Statistical Mechanical Analysis of Semantic Orientations on Lexical Network

Takuma GOTO[*] *Yoshiyuki KABASHIMA*
Hiroya TAKAMURA
Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Japan
takuma510@gmail.com, kaba@dis.titech.ac.jp, takamura@pi.titech.ac.jp

ABSTRACT
Many of the state-of-the-art methods for constructing a polarity lexicon rely on the propagation of polarity on the lexical network. In one of those methods, where the Ising spin model is employed as a probabilistic model, it is reported that the system exhibits the phase transition in the vicinity of the optimal temperature parameter. We provide an analysis of this phenomenon from the viewpoint of statistical mechanics and clarify the underlying mechanism. On the basis of this analysis, we propose a scheme for improving the extraction performance, i.e., by removing the largest eigenvalue component from the weight matrix. Experimental results show that the scheme significantly improves the accuracy of the extraction of the semantic orientations at negligible additional computational cost, outperforming the state-of-the-art algorithms. We also explore the origin of the high classification performance by analyzing eigenvalues of the weight matrix and a linearized model.

KEYWORDS: sentiment analysis, polarity lexicon, spin model, label propagation.

[*] Takuma Goto now works for ACCESS CO., LTD.

Proceedings of COLING 2012: Technical Papers, pages 977–994,
COLING 2012, Mumbai, December 2012.

1 Introduction

A huge amount of semantic information is constantly being produced and accumulated on the Internet by the activities of individuals through, for example, their blog, Twitter, and Facebook postings. The information tends to focus on personal interests but can include generally useful information such as opinions about fashion and comments about new products. This means that extracting and structuralizing such information can be beneficial for both producers and consumers, which led us to focus on the development of methods for handling semantic information.

In general, each word constituting sentences possesses its specific orientation. For example, we usually receive positive impressions for words such as "good", "excellent" and "enjoyable", while "bad", "poor" and "boring" sound negative. Such word-specific orientation of impression is termed *polarity* (or *semantic orientation*). A polarity lexicon is a list of words and phrases that are labeled by their polarity, and is an important resource in extracting semantic information from natural language data. Accordingly, the construction of such lists under various conditions has been a major focus in sentiment analysis research (Hatzivassiloglou and McKeown, 1997; Choi and Cardie, 2009). Many construction methods have been developed so far (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005; Turney and Littman, 2003; Velikovich et al., 2010; Kamps et al., 2004; Rao and Ravichandran, 2009).

Among those methods for polarity lexicon construction, the method proposed by Takamura et al. (2005) is distinctive in terms of emphasizing the utility of probabilities. In their method, the construction of a polarity lexicon is mapped to the Ising spin model of magnetism at a finite temperature. This mapping, in conjunction with the formalism of equilibrium statistical mechanics, yields a probabilistic model for assigning a polarity to each word. The optimal assignment of the polarities is determined by approximately evaluating the averages of the spin variables. Its experimental application to a lexical network of 88,015 words demonstrated its utility. Besides this, we would like to draw attention to the following observations:

- The classification performance is optimized at a finite temperature. This is somewhat counterintuitive, since the cost (energy) function is not minimized unless the temperature vanishes. At a finite temperature, the probability mass is split and given to all states, instead of a single one that minimizes the cost function. As a result, most of the spin averages are hardly biased. This suggests that performance can be optimized by using relatively small signals which are offered by the hardly biased spin averages.
- The experimental data indicates that the spin system exhibits a ferromagnetic phase transition in the vicinity of the optimal temperature.

The purpose of this paper is to propose a simple but effective scheme to improve the performance of the original spin-model-based method by clarifying the mechanism underlying the above observations. We show that the rate of correct word classification is significantly increased simply by removing the largest eigenvalue component from the weight matrix of the lexical network, which is directly related to avoidance of the ferromagnetic phase transition. The computational cost for the removal grows only linearly with network size, which means that the improved method is highly practical and useful. We also show that this scheme of removing the largest eigenvalue component also improves the linearized model, which is almost equivalent to the label propagation (Zhu and Ghahramani, 2002).

Although we focus on a specific problem of constructing the polarity lexicon in this paper, the developed methodology can be employed for general purposes of assessing influences of a few representative nodes in a network via local communications. In fact, network-based semi-supervised models are used in a number of tasks in natural language processing, such as word sense disambiguation (Yu et al., 2011), machine translation (Alexandrescu and Kirchhoff, 2009), query classification (Hu et al., 2009; Li et al., 2008).

2 Related Work

There has been much related work on building polarity lexicons.

One of the earliest studies was done by Hatzivassiloglou and McKeown (1997), who focused on conjoined adjectives in the Wall Street Journal corpus (Marcus et al., 1993). They deduced the polarity of adjectives by using pairs of adjectives appearing with a conjunction in the corpus. For example, pairs of adjectives joined with an "and" tend to have the same semantic orientation (e.g., "simple and well-received") while those joined with a "but" tend to have the opposite semantic orientation (e.g., "simplistic but well-received"). Because of the limited applicability of this method, only adjectives can be entried in a polarity lexicon.

Taking a corpus-based approach, Turney and Littman (2003) built a polarity lexicon by using two algorithms. They used a query such as "word NEAR good" and "word NEAR bad," where "good" and "bad" are seed words (their polarities are known), and obtained the number of hits returned by a search engine. The association strength of the target word with positive (negative) seed words was calculated. In their work, 3596 words from the General Inquirer lexicon (Stone et al., 1966) were used for empirical evaluation. General Inquirer is a list of words with polarity labels such as "Positiv" and "Negativ", which we also used for our evaluation.

Kamps et al. (2004) proposed a thesaurus-based method for adjectives that uses a network constructed by connecting each pair of synonymous words provided by WordNet (Fellbaum, 1998) in which the shortest paths to two seed words, "good" and "bad," are used to obtain the polarity of a word. This method is attractive in terms of computational cost, but the shortest paths are sensitive to local disturbances in the network topology.

Similarly, Velikovich et al. (2010) developed a method that aggregates a huge amount of unlabeled corpus data from the Web and constructs a lexical network. They used a graph propagation algorithm, in which the weighted shortest paths from seed words are used.

Kaji and Kitsuregawa (2007) proposed a method for constructing a Japanese polarity lexicon from Web data. They collected positive (negative) sentences from the Web using structural clues such as HTML tags and then extracted polar phrases from them.

The method proposed by Rao and Ravichandran (2009) increases robustness against network disturbances. It treats polarity detection as a semi-supervised label propagation problem in a graph in which higher order correlations of network topology other than the shortest paths are involved. They showed that the *label propagation* algorithm (Zhu and Ghahramani, 2002) leads to a significant improvement in the accuracy of polarity detection for WordNet-based networks compared to various known heuristics. The label propagation is used for word polarity extraction also in the recent literature (Speriosu et al., 2011; Brody and Diakopoulos, 2011).

The linearized model in Section 5.3 can also be interpreted as a graph kernel, which is used in natural language processing (e.g., Komachi et al. (2008)).

3 Ising spin model

3.1 Overview of Ising spin model

For later analysis, we briefly summarize the basic notation and techniques of Ising spin systems. In general, the Ising spin model is composed of N binary variables termed (*Ising*) *spins*: $\boldsymbol{S} = \left(S_1, S_2, \ldots, S_N\right)$, where $S_i \in \{+1, -1\}$ $(i = 1, 2, \ldots, N)$, for which energy function

$$E(\boldsymbol{S}, \beta) = -\beta \sum_{i>j} J_{ij} S_i S_j - \sum_{i=1}^{N} h_i S_i \tag{1}$$

is defined. Here, J_{ij} represents the efficacy of interactions between two spins, S_i and S_j, and h_i stands for the external field added to S_i, and β is called the inverse temperature. The most fundamental assumption of equilibrium statistical mechanics is that, when a system is characterized by energy function $E(\boldsymbol{S})$, the probability that microscopic state $\boldsymbol{S}$ is generated in equilibrium at temperature T $(= \beta^{-1} > 0)$ is provided by the Boltzmann-Gibbs distribution:

$$P(\boldsymbol{S}) = \frac{e^{-E(\boldsymbol{S}, \beta)}}{Z}, \tag{2}$$

where Z is the normalization factor. This equation assigns a larger probability to microscopic state $\boldsymbol{S}$ if it has a lower value of energy (Equation (1)). This tendency is more/less significant when β is set to a higher/lower value. Given this assumption, the main task of statistical mechanics is to evaluate the averages of various quantities using Equation (2), which is unfortunately computationally difficult.

A family of mean field approximations offers a practical solution for resolving this difficulty (Opper and Saad, 2001). The approximations are calculated on the basis of the Kullback-Leibler divergence between Equation (2) and test distribution $Q(\boldsymbol{S})$: $D(Q||P) = \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \log(Q(\boldsymbol{S})/P(\boldsymbol{S})) = \log(Z) + F[Q, \beta]$, where

$$F[Q, \beta] = \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) E(\boldsymbol{S}, \beta) + \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \log Q(\boldsymbol{S}) \tag{3}$$

is termed the variational free energy. $D(Q||P)$ is generally non-negative and vanishes if and only if $Q(\cdot) = P(\cdot)$, which means that minimizing $F[Q, \beta]$ with respect to test distribution $Q(\cdot)$ leads to the correct evaluation of the true distribution, $P(\cdot)$.

A naïve approximation is derived by limiting the test distribution to one in factorizable form:

$$Q(\boldsymbol{S}) = \prod_i \frac{1 + m_i S_i}{2}. \tag{4}$$

Here, m_i denotes the mean of spin S_i with respect to the test distribution, which parameterizes the marginal distribution as $Q_i(S_i) = \sum_{\boldsymbol{S} \backslash S_i} Q(\boldsymbol{S}) = (1 + m_i S_i)/2$. Here, $A \backslash x$ generally denotes exclusion of x from set A. Plugging Equation (4) into Equation (3) and minimizing $F[Q, \beta]$ with respect to m_i $(i = 1, 2, \ldots, N)$ yield the *mean field equation*:

$$m_i = \tanh\left(\beta \sum_j J_{ij} m_j + h_i\right). \tag{5}$$

In many cases, Equation (5) can be solved by iterative substitution. Its computational cost is at most $O(N^2)$ per update, which is much lower than that required for the exact evaluation of the spin average, $O(2^N)$.

3.2 Analytical analysis

Although obtaining the exact solution is technically difficult, one can still handle Equation (5) analytically to a certain extent if $h_i = 0$ $(i = 1, 2, \ldots, N)$ is satisfied. For this, we set $h_i = 0$ $(i = 1, 2, \ldots, N)$ and use a Taylor series, $\tanh(x) = x - x^3/3 + \ldots$, on the right hand side of Equation (5), which yields

$$m_i = \sum_j \beta J_{ij} m_j - \frac{1}{3}\left(\sum_j \beta J_{ij} m_j\right)^3 + \ldots. \tag{6}$$

This guarantees that Equation (5) possesses the trivial solution[1] $m_i = 0$ $(i = 1, 2, \ldots, N)$. When T (β) is sufficiently high (low), this solution minimizes Equation (3) under the constraint of Equation (4) since the second term (the negative entropy part) of Equation (3) is dominant and provides an appropriate approximation. Let us decompose (symmetric) matrix $\boldsymbol{J} = (J_{ij})$ as $\boldsymbol{J} = \sum_{\mu=1}^{N} \lambda_\mu \boldsymbol{x}_\mu (\boldsymbol{x}_\mu)^{\mathrm{tr}}$, where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N$, and $\boldsymbol{x}_\mu$ $(\mu = 1, 2, \ldots, N)$ are eigenvalues of $\boldsymbol{J}$ and the corresponding unit eigenvectors, respectively, and tr denotes the matrix transpose. As explained in Appendix A, since the eigenvalues of the Hessian of the mean field free energy (Equation (16)) are given as $\beta^{-1} - \lambda_\mu$ $(\mu = 1, 2, \cdots, N)$, the trivial solution becomes unstable in the direction of $\boldsymbol{x}_1$ when

$$1 - \beta\lambda_1 = 0, \tag{7}$$

breaking the stability condition that all eigenvalues of the Hessian are positive. When no solutions other than the trivial one coexist, which is experimentally confirmed in our system shown later, this signals the onset of a phase transition to a non-trivial solution. In particular, when $J_{ij} \geq 0$ holds for all spin pairs, which is the case for models of ferromagnetism, every element of the critical eigenvector $\boldsymbol{x}_1$ has an identical sign. As a consequence, the non-trivial solution is characterized by a non-vanishing value of *magnetization*:

$$m = \frac{1}{N}\sum_i m_i, \tag{8}$$

whose absolute value is kept non-negligible as N becomes large if and only if m_i $(i = 1, 2, \ldots, N)$ mostly have an identical sign being dominated by the components of $\boldsymbol{x}_1$. This is considered to correspond to the emergence of spontaneous magnetization in ferromagnetic materials, which is particularly referred to as the *ferromagnetic phase transition*.

More precisely, the analysis presented above means that, when the trivial solution is perturbed by weak fields $\boldsymbol{h} = (h_i)$ for $\beta < \lambda_1^{-1}$, the spin averages depend linearly on $\boldsymbol{h}$ (to the first order):

$$\boldsymbol{m} \simeq (\boldsymbol{I} - \beta\boldsymbol{J})^{-1}\boldsymbol{h}, \tag{9}$$

where $\boldsymbol{I}$ is the identity matrix. This can be derived from Equation (6) by ignoring the $O(\beta^3)$ terms, adding $\boldsymbol{h}$, and solving $\boldsymbol{m} \simeq \beta\boldsymbol{J}\boldsymbol{m} + \boldsymbol{h}$. An expression of eigenvalue decomposition, $(\boldsymbol{I} - \beta\boldsymbol{J})^{-1} = \sum_{\mu=1}^{N} (1 - \beta\lambda_\mu)^{-1} \boldsymbol{x}_\mu (\boldsymbol{x}_\mu)^{\mathrm{tr}}$, suggests that the phase transition signaled by Equation (7) is brought about by divergence of the sensitivity of the spin averages with respect to the addition of infinitesimal external fields that are proportional to $\boldsymbol{x}_1$.[2]

[1] Trivial solution is called *paramagnetic solution* in statistical mechanics.

[2] $\boldsymbol{x}_\mu (\boldsymbol{x}_\mu)^{\mathrm{tr}}$ works as an operator that projects a vector $\boldsymbol{h}$ to the direction of $\boldsymbol{x}_\mu$ when multiplied as $\boldsymbol{x}_\mu (\boldsymbol{x}_\mu)^{\mathrm{tr}} \boldsymbol{h}$.

4 Original Method

The original method (Takamura et al., 2005) is based on a lexical network constructed from three types of resources: a dictionary, a corpus, and a thesaurus. Although it is not easy to directly infer word polarities from those resources, it is relatively easy to obtain the tendency that two words have the same polarity as described in the next paragraph. One of the simplest probabilistic models that connect such tendency and the polarity assignment is the Ising spin model.

First, two words taken from the dictionary are linked if one of them appears in the gloss of the other. Each link is classified as either *same orientation SL* or *different orientation DL*. Next, synonyms, antonyms, and hypernyms taken from the thesaurus are connected by the link. Only antonym links are categorized as *DL*. Two adjectives are connected if they appear in a conjunctive form in the corpus (Hatzivassiloglou and McKeown, 1997).

After the links are provided, the weight of each link is set:

$$J_{ij} = \begin{cases} \frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in SL) \\ -\frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in DL), \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

in order to introduce interactions into the spin model. Here, l_{ij} denotes the link between words i and j, and $d(i)$ denotes the degree of word i, i.e., the number of words linked directly to word i.

We assume that a small set of seed words, for which the correct polarities are known, is available. This assumption can be implemented into the spin model by imposing sufficiently large external fields on the spins of the initial word set:

$$E(\boldsymbol{S}, \beta) = -\beta \sum_{i>j} J_{ij} S_i S_j - \alpha \sum_{i \in L} a_i S_i, \tag{11}$$

where L denotes the initial word set (set of seed words), a_i denotes polarity (± 1) of a seed word i, and α denotes the strength of the external field of the seed words.

Substituting Equation (11) into Equation (2) yields the joint probability that all words simultaneously have semantic orientations labeled by $\boldsymbol{S} = (S_i)$ and $P(\boldsymbol{S})$. $P(\boldsymbol{S})$ can be used to evaluate marginal distribution $P(S_i) = \sum_{S \setminus S_i} P(\boldsymbol{S})$, which stands for the probability that word i has a polarity of S_i.

Given the marginal probabilities, the Bayesian framework guarantees that assigning a polarity $\sigma_i = \mathrm{argmax}_{S_i}\{P(S_i)\}$ to word i maximizes the rate of correct classification (Iba, 1999). In practice, this can be approximated as $\sigma_i = \mathrm{sign}(m_i) = m_i/|m_i|$ by solving the mean field equation (Equation (5)).

The performance of the method was evaluated for a lexical network constructed from WordNet (Fellbaum, 1998), the Wall Street Journal and Brown corpora of the Penn Treebank (Marcus et al., 1993), partially using TreeTagger (Schmid, 1994). The network was composed of 88,015 nodes (words) that were sparsely connected as they were characterized by a power-law-type degree distribution with an average of 18.94 (Figure 1). Only 5.63% of the weights were negative, which suggests that the spin system had a tendency to exhibit the

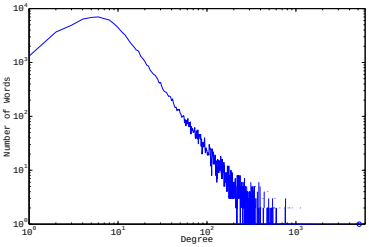

Figure 1: Degree distribution of constructed lexical network (log-log plot). Average degree was 18.94; maximum was 5144 (circle on the horizontal axis).

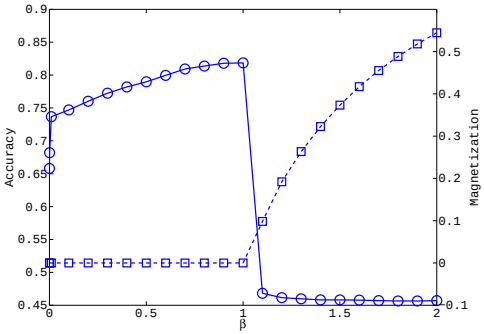

Figure 2: Temperature dependence of classification accuracy (circles) and magnetization (squares) for original method for 14 seed words.

ferromagnetic phase transition described above. The performance was assessed by using the General Inquirer labeled word list (Stone et al., 1966) as the gold standard. Of the 88,015 words in the network, 3596 were included in this list. Of these words, 1616 were positive, and 1980 were negative.

Testing was performed by varying β from 0.1 to 2.0 by 0.1 and setting α to 1000. The number of fixed seed words ranged from 2 to 14: {good, bad}, {good, superior, bad, inferior}, and {good, nice, excellent, positive, fortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior}.

Figure 2 shows how the classification accuracy (rate of classifying words correctly) and magnetization depended on temperature for the case of 14 seed words. Classification accuracy was maximized at $\beta \simeq 1.0$, where non-zero magnetization appeared, signaling the ferromagnetic phase transition, but deteriorated drastically as soon as β was raised to $\beta > 1.0$. Similar behavior was observed for four and two seed words.

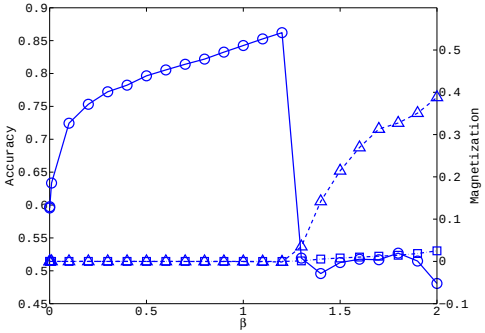

Figure 3: Classification accuracy and magnetization with improved method for 14 seed words. Circles ∘ denote classification accuracy. Squares □ and triangles △ denote magnetization evaluated using all words and only the 3596 labeled words, respectively.

5 Improved Method

5.1 Attention to Largest Eigenvalue

The setting and results described above naturally led to the following considerations.

1. Because external fields are added to only a small number of seed words, a spin system operating at high temperature (small β) can be handled as a perturbed state from the trivial solution.

2. When the temperature is lowered from a sufficiently high value, classification accuracy monotonically improves until non-vanishing magnetization appears. This suggests that the ferromagnetic phase transition is the main cause of the drastic performance deterioration.

The second consideration suggests that the classification accuracy can be improved by preventing the ferromagnetic phase transition. Equations (7) and (9), in conjunction with the first consideration, imply that the phase transition is caused by divergence of the sensitivity matrix, $(\boldsymbol{I}-\beta\boldsymbol{J})^{-1}=\sum_{\mu=1}^{N}(1-\beta\lambda_\mu)^{-1}\boldsymbol{x}_\mu(\boldsymbol{x}_\mu)^{\mathrm{tr}}$, in the direction of $\boldsymbol{x}_1$. This means that a possible way to prevent this transition is to simply expurgate the λ_1 component from weight matrix $\boldsymbol{J}=(J_{ij})$:

$$\boldsymbol{J}'=\boldsymbol{J}-\lambda_1\boldsymbol{x}_1(\boldsymbol{x}_1)^{\mathrm{tr}}. \tag{12}$$

Figure 3 shows the profiles of the classification accuracy and magnetization versus β for the modified weight (Equation (12)) for 14 seed words. Similar profiles were obtained for the other two cases. Note that solving the mean field equation for $\boldsymbol{J}'$ does not increase the computational cost significantly; $\boldsymbol{x}_1$ can be obtained using the power method, for which the computational cost is similar to that of solving the original mean field equation. The $\lambda_1\boldsymbol{x}_1(\boldsymbol{x}_1)^{\mathrm{tr}}$ term requires only $O(N)$ computations per update in the iterative substitution scheme.

Seeds	SP	LP	Original	Improved	Improved (II)
2	70.0	74.8(0.6)	75.2(0.8)	84.5(1.2)	84.4(1.2)
4	70.0	74.2(0.7)	74.4(0.6)	83.5(1.2)	83.7(1.1)
14	74.2	81.6(0.9)	81.9(1.0)	86.2(1.2)	86.2(1.2)

Table 1: Optimal classification accuracy (%) for 2, 4, 14 seed words, and the cross-validation setting. "SP" corresponds to the method based on shortest path. "LP" corresponds to label propagation. "Original" and "Improved" correspond to original method (Takamura et al., 2005) and one based on Equation (12). "Improved (II)" is same as "Improved" except that second largest eigenvalue component, λ_2, is also removed from Equation (12). Values in parentheses are β value at which accuracy was optimized.

As we speculated, the ferromagnetic phase transition was prevented, as shown by the magnetization for "all words" (squares) in Figure 3. As a consequence, classification accuracy was improved beyond $\beta \simeq 1.0$. Classification performance was evaluated on General Inquirer as in the previous work (e.g., (Turney and Littman, 2003)). Table 1 shows the optimal classification accuracy achieved for the original method (Original) and two improved methods, together with two existing state-of-the-art algorithms (SP and LP). SP is the method based on the shortest-path from seed words on the network (Velikovich et al., 2010). LP is the label propagation (Rao and Ravichandran, 2009). Note that, since we are interested in the impact made by the choice of algorithms, both of SP and LP were test on the lexical network that we used for our method, except that the edges with negative weights are removed because SP and LP cannot work properly with negative weights.[3] Also, although the label propagation by Rao and Ravichandran (2009) did not have the parameter β, we introduced it to LP for the purpose of fair comparison. Its value is optimized on the test set.

The performance of Original was improved in all cases by using the proposed scheme (Improved). All the differences were statistically significant in the sign test with significance level of 1%. The result also shows that the improved methods are significantly better than the shortest-path based method and the label propagation.[4] Note that the increase in accuracy compared with the values previously reported in some other papers (e.g., 82.8% (Turney and Littman, 2003) and 82.2% (Esuli and Sebastiani, 2005)) is substantial.

The results with a few seed words (e.g., 2, 4, and 14) are more important since semi-supervised methods should be applied to resource-scarce languages or new domains where creating a large amount of seed words is not very practical. However, in order to examine the behavior of our method when we have many seed words, we employed 10-fold cross validation (i.e., approximately 3,200 seed words). The accuracy of 91.5% (Original) was slightly improved to 91.8 (Improved) and 91.9 (Improved (II)).

5.2 Removal of more eigenvalue components

Figure 3 also shows that the performance still deteriorated at a higher β value (1.3). To clarify the reason for this, we also plotted "selected magnetization" (triangles), which was evaluated

[3] The label propagation is not guaranteed to converge in the presence of negative weights.

[4] Yu et al. (2011) used the shortest-path based method on an extended lexical network, and compared it with the label propagation on a non-extended lexical network, resulting in a better performance of the former method. The combination of their extended network with our algorithm would be a promising piece of future work.

using only the 3596 labeled words for checking a possibility that a certain phase transition relevant to only the labeled words brings about the deterioration. The plot indicates that the selected magnetization bifurcates to a finite value for $\beta \simeq 1.3$. This indicates that another phase transition occurred due to the second largest eigenvalue component, λ_2.

However, this indication is only partially correct. Figure 4 plots the profiles of the first and second eigenvectors. While the components of the first eigenvector (top) are evenly spread across almost all sites, those of the second one (bottom) are significantly localized at several sites. This "localization" feature is actually quite common for the eigenvectors of many other relatively large eigenvalues. Figure 5 plots the 30 largest eigenvalues and the inverse participation ratio (IPR), which is defined as

$$\mathrm{IPR} = \frac{\sum_i v_i^4}{(\sum_i v_i^2)^2}, \tag{13}$$

for a real vector, $\boldsymbol{v} = (v_i)$, of their eigenvectors (Biroli and Monasson, 1999). This quantity IPR takes a value between 0 and 1. In particular, as the dimensionality tends to infinity, it remains positive for localized vectors but vanishes for spread ones, so this quantity is widely used as a standard measure for characterizing the localization property of high dimensional vectors.

The plots in Figure 5 show that, although λ_1 is isolated, many other eigenvalues are distributed in a rather degenerated manner and are accompanied by localized eigenvectors. The localized eigenvectors have non-negligible values only for a few elements. Therefore, removal of one of such components does not provide significant effects for most spins. This suggests that the bifurcation of the selected magnetization is caused by not only the λ_2 component but also by many other components that are simultaneously excited at $\beta \simeq 1.3$. Accordingly, little improvement in the classification accuracy can be gained by removing the components of the second largest eigenvalue component from the weight matrix, as shown in the rightmost column of Table 1. The performance was not improved significantly by further removing the third and the fourth eigenvalue components.

Note that most of the edges in the current network have positive weights, i.e., very biased to the same sign. This is the reason why the first eigenvector is not localized and the first eigenvalue is much larger than the others. Also, the localization of eigenvectors with large eigenvalues, which is observed except for the first eigenvalue in the current network, often happens in *complex network*. The theoretical background can be found in the paper by Kabashima and Takahashi (2012). Therefore, the characteristics of the network is not completely accidental.

5.3 What is Essential?

The analysis presented so far indicates that high classification performance is achieved for a relatively small β, at which no magnetization appears unless external fields are imposed. Figure 6 plots the average of spins obtained using the improved method for $\beta = 1.2$ and 1.3 for 14 seed words. Most of the absolute values of the spin averages were rather small when a high classification accuracy was gained ($\beta = 1.2$, top) and became much larger after the performance deteriorated due to a phase transition ($\beta = 1.3$, bottom). When the spin averages have small absolute values, one can handle the system as a state produced by slightly perturbing the trivial solution by external fields imposed on only the seed word spins. This suggests a possibility that the linear response for the external fields plays a key role for the

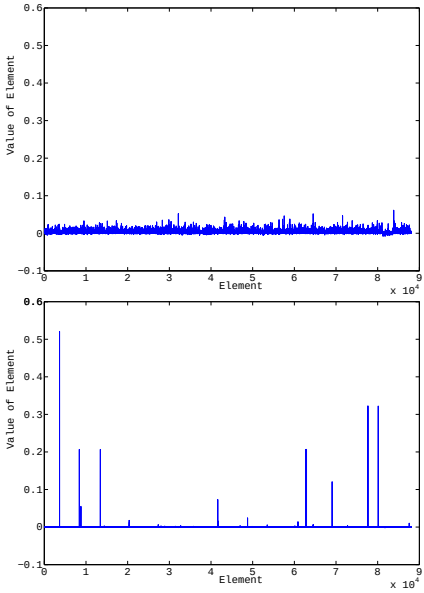

Figure 4: Eigenvectors of λ_1 (top) and λ_2 (bottom) of weight matrix $\boldsymbol{J}$.

Seeds	14	4	2
Accuracy	81.6(0.9)	74.2(0.7)	74.8(0.7)

Table 2: Optimal classification accuracy (%) of linearized model. Values in parentheses are β at which accuracy was optimized.

high classification accuracy. For checking this possibility, we examined the performance of a simplified model defined by the linear approximation of Equation (9):

$$\boldsymbol{m} = (\boldsymbol{I} - \beta \boldsymbol{J}')^{-1} \boldsymbol{h}^0, \tag{14}$$

where $\boldsymbol{h}^0 = (h_i^0)$ is provided as $h_i^0 = \alpha a_i$ if i is included in the set of seed words and vanishes otherwise. A power series expression, $(\boldsymbol{I} - \beta \boldsymbol{J}')^{-1} = \sum_{n=0}^{\infty} (\beta \boldsymbol{J}')^n$, indicates that this can be practically assessed by iterating the recursive equations:

$$\boldsymbol{m}^{t+1} = \boldsymbol{m}^t + \boldsymbol{u}^t \quad \text{and} \quad \boldsymbol{u}^{t+1} = (\beta \boldsymbol{J}') \boldsymbol{u}^t, \tag{15}$$

a sufficient number of times by setting the initial conditions as $\boldsymbol{m}^0 = \boldsymbol{0}$ and $\boldsymbol{u}^0 = \boldsymbol{h}^0$. This can be carried out with a computational cost similar to that of the naive mean field method.

Table 2 shows the optimal classification accuracy obtained with the linearized model. These results, in conjunction with those in Table 1, indicate that the performance of the linearized model was worse than that of the improved method, but similar to those of the original

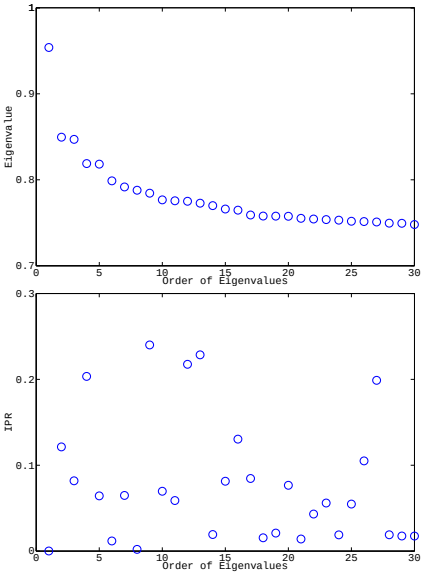

Figure 5: 30 largest eigenvalues of $\boldsymbol{J}$ (top) and inverse participation ratio (IPR) for corresponding eigenvectors (bottom). Larger IPR suggests that the corresponding eigenvector is more localized.

one (Takamura et al., 2005). This suggests that much of the information needed to correctly classify words is included in the linear response of the spin averages to the polarities of the seed words at high temperature.

As shown in Figure 5, the eigenvectors of large eigenvalues, which are emphasized at low temperature, are mostly localized and could contain relevant classification information only for a few words that corresponds to non-negligible values of elements. Therefore, they are individually insufficient for correctly classifying most other words corresponding to negligible elements. This means that, assigning polarities at high temperature so that information for all spin alignments is summed up with moderate probabilities is essential in the current spin-model-based method. This is achieved by assessing the linear response of the trivial solution to external fields representing the polarities of the seed words in the simplified scheme of Equation (14), and employment of the mean field equation (Equation (5)) offers a further gain under favor of the nonlinearity effect of tanh(·). This is in contrast to other approaches in which a single state that (approximately) optimizes a certain cost function is used for determining polarities (Blum and Chawla, 2001; Blum et al., 2004; Choi and Cardie, 2009).

A question that arises would be whether or not the removal of the largest eigenvalue component improves the linearized model. To answer this question, we conducted more experiments with the linearized model with the largest eigenvalue component being removed. The result

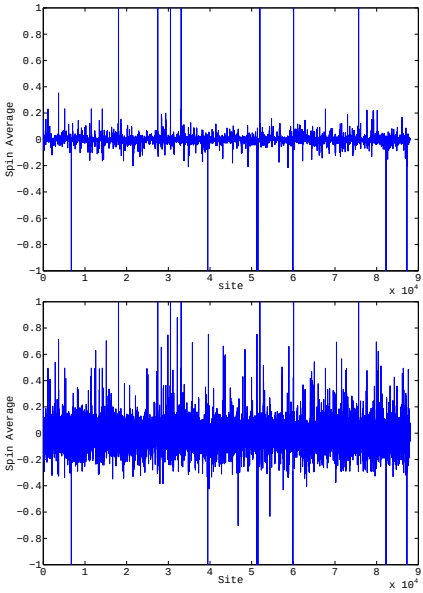

Figure 6: Spin averages m_i for $\beta = 1.2$ and 1.3 for 14 seed words. The mean of the absolute value of $|m_i|$ increased more than ten times, from 0.0028 ($\beta = 1.2$, top) to 0.0310 ($\beta = 1.3$, bottom).

is shown in Table 3. The comparison with the result in Table 2 suggests that the removal of the largest eigenvector also improves the linearized model. We note that the linearized model is almost equivalent to the label propagation (see the Taylor series in (Equation (6)); the difference is simply the presence of the normalization step. In fact, the result of the linearized model (Table 2) is almost the same as that of the label propagation (Table 1). The experimental result in Table 3 suggests that an idea similar to the removal of eigen components might also improve the label propagation, although we need to overcome the difficulty that the label propagation is not guaranteed to work properly if some edge weights are negative.

Conclusion

We have provided an analytical analysis of the behavior of a previously proposed spin-model-based method for constructing a polarity lexicon from the viewpoint of statistical mechanics.

Seeds	14	4	2
Accuracy	85.3(1.1)	83.8(1.1)	83.8(1.1)

Table 3: Optimal classification accuracy (%) of linearized model with the largest eigenvector being removed. Values in parentheses are β at which accuracy was optimized.

On the basis of this analysis, we proposed a scheme for improving the performance of polarity lexicon extraction, i.e., removing the largest eigenvalue component from the weight matrix of the lexical network, the result is quite significant. For example, classification accuracy was increased from 75.2 to 84.5% for the case of two seed words without significantly increasing computational cost. This scheme also improves the linearized model.

We also examined the possibility of improving the performance further by removing more eigenvalue components. However, the resulting degeneracy of the eigenvalues in the weight matrix, which is accompanied by eigenvector localization, minimizes the gain improvement. In addition, we investigated a linearized model to characterize the classification performance and found that the linear response to the polarities of the seed words at high temperature contains essential information. While many methods have been proposed for binary classification, apparently most of them are based on optimization of a certain cost function or on achievement of the low-temperature state of the Boltzmann-Gibbs distribution. In general, high-temperature states are technically easier to handle than low-temperature ones because a greater variety of perturbative techniques can be used. The utility of the (linear) response in the high-temperature state shown here offers a novel promising approach to generic classification when labels are provided for a small fraction of representative instances.

The developed methodology can be employed for general purposes of assessing influences of a few representative nodes in a network via local communications. The Ising spin model or similar models including its linearized model are used in a number of tasks in natural language processing.

Future work includes more use of language data and development of applications using this polarity lexicon construction method as well as use of other approximation schemes such as advanced Markov chain Monte Carlo methods in which equilibration is significantly accelerated by using extended ensembles (Iba, 1999).

Acklowdedgments

This work was partially supported by JSPS KAKENHI No. 22300003 and Mitsubishi Foundation (YK).

A Stability of Trivial Solution

Inserting Equation (4) into Equation (3) and setting h_i to zero ($i = 1, 2, \ldots, N$) yields an expression of the *mean field free energy*:

$$F_{\mathrm{MF}}(\boldsymbol{m}) \quad = \quad -\beta \sum_{i>j} J_{ij} m_i m_j + \sum_{i=1}^{N} \sum_{S_i=\pm 1} \frac{(1+m_i S_i)}{2} \log \frac{(1+m_i S_i)}{2}. \tag{16}$$

To examine the local stability of the paramagnetic solution, $m_i = 0$ ($i = 1, 2, \ldots, N$), we evaluated the Hessian of $F_{\mathrm{MF}}(\boldsymbol{m})$:

$$\boldsymbol{H} = \left(\left. \frac{\partial^2 F_{\mathrm{MF}}(\boldsymbol{m})}{\partial m_i \partial m_j} \right|_{\boldsymbol{m}=\boldsymbol{0}} \right) = -\beta \boldsymbol{J} + \boldsymbol{I}. \tag{17}$$

The solution is locally stable if and only if $\boldsymbol{H}$ has no negative eigenvalues. Equation (17) indicates that the eigenvalues of $\boldsymbol{H}$ are given as $\beta^{-1} - \lambda_\mu$ ($\mu = 1, 2, \ldots, N$) using the eigenvalues of $\boldsymbol{J}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. Thus, as β increases from a very low value, the stability condition that all eigenvalues are positive is broken when $\beta^{-1} - \lambda_1 = 0$ holds, i.e., Equation (7).

References

Alexandrescu, A. and Kirchhoff, K. (2009). Graph-based learning for statistical machine translation. In *Proceedings of Human Language Technologies and the Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2009*, NAACL '09, pages 119–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Biroli, G. and Monasson, R. (1999). A single defect approximation for localized states on random lattices. *Journal of Physics A: Mathematical and General*, 32(24): pages L255–L261.

Blum, A. and Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In Brodley, C. and Danyluk, A., editors, *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*, pages 19–26, Williams College, US. Morgan Kaufmann Publishers, San Francisco, US.

Blum, A., Lafferty, J., Rwebangira, M. R., and Reddy, R. (2004). Semi-supervised learning using randomized mincuts. In *Proceedings of the 21st international conference on Machine learning (ICML '04)*, pages 97–104, New York, NY, USA. ACM.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Choi, Y. and Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for dom ain-specific sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 590–598.

Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series*. MIT Press.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

Hu, J., Wang, G., Lochovsky, F., tao Sun, J., and Chen, Z. (2009). Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 471–480, New York, NY, USA. ACM.

Iba, Y. (1999). The Nishimori line and Bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21): pages 3875–3888.

Kabashima, Y. and Takahashi, H. (2012). First eigenvalue/eigenvector in sparse random symmetric matrices: influences of degree fluctuation. *Journal of Physics A: Mathematical and Theoretical*, 45(32): 325001 (19pp).

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083.

Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (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.

Komachi, M., Kudo, T., Shimbo, M., and Matsumoto, Y. (2008). Graph-based analysis of semantic drift in espresso-like bootstrapping algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1011–1020.

Li, X., Wang, Y.-Y., and Acero, A. (2008). Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 339–346, New York, NY, USA. ACM.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Opper, M. and Saad, D. (2001). *Advanced Mean Field Methods: Theory and Practice*. MIT Press.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 685–682.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 53–63.

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

Takamura, H., Inui, T., and Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. (2010). The viability of web-derived polarity lexicons. In *Proceedings of Human Language Technology and the Annual Conference of the North American Chapter of the Association for Computational Linguistic (HLT-NAACL 2010)*, pages 777–785.

Yu, M., Wang, S., Zhu, C., and Zhao, T. (2011). Semi-supervised learning for word sense disambiguation using parallel corpora. In *Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 1490–1494.

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.