Identifying Sentiments in N-grams

Noriaki Kawamae
NTT Comware
1-6 Nakase Mihama-ku Chiba-shi, Chiba 261-0023 Japan
kawamae@gmail.com

ABSTRACT

Our proposal, identifying sentiments in N-grams (ISN), focuses on both word order and phrases, and the interdependency between specific ratings and corresponding sentiments in texts to detect subjective information.

Categories and Subject Descriptors

I.7 [**DOCUMENT AND TEXT PROCESSING**]: Document analysis

General Terms

Algorithms, experimentation

Keywords

Sentiment Analysis, N-gram topic model

1. INTRODUCTION

ISN aims to capture both ratings and their corresponding sentiment phrases from reviews in an integrated manner; other models detect them in separate steps. This model assumes that N-gram phrases can carry more meaning than the sum of the individual words, convey different meaning depending on their context, and express sentiment jointly with the given rating value. ISN extends topic models such that topic discovery is influenced by not only word co-occurrence but also rating information. ISN automatically removes noise words (such as typos and jargon) as document-specific words, and groups synonym terms into topics without any human supervision or dictionaries, and so, unlike the conventional two stage approaches, avoids the limitation of depending on these supports.

2. IDENTIFYING SENTIMENT PHRASES

Table 1 shows the notations used in this paper; Figure 1 shows the graphic model of ISN. ISN extends sTOT [1] to form N-grams through the concatenation of consecutive topics. This model incorporates v instead of t, and, in each token, empowers r to handle more status in word generation and connect the current topic with the previous topic.

The innovations are to use the document-specific word distribution for word selection, and deciding whether to form a topic bigram by concatenating the current topic with the previous topic, and then selecting the previous work-specific word distribution in each token. For distinguishing these differences in word tokens, we define r as a switch for handling more kinds of statuses as follows. If r_i=0 (1), ISN generates word w_i from the background word distribution ϕ_b (the document-specific word distribution ϕ_d). If r_i=2 (3), ISN selects topic z_i from the document-specific topic distribution θ_d (the previous topic specific topic distribution $\psi_{\bar{z}}$), and then generates word w_i from topic-specific word distribution ϕ_z (this current topic and previous word-specific word distribution $\varphi_{w_{i-1}z_i}$). These approaches allow ISN to predict the absolute rating value of a review article, and, conversely, the word/phrase distribution given a rating.

ISN can be inferred by Gibbs sampling in the same way used for previous models without loss of generalization. For each token in the Gibbs sampling procedure, we use the chain rule and then obtain the predictive distribution of adding word w_{di} in document d to the topic z_{di} as $p(r_{di}, z_{di} | z_{d(i-1)} = j, w_{d(i-1)} = u, \mathbf{z}_{\setminus di}, \mathbf{r}_{\setminus di}, \mathbf{w}, \alpha, \beta, \gamma, \epsilon)$; it is written

Table 1: Notations used in this paper

SYMBOL	DESCRIPTION
D (Z, W)	number of documents (topics, words)
N_d	number of word tokens in document d
v_d	the rating associated with document d
r_i	the switch associated with the ith token
z_i	the topic associated with the ith token
w_i	the ith token
θ_d	the document d specific multinomial distribution of topics ($\theta \vert \alpha \sim$ Dirichlet(α))
$\phi_{z(b,d)}$	the topic z (background b, document d) specific multinomial distribution of words ($\phi_{z(b,d)} \vert \beta \sim$ Dirichlet(β))
$\varphi_{\bar{w}z}$	the previous word w and current topic z specific multinomial distribution of next words ($\varphi_{wz} \vert \gamma \sim$ Dirichlet(γ))
$\psi_{\bar{z}}$	the previous topic z specific multinomial distribution of next topics ($\psi_{\bar{z}} \vert \delta \sim$ Dirichlet(δ))
μ_d	the document d specific multinomial distribution of r_{di}($\mu_d \vert \epsilon \sim$ Dirichlet(ϵ))
λ_z	the topic z specific beta distribution of v
$\alpha, \beta, \gamma, \delta, \epsilon$	the fixed parameters of symmetric Dirichlet priors of ($\theta, \phi, \varphi, \psi, \mu$)

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

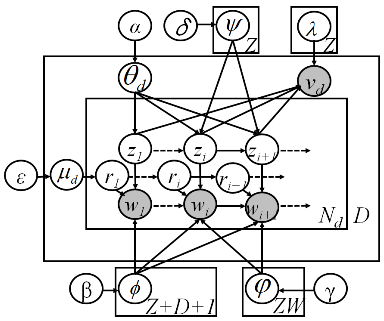

Figure 1: Graphic Model of ISN: In this figure, shaded and hollow variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and stacked panes indicate repeated sampling with the iteration number shown.

Table 2: Details of data sets:

	DVD	Books	Music
# reviews	4360	4269	4158
# items	12	15	15
# words	11875	12512	13523

as

$$p(r_{di}, z_{di}|\cdots) \propto$$

$$\begin{cases} (n_{d0\backslash di}+\epsilon_0)\frac{(n_{bw_{di}\backslash i}+\beta_{w_{di}})}{\sum_w^W (n_{bw\backslash di}+\beta_w)}, \text{if} \quad r_{di}=0, \\ (n_{d1\backslash di}+\epsilon_1)\frac{(n_{dw_{di}\backslash i}+\beta_{w_{di}})}{\sum_w^W (n_{dw\backslash di}+\beta_w)}, \text{if} \quad r_{di}=1, \\ (n_{d2\backslash di}+\epsilon_2)\frac{n_{dk\backslash di}+\alpha_k}{\sum_z^Z n_{dz\backslash di}+\alpha_z}\frac{(n_{kw_{di}\backslash i}+\beta_{w_{di}})}{\sum_w^W (n_{kw\backslash di}+\beta_w)}\frac{(1-v_d)^{\lambda_{k1}-1}v_d^{\lambda_{k2}-1}}{B(\lambda_{k1},\lambda_{k2})}, \\ \text{if} \quad r_{di}=2 \quad and \quad z_{di}=k, \\ (n_{d3\backslash di}+\epsilon_3)\frac{n_{jk\backslash di}+\alpha_k}{\sum_z^Z n_{jz\backslash di}+\alpha_z}\frac{(n_{ukw_{di}\backslash i}+\delta_{w_{di}})}{\sum_w^W (n_{ukw\backslash di}+\delta_w)}\frac{(1-v_d)^{\lambda_{k1}-1}v_d^{\lambda_{k2}-1}}{B(\lambda_{k1},\lambda_{k2})}, \\ \text{if} \quad r_{di}=3 \quad and \quad z_{di}=k, \end{cases}$$

where $n_{d0(1,2,3)\backslash di}$ represents the number of switches that have been assigned to the background (document specific, current, current topic given the previous) topic in document d, except d_i, $n_{dk\backslash di}$ represents the number of tokens assigned to topic k in document d, except di, and B is the beta function with k specific shape parameters λ_{k1} and λ_{k2}, and $n_{bw(dw,kw,ukw)_{di}\backslash di}$ represents the number of word w_{di} in background topic (the document specific topic, current topic k, current topic k given previous word $\bar{w}_{d(i-1)}=u$), except d_i. Only a unigram is allowed at the beginning of a document, as this model can always generate a bigram depending on nearby context. Therefore, we constrain topic assignment to be considered from the next word (di>1).

3. EXPERIMENTS

To evaluate the proposed model we use Amazon review data[1]: We normalized rating scores to the range [0,1] and then assigned these scores as v to each article, and selected 3 categories based on the number of reviewed products, and then split this data set into three data sets according to product type. The data sets were tokenized automatically without using a stop word list. Details of the sets are shown in Table 2. In our evaluation, the smoothing parameters α, β, γ, δ and ϵ were set to 1/Z, 0.1, 0.1, 1/Z and 0.25, respectively (all weak symmetric priors following previous work). The number of topics, $|Z|$, was set to 12 (DVD), 15 (Book), and 15 (Music); a preliminary experiment confirmed that just one topic is enough for generating each item specific word. Additionally, we doubled the number of topics so that each topic with a high rating corresponds to a positive topic and one with a low score to a negative topic.

[1]Amazon Product Review Data (Huge): http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

Table 3: MAE (rating) comparison of NB, TNG, sTOT and ISN: TNG, sTOT and ISN were trained using the number of topics Z set at 24 (DVD) and 30 (Music, DVD). Results that differ significantly, t-test $p < 0.01$, $p < 0.05$, from ISN are marked with '', '*' respectively.**

	NB	TNG	sTOT	ISN
DVD	0.317	0.263	0.225	0.208^{**}
Book	0.298	0.253	0.206	0.193^{*}
Music	0.321	0.273	0.228	0.217^{*}

We evaluate the predictive power given the words/sentiment words in a review. This evaluation aims to compare which model more precisely infers the rating score from just the word distributions. Given a review, we predict its rating by choosing the discretized rating that maximizes the posterior, which is calculated by multiplying the rating probability of all word tokens/N-grams from a topic-wise Beta distribution over rating $\prod_{i=1}^{n_d} p(v|\lambda_{z_{di}})$. As the baseline methods, we prepared Naive Bayes (NB), Topical N-grams (TNG) [2] model and sTOT, and then measured the difference between the predicted score and the correct rating score. Although original TNG does not output any rating score, we inserted this value in each token of TNG as the observed score v_d that is conditioned on each topic z and can be sampled from the topic specific Beta distribution ($v_d \sim \lambda_z$), like sTOT and ISN, so that TNG predicts the score. As shown in Table 3, ISN provides an average reduction in MAE relative error of 6.2%. This result also indicates that the manipulation of background topic is essential for describing the generative process of review articles. Most background topic words are reused over almost all reviews regardless of content, referred item, and corresponding score. TNG assigns background topic words to topics like other sentiment specific words. ISN detected the positive sentiment phrase “not disappointed”. On the contrary, previous models judged “disappointed” as a negative word, or this phrase is overwhelmed by the plurality of words generated under the bag of words assumption; this defect weakens predictive performance. This result confirms that ISN can realize automatic rating annotation and so offers rating-based item retrieval.

4. CONCLUSION

This paper introduced a generative model that detects both sentiments and the corresponding rating, simultaneously. Future work is to detect each aspect and then discover each individual reviewer's latent attitude with regard to to each aspect.

5. REFERENCES

[1] N. Kawamae. Trend analysis model: Trend consists of temporal words, topics, and timestamps. In *WSDM*, pages 317–326, 2011.

[2] X. Wang, A. McCallum, and X. Wei. Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702, 2007.