A translation approach for unstructured online reviews

Jongho Im, Taikgun Song, Jewoo Kim, and Youngsu Lee
Iowa State University, Department of Statistics, Ames, IA 50011

Introduction

Topic modeling techniques including probabilistic Latent Semantic Indexing (pLSI) by Hofmann (1999), Latent Dirichlet Allocation (LDA) by Blei et al.(2003), and other Bayesian analysis have been developed to analyze unstructured text data. However, such topic modeling techniques may not be as effective when it comes to online user-generated contents (UGCs) due to characteristics of the short text (data sparsity) or if there is a pre-selected topics to consider. In this poster, we propose Confirmatory Topic Modeling (CTM) for modeling short texts using paired biterms, common expressions, and pre-assigned topics.

Confirmatory Topic Modeling

1. Conventional Topic Model (Latent Dirichlet Allocation. Conceptually similar to explanatory factor analysis)

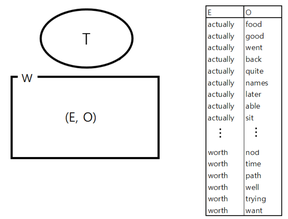
$$p(Word|Topic)p(Topic|\cdot)$$

   LDA treats topic as latent variable with Dirichlet prior and is interested in estimating $p(Word|Topic)$

2. Confirmatory Topic Modeling (Conceptually similar to confirmatory factor analysis)
   Assume where topics are pre-spcified and the data is short text. Unlike LDA, we are interested in p(Topic|Word) instead of p(Word|Topic). Words are classified into to groups: Evaluation words (E) and Objective words (O). Objective words (O) are directely related to the topic, however, Evaluation words (E) are related to the topics only through O words. $p(Topic, Word) \approx p(T(O)|E)$

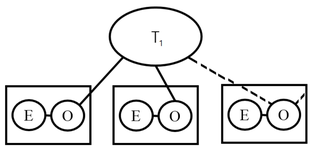
   - Step 0: Parse/split sentence (phrase) process where each sentence/phrase contains E and O.
   - Step 1: Given E, fit multivariate normal to $p(O|E)$.
     Select common expression where the cell probability $p(O|E) > p_0$, $p_0 \geq 0$
   - Step 2: Manually match selected biterms to pre-specified topics.
   - Step 3: Match unselected biterms to selected biterms from Step 1 using external information (e.g. online dictionary) or supervised machine learning. All unmatched biterms left after Step 3 is discarded.

Figure of Step 1 and its Example



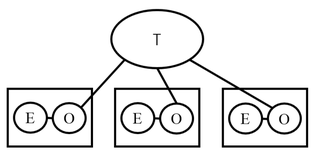
E	O
actually	food
actually	good
actually	went
actually	back
actually	quite
actually	names
actually	later
actually	able
actually	sit
⋮	⋮
worth	nod
worth	time
worth	path
worth	well
worth	trying
worth	want

Figure of Step 2 and its Example (alpha=0.1)



	E	O
1	actually	food
1	actually	good
1	actually	not
2	almost	every
2	almost	like
2	almost	not
3	amazing	not
3	amazing	place
3	amazing	service
⋮	⋮	⋮
203	worth	price
203	worth	trip
203	worth	try
203	worth	visit
203	worth	wait
203	worth	well

Figure of Step 3 and its Example



E	O
⋮	⋮
amazing	fish
amazing	fixings
amazing	flavor
amazing	flavors
amazing	flavours
amazing	flying
amazing	food
amazing	fool
amazing	found
amazing	free
amazing	french
amazing	fresh
amazing	fried
amazing	friendly
⋮	⋮

E	O
⋮	⋮
amazing	fillings
amazing	fires
amazing	fireworks
amazing	first
amazing	fixings
amazing	flavours
amazing	flying
amazing	food
amazing	fool
amazing	found
amazing	free
amazing	french
amazing	fresh
amazing	fried
⋮	⋮

Figure: Graphical display of the CTM process and its example

Application

- 7222 user generated Honolulu Restaurant reviews from Trip Advisor was used.
- LDA method and CTM method were conducted for comparison and the result follows.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	food	place	good	good	food	good
2	one	great	food	food	great	service
3	like	food	service	restaurant	good	great
4	time	just	menu	get	place	place
5	nice	good	best	back	menu	restaurant
6	restaurant	restaurant	chicken	really	best	one
7	back	lunch	delicious	like	one	get
8	service	little	restaurant	just	time	back
9	will	get	great	staff	delicious	went
10	excellent	dont	like	will	breakfast	best

Table 1: Example of LDA with 6 topic categories
Note topice specification could be difficult

Environment		Food		Price		Service	
E	O	E	O	E	O	E	O
amazing	cafe	amazing	dining	amazing	price	amazing	service
amazing	comfortable	amazing	cocktails	cheap	budget	amazing	dining
amazing	view	amazing	filet	cheap	dollars	amazing	server
amazing	watch	bad	food	cheap	prices	bad	wait
bad	environment	bad	nobu	good	price	bad	services
bad	place	best	food	great	price	beautiful	service
bad	restaurant	cheap	food	large	price	busy	server
beautiful	environment	cold	broth	very	price	busy	waiting
busy	place	delicious	food	reasonable	price	cheap	service
cheap	plastic	delicious	ribs	huge	price	friendly	service
cheap	glasses	perfect	food	worth	price	friendly	explains

Table 2: Selected example of CTM with 4 categories

Conclusions

Our proposed Confirmatory Topic Modeling method have advantage over conventional method for short text with pre-specified topics. Moreover, the natural structure of our method enables further access to sentiment analysis. However, limitation of our method is that it requires initial matching and the quality depends on the biterm decomposition