View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Waseda University Repository

# Chinese Sentiments on the Clouds:

## A Preliminary Experiment on Corpus Processing and Exploration on Cloud Service

**Shu-Kai Hsieh, Yu-Yun Chang, Meng-Xian Shih**
Graduate Institute of Linguistics, National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
shukaihsieh@ntu.edu.tw, {june06029;simon.xian}@gmail.com

## Abstract

This study aims to propose a novel pipeline architecture in building and analyzing large-scaled linguistic data on the cloud-based environment, an experimental survey on Chinese Polarity Lexicon will be taken as an example. In this experiment, data are evaluated and tagged by applying crowd sourcing approach using online Google Form. All the data processing and analyzing procedures are completed on-the-fly with free cloud services automatically and dynamically.The paper shows the advantages of using cloud-based environment in collecting and processing linguistic data which can be easily scaled up and efficiently computed. In addition, the proposed pipeline architecture also brings out the potentials of merging with mashups from the web for representing and exploring corpus data of various types.

## 1 Introduction

With the emergence of huge amount of web data available in recent years, corpus linguistics as well as other related empirical fields such as the collecting and processing of language resources, and their evaluation are facing with the greatest challenges ever. The spread of corpus and lexical resources in linguistics has been led to a great level of theoretical survey and enhanced the empirical foundation, not only with respect to sampling and annotation, but also with exploratory data analysis. However, more recently there have been long discussions about what the current state of art in corpus linguistics fails to do, which can be pinpointed at least in two respects: (1) the lack of socio-cultural (meta-) information reflected in the data, is incompetent for pragmatic usages and discourse analysis; (2) rather skewed with data in the public domain, heterogeneity of (individualized) language usages and development is not able to be traced.

With the advanced technological progress in data availability with storage and computing ability, the issues mentioned can be tackled to a great extent. We take it as the turning point for *corpus-based* linguistics to transform into a data-intensive and *cloud-based* linguistics. In light of that, we want to explore the transformation viability in this paper. As a first step, we present a novel pipeline architecture to build Chinese Polarity Lexicon on the cloud environment by taking the data from the web as resource. Polarity lexicon contains sentiment-bearing words and phrases, encoded with polarities to each word or phrase, usually either assigned as positive or negative. The study of polarity lexicon has attracted much attention in recent years for classifiers to train on the lexical dataset, and is becoming important for applications such as Sentiment Analysis and Opinion Mining.

For the purpose of constructing automatic identifying and classifying polarity lexicon systems, a lot of (semi-) unsupervised machine learning methods for recognizing polarities of words and phrases have been proposed. In terms of language resources, these approaches either consider the information provided from the synonyms or glosses of a thesaurus or WordNet (Hu and Liu, 2004; Kamps et al., 2004; Kim and Hovy, 2004; Esuli and Sebastiani, 2005), or based on the co-occurrence relationship



Copyright 2012 by Shu-Kai Hsieh, Yu-Yun Chang, and Meng-Xian Shih
26th Pacific Asia Conference on Language,Information and Computation pages 491–497

messages derived from the corpus (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Kanayama and Nasukawa, 2006) to assign and determine the word polarity.

Notwithstanding their significant success in achieving accuracy rate, in this paper, we will argue that current approaches to the problem might face with the *methodological* drawbacks due to the lack of **scalability** on the one hand, and indifference to the **individual sentimental varieties** on the other hand. First, referring to the lack of scalability, it is rather difficult to handle out-of-vocabulary (OOV) issue on the lexical and corpus resources, in particular, those OOV words and phrases (or called as neologisms) often carried with popular usage meanings generated from the social network, and given with explicit polarities; and secondly, regarding the individual sentimental varieties, which may correspond to the linguistic varieties, subjectivities and sentiments, are largely *ad-hoc*, that is with whom s/he chats and temporal, geographical, and communication situations, etc. will have influence on her/his sentiment. Those heterogeneous properties are not properly embodied in lexical and corpus resources.

## 2 Cloud-based vs Corpus-based Linguistics

To track the essentially emergent, ever-changing, and large-scaled lexicalized sentimental social web data, we argue that corpus statisticians and linguists will need to tap into the opportunities that cloud computing environment offers. In this paper, rather than corpus-based, we propose a novel *crowd-aided cloud-based* methodology for constructing Language Resource and its Evaluation (LRE), with an experiment on Chinese Polarity Lexicon as example. The advantages of connecting LRE with cloud computing environment are multi-fold:

1. [**Easy and multi-sourced online data collection, management, integration and collaboration**] Linguistic and sentimental data can be gathered online easily using Web as Corpus (WaC), and further powered by the increasing evaluating possibilities through crowdsourcing and the enlarging of cloud storage space for reserving large-scaled data.

2. [**Seamless data preprocessing and exploratory data analysis**] The collected WaC data in the cloud storage, can be processed seamlessly online (without downloading the data) for the preliminary data preprocessing (e.g., Chinese segmentation and POS tagging), and may further apply to early data introspection with online preprocessing statistical analysis and data visualization. These techniques could be accomplished by using various application programming language interfaces (e.g., APIs for R and Python). By hosting a web interface could even facilitates the scattered tasks that used to be.

3. [**Mashup for data and models**] Once the data is collected and processed, the owner can adapt the data and mashup with others (textual, pictures or videos) to make the resource even more creative and full of varieties, which is in line with emerging trend of 'web of data' ('linked data web') proposed by Tim Berners-Lee (Tim, 2009) recently. In addition, the data can be taken as seeds and fed up the prediction models processing on the clouds, which is so-called (dynamically) stream learning .

We believe the proposed architecture above will unlock the potentials and values of linguistic data instantiated by the web. In the following, we focus on the preliminary survey of Chinese Polarity Lexicon as an example adapting the *cloud-based* methodology.

## 3 Review of Polarity Lexicons

This session explores different paradigms for how to build and evaluate polarity lexicons.

Words had been discovered with three main factors, which were evaluative factor, potency factor and activity factor, as described by Osgood et al. (1957). Within the three factors, what many researchers generally mentioned is the evaluative factor. The evaluative factor, also known as **polarity** or **semantic orientation** called by Hatzivassiloglou and McKeown (1997), which can present the intensity and the positive or negative of a word. Some researchers have found that most antonyms can be assigned with relevant polarities (e.g. *happy* can be

assigned as positive; and its antonym *sad* as negative).

Learning the polarity of words can be helpful for an amount of applications, in addition, the synonyms in the data could be further refined as well. Hatzivassiloglou and McKeown (1997) had taken the polarity of words into a system, and tried to investigate antonyms from the collected corpus and also to disambiguate the synonyms automatically. Also, Turney and Littman (2003) mentioned that an automated system containing polarity information, could be applied to text classification, analysis of survey response, filtering, tracking online opinions, and even generating chatbots.

There are a lot of ways for collecting and detecting word polarity from the text or corpora. For collecting data, Turney and Littman (2003), and Rao and Ravichandran (2009) had taken the General Inquirer lexicon (Stone et al., 1966) as their reference data, which the word polarity list was already constructed by manually tagged and evaluated via a group of people. In addition, other papers used different methods for collecting data, such as taking the 1987 Wall Street Journal with tagged data as corpus (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000), WordNet (Rao and Ravichandran, 2009; Wiebe, 2000) and via crowd sourcing (Mohammad, 2011).

As for detecting polarity, Hatzivassiloglou and McKeown (1997) introduced using conjunctions of adjectives to train the model and then labeled an orientation to each adjective through clustering. Also, Rao and Ravichandran (2009) tried using three graph-based semi-supervised learning methods to detect the word polarity, which were Mincuts, Randomized Mincuts, and Label Propagation.

Most previous papers chose to use the existed large databases for their experimental usages, and followed with different training approaches to extract or detect word polarities. Since our goal in this paper does not focus on the machine learning performance in this experimental task, we would rather demonstrate the data collection *on the fly*, so we use a naive PMI method enriched with emoticon information to dynamically and semi-automatically detect Chinese word polarity based on Plurk API (Chen et al., 2010), [1] by which all the training and testing tasks are constructed and pipelined to Google Form.

## 4 Pipelining Cloud and Crowd Computing in Lexicon Resource Development

This session explains the proposed framework, generally speaking, the data retrieved from Plurk API and preprocessed (segmented and POS-tagged) by other Chinese NLP APIs, is sent to the collaborative platform of Google called Google Form for evaluation, once evaluated they are sent to Google Fusion Table for data exploration and visualization, and stored in Google Cloud Storage with Google BigQuery. For leaning purpose, the corpus is also sent to Google Prediction model with stream machine learning method. Detailed procedures are explained in the following.

### 4.1 General framework

Once the data are collected, many of the typical preprocessing tasks can be done in a pipeline, like the tools adapted from openNLP [2]. In this paper, we propose to pipeline the processing tasks in the cloud and crowd computing environment schematized as follows, taking the extracted Plurk data as example:

We used third-party APIs (e.g. Plurk API) to get the training data from the Mood classifier (Chen et al., 2010), and evaluated two types of resulting data given the testing data: automatically tagged by Mood classifier (Chen et al., 2010) against the crowd tagged data collected from Google Form[3] (Crowd sourcing). Only the data that have the same evaluated results from the two types are considered and sent to Google Fusion Table [4] for visualizing introspection, and once the coming data scaled up, it was sent to the backend of Google Cloud Storage[5] with

[1] Plurk, like Twitter, is the most popular social microblogging system in Taiwan, we focused on it because of the advantages of tracking attitudes by mining prevalent language usages, the thus constructed Plurk Corpus can be browsed at `lope.linguistics.ntu.edu.tw/plurk/`

[2] `opennlp.sourceforge.net`

[3] `http://www.google.com/google-d-s/forms/`

[4] `http://www.google.com/fusiontables/public/tour/index.html`

[5] `https://developers.google.com/storage/`


Figure 1: Proposed pipeline framework

Google BigQuery[6]. Then finally, it was sent forward to prediction model with stream machine learning.

### 4.2 Formulate seed-sets from emotion-tagged Plurk Corpus

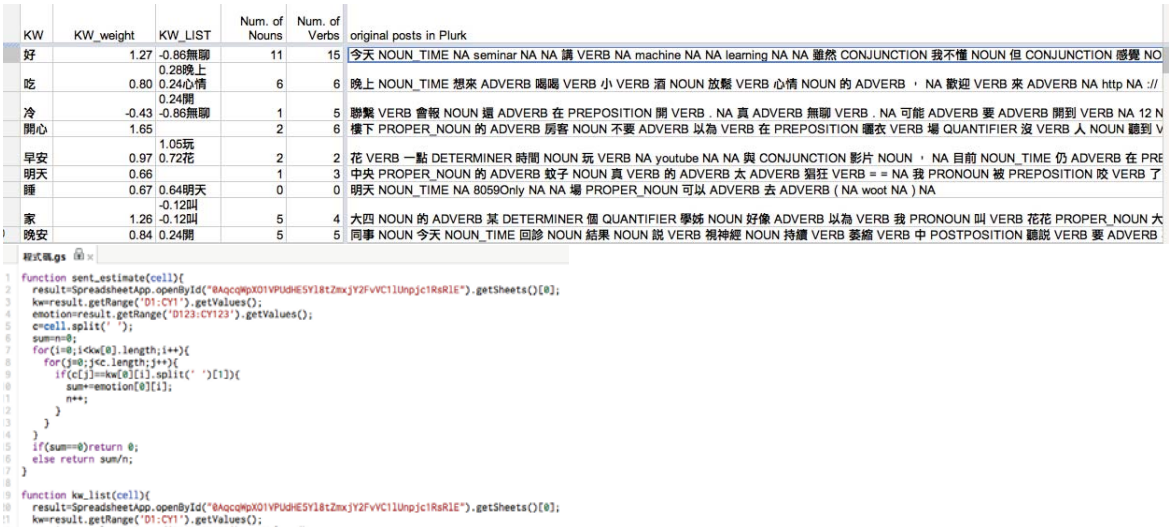
We have collected a total of 13534 Chinese posts extracted from Plurk corpus. The data have been segmented and POS-tagged by importing yahoo! Chinese Segmentator and Tagger [7], for which the yahoo! segmentation system is powerful for its lexicon extension on new emerged words from the social web (e.g. trendy words and code-mixing words). From the previous research (Chen et al., 2010), a Mood classifier had built up by training the target text with the keywords generated from Anctconc using log-likelyhood feature selection method (Kilgarriff, 2001; Anthony, 2004). Also, a manually classified result was used to evaluate the accuracy of Mood classifier. In this paper, in order to construct the word polarity prediction model more specifically, the Plurk posts are collected and fed on only if the posts have identical resulting results from Mood classifier and manual evaluation.

After the chosen posts are selected and segmented, a list of 100 seed words (all seed words are content words, including nouns, adjectives and verbs) are chosen based on the following three criteria: the balance in corpus frequency distribution, opinion relevance, and wordnet POS representatives. The first criterion is used to create a corpus frequency word list, and the 100 seed words are extracted in balance according to the corpus frequency distribution. Then the last two criteria are applied to confirm whether each seed word has opinion expressions in meaning and an relevant description from wordnet. The above seed word selecting elements are integrated to ensure the seed words have apparent meaning descriptions before adapting to the prediction model.

An online survey is created by using Google Form, which we ask the participants to evaluate the 100 seed words with a finer granularity of 5 star rating (scaling from 1 (Extremely Positive) to 5 (Extremely Negative)). Instead of taking a readily prepared corpus as previous studies (which the word polarity corpora are manually tagged by a specific group of people), for example as using the General Inquirer corpus, we let the participants to decide the polarities of each seed word, and then assign different weights to each seed word based on the overall survey statistical results. Therefore, a list of seed words with polarities is created, which only the seed words given with positive weights are tagged with positive polarity, and reversely, assigned with negative polarity.

So far, we already have 100 people complete this evaluation survey. The advantages of using Google Form to construct this word polarity evaluation survey, are that the scale of the investigation can be easily expanded and its statistical results can be renewed automatically and rapidly online, whenever there are more participants join this task. Figure 2

[6] https://developers.google.com/bigquery/
[7] http://tw.developer.yahoo.com/cas/



| KW | KW_weight | KW_LIST | Num. of Nouns | Num. of Verbs | original posts in Plurk |
|---|---|---|---|---|---|
| 好 | 1.27 | -0.86無聊 | 11 | 15 | 今天 NOUN_TIME NA seminar NA NA 講 VERB NA machine NA NA learning NA NA 雖然 CONJUNCTION 我不懂 NOUN 但 CONJUNCTION 感覺 NO |
| 吃 | 0.80 | 0.28晚上 0.24心情 | 6 | 6 | 晚上 NOUN_TIME 想來 ADVERB 喝喝 VERB 小 VERB 酒 NOUN 放鬆 VERB 心情 NOUN 的 ADVERB ， NA 歡迎 VERB 來 ADVERB NA http NA ://  |
| 冷 | -0.43 | 0.24開 -0.86無聊 | 1 | 5 | 聯繫 VERB 會報 NOUN 還 ADVERB 在 PREPOSITION 開 VERB . NA 真 ADVERB 無聊 VERB . NA 可能 ADVERB 要 ADVERB 開到 VERB NA 12 N |
| 開心 | 1.65 | | 2 | 6 | 樓下 PROPER_NOUN 的 ADVERB 房客 NOUN 不要 ADVERB 以為 VERB 在 PREPOSITION 曬衣 VERB 場 QUANTIFIER 沒 VERB 人 NOUN 聽到 V |
| 早安 | 0.97 | 1.05玩 0.72花 | 2 | 2 | 花 VERB 一點 DETERMINER 時間 NOUN 玩 VERB NA youtube NA NA 與 CONJUNCTION 影片 NOUN ， NA 目前 NOUN_TIME 仍 ADVERB 在 PRE |
| 明天 | 0.66 | | 1 | 3 | 中央 PROPER_NOUN 的 ADVERB 蚊子 NOUN 真 VERB 的 ADVERB 太 ADVERB 猖狂 VERB = = NA 我 PRONOUN 被 PREPOSITION 咬 VERB 了 |
| 睡 | 0.67 | 0.64明天 | 0 | 0 | 明天 NOUN_TIME NA 8059Only NA NA 場 PROPER_NOUN 可以 ADVERB 去 ADVERB ( NA woot NA ) NA |
| 家 | 1.26 | -0.12叫 -0.12叫 | 5 | 4 | 大四 NOUN 的 ADVERB 某 DETERMINER 個 QUANTIFIER 學姊 NOUN 好像 ADVERB 以為 VERB 我 PRONOUN 叫 VERB 花花 PROPER_NOUN 大 |
| 晚安 | 0.84 | 0.24開 | 5 | 5 | 同事 NOUN 今天 NOUN_TIME 回診 NOUN 結果 NOUN 說 VERB 視神經 NOUN 持續 VERB 萎縮 VERB 中 POSTPOSITION 聽說 VERB 要 ADVERB |

程式碼.gs

```
function sent_estimate(cell){
  result=SpreadsheetApp.openById("0AqcqWpXO1VPUdHE5Yl8tZmxjY2FvVC1lUnpjc1RsRlE").getSheets()[0];
  kw=result.getRange('D1:CY1').getValues();
  emotion=result.getRange('D123:CY123').getValues();
  c=cell.split(' ');
  sum=n=0;
  for(i=0;i<kw[0].length;i++){
    for(j=0;j<c.length;j++){
      if(c[j]==kw[0][i].split(' ')[1]){
        sum+=emotion[0][i];
        n++;
      }
    }
  }
  if(sum==0)return 0;
  else return sum/n;
}

function kw_list(cell){
  result=SpreadsheetApp.openById("0AqcqWpXO1VPUdHE5Yl8tZmxjY2FvVC1lUnpjc1RsRlE").getSheets()[0];
  kw=result.getRange('D1:CY1').getValues();
```

Figure 2: Score form with functions written online

shows a snapshot of this scored form.

### 4.3 Preprocessing and Exploratory Data Analysis

In order to compute the model training more efficiently, we use Google API to write our own functions and apply them to the data in Google Form. This is a convenient way for training and analyzing the data online without needing to run the whole programs on our own devices. To be even better, once the data is renewed, the programs will run the functions automatically in an instant and no need to execute the programs manually. Through this small experiment of this paper, we hope to provide a practical method for linguists to deal with data in a more skilled and dynamic way.

With the data imported into Google Fusion Table, it is convienient to use a variety of Visualize plots as shown in Figure 3. The Visualize function contains table, map, line, bar, pie and other plotting tools to help quickly analyze the data. In addition, it allows us to add some specific conditions while plotting. By applying the line chart, we can quickly find out which city has the greatest Plurk population. Visualize function has also implanted the Geocode (geography codes) which takes the location data to tag places on a map or colored up the related regions. Also, once clicking on the tagged places or colored regions, the related meta information will be presented. For investigating data distribution (for example gender, location, and age), the pie chart can be used which provides results calculated in percentages. On the other hand, for more detailed distribution analysis, bar chart is shown to be better presenting the statistic results between two variables, such as location and age.

### 4.4 Prediction Model

The prior polarity lexicons thus constructed are used as training data for cloud-based prediction model in extracting more polar words and determining the overall sentiment of Plurk texts. As the first attempt, we are using the Google-hosted prediction model as a black-box for primer experiment,[8] and a Predictive Model Markup Language (PMML)-based[9] *adaptive* prediction model is envisioned which combines Chinese wordnet-based sense/sentiment propagation approach (by assuming that sentiment and lexical relatedness are linked) with bootstrapped individualized parameters.

### 4.5 Limitations

With its versatility in processing and exploring corpus data, the limitations we encountered with this framework so far lie in two aspects: First, due to the free service provided by Google, the experiment is largely dependent on the services provided by the

[8] http://lope.linguistics.ntu.edu.tw/wordpola/iosubscribe.html

[9] http://dmg.org/pmml-v4-0-1.html

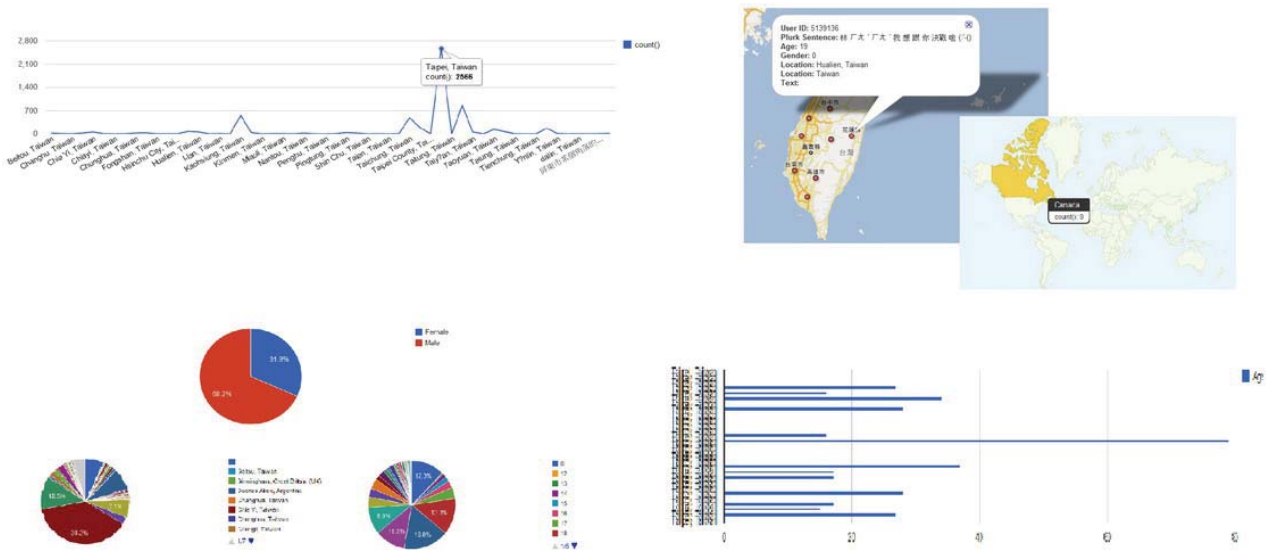



Figure 3: Corpus data exploratory analysis using Fusion Table APIs

commercial company, which could be a weak point in terms of stability in the future; and (2) what is also at stake here is indeed the embedded use of other tools and APIs can be a problem due to their compatibility with Google services. However, with the downside mentioned, we believe that much more can be improved with respect to the likelihood of future convergence for the open collaboration between academic and commercial fields.

## 5 Conclusion

In conclusion, we showed a novel architecture of language resource construction and evaluation on the cloud computing environment, and illustrated it with the experiment on Chinese Polarity Lexicon. We believe that this approach will open up many possibilities to be explored. This mixed scenario of folksonomy and cloud computing allow us to not only detect how different groups of people recognize prior polarities and their weights from the contextual clues, but also understand further which parameters should be modeled as patterns for polarity detection. The compiled lexicon can be served as a dynamic input for the cloud-based streaming prediction model(s) for the maximum performance.

In future work, we will apply the proposed architecture to augment the newly released Chinese Wordnet[10] by polarity classification of synsets instead of lemma, since the current way is not able to capture the fact that a word with various senses could have different polarities. In addition, although these methods can be applied on Chinese words, word sentiment is in fact a function of the composite characters and the way as how people process an ideogram while encountering a new word. In the future, we will consider running an experiment on Character Sentiment in parallel.

## References

L. Anthony. 2004. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pages 7–13.

M.-Y. Chen, H.-N. Lin, C.-A. Shih, Y.-C. Hsu, P.-Y. Hsu, and S.-K. Hsieh. 2010. Classifying Mood in Plurks. In *The 22th Conference on Computational Linguistics and Speech Processing*. Chi-Nan University, Taiwan.

A. Esuli and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM05)*, pages 617–624. Bermen, DE.

[10] http://lope.linguistics.ntu.edu.tw/cwn

V. Hatzivassiloglou and K. R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 174–181.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177. ACM.

J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, volume IV, pages 1115–1118. Lisbon, Portugal.

H. Kanayama and T. Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–363.

A. Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. Geneva.

S. Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairytales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.

C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. 1957. *The measurement of meaning*. Urbana, USA: University of Illinois Press.

D. Rao and D. Ravichandran. 2009. Semi-Supervised Polarity Lexicon Induction. In *The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

B.-L. Tim. 2009. Linked data. In *TED 2009 conference*. "The Great Unveiling" in Long Beach, CA, USA.

P. D. Turney and M. L. Littman. 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

P. Turney. 2002. Thumbs up or thumbs down? sentiment orientation applied to unsupervised classication of reviews. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*.