Representation learning for aspect category detection in online reviews. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. pp. 417–423. Texas, United States (2015)