Classification based on aspect extraction

Lukas Hirsch

Contributing authors: lukas.hirsch@studenti.unimi.it;

Abstract

This study explores the use of semantic parsing techniques to classify user comments on board games. Using a graph-based approach and *ConceptNet* embeddings, the proposed method aims to categorize comments into predefined aspects. The performance of this method is compared to OpenAI's *GPT-4o* model by manually counting classification errors. The results show that while the *GPT-4o* model achieved higher accuracy in correct classifications, the presented method extracted a significant number of concepts, although many of them lacked semantic relevance. This research highlights the strengths and limitations of both approaches and underscores the challenge of accurately classifying aspects based on natural language.

1 Introduction

In the field of *natural language processing (NLP)*, the ability to semantically parse and understand text is critical for a variety of applications, including sentiment analysis and reasoning. Semantic parsing, which involves breaking down text into multi-word concepts, improves the understanding and interpretation of textual data. This can be achieved through phrase structure grammars or statistical, training-based algorithms. Recent advances, such as transformational models like *BERT* [1], have shown great promise in this area.

This project aims to classify user comments on the board game *Mighty Empires* [2] into predefined categories using a graph-based technique combined with embeddings from the multilingual knowledge graph *ConceptNet* [3]. To evaluate the effectiveness of this approach, it will be compared to the *GPT-4o* [4] model from OpenAI. The dataset obtained from *BoardGameGeek* [5] contains 100 user comments on *Mighty Empires*. The predefined categories are based on the Italian definition of *Goblinpedia* [6] and include various aspects such as luck, accounting, downtime, interaction, and more. The primary objective is to compare the performance of the unsupervised

method presented in this study with that of the *GPT-4o* model, highlighting the strengths and limitations of each.

The experimental results show that the *GPT-4o* model significantly outperforms the proposed method in terms of correct classifications. In particular, the *GPT-4o* model achieved a higher number of correct classifications with fewer errors, demonstrating the advantages of training-based algorithms. In contrast, while the unsupervised method extracted a significant number of concepts, many lacked semantic relevance, resulting in incorrect classifications. These results underscore the challenge of accurately classifying aspects based on user comments and highlight the need for further refinement of semantic parsing techniques.

2 Dataset

The dataset used is provided by *BoardGameGeek (BGG)* [5], a comprehensive database of board games and the global community of gamers. Users contribute data, including statistics and ratings, which assess the popularity of each game based on several criteria. These criteria include an overall rating, the number of users who voted for the game, and the community's opinion of the game's playability with different numbers of players. In addition, BGG allows access to user comments on games, which is the only part used for this project. Data was retrieved using the *BGG* API.

3 Methodology

The primary goal of this project is to classify comments on board games into predefined categories using semantic parsing techniques. The predefined categories are based on the Italian definition of Goblinpedia:

1. **luck or alea:** all those game elements independent of player intervention, introduced by game mechanics outside the control of the players.
2. **bookkeeping**: manual recording of data and potentially automatic or semi-automatic game processes, including also the need of continuosly accessing the rulebook for reference.
3. **downtime**: unproductive waiting time between one player turn and the next. By unproductive we mean not only having nothing (or little) to do, but also nothing (or little) to think about.
4. **interaction**: the degree of influence that one player's actions have on the actions of the other participants.
5. **bash the leader**: when, to prevent the victory of whoever is first, the players are forced to take actions against him, often to the detriment of their own advantage or in any case without gaining anything directly. At the table, the unfortunate situation can arise whereby one or more must "sacrifice" themselves to curb the leader and let the others benefit from this conduct.
6. **complicated** vs complex: A game is complicated the more the rules are quantitatively many and qualitatively equipped with exceptions. Once you understand and learn all the variables, a game (that is only) complicated is not difficult to

master. In a complicated game, solving a problem leads to immediate, certain and predictable results.
A game is as complex as the repercussions of one's actions are difficult to predict and master. Even once you understand and learn all the variables, a complex game is still difficult to master. In a complex game, solving one problem leads to other problems.

A key goal is to perform a comprehensive comparison between the unsupervised, non-learning-based method proposed in this study and the *GPT-4o* model developed by OpenAI. Highlight the strengths and limitations of the presented methodology in relation to the advanced capabilities of the *GPT-4o* model. To achieve this goal, the process starts by parsing the comments to identify their structure, including the different sentences, which are then further parsed to identify the verbs and nouns. Verbs are essential because they typically denote actions or events. Each verb is analyzed in its base form, via lemmatization, which helps to understand the core action being described. Noun phrases within the sentence are also identified. These phrases often serve as the subject or object associated with the verbs. By linking these noun phrases to their corresponding verbs, potential event concepts are formed. In addition, any adjectives within the noun phrases are included to provide more context for the concept. In cases where noun phrases are associated with auxiliary verbs, which are verbs used in conjunction with a main verb to express tense, mood, or voice, this is accounted for by forming concepts based on the structure of the auxiliary verb. Resulting in a list of unique concepts that encapsulate the events described in the sentence. These concepts are then compiled into a comprehensive list after processing each sentence individually. In this way, the slightly modified method from [7], ensures that all potential event concepts are captured and organized effectively, providing a structured representation of the information contained in the text.

To further simplify the extracted concepts, pre-trained word embeddings from *ConceptNet* are used. These embeddings provide a numerical representation of words, capturing semantic similarities based on their contextual usage in a large corpus of text. To represent complex concepts, which are often multi-word expressions, an embedding is computed for each concept by averaging the embeddings of the individual words that make up the concept. This process ensures that each concept is represented as a single, coherent vector in the embedding space. Concepts without valid embeddings (i.e., those not found in the pre-trained embeddings) are excluded from further analysis to maintain the integrity of the representation. The next step is to cluster these concept embeddings to identify groups of semantically related concepts. This is done using the *K-Medoids* algorithm [8], a clustering technique particularly suited to scenarios where representative examples (medoids) from the dataset are preferable to mean-based representatives (centroids). The number of clusters is three, with a dynamic determination when the input number of concepts is less than three, assuming that no more aspects are mentioned in a comment. The clustered concepts are then used for further processing.

To classify the concepts into the given classes, an embedding is created by averaging

the embeddings of the words that make up the concept. Similarly, each class is represented by an embedding computed from the keywords associated with that class. To determine the similarity between a concept and a class, cosine similarity is used. If the embedding of a concept is sufficiently similar to the embedding of a class (exceeding a specified threshold), the class is considered a potential match for the concept. The classes are then ranked based on their similarity to the concept, and the top matches are identified. For each concept, the class with the highest similarity score is selected, ensuring that each concept is assigned to the class that best represents its meaning.

4 Results

4.1 Experimental Methodology

The data used for the experiment consisted of the 100 user comments of the game *Mighty Empires* [2]. To simplify the distance calculation and the resulting classification of the concepts in a comment, the above classes were simplified and reduced to descriptive keywords. The metric used to evaluate and compare the performance of the *GPT-4o* model and the method presented in this study involves manually counting errors. An error is defined as an obviously incorrectly classified aspect, as well as a missing aspect that is present in the comment. This approach allows a direct and clear comparison between the two methods. To capture the aspects generated by the *GPT-4o* model, a simple *CSV-file* is uploaded to the chat. Immediate prompt engineering ensures that the output produced by the *GPT-4o* model mirrors that of the implemented unsupervised method. This output is a combination of the extracted concepts and their classified aspects.

4.2 Experimental Results

Due to the limitation to one board game, it is not expected that the data represent the aspects equally. It is noticeable that the result of the *GPT* model largely consists of no assignment for the comments. The presented method extracts for most of the comments a significant amount of concepts, most of which are not assigned to an aspect by the similarity calculation. This can be explained by the lack of meaning in some concepts. Regarding the similarity between concepts and aspects, in most cases it does not exceed 0.3. The manual evaluation of the comments revealed a total of 23 mapped aspects to the comments, mostly "luck or alea" and "downtime". A key

Metrics	**Method**	**GPT-4o**
Errors (wrong classified)	20	21
Errors (missing)	18	6
Correct classifications	5	17

Table 1: Comparison of the proposed method and *GPT-4o* aspect classifications

observation is that, as expected, the *GPT-4o* model performed more accurately with a

significantly higher number of correct classifications. In contrast, the presented unsupervised method managed to extract a significant number of concepts from most of the comments, although many of them were incorrect. This indicates a potential problem with the semantic relevance of the extracted concepts, highlighting the challenge of accurately classifying aspects based on the given comments. It is also noteworthy that the correct mappings to the aspects are sometimes not based on the meaning of the concepts, but rather on the chosen similarity comparison, e.g. the concept "play I" is considered similar to the aspect "luck or alea".

5 Conclusion

The experimental results show that the *GPT-4o* model outperforms the unsupervised method in terms of correct classifications. The superior accuracy of the *GPT-4o* model underlines the progress of training-based algorithms for natural language processing tasks. The significance of the results of the presented approach is generally questionable. However, the ability of the unsupervised method to extract numerous concepts suggests potential for refinement in semantic relevance and aspect classification. Also, the manual evaluation of the comments is prone to error, as it is interpretable whether an aspect is mentioned in a comment, as well as the definition of the aspects themselves. Future work could focus on improving the semantic parsing accuracy by integrating more sophisticated similarity measures or hybrid approaches that combine supervised and unsupervised techniques. In addition, expanding the dataset to include different board games could provide a more comprehensive evaluation of the performance of the methods in different contexts.

References

[1] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018) 1810.04805

[2] Games Workshop: Mighty Empires. https://boardgamegeek.com/boardgame/52/mighty-empires. board game (1990)

[3] Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. CoRR **abs/1612.03975** (2016) 1612.03975

[4] OpenAI: GPT-4o: Generative Pre-trained Transformer 4 - Optimized. Available online at https://www.openai.com/gpt-4o. Accessed: 2024-07-22 (2024)

[5] BoardGameGeek: BoardGameGeek: The BGG API. Accessed: 2024-07-21 (2000). https://www.boardgamegeek.com

[6] Goblinpedia: Goblinpedia: The Italian Board Game Encyclopedia. https://www.goblins.net/goblinpedia. Accessed: 2024-07-26 (2024)

[7] Rajagopal, D., Cambria, E., Olsher, D., Kwok, K.: A graph-based approach to commonsense concept extraction and semantic similarity detection. In: Proceedings of the 22nd International Conference on World Wide Web. WWW '13 Companion, pp. 565–570. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2487788.2487995 . https://doi.org/10.1145/2487788.2487995

[8] Park, H.-S., Jun, C.-H.: A simple and fast algorithm for k-medoids clustering. Expert Systems with Applications **36**(2, Part 2), 3336–3341 (2009) https://doi.org/10.1016/j.eswa.2008.01.039