Movie Review Ontology

Yining Feng

February 28, 2021

Abstract

Ontology itself is an explicitly defined reference model of application domains with the purpose of improving information consistency and knowledge sharing. It describes the semantics of a domain in both human-understandable and computer-processable way. In this study, an ontology-based approach is proposed for opinion mining. In general, opinion mining is quite context-sensitive, and, at a coarser granularity, quite domain dependent. This study introduces a fine-grain approach for opinion mining, which uses the ontology structure as an essential part of the feature extraction process, by taking account the relations between concepts. The experiment result shows the benefits of exploiting ontology structure to opinion mining. In this study, 42 movie reviews collected from 12 popular and authoritative sources for film ratings and critical reviews are analyzed, respectively.

1. Introduction & Problem Statement

The main goal is to improve feature level opinion mining by employing ontology. The project will be broken down into five steps/stages: word segmentation, part-of-speech tagging, ontology development, feature identification, and sentiment classification (positive, neutral, negative). Specifically, for polarity identification step, polarity measurement will be developed based on a lexicon of tagged positive and negative sentiment terms which are used to quantify positive/negative sentiment. In this part, SentiWordNet 3.0 will be used as it provides a readily interpretable positive and negative polarity value for a set of "affective" terms.

This study will focus on two tasks for sentiment analysis:

(1) convert the polarity obtained from the polarity identification step to more precise one by analyzing contextual features, such as "negation" rules.

(2) with the help of ontology again, make it possible to compute the polarities of the nodes through the hierarchy relationship.

2. Research Design and Modeling Method

2.1 Data Description and Processing

The movie review dataset used for sentiment analysis is comprised of 42 movie review documents collected from New York Times, Washington Post, RogerEbert.com, The Seattle Times, Vanity Fair, IGN, Hollywood Reporter, Breathe Dream Go, The Guardian, Variety, NPR, and Cinemajam. Each movie review document is trimmed to 500 words and contains the movie title, plot summary, release date, actors /actresses, directors, movie review texts and so on.

All 42 movie review texts would be cleaned and pre-processed by using special natural language processing features, such as: *bags of words* in combination of *n*-grams; *segmentation* by separating each single word with punctuation or white space, removing all stop words, such as *a* and *the*, or by making all capital letters a lower case; *stemming* by reducing words to their base or root forms; *term frequency* by counting the frequency of words which helps identify how important a word is to a document in a corpus; *word embedding* is transformation of words to an array of numeric values of semantic or contextual information that computer can understand. The resulting dataset includes 42 unlabeled movies reviews trimmed from 21,896 terms to 9,835 terms.

2.2 Methodology

Movie Review Ontology

Semantic modeling and Ontology are the most common techniques for inferring and modeling contextual information from users' data. Expressing context values using ontology is advantageous because ontology can reveal various characteristics and properties of the context. Ontologies represent the systematically classification of items. The concepts used in the ontology proposed in this study are categorized and related to movie, genre, themes, and geographic region. The nouns are considered as objects and the verbs as object properties. These classes, objects and object properties' information are used to build the ontology model. The Stanford's Protégé software is used to build the ontology model. Class, object and object property are identified as entity, individual, and object property in the ontology model, respectively. The relations between classes, objects and object properties were derived manually as per the human understanding of a sentence. The ontology would use Web Ontology Language (OWL) for structural specifications and 42 movie review texts as the data source.

In an ontology, concepts are classes and subclasses of a domain, object properties represent relationships among various objects, and data properties represent attributes of the

objects. To build a movie ontology, first of all, all the entities included in movie reviews are identified (Figure 1), such as persons, places, and concepts, etc. The top class is “owl: Thing”. There are three subclasses: Concept, Movie Thing, and NewsStation. While instances of “Concept” are derived from tokens and phrases extraced from movie reviews, there is only one instance of “NewsStation” since all my movie reviews are downloaded from New York Times. Next, the subclass “Movie Thing” is broken down into five subclasses: Genre, Movie, Person, Place, and Review. All 42 movies fall under the subclass “Movie”. Each movie is assigned with 1 data property (“hasReleaseYear”) and 8 object properties: “hasReviewby”, “hasActor”, “happensinPlace”, “hasBiographicalReference”, “hasCharacter”, “hasConcept”, “hasDirector”, “hasGenre”. On the other hand, 6 categories are created for the “Person” subclass: Actor, Author, Character, Director, BiographicalReference, Role. Each instance of “Person” is assigned with one object property “hasRoleas”. In addition, each instance of “Person” (except those in the “Role” category representing occupations) is assigned with two data properties “hasName” and “hasGender”. Similarly, the “Place” subclass, which refers to the narrative place in each movie, is divided into “City” and “Country” categories. Meanwhile, 15 instances are created for the subclass “Genre”: Action, Adventure, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Political, Romance, Science Fiction, Thriller, War, Western.

In this study, the goal of the movie ontology is to extract four different item-based contextual features – story, direction, cast performance, visual effects, and respective sentiment polarity for all movies. All 42 movie reviews related to fifteen different genres from the movie review datasets would be manually analyzed to generate a list of seed words for mapping movies to their respective concepts. The table below presents a partial list of identified seed words corresponding to all four movie contextual features mentioned above.

Contextual Features	Seed words
Story	story, concept, plot, sub-plot, screenplay, script, storyline, ending, climax, portrayal, dialog, storytelling, writer, writing
Direction	directing, direct, directed, directs, direction, film maker, directional, directional debut, directorial, film making
Cast Performance	acting, performance, character, actor, support cast, acted, played, acts, portray, debutant, villain, performed, lead, actress, artist, role, hero, heroine, star, cast performances

Visual-Effects	visually, animatronic, visual effects, animation, CGI, visual, graphics, animation, animate, animated, digital effect, stunt

Table 1. *List of seed words corresponding to all eight interactional contextual features.*

Polarity Identification

SentiWordNet introduces a semi-automatic approach to derive a version of WordNet where word senses are bearing polarities. Table 2 shows the entry of 'unimaginative' in SentiWordNet. The first column indicates the word's POS tagger, "a" represents adjective. The numbers under "synset-offset" indicate the word's position in the dictionary. The third and fourth columns with polarity tags (pos, neg) represent the word's polarity strength (the value of maximal strength equals to 1). Column "sense" indicates word's polarity, 1 and 3 represent negative while 2 represents neutral. Since the entry does not give the neutral score directly, the following formula is used to obtain the neutral score:

$$score(obj)=1 - score(pos) - score(neg)$$

So the neutral scores of 'unimaginative' are 1.0, 0.375, 0.375 respectively.

POS	synset-offset	pos	neg	word	sense
a	1775641	0.0	0.0	unimaginative a	2
a	580160	0.125	0.5	unimaginative a	3
a	614153	0.0	0.625	unimaginative a	1

Table 2. *SentiWordNet: 'unimaginative'.*

3. Results

3.1 Overall Result

There are four of my movie reviews that are classified as positive reviews, whereas three of my movie reviews (YF_Doc3_Argo, YF_Doc4_Snowden, YF_Doc7_The_Fifth_Estate) are categorized as negative reviews. This experimental result seems pretty close to my manually labeled results because I also identified three movie review documents (YF_Doc2_The_Imitation_Game, YF_Doc4_Snowden, YF_Doc7_The_Fifth_Estate) that

contain negative sentiments based on the sentiment polarity scores associated with all four different item-based contextual features – story, direction, cast performance, visual effects.

3.2 Analysis and Interpretation

The inconsistency between the experiment results and the manually labeled results suggests that sentiment lexicon-based methods tend to put more weights on identifying negative words for predicting sentiment polarity regardless of contexts. Since *Argo* is also a political thriller, so the plots described in the movie review inevitably contains a lot of words that seem to indicate negative sentiments. Nevertheless, in fact, these words do not represent any negative opinion of the movie critics who evaluates the film. In contrast, although the movie *The Imitation Game* does not have dark and gloomy plots, but this movie's review reveals the movie critics' negative opinion of the film, as evidenced by the word "dull" that is used by the critics to describe the movie. Overall, however, the experimental accuracy (57.14% positive) indeed improved with the ontology incorporated compared to the prediction accuracy (100% positive) obtained without the ontology.

4. Conclusions

The experimental results imply that sentiment lexicon-based methods are domain specific and are typically based on bag-of-words models which ignore the semantic composition problem. Because ontology aims to provide knowledge about specific domains that are understandable by both developers and computers. The experiment is carried out effectively, and the result is good. Therefore, it is rational and effective to employ ontology to opinion mining.

For future works, the same experiments can be carried out on different datasets like Wikitology, Wikipedia, Word Net, Open Project Directory (OPD) etc. and other DNN-based classification methods which are data dependent and can learn high-level interactions among deep latent features. In-depth analysis and comparison can be holding on the diverse datasets and sentiment polarity scoring techniques.

Appendix

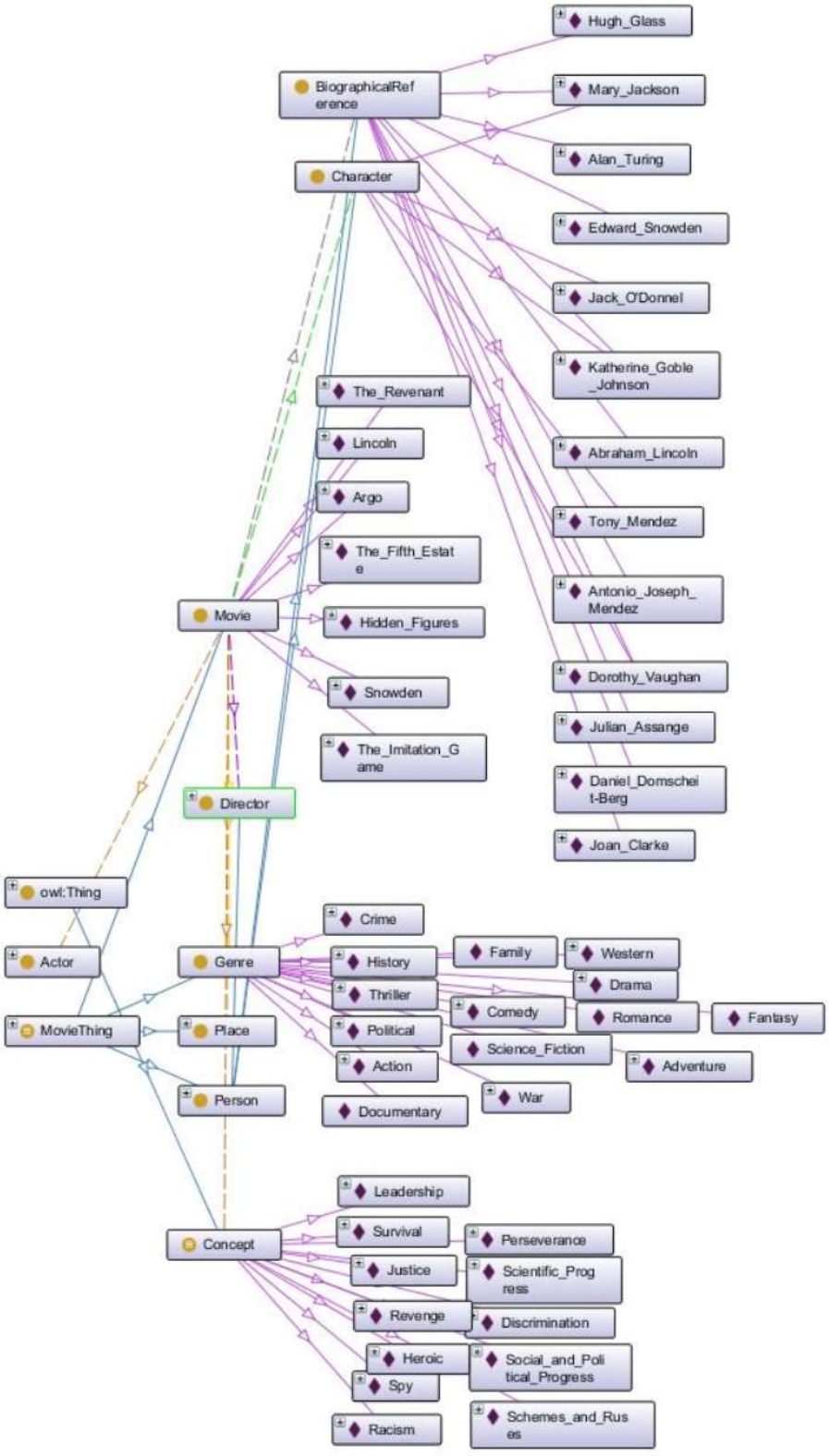
Hugh_Glass
BiographicalReference
Mary_Jackson
Alan_Turing
Character
Edward_Snowden
Jack_O'Donnel
Katherine_Goble_Johnson
The_Revenant
Lincoln
Abraham_Lincoln
Argo
Tony_Mendez
The_Fifth_Estate
Antonio_Joseph_Mendez
Movie
Hidden_Figures
Dorothy_Vaughan
Snowden
Julian_Assange
The_Imitation_Game
Daniel_Domscheit-Berg
Director
Joan_Clarke
owl:Thing
Crime
Actor
Genre
History
Family
Western
Drama
Thriller
Comedy
Romance
Fantasy
MovieThing
Place
Political
Science_Fiction
Action
Adventure
War
Person
Documentary
Leadership
Concept
Survival
Perseverance
Justice
Scientific_Progress
Revenge
Discrimination
Heroic
Social_and_Political_Progress
Spy
Racism
Schemes_and_Russes

Figure 1. *OWL ontology of movie entities*

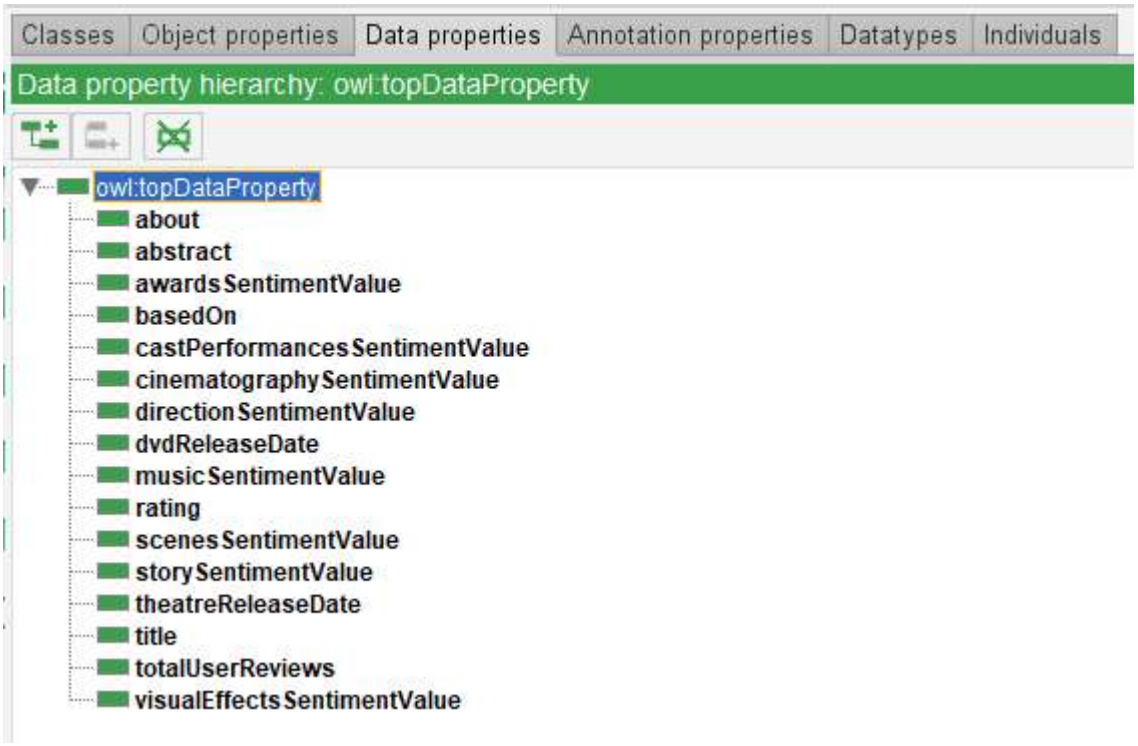

Figure 2. *OWL data property hierarchy of movie review ontology*

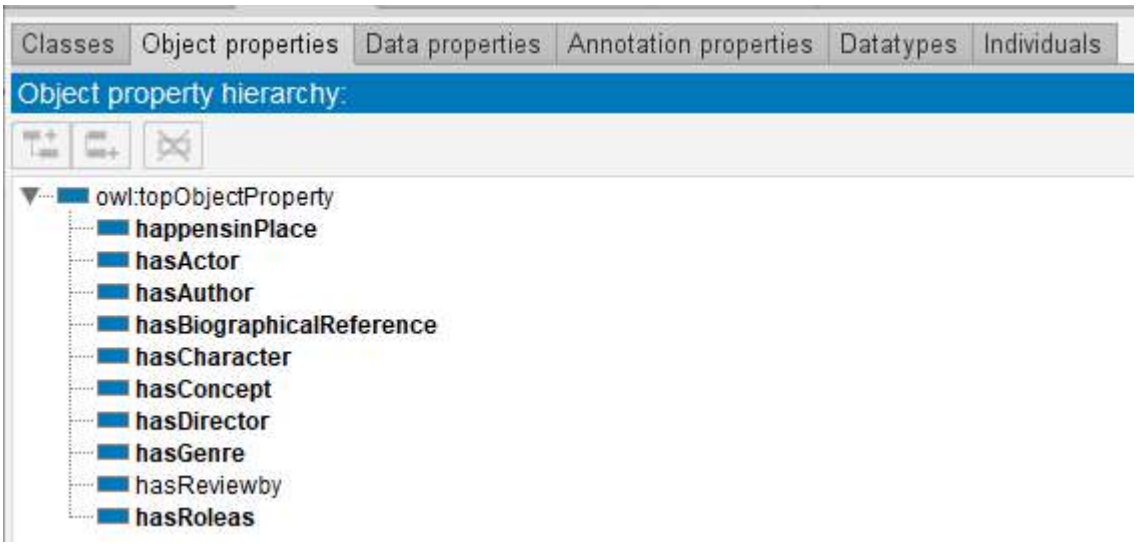

Figure 3. *OWL object property hierarchy of movie review ontology*

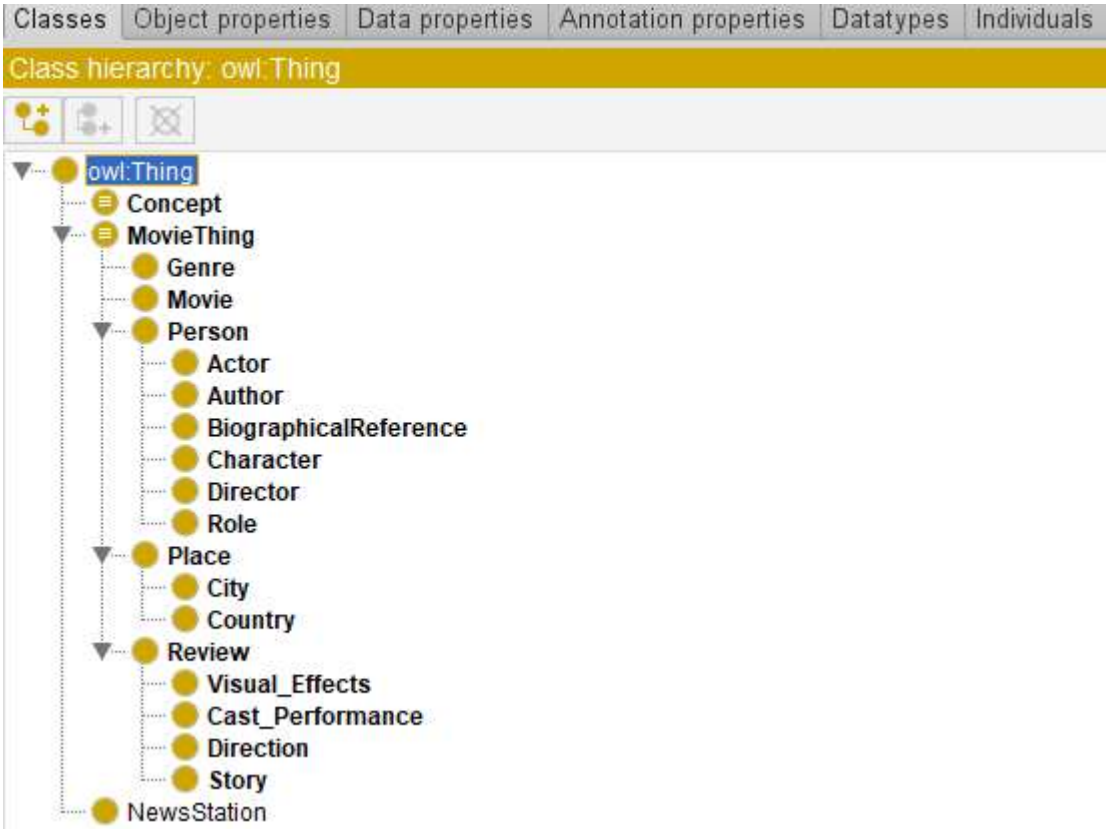

Figure 4. *OWL class hierarchy of movie review ontology*