Clustering and hierarchical organization of opinion aspects: a corpus study

Francielle Alves Vargas, Thiago Alexandre Salgueiro Pardo

[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
francielleavargas@usp.br, taspardo@icmc.usp.br

Abstract. *This paper consists of an empirical study on the problem of clustering and hierarchically organizing opinion aspects in product reviews in order to support aspect-based opinion mining applications. We performed a corpus study for characterizing and understanding the involved tasks, looking for linguistic patterns and convergences and divergences across domains. The process has been manually performed and resulted in reference data for future developments and evaluation of automatic methods in the area.*

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 [Pang et al. 2002]. According to [Zhao and Li 2009], this area is different from the traditional text mining area, which is mostly based on objective topics rather than on subjective perceptions.

Opinion mining, according to [Liu 2012], represents a "delicious challenge". Natural languages are very rich and allow to express subjectivity in different ways. Not every opinion is directly expressed and not every aspect appears in a explicit way in the text. For example, in *The camera is expensive*, the evaluated aspect is "price", but it is implicit, not being explicitly said in the sentence and, therefore, must be inferred from the context. Another challenge consists in identifying different aspects that refer to the same object attribute/property. Users recurrently make reference to services or products attributes/properties using different terms. For example, consumers may use the terms "value" and "cost" to designate the price of a product, or use the terms "screen" and "display" to qualify the same smartphone property. Another significant challenge for opinion mining, according to [Yu et al. 2011], is that the product reviews are numerous and disorganized. For example, at the Buscape website, the product *Smartphone Samsung Galaxy J5 SM-J500M* has 931 reviews[1] and, for each review, several aspects are evaluated. Thus, consumers will hardly learn all the other consumers' opinions about the product. According to the author, the hierarchical organization of aspects in product reviews would allow for a better structuring of this data, so that it becomes intelligible for both machines and humans.

[1] According to access at 16 February 2017.

In this scenario, due to all the above challenges, we present a corpus study of opinion aspects, regarding both their clustering and hierarchical organization. We analyze reviews for books and electronic products, looking for linguistic patterns and convergences and divergences across these domains. We expect that such investigation may help characterizing the involved tasks and provide valuable reference data for future developments and evaluations in the area.

The rest of this paper is organized as follows. Section 2 describes the dataset and the analysis method. Section 3 presents the achieved results and learned lessons. Some final remarks are made in Section 4.

2. Corpus study

2.1. The dataset

The dataset overview is shown in Table 1. According to [Zhao and Li 2009], most of the existing opinion mining initiatives are based on product reviews because reviews usually focus on specific products and contain little irrelevant information. Therefore, we randomly selected 60 smartphone and 60 camera reviews from the Buscapé corpus [Hartmann et al. 2014] and 60 book reviews from the ReLi corpus [Freitas et al. 2012]. We manually analyzed the data behavior in each domain.

The Buscape corpus is composed of product reviews in Portuguese language for cameras, notebooks, telephones, TVs, etc. In this corpus, the reviews are partially structured, with sections for “overall impression”, “what I liked” and “what I did not like”. For example, see the following camera review (kept in Portuguese, the original language): *“Excelente, ate hoje todo mundo se impressiona com o tamanho e a beleza dela, alem de fotos perfeitas que podem ser tiradas até 6.3 megapixels! o que gostei: Fina, pratica e leve. o que não gostei: Nenhum!”*. One may note that several aspects were evaluated in this review, but some are not explicit. For example, the terms “beleza”, “fina”, “pratica” and “leve” are clues that indicate the implicit aspects “design”, “size”, “usability” and “weight”, respectively.

The ReLi corpus consists of book reviews, that are also in Portuguese language. As an example, one may find the following review: *“Ótimo livro, bem diferente do que eu imaginava. Apesar de antigão, é uma leitura gostosa, com a linguagem bem moderna. Um livro adolescentes, de aqueles momento foda-se*”. In general, it is possible to notice the challenges in dealing with such texts. They are usually marked by orality and informality, orthographic and grammar errors, and bad language occurrences.

Table 1. Overview of the dataset

domain	reviews	tokens	types
book	60	35,771	1,577
smartphone	60	6,077	1,496
camera	60	3,887	1,060

2.2. The analysis method

In this work, the main purpose is to investigate the clustering and hierarchical organization of opinion aspects. Figure 1 illustrates the clustering process, which was manually performed.

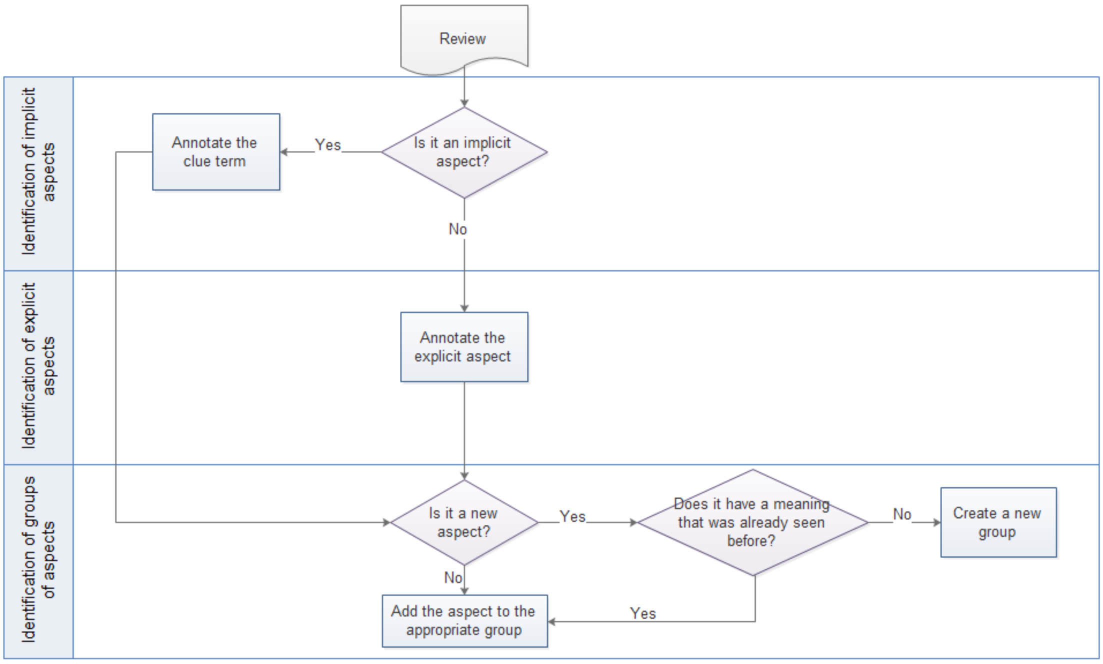
Review
Identification of implicit aspects
Annotate the clue term
Yes
Is it an implicit aspect?
No
Identification of explicit aspects
Annotate the explicit aspect
Identification of groups of aspects
Is it a new aspect?
Yes
Does it have a meaning that was already seen before?
No
Create a new group
No
Add the aspect to the appropriate group
Yes

Figure 1. Clustering product aspects

For the 180 reviews, a human labeled the implicit and explicit aspects. In the identification of implicit aspects were labeled the clue terms that indicated the aspects. For example, in *This camera is expensive*, the evaluated aspect is “price”, but it is implicit. The term “expensive” is the *clue term*. The identification of explicit aspects were directly labeled the aspects. For example, in *The history of the book is bad*, “history” is an explicit aspect. In the last stage, the aspects were clustered the that had similar meaning but with different wording, in order to identify groups. For example, the “cost”, “value”, “price” and “investment” aspects form an unique group. We also modeled the progression of this process of clustering product aspects, looking for a “learning curve”. Once clustering was ready, the obtained groups were manually organized in hierarchies (one for each domain). We compared our obtained hierarchies with other available hierarchies in the area.

3. Results

As explained before, we manually analyzed the product reviews and could observe some very interesting things. The results demonstrated that product reviews may contain portions of irrelevant information, i.e., information that is not directly related to the opinions about the products. The book domain showed 47.98% of irrelevant content, when users comment about the books but do not express any opinion or sentiment. However, for smartphone and camera domains, there was no significant value of irrelevant content.

We could notice that the user profile influences the review informational status[2]. We observed that the smartphone and camera domains present more aspects and groups of aspects than the book domain, as shown in Table 2. Smartphones and cameras are popular technological products and their aspects are more easily identified by non-expert users.

[2] According to [Koch 2009], the informativeness of a text is associated to its ability to present new and unexpected information.

In the book domain, the users often are "just readers" and non-expert users in literature or literary critic. Therefore, they usually do not care about the book technical aspects (such as "size" or "paper type"). These users have been able to evaluate a limited number of product aspects, generally *prototypical aspects* of the books. It is also interesting that the vocabulary in book reviews are not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level. More "adult" books have more sophisticated reviews, with better language, while the reviews of "teenager" books are more often marked by the orality and informal language.

Table 2. Number of aspects

	smartphone	camera	book	average
total number of aspects	459	342	323	**374.66**
unique aspects	180	132	103	**138.33**
explicit aspects	392	289	298	**326,33**
implicit aspects	67	53	25	**48.33**

Overall, 87.08% of the aspects are explicit and 12.91% are implicit in the domains. Furthermore, a product review is composed of, on average, 6 aspects, and it may have at least 1 implicit aspect (see Table 3). We also identified product reviews with the maximum of 20 aspects and the maximum of 5 implicit aspects.

Table 3. Average of aspects

	smartphone	camera	book	average
average number of aspects	7.65	5.70	5.38	**6.24**
average number of explicit aspects	6.53	4.81	4.96	**5.43**
average number of implicit aspects	1.11	0.88	0.41	**0.80**
maximum number of aspects	20	20	15	**18.33**

Regarding the clustering step, we identified, on average, 3.08 explicit aspects and 0.77 implicit aspects per group. Some groups presented the maximum of 19 aspects, as shown in Table 4. In these groups (those that are not unitary, i.e., that contain more than one aspect), the predominant relation between 2 aspects is of the *is-a / hypernym* (or *hyponym*, dependending of the direction of the relation) type (e.g., between the aspects "aparelho" and "produto"), followed by *synonym* ("preço" and "custo") or *identity* (when there is a single aspect without a direct corresponding synonym in the group), *part-of / metonym* (or *holonym*) ("tecla" and "teclado"), *deverbal construction* ("refletir" and "reflexão") and *coreference* ("fabricante" and "marca"). The remaining cases are formed by unitary groups, with only one aspect (without relations, therefore). Table 5 shows the distribution of these relations.

We found several challenges in the analysis: (i) the *inherent ambiguity of the natural languages*, occurring, for example, for the terms "function", "resource" and "application", that are used to refer to the same smartphone application; (ii) the *specificities of the domain*, as each domain requires specific background knowledge; (iii) the *implicit aspects*, as the implicit aspect identification task is not always intuitive; (iv) the *aspects outside the domain*, as the terms "delivery", "technical assistance" and "SAC", which, although have been evaluated, are not directly related to the products.

Table 4. Results of the clustering step

	smartphone	**camera**	**book**	**average**
number of groups of aspects	48	37	24	**36.33**
avg number of aspects in a group	3.75	3.56	4.29	**3.86**
avg number of explicit aspects in a group	2.85	2.78	3.62	**3.08**
avg number of implicit aspects in a group	0.89	0.78	0.66	**0.77**
maximum number of aspects in a group	15	19	17	**17**

Table 5. Relations among aspects

relation	**smartphone**	**camera**	**book**	**average**
is-a / hypernym	45.00%	37.12%	46.60%	**42.90%**
synonym / identity	23.88%	18.93%	26.21%	**23.00%**
part-of / metonym	8.88%	15.90%	7.76%	**10.84%**
deverbal construction	5.55 %	6.81%	9.70%	**7.35%**
coreference	6.66%	8.33%	0.00%	**4.99%**
no relation (unitary groups)	10.00%	13.63%	9.70%	**11.11%**

Our study also showed that it is necessary the analysis of 40 reviews, on average, to learn/identify most of the relevant aspects in a given domain. The "learning curves" (shown in Appendix 1), represent the learning behavior of groups of aspects for the analyzed domains, that is the amount of new groups of aspects learned at each review. We also hierarchically organized the identified groups of aspects (see in Appendix 2) and compared our hierarchies with the hierarchies proposed by [Condori 2014], [Aciar et al. 2006] and [Goulart and Montardo 2007]. In the hierarchies in the literature, the relations of the type *is-a* are more often used. However, we observed that reviews are predominantly composed by *part-of* relations. Furthermore, the hierarchies in the literature do not represent all the domain specificities.

4. Final remarks

As shown above, clustering product aspects and building their hierarchical organizations are not simple tasks. There are several challenges to overcome. The results demonstrated that product reviews may contain a significant portion of irrelevant content and that informational status may be influenced by the user profile. The vocabulary in book reviews is not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level, which results in varied writing behavior. In addition, it was found that, for a good domain coverage, at least 40 reviews are required, on average. We also observed that, on average, some domains may have more identifiable aspects. The aspect groups and the hierarchies will be made available for research purposes. We expect that automatic methods for opinion mining may be trained and/or evaluated over such datasets.

Acknowledgments

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

References

Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2006). Recommender system based on consumer product reviews. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 719–723, Washington, DC, USA. IEEE Computer Society.

Condori, R. E. L. (2014). Sumarização automática de opiniões baseada em aspectos. Master's thesis, São Carlos, SP, Brazil.

Freitas, C., Motta, E., Milidiú, R., and Cesar, J. (2012). Vampiro que brilha... rÁ! desafios na anotação de opinião em um córpus de resenhas de livros. In *Anais do XI Encontro de Linguística de Corpus*, São Carlos, SP, Brazil.

Goulart, R. R. V. and Montardo, S. P. (2007). Os mecanismos de busca e suas implicações em comunicação e marketing. In *Anais do V Congresso Nacional de História da Mídia*, pages 478–514, São Paulo, SP, Brazil.

Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M. D. G. V., Pardo, T., and Aluísio, S. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3865–3871, Reykjavik, Iceland. European Language Resources Association.

Koch, I. G. V. (2009). *Introdução à Linguística Textual*. Martins Fontes, São Paulo, SP, Brazil, 2nd edition.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 1st edition.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu, J., Zha, Z., Meng, W., Wang, K., and Chua, T. (2011). 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*, pages 140–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhao, L. and Li, C. (2009). Ontology based opinion mining for movie reviews. In *Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management*, pages 204–214, Berlin, Heidelberg. Springer-Verlag.

Appendix 1

We present below the learning curves for the identification of groups of aspects. As an illustration of how to interpret these graphics, in Figure 2, one may see that, after have analyzed 2 smartphone reviews, we could identify 10 groups of aspects; after 60 reviews, we end up with 48 groups.

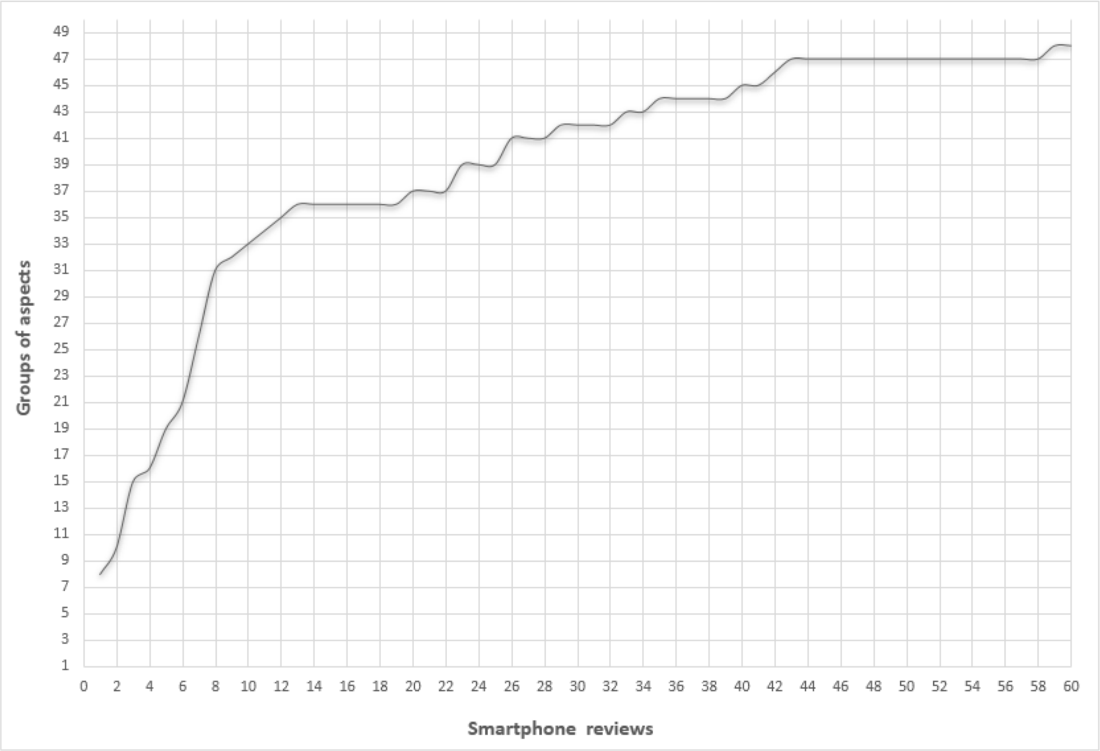

Figure 2. Learning curve for the smartphone domain

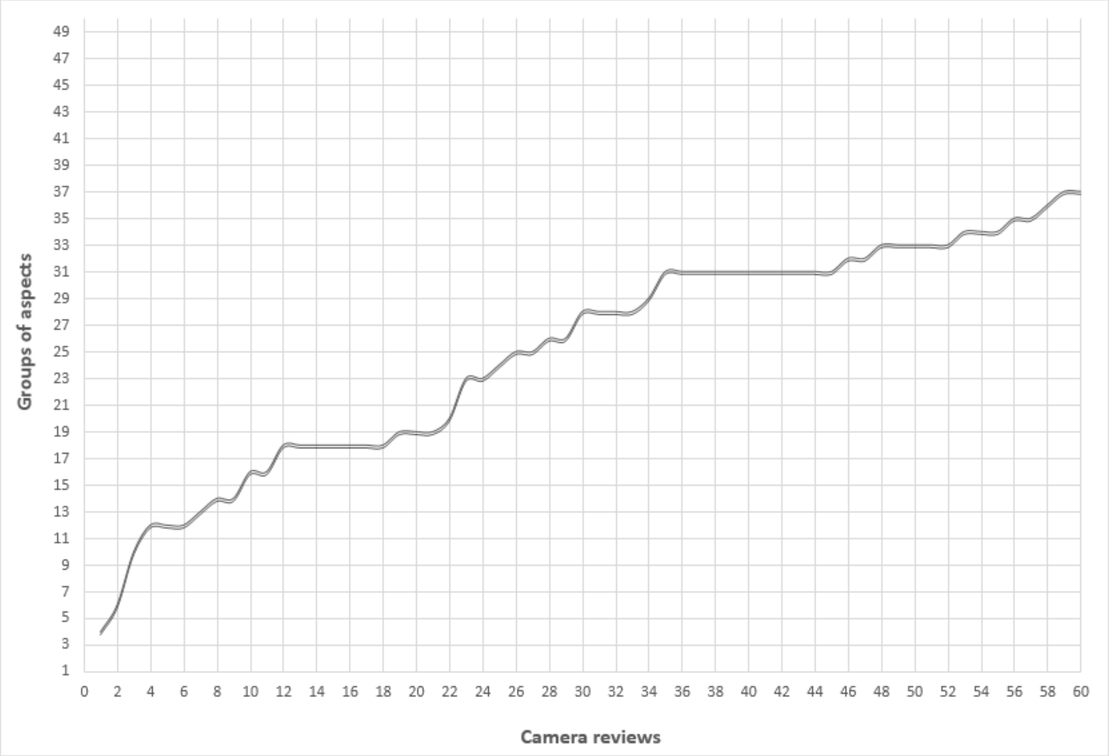

Figure 3. Learning curve for the camera domain

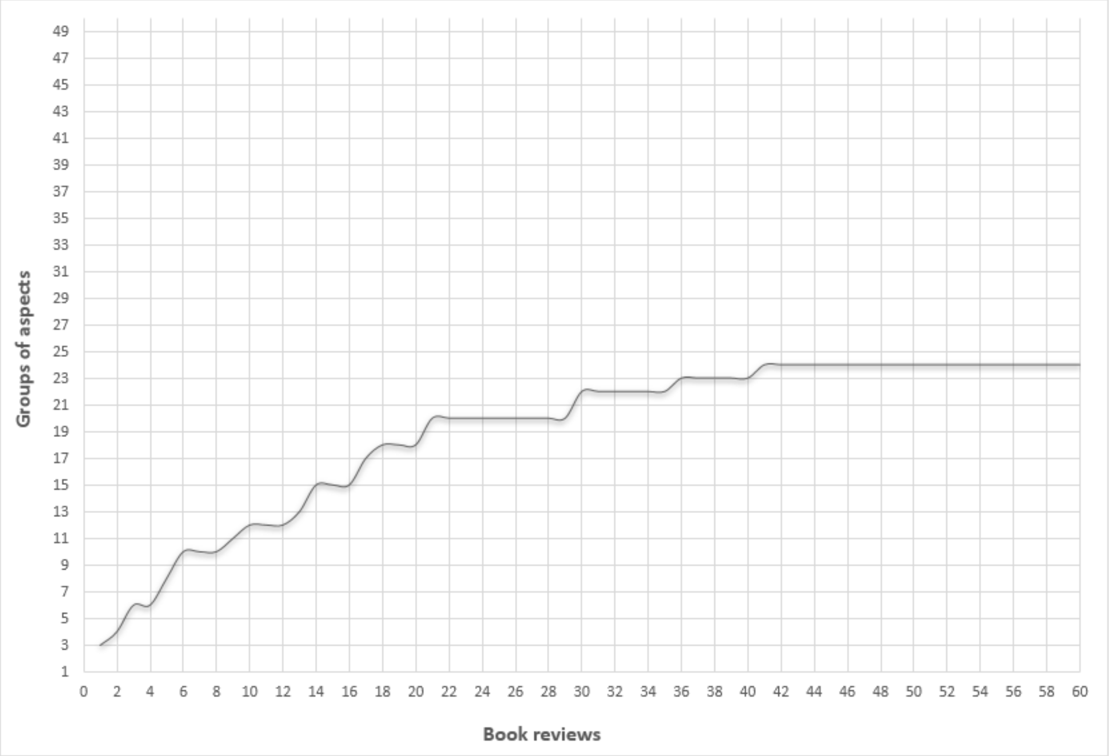

Figure 4. Learning curve for the book domain

Appendix 2

We present below the hierarchies obtained for the smartphone, camera and book domains, where each circle represents a group of aspects. For each group, we show only the most representative word. We show them in Portuguese because the corpus is in this language.

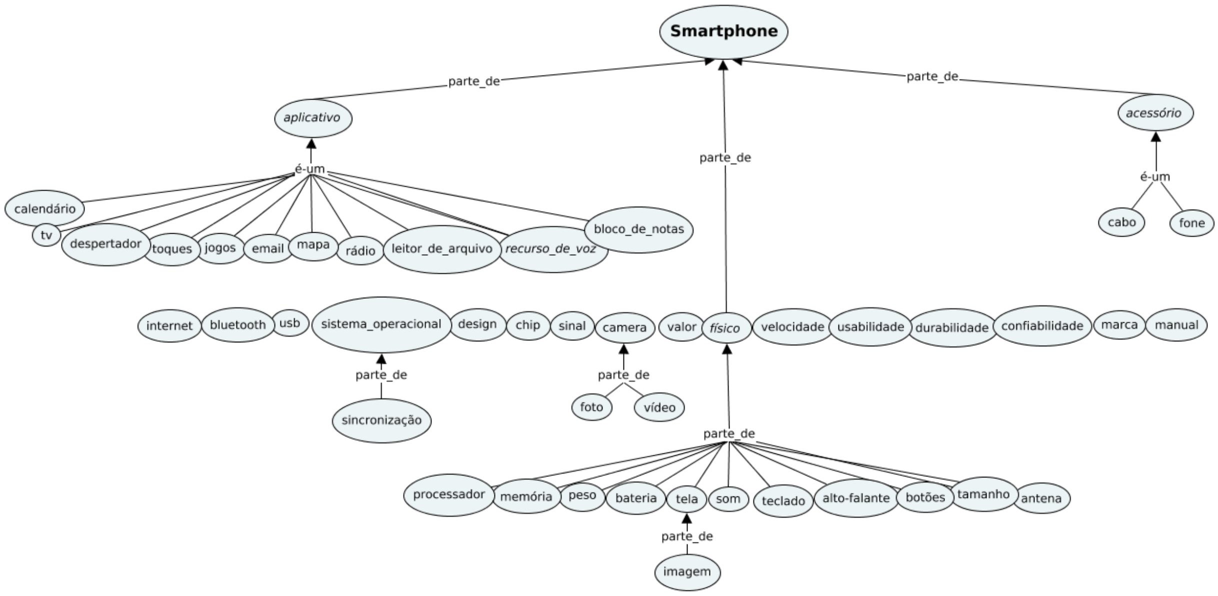

Figure 5. Hierarchy for the smartphone domain

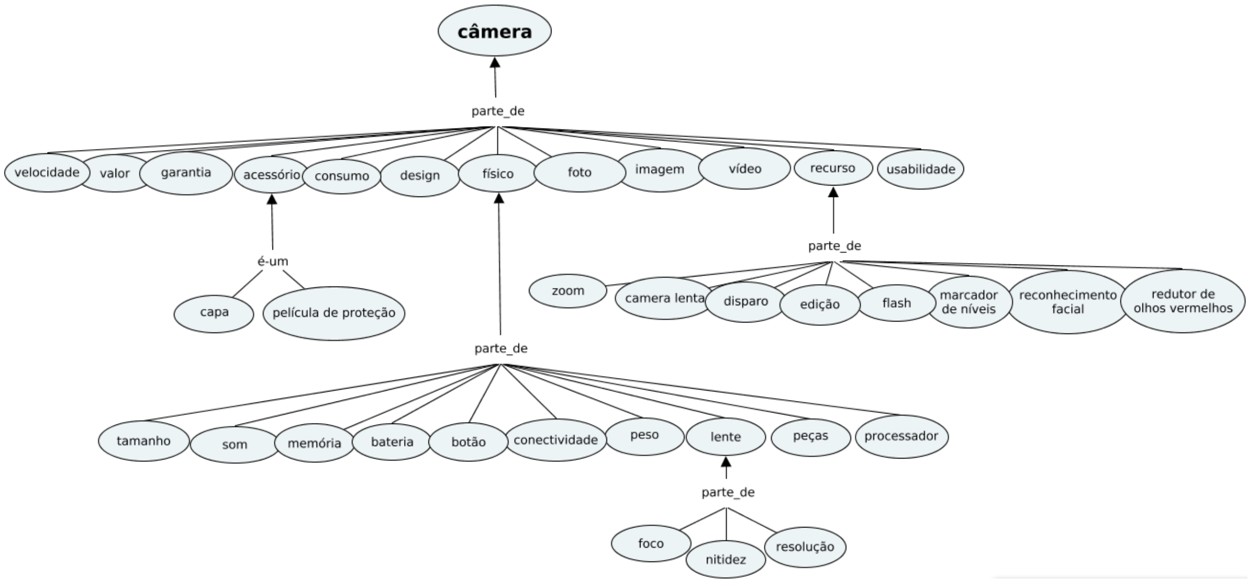

Figure 6. Hierarchy for the camera domain

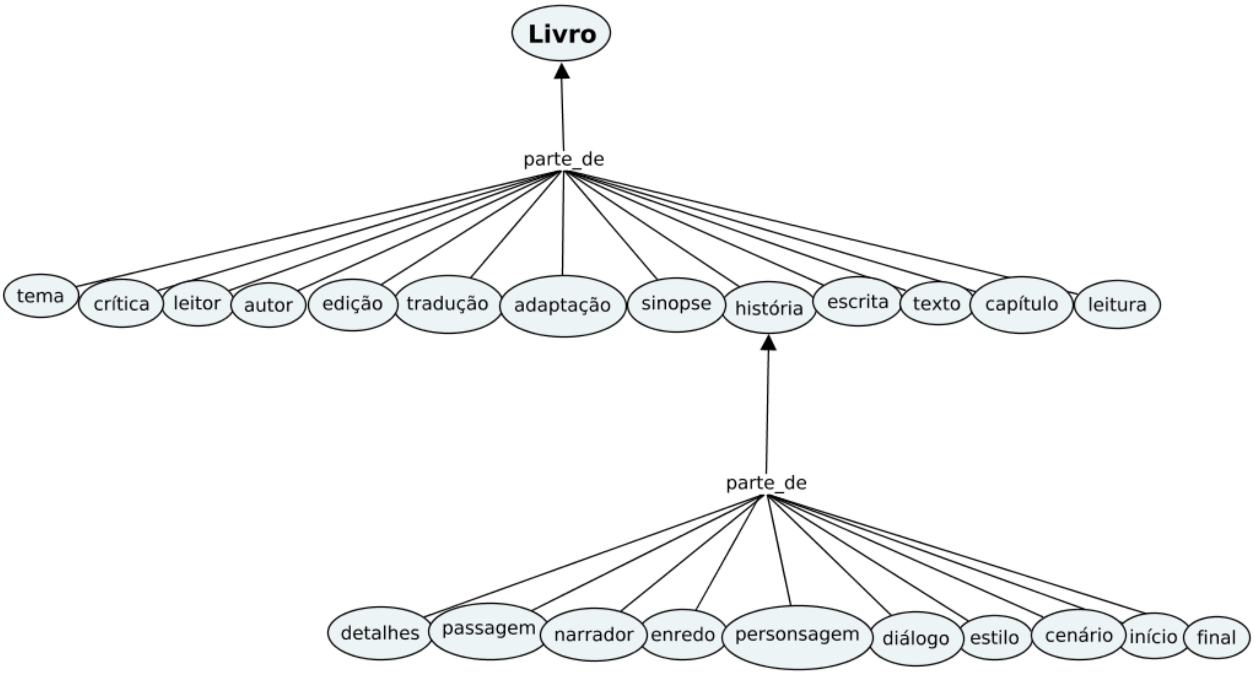

Figure 7. Hierarchy for the book domain