Wordnet and SUMO for Sentiment Analysis

Adam Pease, John Li, Karen Nomorosa
Rearden Commerce
Foster City, CA, USA
[adam.pease | john.li | karen.nomorosa]@reardencommerce.com

Abstract

In this paper, we show how Princeton's WordNet and associated resources can be used as part of an integrated system for sentiment analysis, called SigmaSentiment. We discuss the development of a system for sentiment analysis and concept extraction. We first provide an introduction to the user experience to motivate our work. We begin technical discussion with some background on the Suggested Upper Merged Ontology and WordNet, as well as their relation and the associated research that makes this work possible. We describe system components and data flow in detail, and some detail about the deployment architecture in the fielded system.

1 Introduction

Sentiment analysis refers to the assessment of the emotional content of text. Our goal is to support a more personalized travel experience in searching for a hotel. In this pilot project we combined techniques in computational linguistics with concept extraction with respect to an ontology, as well as some initial numerical analysis of the resulting statistics. We should note up front that this is just a pilot project and the computational linguistic method used is really basic, not state of the art. However, even with the simple methods applied, the relationship to formal ontology is relatively novel, and the results appear to be sufficient to provide practical utility in support of a travel application.

Most current travel applications, as opposed to web search, have a limited search structure for the features of hotels. Most search on price and location, as well as a few categories defined by the search provider. A few also expose “amenities” that are self-reported according to standard industry lists. But hotel features that travelers care about can be almost anything stated in natural language. Fine grained search by features would be advantageous. But language is so flexible, some method is also needed to standardize concepts, so that sentiment expressed on the same features can be combined across a large number of reviews.

There are many publications that rate hotels. Each has its own rating approach and scale. If sentiment can be extracted from reviews, those scales can be normalized according to the source data, rather than the reported summary scores.

Sentiment that is linked to concept extraction has the potential to provide a much finer-grained assessment of hotel quality, with respect to the features that each traveler cares about. Not all travelers are the same in their concerns and preferences. More information about the quality of hotel features should considerably improve customer satisfaction.

2 Background: Wordnets

Since Princeton's WordNet (PWN) is well-known, it may be sufficient simply to refer the reader to (Fellbaum, 1998). For the purposes of this paper, it bears mentioning that there are several features of WordNet that make it an essential product to link to.

- PWN is a mature product, having been started over two decades ago (Miller, 1985)
- It is very comprehensive, with over 115,000 word senses, making it the largest wordnet in existence
- It has been free since the project's inception
- It is richly interconnected as a semantic network
- Many other languages have linked their wordnet projects to it manually

3 Background: Suggested Upper Merged Ontology

We had previously mapped all of PWN to a formal ontology (Niles & Pease, 2003), the Suggested Upper Merged Ontology (Niles & Pease, 2001).

Synsets map to a general SUMO term or a term that is directly equivalent to the given synset (Figure 1). New formal terms created for any particular domain will be defined to cover a greater number of equivalence mappings, and the definitions of the new terms will in turn depend upon existing fundamental concepts in SUMO. The process of formalizing definitions will generate feedback as to whether word senses in WN need to be divided or combined and how the glosses may be clarified. Since many wordnets in other languages are already linked by synset number, this work benefits wordnets in other languages as well.

The Suggested Upper Merged Ontology (SUMO) (Pease, 2011), (Pease&Niles, 2002), (Niles&Pease, 2001) is a freely available, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format (SUO-KIF) (Pease, 2000), and also has a necessarily lossy translation into the OWL semantic web language. It has undergone nine years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. SUMO has been extended with a number of domain ontologies, which are also public, that together number some 20,000 terms and 80,000 axioms. SUMO has been mapped by hand to the WN lexicon of over 115,000 noun, verb, adjective and adverb senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list. SUMO is the only formal ontology that has been mapped to all of WN, and the only formal upper ontology that has been extended with a number of domain ontologies that are also open source. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUMO-KIF and SUMO to be expressed in multiple natural languages. These include English, German, Arabic, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin).

4 Project Description

We'll first describe the process and then describe the technical details of how it works.

Take for example the following (slightly fictionalized) hotel reviews

> Meadowland Resort, Vineland, CA
>
> "*In recent years the elegant but unstuffy dining room has won rave reviews, becoming a destination restaurant.*"
>
> Crystal Lake Lodge and Resort, CO
>
> "*Not to mention it is very expensive and located in a place that doesn't get much sun so it's icy and cold; and the maintenance of roads is terrible in winter.*"

The first review is very positive. We extract SUMO concepts from the sentence, and associate them with a positive score. In this case, we assert that the concept of Restaurant has a sentiment of +10. The second review is quite negative. We assert that the concept of Roadway has a -8 sentiment.

We gather reviews for all the hotels we can, extract SUMO concepts and associate them with sentiment scores. We then total the scores for each concept with regard to a particular hotel to create a matrix of hotels and the total sentiment for every concept associated with each hotel.

	Restaurant	Breakfast	Walking	Bed	Fireplace	City
Hotel 1		1		5		
Hotel 2	4					
Hotel 3						
Hotel 4	10	10				
Hotel 5	6	6				
Hotel 6			15		11	5
Hotel 7	1					
Hotel 8	-3					-23

In this section of the resulting matrix with see that Hotel 4 has positive sentiment associate with its breakfast and restaurant. In contrast, Hotel 8 has a very negative sentiment associated with the city (or likely the section of the city) in which it is located.

An open question is how to normalize the sentiment scores. Currently, they are just totaled, which gives greater weight to those cases where there are a large number of reviews. But reviews typically do not mention all the same concepts. One can make the case the frequent mention does legitimately make a statement about the strength of sentiment, since neutral concepts are not likely to be mentioned in reviews.

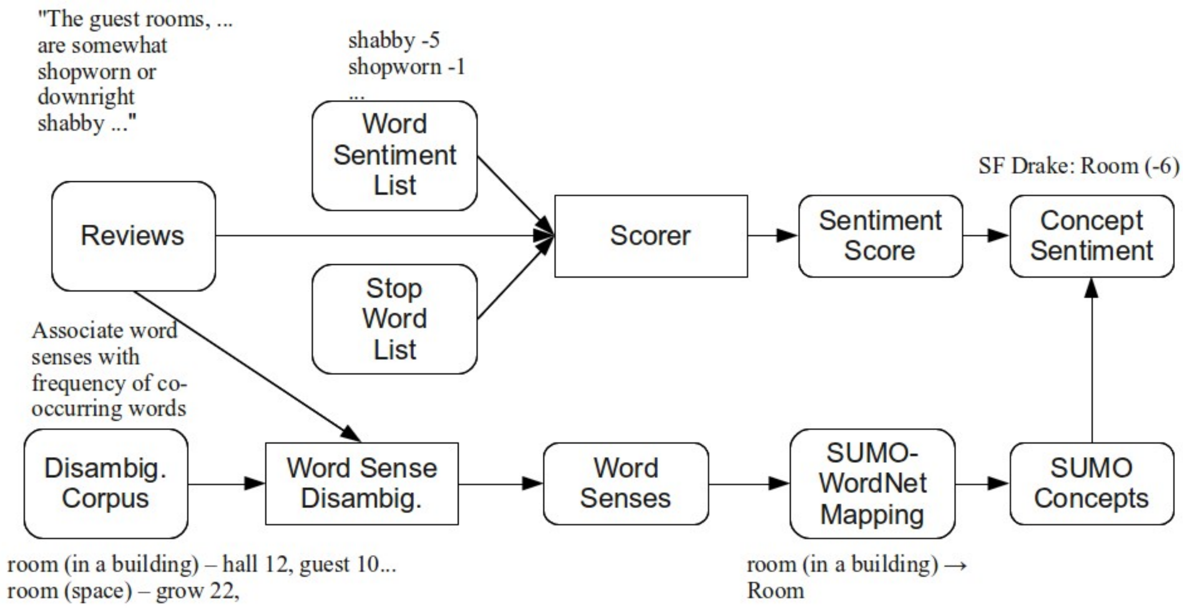

Figure 1: SigmaSentiment data flow

5 Architecture and Data Flow

We now describe all the different elements of the approach as diagrammed in Figure 1. The first step when processing reviews is to determine the sense of each word with respect to WordNet. To accomplish this, we employ WordNet SemCor – a manual markup of word senses in the Brown Corpus. We created a matrix of statistics that associates each word sense with the non-disambiguated words which it co-occurs with. We can then process each sentence, looking at each polysemous word and all other words in the sentence. The sense in SemCor that has the most words in common with the given sentence is the one that “wins” and is selected. This is not a particularly advanced method, and a much larger corpus would be desirable to get better statistical significance, but it is the best that we have found that is also available open source.

For example, consider the word “bed” in the context of the sentence “The bed was so comfortable I had a great night's sleep.” An amended list of SemCor's associations between sense and words (also omitting word counts) shows:

Bed sense	**Co-occuring words**
1	air_mattress curtain sleep sleeping_bag slipper
2	compost decayed manure pansy spade spread_over yard
3	dry face homely river tilt

“sleep” is the only word in this simple example that appears in the word lists, and so the highest associated score is with sense 1.

We improve the statistical significance of our word sense discrimination by combining those WordNet senses that map to the same SUMO term. While, SUMO is large, it is not as large as WordNet, and many WordNet senses map to the same SUMO concept. In addition, there has been some discussion that WordNet's sense may present distinctions that are arguably too fine, and so may not reliably be discriminated in actual usage.

Once we have determined the WordNet sense, then we have a mapping to the appropriate SUMO term.

In parallel with determination of the SUMO terms, we need to calculate a sentiment score for each sentence. The OpinionFinder team has released a file of WordNet senses that have been manually marked with a sentiment score (Wiebe&Mihalcea 2006). Each sense is marked positive or negative, and whether that sentiment is weak or strong.

type	**word**	**POS**	**polarity**
weak	abandon	verb	negative
weak	abate	verb	negative
weak	abdicate	verb	negative
strong	aberration	adj	negative
weak	able	adj	positive

We arbitrarily assign the value of +/-5 to strong sentiment and +/-1 to weak sentiment.

Next we tag every SUMO concept extracted from each sentence with the total sentiment score for that sentence. All the scores are totaled for all reviews for a given hotel to arrive at a total sentiment for each SUMO concept associated with each hotel.

6 Ontology and Lexicon Development

Since April of 2011, we have been expanding SUMO and the SUMO-WordNet mappings to cover topics in travel and tourism in much greater detail. Significant, new ontologies of Dining, Food and TransportationDetail have been created. Many other ontologies are under development and are available as "beta" ontologies that extend SUMO. These include, Biography, Catalog, Contract, LoyaltyProgram, Pricing and TravelPolicy. Each ontology may have hundreds of terms and formal axioms. Over the same period there have been hundreds of revisions and corrections to existing SUMO ontologies as well. Roughly 2300 SUMO-WordNet mappings have been changed to map to the new more specific concepts now available in SUMO.

Using SUMO terms as the structure to which we attach sentiment scores has several attributes. SUMO is a consistent logical theory, so we are guaranteed that if we follow transitive links, such as subclass, that the results are still inferentially valid. Using a language-independent formal ontology also supports future extensions such as presenting sentiment results in any target language supported by the SUMO language mappings, regardless of the fact that reviews may be processed from English. Lastly, the current work is just the beginning. We plan on moving from concept extraction to statement extraction, and then associating sentiment with entire logical statements, rather than just a mention of a particular concept.

7 Evaluation

We tested our algorithm against a standard test corpus for sentiment analysis (Pang et al 2002). This corpus consisted of roughly 10,000 sentences taken from movie reviews from RottenTomatoes.com using the "fresh" and "rotten" scores supplied by the reviewers as a proxy for a positive or negative sentiment ground truth rating.

We also compared our simple algorithm to OpinionFinder (Wiebe et al 2005), which is a system that assesses subjectivity in text. Because it does not have the same simple goal of just assessing sentiment, comparison is somewhat unfair, but it did provide a useful baseline for evaluation.

OpinionFinder does not rank polarity at the level of sentences. We used OpinionFinder's polarity marks on individual words in the text. We used OpinionFinder as a tool but the final scores at the sentence level are the results of our approach. We sum up the scores that OpinionFinder provides at the word level to give an overall assessment of sentiment at the sentence level.

Pang et al have shown results of up to 80% correctness for sentiment scoring on machine-learning based systems.

% correct	Opinion Finder	Sigma Sentiment
+ reviews	15.04%	64.15%
- reviews	50.78%	49.39%

Figure 2 shows how well SigmaSentiment does with respect to the test corpus of positive reviews, with the scores broken out by sentiment score. It shows the difficulty of determining sentiment when a sentence is mild or ambiguous, but that strong sentiment is relatively easy to determine correctly.

We should note that this problem is quite difficult for humans also. Take for example the following movie review excerpt.

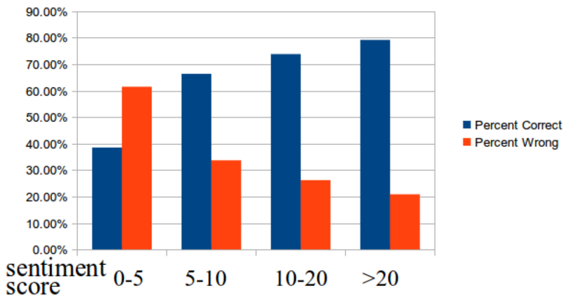

Figure 2: SigmaSentiment incorrect and correct scores on positive reviews

> *"A disturbing and frighteningly evocative assembly of imagery and hypnotic music composed by Philip Glass."*

Is this review positive or negative? It really depends on whether one finds movies that can be described as "disturbing" to be thought-provoking and therefore interesting and enjoyable, or just unpleasant.

In fact, human raters typically agree about 80% of the time (Wilson et al 2005), where the task was slightly easier in that it allowed raters to agree on "neutral" ratings that correspond to the most difficult cases where sentiment is not strongly present. As a result, we believe the 80% score should be compared to the portion of our tests performed on sentiment scores above 5 (see Figure 2). A 80% accurate program is doing as well as humans, and our scores of 65%-79% on sentences with strong sentiment may be considered encouraging.

We have also been collecting example sentences from our domain of hotel reviews to calibrate the results we can expect. We place them into three categories. The category of "regular difficulty" contains those sentences with clear sentiment resulting from adjectives. For example,

> *"The paint of the room is ugly."*
> *"We had a pleasant stay there."*

The second category is "medium difficulty", which includes sentences with contrasts or those where some implication or inference is needed. For example,

> *"The faucet was leaking and making noise whole night."*
> *"The bathroom is gorgeous but the shower doesn't work properly."*

In the first sentence it is possible that "leaking" and "noise" could be scored as negative in isolation, although it would be better to know that noise is bad in the context of sleep, which is a primary purpose of a hotel stay. In the second case, some level of analysis is needed to separate the first positive part of the sentence from the second, negative part.

The last category is "most difficult, may need human intelligence." For example,

> *"The warm welcome atmosphere disappeared right after I checked into the room."*
> *"I found that the mattress is no younger than my age although we're told the rooms were completely renovated last year."*

In these cases, one needs to apply significant common sense knowledge to understand the sentence.

8 Conclusions and Future Work

Language is very flexible, and there are many ways for the algorithm to make the wrong guess at the sentiment associated with a given concept. It appears that given the large volume of reviews we have available, and given that the algorithm gets the sentiment right in a majority of the cases with unambiguous sentiment, that the errors are overwhelmed. More work is needed to see if there are a statistically significant set cases of errors that can be reliability corrected.

One common error case is in sentences that are divided into positive and negative sentiment, such as the pattern "I liked the bed, but didn't like the cleanliness of the bathroom." If we apply the Stanford parser (Klein & Manning 2003) to this sentence, we get the following parse tree:

```
(ROOT
  (S
    (NP (PRP I))
    (VP
      (VP (VBD liked)
        (NP (DT the) (NN bed)))
      (, ,)
      (CC but)
      (VP (VBD did) (RB n't)
        (VP (VB like)
          (NP
            (NP (DT the) (NN cleanliness))
            (PP (IN of)
              (NP (DT the) (NN bathroom)))))))
    (. .)))
```

We then assign sentiment

```
(ROOT
  (S
    (NP (PRP I))
    (VP
      (VP (VBD [liked +5])
        (NP (DT the) (NN bed)))
      (, ,)
      (CC but)
      (VP (VBD did) (RB n't)
        (VP (VB [like +5])
          (NP
            (NP (DT the) (NN [cleanliness +1]))
            (PP (IN of)
              (NP (DT the) (NN bathroom)))))))
    (. .)))
```

And flip the polarity of the sentiment under a negation in the parse tree

```
(ROOT
 (S
  (NP (PRP I))
  (VP
   (VP (VBD [liked +5])
    (NP (DT the) (NN bed)))
   (, ,)
   (CC but)
   (VP (VBD did) (RB n't)
    (VP (VB [like -5])
     (NP
      (NP (DT the) (NN [cleanliness -1]))
      (PP (IN of)
       (NP (DT the) (NN bathroom)))))))
  (. .)))
```

There are of course many exceptions where this approach does not work. However, all we need to is improve the number of correct interpretations. We will be testing against the movie review corpus to see whether this is the case.

Another area of effort is in capturing user preferences. We cannot expect busy users to go through long lists of concepts, specifying that this or that concept is of particular interest or concern. The existing Rearden Commerce hotel search has the capability of the user to select some concepts such as "best for business" or "good for families" in order to influence the ranking of hotels. The ranking is currently done on the basis of amenities that hotels have self-reported, and a matrix in which company developers have made a judgement that particular amenities are relevant to those categories. While the current approach allows only for the presence or absence of those amenities, sentiment analysis allows us to rate those amenities, so for example, even if a pool is desired by families, a bad or dirty pool should not improve the ranking of a hotel in that category.

Another possibility is to ask users to write a short description of their ideal hotel, and use the same concept extraction process used in the reviews, and then match the preferences of the ideal hotel with those hotels that have positive sentiment for those items.

Acknowledgments

We would like to thank Rearden Commerce for supporting this work, and the OpinionFinder team at U. Pittsburgh for making their work available open source.

References

Fellbaum, C., (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Klein, D., and Manning, C., (2003). Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Miller, G., (1985) "WordNet: a dictionary browser." In Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo.

Morato, J., Marzal, M.A., Llorens, J., & Moreiro, J (2004). WordNet Applications. In Proceedings of the Second Global WordNet Conference (GWC-2004). Brno, Czech Republic.

Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.

Niles, I., and Pease, A. (2001). Towards a Standard Upper Ontology. In: Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9. See also http://www.ontologyportal.org

Pang, B., Lee, L., and Vaithyanathan, S., (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.

Pease, A., (2003). The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.

Pease, A., (2011). Ontology: A Practical Guide. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.

Wilson, T., Wiebe, J., and Hoffman, P., (2005). "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,"

Wiebe, J., Wilson, T., and Cardie, C., (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.

Wiebe, J., Mihalcea, R., (2006). Word Sense and Subjectivity. Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. (COLING-ACL 2006).