

Extracting Semantic Orientations of Words using Spin Model

Hiroya Takamura Takashi Inui Manabu Okumura

Precision and Intelligence Laboratory

Tokyo Institute of Technology

4259 Nagatsuta Midori-ku Yokohama, 226-8503 Japan

{takamura, oku}@pi.titech.ac.jp,

tinui@lr.pi.titech.ac.jp

Abstract

We propose a method for extracting semantic orientations of words: desirable or undesirable. Regarding semantic orientations as spins of electrons, we use the mean field approximation to compute the approximate probability function of the system instead of the intractable actual probability function. We also propose a criterion for parameter selection on the basis of magnetization. Given only a small number of seed words, the proposed method extracts semantic orientations with high accuracy in the experiments on English lexicon. The result is comparable to the best value ever reported.

1 Introduction

Identification of emotions (including opinions and attitudes) in text is an important task which has a variety of possible applications. For example, we can efficiently collect opinions on a new product from the internet, if opinions in bulletin boards are automatically identified. We will also be able to grasp people's attitudes in questionnaire, without actually reading all the responds.

An important resource in realizing such identification tasks is a list of words with semantic orientation: positive or negative (desirable or undesirable). Frequent appearance of positive words in a document implies that the writer of the document would

have a positive attitude on the topic. The goal of this paper is to propose a method for automatically creating such a word list from glosses (i.e., definition or explanation sentences) in a dictionary, as well as from a thesaurus and a corpus. For this purpose, we use *spin model*, which is a model for a set of electrons with spins. Just as each electron has a direction of spin (up or down), each word has a semantic orientation (positive or negative). We therefore regard words as a set of electrons and apply the mean field approximation to compute the average orientation of each word. We also propose a criterion for parameter selection on the basis of magnetization, a notion in statistical physics. Magnetization indicates the global tendency of polarization.

We empirically show that the proposed method works well even with a small number of seed words.

2 Related Work

Turney and Littman (2003) proposed two algorithms for extraction of semantic orientations of words. To calculate the association strength of a word with positive (negative) seed words, they used the number of hits returned by a search engine, with a query consisting of the word and one of seed words (e.g., “*word* NEAR good”, “*word* NEAR bad”). They regarded the difference of two association strengths as a measure of semantic orientation. They also proposed to use Latent Semantic Analysis to compute the association strength with seed words. An empirical evaluation was conducted on 3596 words extracted from General Inquirer (Stone et al., 1966).

Hatzivassiloglou and McKeown (1997) focused on conjunctive expressions such as “simple and

well-received” and “simplistic but well-received”, where the former pair of words tend to have the same semantic orientation, and the latter tend to have the opposite orientation. They first classify each conjunctive expression into the same-orientation class or the different-orientation class. They then use the classified expressions to cluster words into the positive class and the negative class. The experiments were conducted with the dataset that they created on their own. Evaluation was limited to adjectives.

Kobayashi et al. (2001) proposed a method for extracting semantic orientations of words with bootstrapping. The semantic orientation of a word is determined on the basis of its gloss, if any of their 52 hand-crafted rules is applicable to the sentence. Rules are applied iteratively in the bootstrapping framework. Although Kobayashi et al.’s work provided an accurate investigation on this task and inspired our work, it has drawbacks: low recall and language dependency. They reported that the semantic orientations of only 113 words are extracted with precision 84.1% (the low recall is due partly to their large set of seed words (1187 words)). The hand-crafted rules are only for Japanese.

Kamps et al. (2004) constructed a network by connecting each pair of synonymous words provided by WordNet (Fellbaum, 1998), and then used the shortest paths to two seed words “good” and “bad” to obtain the semantic orientation of a word. Limitations of their method are that a synonymy dictionary is required, that antonym relations cannot be incorporated into the model. Their evaluation is restricted to adjectives. The method proposed by Hu and Liu (2004) is quite similar to the shortest-path method. Hu and Liu’s method iteratively determines the semantic orientations of the words neighboring any of the seed words and enlarges the seed word set in a bootstrapping manner.

Subjective words are often semantically oriented. Wiebe (2000) used a learning method to collect subjective adjectives from corpora. Riloff et al. (2003) focused on the collection of subjective nouns.

We later compare our method with Turney and Littman’s method and Kamps et al.’s method.

The other pieces of research work mentioned above are related to ours, but their objectives are different from ours.

3 Spin Model and Mean Field Approximation

We give a brief introduction to the spin model and the mean field approximation, which are well-studied subjects both in the statistical mechanics and the machine learning communities (Geman and Geman, 1984; Inoue and Carlucci, 2001; Mackay, 2003).

A spin system is an array of N electrons, each of which has a spin with one of two values “+1 (up)” or “−1 (down)”. Two electrons next to each other energetically tend to have the same spin. This model is called *the Ising spin model*, or simply *the spin model* (Chandler, 1987). The energy function of a spin system can be represented as

$$E(\mathbf{x}, W) = -\frac{1}{2} \sum_{ij} w_{ij} x_i x_j, \quad (1)$$

where x_i and x_j ($\in \mathbf{x}$) are spins of electrons i and j , matrix $W = \{w_{ij}\}$ represents weights between two electrons.

In a spin system, the variable vector \mathbf{x} follows the Boltzmann distribution :

$$P(\mathbf{x}|W) = \frac{\exp(-\beta E(\mathbf{x}, W))}{Z(W)}, \quad (2)$$

where $Z(W) = \sum_{\mathbf{x}} \exp(-\beta E(\mathbf{x}, W))$ is the normalization factor, which is called *the partition function* and β is a constant called *the inverse-temperature*. As this distribution function suggests, a configuration with a higher energy value has a smaller probability.

Although we have a distribution function, computing various probability values is computationally difficult. The bottleneck is the evaluation of $Z(W)$, since there are 2^N configurations of spins in this system.

We therefore approximate $P(\mathbf{x}|W)$ with a simple function $Q(\mathbf{x}; \theta)$. The set of parameters θ for Q , is determined such that $Q(\mathbf{x}; \theta)$ becomes as similar to $P(\mathbf{x}|W)$ as possible. As a measure for the distance between P and Q , the variational free energy F is often used, which is defined as the difference between the mean energy with respect to Q and the entropy of Q :

$$F(\theta) = \beta \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) E(\mathbf{x}, W)$$

$$-\left(-\sum_{\mathbf{x}} Q(\mathbf{x}; \theta) \log Q(\mathbf{x}; \theta)\right). \quad (3)$$

The parameters θ that minimizes the variational free energy will be chosen. It has been shown that minimizing F is equivalent to minimizing the Kullback-Leibler divergence between P and Q (Mackay, 2003).

We next assume that the function $Q(\mathbf{x}; \theta)$ has the factorial form :

$$Q(\mathbf{x}; \theta) = \prod_i Q(x_i; \theta_i). \quad (4)$$

Simple substitution and transformation leads us to the following variational free energy :

$$F(\theta) = -\frac{\beta}{2} \sum_{ij} w_{ij} \bar{x}_i \bar{x}_j - \sum_i \left(-\sum_{x_i} Q(x_i; \theta_i) \log Q(x_i; \theta_i) \right). \quad (5)$$

With the usual method of Lagrange multipliers, we obtain the *mean field equation* :

$$\bar{x}_i = \frac{\sum_{x_i} x_i \exp\left(\beta x_i \sum_j w_{ij} \bar{x}_j\right)}{\sum_{x_i} \exp\left(\beta x_i \sum_j w_{ij} \bar{x}_j\right)}. \quad (6)$$

This equation is solved by the iterative update rule :

$$\bar{x}_i^{new} = \frac{\sum_{x_i} x_i \exp\left(\beta x_i \sum_j w_{ij} \bar{x}_j^{old}\right)}{\sum_{x_i} \exp\left(\beta x_i \sum_j w_{ij} \bar{x}_j^{old}\right)}. \quad (7)$$

4 Extraction of Semantic Orientation of Words with Spin Model

We use the spin model to extract semantic orientations of words.

Each spin has a direction taking one of two values: up or down. Two neighboring spins tend to have the same direction from a energetic reason. Regarding each word as an electron and its semantic orientation as the spin of the electron, we construct a lexical network by connecting two words if, for example, one word appears in the gloss of the other word. Intuition behind this is that if a word is semantically oriented in one direction, then the words in its gloss tend to be oriented in the same direction.

Using the mean-field method developed in statistical mechanics, we determine the semantic orientations on the network in a global manner. The global optimization enables the incorporation of possibly noisy resources such as glosses and corpora, while existing simple methods such as the shortest-path method and the bootstrapping method cannot work in the presence of such noisy evidences. Those methods depend on less-noisy data such as a thesaurus.

4.1 Construction of Lexical Networks

We construct a lexical network by linking two words if one word appears in the gloss of the other word. Each link belongs to one of two groups: the same-orientation links SL and the different-orientation links DL . If at least one word precedes a negation word (e.g., not) in the gloss of the other word, the link is a different-orientation link. Otherwise the links is a same-orientation link.

We next set weights $W = (w_{ij})$ to links :

$$w_{ij} = \begin{cases} \frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in SL) \\ -\frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in DL) \\ 0 & otherwise \end{cases}, \quad (8)$$

where l_{ij} denotes the link between word i and word j , and $d(i)$ denotes the degree of word i , which means the number of words linked with word i . Two words without connections are regarded as being connected by a link of weight 0. We call this network *the gloss network* (G).

We construct another network, *the gloss-thesaurus network* (GT), by linking synonyms, antonyms and hypernyms, in addition to the the above linked words. Only antonym links are in DL .

We enhance the gloss-thesaurus network with cooccurrence information extracted from corpus. As mentioned in Section 2, Hatzivassiloglou and McKeown (1997) used conjunctive expressions in corpus. Following their method, we connect two adjectives if the adjectives appear in a conjunctive form in the corpus. If the adjectives are connected by “and”, the link belongs to SL . If they are connected by “but”, the link belongs to DL . We call this network *the gloss-thesaurus-corpus network* (GTC).

4.2 Extraction of Orientations

We suppose that a small number of seed words are given. In other words, we know beforehand the semantic orientations of those given words. We incorporate this small labeled dataset by modifying the previous update rule.

Instead of $\beta E(\mathbf{x}, W)$ in Equation (2), we use the following function $H(\beta, \mathbf{x}, W)$:

$$H(\beta, \mathbf{x}, W) = -\frac{\beta}{2} \sum_{ij} w_{ij} x_i x_j + \alpha \sum_{i \in L} (x_i - a_i)^2, \quad (9)$$

where L is the set of seed words, a_i is the orientation of seed word i , and α is a positive constant. This expression means that if x_i ($i \in L$) is different from a_i , the state is penalized.

Using function H , we obtain the new update rule for x_i ($i \in L$) :

$$\bar{x}_i^{new} = \frac{\sum_{x_i} x_i \exp\left(\beta x_i s_i^{old} - \alpha(x_i - a_i)^2\right)}{\sum_{x_i} \exp\left(\beta x_i s_i^{old} - \alpha(x_i - a_i)^2\right)}, \quad (10)$$

where $s_i^{old} = \sum_j w_{ij} \bar{x}_j^{old}$. \bar{x}_i^{old} and \bar{x}_i^{new} are the averages of x_i respectively before and after update. What is discussed here was constructed with the reference to work by Inoue and Carlucci (2001), in which they applied the spin glass model to image restoration.

Initially, the averages of the seed words are set according to their given orientations. The other averages are set to 0.

When the difference in the value of the variational free energy is smaller than a threshold before and after update, we regard computation converged.

The words with high final average values are classified as positive words. The words with low final average values are classified as negative words.

4.3 Hyper-parameter Prediction

The performance of the proposed method largely depends on the value of hyper-parameter β . In order to make the method more practical, we propose criteria for determining its value.

When a large labeled dataset is available, we can obtain a reliable *pseudo leave-one-out error rate* :

$$\frac{1}{|L|} \sum_{i \in L} [a_i \bar{x}'_i], \quad (11)$$

where $[t]$ is 1 if t is negative, otherwise 0, and \bar{x}'_i is calculated with the right-hand-side of Equation (6), where the penalty term $\alpha(\bar{x}_i - a_i)^2$ in Equation (10) is ignored. We choose β that minimizes this value.

However, when a large amount of labeled data is unavailable, the value of pseudo leave-one-out error rate is not reliable. In such cases, we use *magnetization* m for hyper-parameter prediction :

$$m = \frac{1}{N} \sum_i \bar{x}_i. \quad (12)$$

At a high temperature, spins are randomly oriented (*paramagnetic phase*, $m \approx 0$). At a low temperature, most of the spins have the same direction (*ferromagnetic phase*, $m \neq 0$). It is known that at some intermediate temperature, ferromagnetic phase suddenly changes to paramagnetic phase. This phenomenon is called *phase transition*. Slightly before the phase transition, spins are locally polarized; strongly connected spins have the same polarity, but not in a global way.

Intuitively, the state of the lexical network is locally polarized. Therefore, we calculate values of m with several different values of β and select the value just before the phase transition.

4.4 Discussion on the Model

In our model, the semantic orientations of words are determined according to the averages values of the spins. Despite the heuristic flavor of this decision rule, it has a theoretical background related to maximizer of posterior marginal (MPM) estimation, or ‘finite-temperature decoding’ (Iba, 1999; Marroquin, 1985). In MPM, the average is the marginal distribution over x_i obtained from the distribution over \mathbf{x} . We should note that the finite-temperature decoding is quite different from annealing type algorithms or ‘zero-temperature decoding’, which correspond to maximum a posteriori (MAP) estimation and also often used in natural language processing (Cowie et al., 1992).

Since the model estimation has been reduced to simple update calculations, the proposed model is similar to conventional spreading activation approaches, which have been applied, for example, to word sense disambiguation (Veronis and Ide, 1990). Actually, the proposed model can be regarded as a spreading activation model with a specific update

rule, as long as we are dealing with 2-class model (2-Ising model).

However, there are some advantages in our modelling. The largest advantage is its theoretical background. We have an objective function and its approximation method. We thus have a measure of goodness in model estimation and can use another better approximation method, such as Bethe approximation (Tanaka et al., 2003). The theory tells us which update rule to use. We also have a notion of magnetization, which can be used for hyper-parameter estimation. We can use a plenty of knowledge, methods and algorithms developed in the field of statistical mechanics. We can also extend our model to a multiclass model (*Q-Ising model*).

Another interesting point is the relation to maximum entropy model (Berger et al., 1996), which is popular in the natural language processing community. Our model can be obtained by maximizing the entropy of the probability distribution $Q(\mathbf{x})$ under constraints regarding the energy function.

5 Experiments

We used glosses, synonyms, antonyms and hypernyms of WordNet (Fellbaum, 1998) to construct an English lexical network. For part-of-speech tagging and lemmatization of glosses, we used Tree-Tagger (Schmid, 1994). 35 stopwords (quite frequent words such as “be” and “have”) are removed from the lexical network. Negation words include 33 words. In addition to usual negation words such as “not” and “never”, we include words and phrases which mean negation in a general sense, such as “free from” and “lack of”. The whole network consists of approximately 88,000 words. We collected 804 conjunctive expressions from Wall Street Journal and Brown corpus as described in Section 4.2.

The labeled dataset used as a gold standard is General Inquirer lexicon (Stone et al., 1966) as in the work by Turney and Littman (2003). We extracted the words tagged with “Positiv” or “Negativ”, and reduced multiple-entry words to single entries. As a result, we obtained 3596 words (1616 positive words and 1980 negative words)¹. In the computation of

¹Although we preprocessed in the same way as Turney and Littman, there is a slight difference between their dataset and our dataset. However, we believe this difference is insignificant.

Table 1: Classification accuracy (%) with various networks and four different sets of seed words. In the parentheses, the predicted value of β is written. For cv, no value is written for β , since 10 different values are obtained.

seeds	GTC	GT	G
cv	90.8 (—)	90.9 (—)	86.9 (—)
14	81.9 (1.0)	80.2 (1.0)	76.2 (1.0)
4	73.8 (0.9)	73.7 (1.0)	65.2 (0.9)
2	74.6 (1.0)	61.8 (1.0)	65.7 (1.0)

accuracy, seed words are eliminated from these 3596 words.

We conducted experiments with different values of β from 0.1 to 2.0, with the interval 0.1, and predicted the best value as explained in Section 4.3. The threshold of the magnetization for hyper-parameter estimation is set to 1.0×10^{-5} . That is, the predicted optimal value of β is the largest β whose corresponding magnetization does not exceeds the threshold value.

We performed 10-fold cross validation as well as experiments with fixed seed words. The fixed seed words are the ones used by Turney and Littman: 14 seed words {good, nice, excellent, positive, fortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior}; 4 seed words {good, superior, bad, inferior}; 2 seed words {good, bad}.

5.1 Classification Accuracy

Table 1 shows the accuracy values of semantic orientation classification for four different sets of seed words and various networks. In the table, cv corresponds to the result of 10-fold cross validation, in which case we use the pseudo leave-one-out error for hyper-parameter estimation, while in other cases we use magnetization.

In most cases, the synonyms and the cooccurrence information from corpus improve accuracy. The only exception is the case of 2 seed words, in which G performs better than GT. One possible reason of this inversion is that the computation is trapped in a local optimum, since a small number of seed words leave a relatively large degree of freedom in the solution space, resulting in more local optimal points.

We compare our results with Turney and

Table 2: Actual best classification accuracy (%) with various networks and four different sets of seed words. In the parenthesis, the actual best value of β is written, except for cv.

seeds	GTC	GT	G
cv	91.5 (—)	91.5 (—)	87.0 (—)
14	81.9 (1.0)	80.2 (1.0)	76.2 (1.0)
4	74.4 (0.6)	74.4 (0.6)	65.3 (0.8)
2	75.2 (0.8)	61.9 (0.8)	67.5 (0.5)

Littman’s results. With 14 seed words, they achieved 61.26% for a small corpus (approx. 1×10^7 words), 76.06% for a medium-sized corpus (approx. 2×10^9 words), 82.84% for a large corpus (approx. 1×10^{11} words).

Without a corpus nor a thesaurus (but with glosses in a dictionary), we obtained accuracy that is comparable to Turney and Littman’s with a medium-sized corpus. When we enhance the lexical network with corpus and thesaurus, our result is comparable to Turney and Littman’s with a large corpus.

5.2 Prediction of β

We examine how accurately our prediction method for β works by comparing Table 1 above and Table 2 below. Our method predicts good β quite well especially for 14 seed words. For small numbers of seed words, our method using magnetization tends to predict a little larger value.

We also display the figure of magnetization and accuracy in Figure 1. We can see that the sharp change of magnetization occurs at around $\beta = 1.0$ (phrase transition). At almost the same point, the classification accuracy reaches the peak.

5.3 Precision for the Words with High Confidence

We next evaluate the proposed method in terms of precision for the words that are classified with high confidence. We regard the absolute value of each average as a confidence measure and evaluate the top words with the highest absolute values of averages.

The result of this experiment is shown in Figure 2, for 14 seed words as an example. The top 1000 words achieved more than 92% accuracy. This result shows that the absolute value of each average

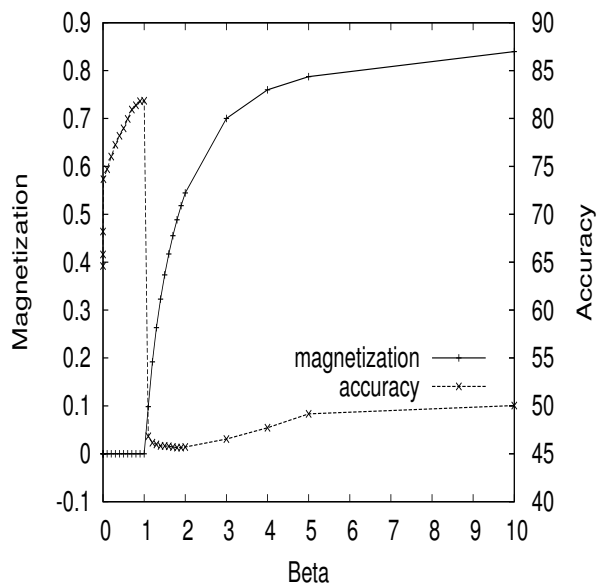


Figure 1: Example of magnetization and classification accuracy(14 seed words).

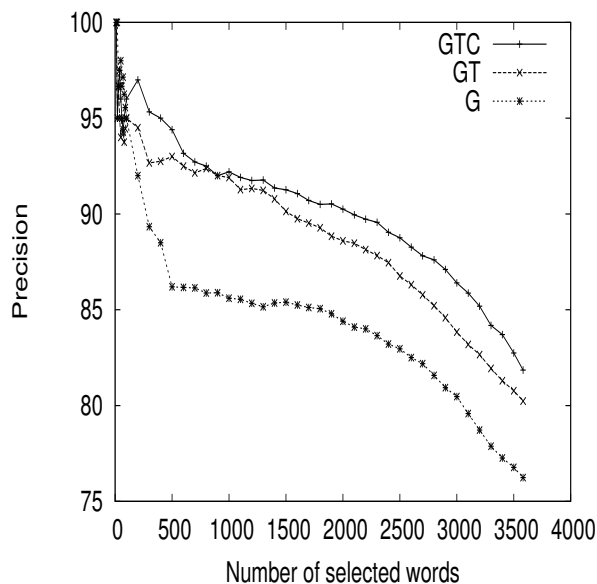


Figure 2: Precision (%) with 14 seed words.

Table 3: Precision (%) for selected adjectives. Comparison between the proposed method and the shortest-path method.

seeds	proposed	short. path
14	73.4 (1.0)	70.8
4	71.0 (1.0)	64.9
2	68.2 (1.0)	66.0

Table 4: Precision (%) for adjectives. Comparison between the proposed method and the bootstrapping method.

seeds	proposed	bootstrap
14	83.6 (0.8)	72.8
4	82.3 (0.9)	73.2
2	83.5 (0.7)	71.1

can work as a confidence measure of classification.

5.4 Comparison with other methods

In order to further investigate the model, we conduct experiments in restricted settings.

We first construct a lexical network using only synonyms. We compare the spin model with the shortest-path method proposed by Kamps et al. (2004) on this network, because the shortest-path method cannot incorporate negative links of antonyms. We also restrict the test data to 697 adjectives, which is the number of examples that the shortest-path method can assign a non-zero orientation value. Since the shortest-path method is designed for 2 seed words, the method is extended to use the average shortest-path lengths for 4 seed words and 14 seed words. Table 3 shows the result. Since the only difference is their algorithms, we can conclude that the global optimization of the spin model works well for the semantic orientation extraction.

We next compare the proposed method with a simple bootstrapping method proposed by Hu and Liu (2004). We construct a lexical network using synonyms and antonyms. We restrict the test data to 1470 adjectives for comparison of methods. The result in Table 4 also shows that the global optimization of the spin model works well for the semantic orientation extraction.

We also tested the shortest path method and the bootstrapping method on GTC and GT, and obtained low accuracies as expected in the discussion in Section 4.

5.5 Error Analysis

We investigated a number of errors and concluded that there were mainly three types of errors.

One is the ambiguity of word senses. For example, one of the glosses of “costly” is “entailing great loss or sacrifice”. The word “great” here means “large”, although it usually means “outstanding” and is positively oriented.

Another is lack of structural information. For example, “arrogance” means “overbearing pride evidenced by a superior manner toward the weak”. Although “arrogance” is mistakenly predicted as positive due to the word “superior”, what is superior here is “manner”.

The last one is idiomatic expressions. For example, although “brag” means “show off”, neither of “show” and “off” has the negative orientation. Idiomatic expressions often does not inherit the semantic orientation from or to the words in the gloss.

The current model cannot deal with these types of errors. We leave their solutions as future work.

6 Conclusion and Future Work

We proposed a method for extracting semantic orientations of words. In the proposed method, we regarded semantic orientations as spins of electrons, and used the mean field approximation to compute the approximate probability function of the system instead of the intractable actual probability function. We succeeded in extracting semantic orientations with high accuracy, even when only a small number of seed words are available.

There are a number of directions for future work.

One is the incorporation of syntactic information. Since the importance of each word consisting a gloss depends on its syntactic role, syntactic information in glosses should be useful for classification.

Another is active learning. To decrease the amount of manual tagging for seed words, an active learning scheme is desired, in which a small number of *good* seed words are automatically selected.

Although our model can easily extended to a

multi-state model, the effectiveness of using such a multi-state model has not been shown yet.

Our model uses only the tendency of having the same orientation. Therefore we can extract semantic orientations of new words that are not listed in a dictionary. The validation of such extension will widen the possibility of application of our method.

Larger corpora such as web data will improve performance. The combination of our method and the method by Turney and Littman (2003) is promising.

Finally, we believe that the proposed model is applicable to other tasks in computational linguistics.

References

- Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- David Chandler. 1987. *Introduction to Modern Statistical Mechanics*. Oxford University Press.
- Jim Cowie, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th conference on Computational linguistics*, volume 1, pages 359–365.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series*. MIT Press.
- Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.
- Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining (KDD-2004)*, pages 168–177.
- Yukito Iba. 1999. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, pages 3875–3888.
- Junichi Inoue and Domenico M. Carlucci. 2001. Image restoration using the q-ising spin glass. *Physical Review E*, 64:036121–1 – 036121–18.
- Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume IV, pages 1115–1118.
- Nozomi Kobayashi, Takashi Inui, and Kentaro Inui. 2001. Dictionary-based acquisition of the lexical knowledge for p/n analysis (in Japanese). In *Proceedings of Japanese Society for Artificial Intelligence, SLUD-33*, pages 45–50.
- David J. C. Mackay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
- Jose L. Marroquin. 1985. Optimal bayesian estimators for image segmentation and surface reconstruction. Technical Report A.I. Memo 839, Massachusetts Institute of Technology.
- Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 25–32.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49.
- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.
- Kazuyuki Tanaka, Junichi Inoue, and Mike Titterton. 2003. Probabilistic image processing by means of the bethe approximation for the q-ising model. *Journal of Physics A: Mathematical and General*, 36:11023–11035.
- Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.
- Jean Veronis and Nancy M. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th Conference on Computational Linguistics*, volume 2, pages 389–394.
- Janyce M. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, pages 735–740.