Room For Improvement: Aspect-Specific Statistical Opinion Mining Of Online Hotel Reviews

Lynd Bacon[1,2]
[1]Loma Buena Associates, 1418 Panorama Drive, Vestavia, AL 35216
[2]Northwestern University School of Professional Studies,
405 Church Street, Evanston, Illinois 60208

Abstract

The availability of publicly accessible product review websites is producing a large and rapidly growing volume of customer feedback data that can be used to improve customer experience and marketing performance. Using this data is not without challenges. It is complex and poorly structured. Much of the potentially important information is in text generated by posters. It is in this language data that the details of customers' opinions about the “aspects,” or characteristics, of their experiences are expressed. We are conducting a research project involving the development and testing of methods for extracting aspect-specific sentiments. We are using over 1.6 million reviews of 12,000 hotel properties. We summarize our application of statistical natural language processing and sentiment lexicon methods, with particular focus on conditional random field models for part of speech tagging and named entity recognition.

1. Introduction

Thanks to the large and growing availability of user-generated content that's online, there are ever growing opportunities for organizations to understand how to create more value for customers. The challenge in doing so is two-fold: accommodating the large amount of data flowing onto and through online properties, and exploiting as effectively as possible the poorly structured data.

Online customer reviews are a widely available and important kind of user-generated content that can be used to understand customers' self reported product and service experiences. Sentiment analysis of various kinds is frequently performed using the language data provided by reviewers. Examples of the very many web properties on which customers offer reviews include amazon.com, TripAdvisor.com, and Edmunds.com.

Analyzing language data to extract meaning is usually a very challenging task. The requisite tasks of natural language processing (NLP) almost always include end of sentence detection (EOS) and part of speech (POS) tagging. A fuller understanding of what text is meant by its creator to communicate will include taking into account slang, sarcasm, negation, coreferencing, and colloquialisms. There are the NLP “evil twins,” polysemy and semantic ambiguity to take into account, as well. For example, consider the following headline:

“McDonald's Fries the Holy Grail for Potato Farmers”

Human natural language users with knowledge about the English language, Western culture, and American (so-called) cuisine will have no difficulty arriving at the correct interpretation. A machine would have much greater difficulty distinguishing the alternative meanings of the word "Fries" without having recourse to extensive context knowledge.

Our team is exploring new methods for extracting from the text of online consumer-generated reviews the dimensions of product experience they mention, and the sentiment they indicate in regard to these "aspects" of their experience. In what follows we summarize our progess to date in applying conditional random field models, a type of statistical sequence learner, to a large collection of customer reviews for the purpose of POS tagging.

2. Methods

2.1 Data

We are using a corpus of approximately 1.6 million reviews of hotel stays obtained from the TripAdvisor.com website. The data were originally downloaded by Wang et al. (2010, 2011). Wang et al. have used this and other data to develop latent models of product evaluation (Ibid.; cf https://www.cs.virginia.edu/~hw5x/dataset.html). They include customer reviews of roughly 12,000 hotel properties. The metadata include hotel names, locations, dates. Other data include customer ratings on various dimensions, like cleanliness, location, and sleep quality. The reviews span dates from 04/13/2001 to 09/15/2012 inclusive.

Here's an example of one reviewer's text comment:

"I stayed here overnight after attending the U2 concert recently. It was really bad! They didn't even have an alarm clock in the room. It's way overpriced. About the only thing good is that the location is quite convenient."

Some of the reviews contain emoticons, and many that include punctuation marks that can complicate end of sentence detection. We needed to identify them as word-like "tokens" in order to treat them as linguistic elements. Rather than develop our own emoticon dictionary, we employed the dictionary used by Hutto and Gilbert (2014) in their implementation of their VADER system for sentiment analysis (Ibid.).

2.2 Data Processing and Analysis Tools

We used tools and methods implemented in the Python 3 language for data manipulation and analysis. We relied heavily on the Natural Language Toolkit (Bird et al., 2009; cf http://www.nltk.org). Other tools employed included the Pandas and TextBlob Python packages. We managed the data using both Python tools and the MongoDB database (www.mongodb.com), a noSQL store, because of its facility for storing un- or poorly structured data.

2.3 Preprocessing the Data for POS tagging

We used the VADER emoticons to detect and replace emoticons that didn't include any alphabetic characters with alphanumeric tags that would could use later that we could use to refer back to their original form. After replacing these emoticons, and after preceeding each single likely sentence-ending punctuation mark like a period, exclamation mark, or a question mark, we counted 1,332,895 distinct "tokens" in our collection of review documents. The most prevalent included the period punctuation mark, and common "stop" words like "and," "a," and "to." Out of all the tokens, 4,924 appeared in the VADER dictionary of 7,517 tokens. Amongst the most prevalent were adjectives like "great," "good", and "nice." The prevalent non-alpha emoticons included "smiley" and "frowny" faces like ":-)" and ":-9".

2.4 Conditional Random Field (CRF) Modeling

There are two basic approaches to POS tagging, rule-based and "learner"-based. We've applied a learner-based method by using conditional random field (CRF) models. CRF models are used for labelling network nodes. Sequences of tokens in text can be represented as a special kind of graph in which the words are nodes in a sequence.

As originally described by Lafferty, McCallum and Pereira (2001), CRFs have a convex loss function, and their parameters can be estimated using various methods. Lafferty et al. (2001) estimated their CRFs using iterative scaling to get maximum likelihood methods. Since then, other methods, including MAP, the Broyden/Fletcher/Goldfarb/Shanno (BFGS) algorithm, various stochastic gradient descent, boosting, and Bayesian methods. have been used. Wallach(2004) provides a basic introduction to CRFs.

A main advantage of CRFs as compared to sequence learners like Hidden Markov Models (HMMs) is that they estimate the joint probability distribution of possible POS tags conditional on the observed tokens of text. This provides a generally easier optimization problem to solve compared to the problem of estimating the joint probability of both POS tags and tokens. Another advantage is that CRFs are not prone to the "label bias" problem like HMMs are (Lafferty et al., 2001; Qi et al, 2005; Wallach, 2004.) Label bias consists of a sequence learner ignoring information in a graph were nodes in sequences representing alternative tags have only one exiting edge.

As originally specified by Lafferty et al. (2001), a CRF consists of the probability of a label sequence conditional on tokens as being an exponential function of transition and state features:

$$p_\theta(y|x) \propto \exp\Big(\sum_{\epsilon\in E,k} \lambda_k f_k(\epsilon, y|_\epsilon, x) + \sum_{\nu\in V,k} \mu_k g_k(\nu, y|_\nu, x)\Big)$$

Here, $p_\theta(.)$ is the probability distribution of POS tags, y, conditional on the observed sequence of tokens, x. $f_k(.)$ and $g_k(.)$ are feature functions that describe alternative possible transitions between tags, and tag states, respectively. λ_k and μ_k are elements of θ, a vector of parameters to be estimated.

Using a learner-based approach for POS tagging requires a tagged "training" collection of text data. We haven't (yet) developed a tagged training data set for this collection of reviews, which would (will) be a very labor-intensive task. Before doing so we thought we'd try training CRF models for POS tagging on a existing, available tagged data set. We selected the Penn Treebank for experimentation. This tagged corpus is frequently used by computational linguists for experimenting with new algorithms and for benchmarking.

We selected an implementation of a CRF tagger in the CRF-Suite (http://www.chokkan.org/software/crfsuite/) that the Natural Language Toolkit provides an API for. We trained and testing our model on the Penn Tree bank. We then used it to tag 600 randomly selected, preprocessed reviews. Of these, we randomly selected 200 to hand verify and to calculate accuracy measures for.

3. Preliminary Results and Conclusions

The results of our initial experiment were as follows. Overall accuracy of tagging was 0.643, a not-stunning level of accuracy, but surprisingly good given that we trained our CRF on a greatly different tagged benchmark corpus.

We calculated precision and recall by aggregating over the noun tag categories and the verb tag categories. For nouns, precision was 0.720, and recall 0.741. F1 was 0.730. For the verbs, 0.683, 0.692, and 00.687 for precision, recall, and F1, respectively.

Although these results are not strong by comparison to the performance of POS tagging sequence learners trained and tested on the same corpus, we think they suggest the notion that tagging for new data collections can be at least adequately accomplished by initially training one or more sequence learners on a different, tagged, corpus, combining their results using an ensemble sort of procedure, perhaps supported by selected active learning participation by some human natural language “experts.” Given the rate at which the nature of casual, online language seems to be evolving (consider, for example, the rapidly growing popularity of emojis), an adaptive and scalable procedure that utilized complementary methods could be quite useful for evolving tagged data collections for research.

In terms of the project at hand, the next step after POS tagging and sentence detection will be the identification of reviewers' experience aspects, the dimensions of their hotel stays that they refer to. We don't seen this as a sequential learning task. Instead, it should lend itself to topic modeling using latent Dirichlet Application models (cf. Alghambi & Alfalqi, 2005), or by topic modeling Lancichinetti et al. (2015).

References

Alghamdi, R. & Alfalqi, K. (2015) “A survey of topic modeling in Text Mining,” Intl. Journal of Advanced Computer Science and Applications, 6(1),147-153.

Bird, Steven, Ewan Klein, and Edward Loper (2009), Natural Language Processing with Python, O'Reilly Media. (See also www.nltk.org, https://github.com/nltk)

Hongning Wang, Yue Lu and ChengXiang Zhai. Latent Aspect Rating Analysis without Aspect Keyword Supervision. The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2011), P618-626, 2011.

Hongning Wang, Yue Lu and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2010), p783-792, 2010.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Lancichinetti, A., Sirer, M., Wang, J., Acuna, D. Kording, K. & Amaral, L. (2015) High-Reproducibility and High-Accuracy Method for Automated Topic Classification," Physical Review X 5, 011007, 1-11.

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, pp. 282–289.

Qi, Y. (Alan), Szummer, M., & Minka, T. (2005) Bayesian Conditional Random Fields, AI & Statistics, 269-276. https://www.cs.purdue.edu/homes/alanqi/papers/Qi-Bayesian-CRF-AIstat05.pdfs

Wallach, H. (2004) Conditional Random Fields: An Introduction. University of Pennsylvania CIS Technical Report MS-CIS-04-21.