Automatic Characteristics Extraction for Sentiment Analysis Tasks

Juan M. Rodríguez ✉[12], Hernán D. Merlino[2], Ramón García-Martínez[2†]

[1]PhD Program in Computer Sciences. National University of La Plata (UNLP). La Plata, Bs. As., Argentina.

[2] Systems Information Research Group. Department of Productive and Technological Development. National University of Lanús. Lanús, Bs. As., Argentina.

jmrodriguez1982@gmail.com, hmerlino@gmail.com

Abstract. The following article proposes the use of Open Information Extraction Methods (OIE), in particular ClausIE, to automatically obtain characteristics from movie reviews. Within automatic summary generation and sentiment analysis frameworks, this approach is compared with other two in which manual steps are used to obtain the characteristics of a service or product. The obtained result shows that ClausIE can be used for the extraction of characteristics in a semi-automatic way. It requires a minimum manual intervention that is explained in the results section.

Keywords. Sentiment analysis, characteristics extraction, knowledge extraction, semantic relations extraction, Open Information Extraction, natural language processing.

1 Introduction

The task of mining film reviews in order to obtain an automatically summary consists mainly in three tasks. The first task is to obtain the pair: characteristic-opinion analyzing one by one the constituent sentences of the review. The second task is to identify the polarity (positive or negative) of each opinion. And the final task consist in building a structured list based on the characteristics and opinions found, calculating the polarity of each characteristic as the average polarity of all opinions in which each characteristic was found [1]. The present work focuses mainly on improving the first of the mentioned tasks, which is, the identification of characteristics and words that express opinions, but mainly in the identification of characteristics.

Characteristics, also called aspects, are individual elements of a larger entity, each of which can be evaluated independently. For instance, a restaurant has the following characteristics: *food*, *atmosphere*, *service* and *price*. Even, if you know that you are talking about a particular restaurant which offers a particular dish such as: “fish tacos with French fries” this dish can be a characteristic.

The main difference between sentiment analysis on reviews and the automatic summary of reviews with sentiment analysis is that in the first case only the global polarity of given text (the review) is calculated, while in the second case the main characteristics are extracted from the text and then the polarity of each characteristic is calculated individually.

Aspects play an important role in sentiment analysis because although it's very valuable to have the general idea of an opinion, the review of aspects (individual characteristics) plays a fundamental role in the decision making process. A classic example is the review of a product, where often a single aspect is decisive for a user to decide to buy it (typically the price and/or the quality).

The focus of this work is the extraction of characteristics. We sought an automatic solution based on the use of a method of Open Information Extraction. In particular, a solution based on ClausIE method [2] was proposed.

1.1 Introduction to Open Information Extraction (OIE)

Knowledge extraction is any technique which allows that an automated process analyze unstructured information sources, such as texts written in natural language and extract the embedded knowledge in order to represent it in a structured way, able to be manipulated in an automatic reasoning processes, for instance: a production rule or a subgraph in a semantic network. The information obtained as output of this type of process is called: *piece of knowledge* [3; 4].

In 2007 Michele Banko introduces a new concept in the field of knowledge extraction, which is called: Open Information Extraction (OIE). It is a paradigm of knowledge extraction where a computer system makes a single pass on the total unstructured information sources in natural language format (called corpus of documents) given as input and extracts a large set of relational tuples without requiring any kind of human participation. In the same work Banko presents a method called TEXT RUNNER, which is the first method that works within this new paradigm [5]. Since this work was published other methods of knowledge extraction were proposed under the paradigm that Banko called Open Information Extraction or just OIE.

Semantic relation extraction methods that work in accordance with the OIE paradigm return a tuple for each semantic relation discovered. The tuple has the form (Entity 1, Relation, Entity 2), where entities are usually well-identified objects, persons, places, companies, dates, etc., and the relationship is the semantic relationship between the two entities, often factual information, such as "Who did what to whom". To illustrate this, consider the following sentence:

Albert Einstein, who was born in Ulm, has won the Nobel Prize.

Extracting the relationships in the sentences and expressing them as a tuple in the form (Entity 1, Relation, Entity 2) should return the following:

- (Albert Einstein, has won, the Nobel Prize)
- (Albert Einstein, was born in, Ulm)

1.2 The selected method: ClausIE

A documentary investigation was carried out in [6] over a few semantic relation extraction methods, which work in accordance with the Open Information Extraction paradigm and it was found that ClausIE was, according to its authors [2] the method that achieved a better precision. This assertion was tested in [7] where a partial publication was made of a comparative evaluation between ClausIE and other similar information extraction methods: ReVerb [8] and OLLIE [9]. A final version of the results is in the process of being published. But these results would be favorable to ClausIE, which is why this method was selected for this work.

2 Related works

Blair-Goldensohn and other used in [10] a hybrid method to extract the characteristics of the reviews, consisting in two methods: a dynamic method and a static extraction method. They searched for nouns or compound nouns, constituted by two or three words that appeared in some phrases that indicated a sentiment load (polarity) or phrases that matched with certain syntactic patterns that were possible indicators of an opinion. They found that the patterns were more accurate than the occurrence of nouns in phrases loaded with sentiments. The most productive pattern they had was looking for sequences of nouns that had an adjective immediately before, so they found, for instance, phrases like "...great fish tacos...", in restaurant reviews. They included "fish tacos" as a characteristic, because this was a very common dish (a characteristic dish) for the restaurants evaluated in the reviews.

For the second approach, the static method for characteristics extraction, they took 1500 random sentences of hotels and restaurants reviews and they manually labeled them indicating the "coarse-grained" characteristics they found there. They called these characteristics "coarse-grained" because they are very general. They are the characteristics that can be found in any hotel or restaurant. These were not as specific as: "fish taco" (which is a "fine-grained" characteristic). The characteristics were the following: *food*, *decor*, *service*, and *value* for restaurants and *rooms*, *location*, *dining*, *service*, and *value* for hotels. They also included a category *other*, to label sentences that did not include any of the previous characteristics. Then they trained a classifier with the set of labeled cases. Finally they used the already trained classifier to detect aspects in any other sentences.

In [1] was carried out an experiment similar to the proposed in this article. An automatic summary of IMDB films reviews was made, focused on finding opinions about the characteristics of a given film. The authors defined a film characteristic as an element (*staging*, *music*, etc.) or as people (*director*, *actor*, etc.) mentioned in an opinion. The authors manually defined a list of main characteristics (called characteristics of type element) that are relevant in a film and for the characteristics associated with people they used the full cast list as it is published in IMDB for a given film.

The element-type characteristics selected manually were the following six:

- OA: general
- ST: script
- CH: character design
- VP: visual effects
- MS: sound and music effects
- SE: special effects

Each feature was associated with multiple keywords, for instance the characteristic *script* was associated to the different keywords: *story*, *plot*, *script*, *storyline*, *dialogue*, *screenplay*, *ending*, *line*, *scene* and *tale*. To obtain these keywords, they worked with a dataset of 1100 IMDB film reviews manually labeled. Then the keywords associated with a characteristic were obtained just filtering the most frequent words.

3 Current problems

Authors in [10] found a fundamental problem with the first approach, the dynamic method; the problem is that the found aspects are only fine-grained. It is not trivial to deduce that *fish soup* and *lobster soup* are part of a larger aspect that could be: *soups*, *entrances* or just *food*.

About the second approach, the classifier achieved a fairly high *precision*. It obtained 86.9% for *service* and 90.3% for *price* in the case of restaurants. For hotels it achieved 83.9% of *precision* for *service* and 83.3% for *price*. The *recall* was little lower, it was between 54.5% and 69.7% for the mentioned cases. However, this method has the disadvantage of needing a set of cases manually labeled.

The main problem associated with the work carried out in [1] is the need to know the set of relevant characteristics before generate the manual labeling.

4 Proposed solution

In order to elaborate experimental tests, a dataset of 2000 film reviews extracted from the IMDB site was used and hand-labeled in two sets: a group of 1000 positive reviews and another of 1000 negative reviews. The data set was originally created by Pang and Lee [11] to train a text classifier to perform tasks of sentiment analysis. Since then the dataset has been available on the web and has been used in other publications.

4.1 Obtaining characteristics

The semantic relation extraction method under OIE paradigm: ClausIE, was executed over the dataset. ClausIE returns for each semantic relation a tuple of the form: (Entity 1, Relation, Entity 2) where "Entity" is any syntactic element that refers to something concrete: a person, a place, a brand, etc. (although it can also be a date or another type of abstract entity), ClausIE uses an entity name detection algorithm for it (NER). It was surmised that the characteristics of a movie should be able to be

detected as entities. And in a fairly large corpus, these would be repeated with a higher frequency than other possible entities. At least the characteristics called fine-grained [10].

The obtained semantic extractions were ordered by the number of times an initial "Entity" was repeated. Then the results were filtered to show only those that start with the article "the", in this way we avoid listing pronouns and other frequently used words. The list obtained is shown in Table 1.

Table 1. Repetitions of the first entity starting with "the"

Entity 1	Repetitions
the film	3538
the movie	1637
the story	683
the plot	501
the audience	396
the script	387
the characters	320
the director	258
the two	234
the filmmakers	197
the acting	192
the actors	184
the camera	147
the world	143
the dialogue	140
the cast	128
the man	123
the ending	114
the music	112
the scenes	101
the result	100
the performances	99
the special effects	99

The list shown in Table 1 corresponds well with a list of characteristics (or keywords that indicate characteristics according to the nomenclature in [1]). However, the generation of this list required two manual steps, so its generation was not completely automatic. These steps were the following:

- An arbitrary cut at 99 repetitions, we didn't take more elements than those that appear up to 99 times.
- Manual elimination of some entities that do not correspond to films characteristics: *the two, the camera, the world, the man* (marked in bold)

Comparing the generated word list, with the list of keywords characteristic that presented in [1], it is observed that there are 12 common words out of 38. However, in the list of Table 1 there are 8 high-frequency words that were not used in the work of Zhuang and others [1]. Finally, it should be noted that with the 12 common words found, the coarse-grained characteristics defined in [1] are all covered, although some groups have only one word. This is shown in Table 2.

Table 2. Coarse-grained characteristics and their associated keywords in [1].

Characteristics	Keywords
OA	**film, movie**
ST	**story, plot, script**, storyline, **dialogue**, screenplay, **ending**, line, **scene**, tale
CH	**character**, characterization, role
VP	**scene**, fight-scene, action-scene, action-sequence, set, battle-scene, picture, scenery, setting, visual-effects, color, background, image
MS	**music**, score, song, sound, soundtrack, theme
SE	**special-effects**, effect, CGI, SFX

The 12 keywords in common are shown in bold. The other keywords found would belong to the coarse-grained characteristics OA and CH, according to the following list:

- **CH:** acting, actors, cast, performances
- **OA:** director, audience, filmmakers, results

4.2 Sentiment analysis of each characteristic

For the following analysis, the list in Table 1 was taken as a list of characteristics (without counting the filtered words) because the goal of this article is obtaining features automatically. For each characteristic, a sentiment analysis task was performed using SentiWordnet 3.0 which is a sentiment lexicon [12].

We proceeded as follows: all the semantic extractions were recovered for a given review, then each extraction was joined in a single sentence concatenated "Entity 01" with "Relationship" with "Entity 02". And if any characteristic in the list appeared in the resulting sentence then it was evaluated using the SentIWordNet 3.0 lexicon. At last, according to the result of the polarity obtained, positive or negative, this characteristic was marked with a 1 (positive) or a -1 (negative) in a final result table.

Finally, for each of the reviews with at least one characteristic, the values of the polarities of each characteristic were sum together in order to obtain a global result or polarity for the review. This last step was carried out in order to compare the analysis of the characteristics, which together should be identical to the global analysis, against the labeled polarity of the review. If the polarities don't match, maybe the characteristics weren't representative of the film or maybe the calculation of the polarity of each characteristic was wrong.

5 Results and Conclusions

The overall *precision* for sentiment analysis (more specifically the obtaining of the polarity), using SentiWordNet 3.0 over the 2000 films reviews is 0.662; There are 1324 reviews ranked correctly. This is the floor on which the characteristics analysis is based, a low floor, especially when comparing the obtained results against supervised classification methods such as those used by Pang and Lee [11].

Only in 1187 reviews was found at least one characteristic to analyze, which is equivalent to 59% of them.

The sum of the positive and negative polarities of each of the characteristics, to obtain the global polarity of the review, gave a *precision* of 0.619, that is, 735 correctly classified of 1187 (the ones that had at least one characteristic). Although it is a low number, it is close to the overall accuracy of SentiWordNet 3.0. On that same segment of reviews, the 1187 that have at least one characteristic, SentiWordNet 3.0 obtained, working directly over the full text of the review, a *precision* of 0.666, which is a total of 790 correctly classified.

However, the average *precision* obtained was greater than that calculated in [1], where the average *precision* of different pairs of characteristics-opinions for different films was calculated and the obtained value was: 0.483. Nevertheless, since the set of used reviews is different (the one used by the authors is not available) and the way of analyzing the polarity is different, the precisions are not directly comparable. It is cited only as a reference.

Finally, the main positive result is the extraction almost automatically (with minimal manual intervention) of the characteristics of a product or service (in this case, films). The characteristics may not be exhaustive, when compared with those used in the work of Zhuang and others [1] but they are representative and undoubtedly used more frequently in the analyzed dataset. The sentiment analysis over individual characteristics does not improve the overall performance of the used method (in this case the sentiment lexicon SentiWordNet) but remains consistent with the *precision* of the method.

6 Future research lines

The revision and comparison of this approach with other extraction methods of automatic characteristics such as SABER [13] is a future work.

References

1. Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 43-50). ACM.
2. Del Corro, L., & Gemulla, R. (2013, May). ClausIE: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web (pp. 355-366). International World Wide Web Conferences Steering Committee.

3. García-Martínez, R. & Britos, P. V. (2004). Ingeniería de sistemas expertos. Nueva Librería. ISBN 987-1104-15
4. Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). Ingeniería del conocimiento. Editorial Centro de Estudios Ramón Areces. ISBN 84-8004-269-9.
5. Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction for the web. In IJCAI (Vol. 7, pp. 2670-2676).
6. Rodríguez, J. M., Merlino, H., García-Martínez, R. (2015). Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web. XXI Congreso Argentino de Ciencias de la Computación (CACIC 2015). Buenos Aires, Argentina.
7. Rodríguez, J. M., Merlino, H. D., Pesado, P., & García-Martínez, R. (2016, August). Performance Evaluation of Knowledge Extraction Methods. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 16-22). Springer International Publishing.
8. Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics.
9. Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 523-534). Association for Computational Linguistics.
10. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008, April). Building a sentiment summarizer for local service reviews. In WWW workshop on NLP in the information explosion era (Vol. 14, pp. 339-348).
11. Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
12. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).
13. Caputo, A., Basile, P., de Gemmis, M., Lops, P., Semeraro, G., & Rossiello, G. (2017). SABRE: A Sentiment Aspect-Based Retrieval Engine. In Information Filtering and Retrieval (pp. 63-78). Springer International Publishing.