Semantic Lexicon Expansion for Concept-based Aspect-aware Sentiment Analysis

Anni Coden, Dan Gruhl, Neal Lewis, Pablo N. Mendes, Meena Nagarajan, Michael Tannenblatt, Steve Welch*

IBM Research, USA

Abstract. We have developed a prototype for sentiment analysis that is able to identify aspects of an entity being reviewed, along with the sentiment polarity associated to those aspects. The prototype relies on a core ontology of the task, augmented by a workbench for bootstrapping, expanding and maintaining semantic assets that are useful for a number of text analytics tasks. The workbench has the ability to start from classes and instances defined in an ontology and expand their corresponding lexical realizations according to target corpora. In this paper we present results from applying the resulting semantic asset to enhance information extraction techniques for concept-level sentiment analysis.

Keywords: concept-based sentiment analysis, aspect-based sentiment analysis, lexicon induction, set expansion

1 Introduction

Detecting the sentiment expressed in text is a challenging task riddled by the inherent ambiguity and contextual nature of human languages. Consider, for a moment, what is the sentiment expressed by the sentence “I had a cold beer in a cold dining room.” Based on common knowledge (which can be location specific), beer is best enjoyed cold, which implies a positive sentiment. But is a cold dining room good or bad? This determination depends on the context of the sentence – e.g. on a very hot and humid summer day one may enjoy a cold room, however when coming into the house from shoveling snow, a warm room would be more desirable.

The above example illustrates that background knowledge and contextual information are important pieces in trying to solve the sentiment analysis puzzle. We propose a core ontology enriched by semantic lexicon expansion to tackle the most trivial sentiment analysis tasks, while alleviating more complex problems such as the aforementioned sentence. The domain model allows the association of concepts and *a priori* polarity information – such as ‘beer’ (a food concept) and ‘cold temperature’ (a temperature concept). Is a ‘cold’ glass of white wine good or should it be served at room temperature? In order to help discover concept

* Authors listed in alphabetical order by last name.

mentions in text for extending the ontology, we used a Semantic Asset Management Workbench to create and expand semantic lexicons. The workbench allows users to expand the ontology's coverage of concept and opinion mentions in text, easing and speeding up the creation of resources to aid in the interpretation of the same text through the eyes of different cultures and contexts.

We have expanded the semantic lexicons associated with our ontology by iterating over several online reviews, microblog posts, encyclopedic text, among other textual sources, and incorporated the results into our sentiment analysis prototype. We have tested the resulting system on the SemEval 2014 Task 4 dataset, as well as the Blitzer Dataset provided by the SemSA and obtained good preliminary results. In future work we plan to expand the evaluations to more datasets with different requirements concerning contextual and background knowledge.

This paper is organized as follows. Section 2 describes the core ontology developed and knowledge bases used. Section 3 describes the semantic lexicon expansion. Section 4 presents the sentiment analysis module. Finally, Section 5 discusses conclusions and future work.

2 Ontology and Knowledge Bases

We have designed an ontology to guide our concept-based aspect-aware sentiment analysis system, hereafter called 'our model' or 'core ontology'.

Our model aims at modeling online reviews – i.e. textual comments provided by a customer with opinions about some entity or aspect of that entity. Each `Review` contains potentially multiple sentences, and each sentence contains 0 to N item reviews (`ItemReview`) and associated opinions (`Opinion`). For example, one review could state that a customer likes the food but dislikes the service. Another might state that the customer likes one food item and dislikes another item in the same category. It is also possible that reviewers provide item reviews with both positive and negative opinions about the same item, in which case we consider that review item as having a polarity `conflict`. Moreover, when contextual knowledge is needed but not present, the system may classify the sentiment as `vague`. For the SemSA challenge, we only output positive and negative polarities.

Each `ReviewItem` refers to a mention of an RDF resource in a sentence – i.e. it represents a surface form or the `rdfs:label` of a resource appearing in a certain position in the textual content of a review. The model is able to include review items that are aspects of other items. Aspects include parts-of, containment, or other characteristics of items. For example, a review may target a shop's floorplan, and offer opinions about the outside seating space (a part of the shop's floorplan). An opinion may also be directed at the review target resource itself, in which case the aspect is the resource itself – e.g. 'the restaurant was great'.

The RDF resources included as instances of our model may come from any number of knowledge bases (KBs). In the current prototype, we have imported instances from DBpedia 3.9 [2], and lexicalizations from the DBpedia Lexicalizations Dataset [3]. We focused on instances relating to Books, DVDs, Electronics, Restaurants, and Kitchen&Housewares. We expanded the lexicalizations through our Semantic Asset Management Workbench (see Section 3). Besides identifying new lexicalizations for existing concepts, this expansion enables the system to detect items or aspects that are in a known category, but that do not have a URI in the imported knowledge bases. Consequently, the system may produce blank (skolemized) nodes when it cannot find a suitable URI in the current KB. This allows for an incremental approach to maintaining and evolving the core ontology used by the system, as new terms can be later added to the KB or new lexicalizations can be associated to their corresponding URIs[1].

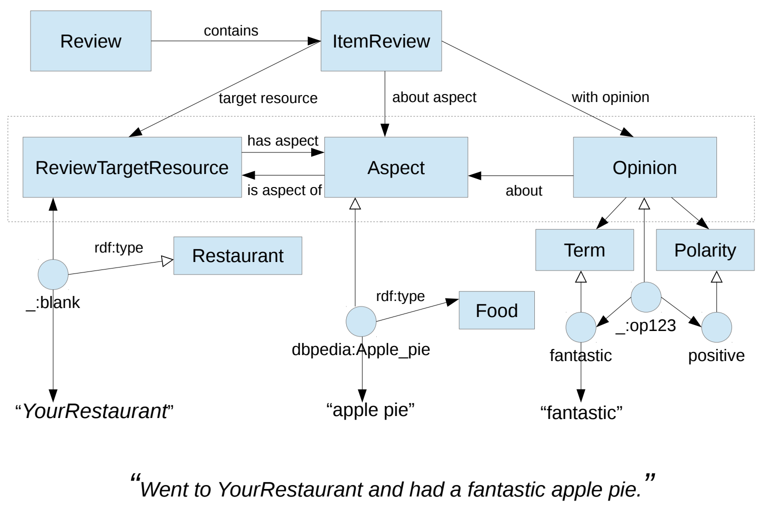

Fig. 1. Concept-based aspect-aware sentiment analysis ontology with examples.

3 Semantic Asset Management Workbench

We have developed a Semantic Asset Management Workbench (SAMW) that allows an analyst to draw on a number of techniques for developing, expanding and refining semantic lexicons. We define "semantic lexicon" as a set of phrases and their associated semantic types, as extracted from an ontology, or as serving the basis for creating one. Starting with a seed set of terms (usually anywhere between 3 and 30), the system finds all occurrences of these terms in a corpus and collects a set of patterns composed of the 0-6 tokens on the left and right of each occurrence of terms in the corpus. It then examines the corpus to find

[1] This process is currently handled manually, but algorithmic support is possible.

other words that match these patterns. The results are scored for confidence, support and prevalence. The users are then prompted to examine the top (up to 100) candidates and select which results to add to the lexicon.The system then iterates, taking these new terms, creating an even larger set of patterns and reprocessing the corpus to find more potential matches. Having the human in the loop helps to contain conceptual drift (e.g., is water a food?) and focus the lexicon on the concepts as necessary for the task at hand. Thus one key characteristic of SAMW is mutual discovery: it draws from user input to discover more terms, and provides output back to the users that prompts them to make new discoveries.

SAMW is a key piece to our solution and goes hand-in-hand with our core ontology, either bootstrapping new concepts or extending existing ones. We started by defining a set of semantic classes of interest and adding them to the core ontology, namely Books, DVDs, Electronics, Restaurants, and Kitchen&Housewares. For the types that existed in DBpedia, we bootstraped SAMW with entity names from DBpedia. Since the main objective in this particular task is to understand user opinions, we also included classes for positive, negative and neutral valence opinion terms. For those classes, we seeded SAMW with 3-5 manually created examples. For each semantic class, we can also define a set of aspect categories. For example, restaurants have aspects categories in ambience, food, price and service[2]. Additionally, valence lexicons were created, negative, positive and true neutral opinions in a food context (which is somewhat suggestive - e.g., is so-so really neutral?).

We ran 5 to 50 iterations per lexicon on a variety of 'open' and 'closed' corpora and acquired between 29 and 1126 terms per category. This let us find rarer terms such as 'sopaipillas' or 'mole sauce' in food, more esoteric opinion terms such as 'exquisite' or 'viable' for positives in food. We note that SAMW identified opinion terms that have the potential to differ by domain. For example, you wouldn't say a food is very compact or blazing fast, nor would you say a laptop is 'flavorful' or 'intimate'. Valence varies by domain too, a 'small' camera is usually a positive opinion while a small car might not be, and SAMW is able to make such distinctions.

4 Sentiment Analysis Component

We have developed a sentiment analysis component that extracts sentiment at the item level (e.g. 'MyRestaurant'), at the aspect level (e.g. 'MyRestaurant's rice'), at the item category level (e.g. `Food` and `Restaurant`), or at the review level – i.e. aggregating opinions of multiple items into a final assessment of the overall sentiment in the review. It computes the sentiment of a sentence based on the sentiments of the concepts expressed within a that sentence. Inference across multiple sentences is planned for our future work.

Therefore, the prototype is able to perform SemSA's Elementary Task (Polarity Detection), Advanced Task #1 (Aspect-Based Sentiment Analysis), and Advanced Task

[2] We used the same categories as the SemEval Task 4.

#3 (Topic Spotting). We plan to tentatively participate on Advanced Task #2 (Semantic Parsing).

In our prototype, each sentence is processed to produce constituency and dependency parses using OpenNLP[3] and ClearNLP[4] [1]. In addition, we use the aforementioned semantic lexicons from our core ontology, therefore considering concepts under the following categories:

Aspects and ReviewTarget Resources (`AR`) – e.g. beer, wine, dining room;

Positive Opinion Terms (`Pos`) express in general a positive sentiment – e.g. like, good, happy;

Negative Opinion Terms (`Neg`) express in general a negative sentiment – e.g. death, bad, unhappy;

Polarity Inversion Terms (`Inv`) used to invert the polarity of a sentiment – e.g. not, cannot, will not, but, however;

Association concepts `AC(concept, opinion, sentiment)` describing the prior polarity for an opinion term given a concept, where `concept` and `opinion` are instances in one of the above lexicons – e.g. `(beer, cold, positive)`. Clearly "negative concepts" can be used in a positive sense; for instance, the phrase "death by chocolate" is used to refer to very rich chocolate desserts delighting many people. Our model is able to capture these cases through the association concepts.

Our algorithm performs the following steps:

1. Extract the concepts $\{c_i, ..., c_n\}$ and opinion terms $\{o_i, ..., o_m\}$ discussed in each sentence based on our semantic lexicons `AR`, `Pos`, and `Neg`;
2. Identify the syntactical association between concepts based on the parse of the sentence. For instance, two concepts are associated if they are the same noun phrase chunk or are connected by the subject-object-verb relation.
3. Query our knowledge base for semantic/sentiment (`AC`) associations. If there is an sentiment from the `AC` lexicon overwrites the sentiment associated with any concepts;
4. Special processing is done to identify lists, parenthesized expressions and hyphened expressions. Example: 'I like beer, wine and vodka cold'. The sentiment specified in `AC` between 'beer' and 'cold' carries over to the other constituents of the list – 'wine' and 'vodka'.
5. Polarity inversion: a. Identify the concepts specified in `Inv`; b. Identify the part of the sentence the polarity inversion applies using syntactic parsing constructs and rules; examples of rules are: determining lists, conjunctions and excluding parenthesized expressions.

5 Conclusion

We have presented a prototype for concept-based aspect-aware sentiment analysis. Our system relies on a core ontology of the task that allows us to model reviews based on the resources that they target, aspects of these resources as well as opinion terms related to these aspects or target resources. The ontology allows the definition of *a priori* concept-based opinion polarity to account for differences in expected polarity when one says 'cold beer' (positive) versus 'cold room' (negative). In order to expand the

[3] http://opennlp.apache.org
[4] http://clearnlp.com

lexical forms in our ontology, we employed a Semantic Asset Management Workbench that empowers users to discover new terms and learns from the discoveries to improve its discovery process. This workbench allowed us to acquire new terms, name variations, as well as specialized opinion terms to particular categories that do not make sense for other categories (e.g. 'flavorful' for food and 'blazing fast' for a laptop). In future work we plan to keep evolving the core ontology and investigate the best vocabularies to reuse for its encoding. Moreover, we plan to expand testing to datasets with different demands with regard to the requirements for background knowledge and contextual information.

References

1. Jinho D. Choi. *Optimization of Natural Language Processing Components for Robustness and Scalability.* PhD thesis, Boulder, CO, USA, 2012. AAI3549172.
2. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, and Sören Auer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2014.
3. Pablo Mendes, Max Jakob, and Christian Bizer. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).