Incorporating Statistical Word Senses in Topic Model

Guoyu Tang

Outlines

- Introduction
- Related Work
- Topic Models Incorporating Statistical Word Senses
- Inference
- Evaluation
- Conclusion

Introduction(1/4)

- Topic model
  - Use document-level co-occurrence information to group semantically related words into a single topic.
- LDA
  - The topic distribution of the document
  - The probability of the topic to emit this word

Introduction(2/4)

- The probability of the topic to the word has some limitations.
  - Traditional LDA treats word as surface string,
- Example:
  - *Robot*
    - Usually mean an electro-mechanical machine
    - In a film review, it may refer to the name of a film
  - In LDA
    - The probability of topic *electronics technology* to emit the word is much higher than the topic *film*.
    - With word sense information
      - Probability of topic *film* to this word sense *film name* is higher than that of topic *electronics technology*

Introduction(3/4)

- We thus hypothesize that, if word senses are incorporated in topic models, a stronger indication of topic will be obtained.
- Topic models with word senses from lexical resources
  - *WordNet* (Boyd-Graber et al., 2007; Chemudugunta et al., 2008; Guo and Diab, 2011).
  - costly and hardly be complete.
- Word Sense Induction (WSI)
  - Discover word senses from unannotated text
  - Have been integrated in information retrieval to resolve senses of query words (Schutze and Pedersen, 1995; Navigli and Crisafulli, 2010).

Introduction(4/4)

- Two manners, i.e., sequential and co-inference, are proposed to incorporate the statistical word senses in the LDA framework.

- Hierarchical Dirichlet Process (HDP) (Teh et al., 2004) to induce statistical word senses from corpora

Related work(1/2)

- Semantic Document Representation Models
  - VSM
    - Ignore sematic relations.
  - Explicit Sematic Representation
    - The lexical ontologies are difficult to construct and can hardly be complete.
  - Latent Sematic Representation(Topic model)
    - Those models treat word as surface string.
    - One word may contain different meanings in different contexts
  - Integrate semantics from lexical resources into topic model framework
    - (Boyd-Graber et al., 2007; Chemudugunta et al., 2008; Guo and Diab, 2011).
    - The coverage issue again leads to performance bottleneck.

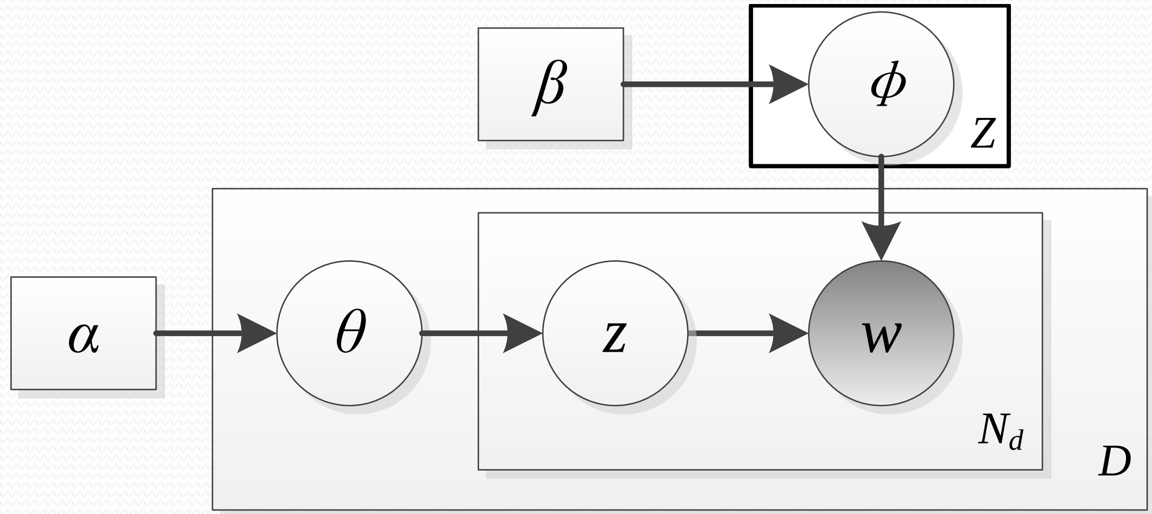
Related work(2/2)

- Word sense disambiguation and word sense induction.
  - The use of word sense
    - Information retrieval (Stokoe, 2003) and text classification (Tufi and Koeva, 2007).
    - Drawbacks:
      - Large, manually compiled lexical resources such as the WordNet database are required.
      - It is hard to decide the proper granularity of the word sense.
  - In this work, word sense induction (WSI) algorithm is adopted in automatically discovering senses of each word in the test dataset.
    - The Bayesian model (Yao and Durme ,2011)
      - HDP: find topic number automatically
      - It outperforms the state-of-the-art systems in SemEval-2007 evaluation (Agirre and Soroa, 2007).
    - Word sense induction algorithms have been integrated in information retrieval (Schutze and J. Pedersen, 1995; Navigli and Crisafulli, 2010).
      - The above researches only consider senses of words and do not investigate connection between words.

Topic Models Incorporating Statistical Word Senses

- Motivation
  - Synonymy
    - different words carrying almost identical or similar meanings.
  - Polysemy
    - one single word carrying two or more senses at the same time.
  - Topic is not able to reflect meaning of word delicately.
  - Incorporating word senses
    - A topic is more directly relevant to a word meaning (i.e., sense) than a word due to polysemy;
    - Word senses are more proper to reflect synonymy than words.

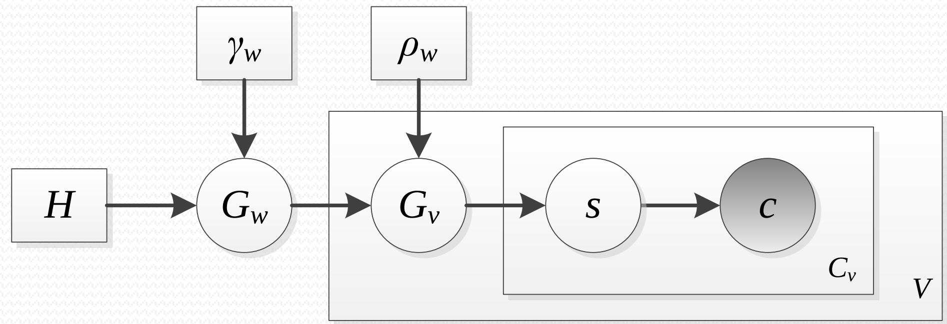
LDA



1. For each topic z:
   a) choose $\phi_z \sim Dir(\beta)$.
2. For each document d_i:
   a) choose $\theta_{d_i} \sim Dir(\alpha)$.
   b) for each word w_{ij} in document d_i:
      i. choose topic $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. choose word $w_{ij} \sim Mult(\phi_{z_{ij}})$.

$$P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{w}) \propto \frac{n^{d_i}_{-ij,z}+\alpha}{n^{d_i}_{-ij}+Z\alpha} \times \frac{n^{w}_{-ij,z}+\beta}{n_{-ij,z}+W\beta}$$

WSI with HDP Algorithm



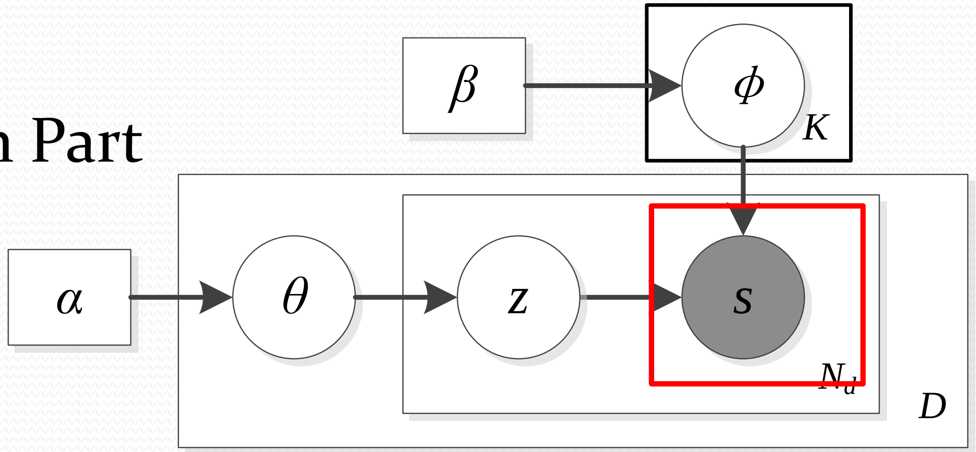
1. Choose $G_w \sim DP(\gamma_w, H)$.
2. For each context window v_i of word w:
   a) choose $G_{v_i} \sim DP(\rho_w, G_w)$.
   b) for each context word c_{ij} of target word w:
      i. choose $s_{ij} \sim G_{v_i}$.
      ii. choose $c_{ij} \sim Mult(\eta_{s_{ij}})$.

Incorporating Statistical Word Senses into Topic Model

- Sequential Approach (SEQ)
- Co-inference Approach (COI)

Sequential Approach (SEQ)

- Word Sense Induction Part
  - Same as HDP



- Document Presentation Part

1. For each topic z, choose $\phi_z \sim Dir(\beta)$.
2. For each document d_i :
   a) choose $\theta_{d_i} \sim Dir(\alpha)$.
   b) For each word w_j in document d_i:
      i. choose topic $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. choose sense $s_{ij} \sim Mult(\phi_{z_{ij}})$.

Example

- Robot
- Topic1 : film
- Topic2: electronics technique

sense *robot#1*
film: 0.159
role: 0.069
performance: 0.019
...

sense *robot#2*
computer: 0.116
system: 0.039
software: 0.026
...

In the end, it's an inspired performance from Robot that keeps the film fresh

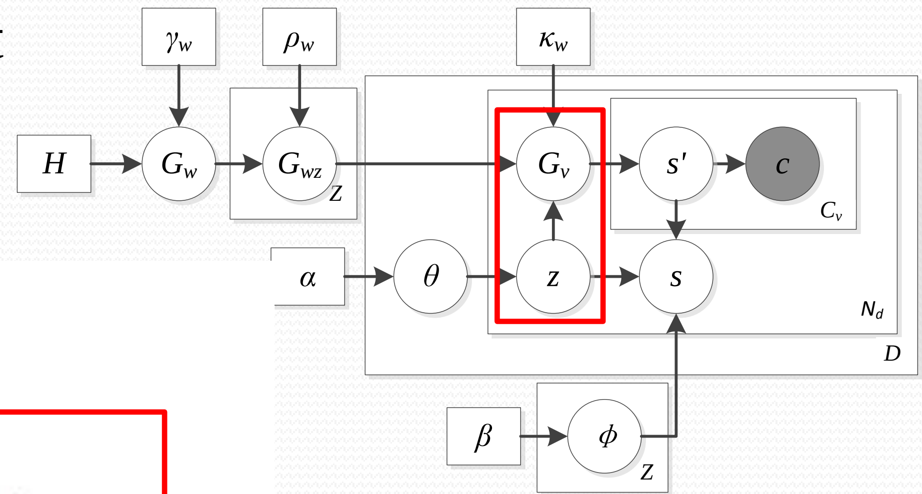
There may be a computer operating system designed mainly for robots

Co-inference Approach (COI)

- Can the topics of words make a positive impact on the indication of senses ?

- Take the topics of words as pseudo feedback and co-infer both topics and senses iteratively.
  - Word *robot* in topic *film* has a higher probability to contain sense *robot#1*.
  - The sense *robot#1* has a higher probability to be assigned topic *film*.

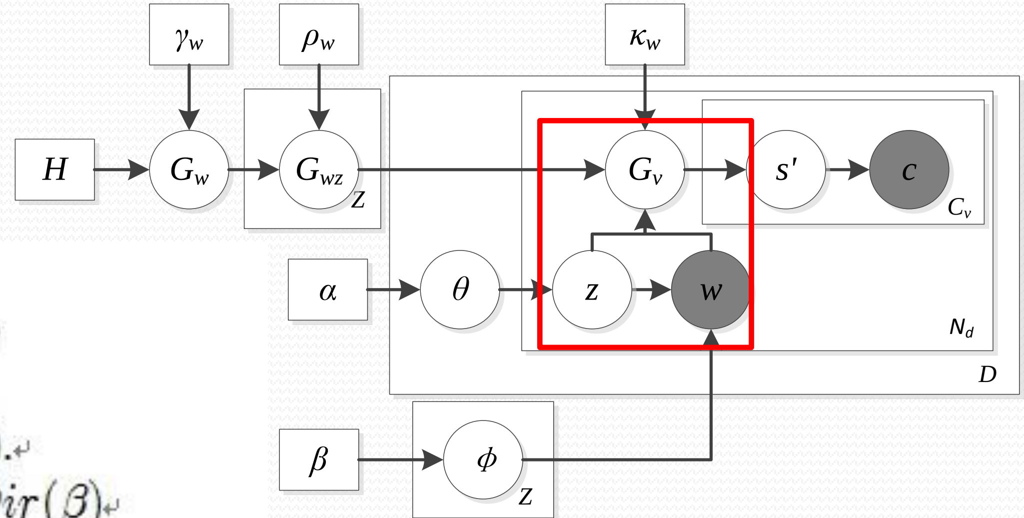
- Document Presentation Part
  - Same as SEQ
- Word Sense Induction Part



1. For each word w:
   a) choose $G_w \sim DP(\gamma_w, H)$.
   b) For each topic z,
      choose $G_{wz} \sim DP(\rho_w, G_w)$.
2. For each document d_i,
   a) For each context v_j of word w_j:
      i. choose $G_{ij} \sim DP(\kappa_{wz}, G_{wz})$.
      ii. For each context word c_k of target word w_j:
         1) choose $s'_{ijk} \sim G_{ij}$.
         2) choose $c_{ijk} \sim Mult(\eta_{s_{ijk}})$
         3) set $s_{ij} = \arg\max_s P(s_{ij}|G_{ij})$

Extended Co-inference Approach (COX)

- The standard COI approach takes the sense with the highest probability as the sense of the target word.
- We now consider the whole sense distribution of the target word in its context
  - COX.
- Three factors are considered to determine the topic of a word:
  - The topic distribution of the document
  - The probability of the topic to emit this word
  - The probability of the word and its topic to generate the sense distribution.
    - reflects the meaning contained by its context.
    - considers the sense distribution of the target word which is more precise.
- Example:
  - In ROBOT, the most important character is an electro-mechanical machine whose software was upgraded to give it the ability to comprehend and generate human emotions
    - The illustrative sense distribution of this context is (0.2, 0.8).
    - In SEQ and COI, the sense will be set as robot#2
    - In COX, it will have a probability of robot#1.



1. For each word w:
   a) choose $G_w \sim DP(\gamma_w, H)$.
   b) For each topic z,
      choose $G_{wz} \sim DP(\rho_w, G_w)$.
2. For each topic z, choose $\phi_z \sim Dir(\beta)$
3. For each document d_i:
   a) choose $\theta_{d_i} \sim Dir(\alpha)$.
   b) For each word w_j in document d_i:
      i. choose topic $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. choose word $w_{ij} \sim Mult(\phi_{z_{ij}})$
      iii. choose $G_{ij} \sim DP(\kappa_{wz}, G_{wz})$.
      iv. For each context word c_k in context v_j of target word w_j:
         1) choose $s'_{ijk} \sim G_{ij}$.
         2) choose $c_{ijk} \sim Mult(\eta_{s_{ijk}})$

Inference

- Sequential Approach

$$P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{s}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{s}_{-ij,z} + \beta}{n_{-ij,z} + S\beta}$$

- Co-inference Approach
  - variables $\boldsymbol{z} = \{z_{ij}\}$assigning words to topics
  - variables $\boldsymbol{s} = \{s_{ijk}\}$ assigning context words of each target word to senses, base distributions of each target word G_w and $\{G_{wz}\}$.
  - COI
    - given the second kind of variables are fixed, the first kind can be sampled using the same scheme as SEQ.
    - Given the first kind of variables are fixed, the second kind can be sampled using the same scheme as described in (Teh et al., 2004)

COX

- Similarly, given the first kind of variables are fixed, the second kind can be sampled using the same scheme as described in (Teh et al., 2004).
- Hence the key issue is how to sample $z = \{z_{ij}\}$ given sense distributions.

$$P(z_{ij} = z | \boldsymbol{z_{-ij}}, \boldsymbol{s}, \boldsymbol{w})$$
$$\propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{ij} + Z\alpha} \frac{n^{w}_{-ij,z} + \beta}{n_{-ij,z} + W\beta} \frac{\prod_{s \in \{s_w\}} \prod_{g=0}^{n^{s-1}_{ij}} (\kappa_{wz}\pi_{zs} + g)}{\prod_{g=0}^{C_{ij}-1} (\kappa_{wz} + g)}$$

Evaluation

Dataset	#doc	#topic	#words	#content words
TDT41	1270	38	18511	5457
TDT42	617	33	11782	3548
Reutes20	9101	20	25748	7454

- Setup
  - Test dataset
    - TDT4 datasets
    - Reuters dataset
  - Evaluation task
    - Document clustering task
    - Evaluation criteria
      - Precision
      - Recall
      - F-Measure

Experiment Result

- Different Word Sense Incorporating Approaches

Method	TDT41	TDT42	Reutes20
LDA	0.735	0.852	0.483
K-Means	0.727	0.843	0.501
SEQ	0.776	0.865	0.491
COI	0.825	0.874	0.597
COX	0.864	0.905	0.612

Conclusion

- In this paper, we present three approaches to incorporating word senses in topic models:
  - SEQ approach
  - COI approach
  - COX approach
- Three conclusions can be drawn from the experimental results.
  - Replacing word surfaces with word senses is helpful in topic modeling.
  - The topics of words can make a positive impact on the indication of word senses thus improve word sense induction.
  - Using the regular sense distribution of the target word can get a better topic indication than that uses merely the definite sense with the highest probability.

Reference(1/2)

- Agirre, E. and Soroa, A. 2007. Semeval-2007 task02: Evaluating word sense induction and discrimination systems. In *SemEval2007*.
- Blei, D.M., Ng, A. Y., and Jordan, M.I. 2003. Latent dirichlet allocation. J. *Machine Learning Research* (3):993-1022.
- Body-Graber, J., Blei, D.M. and Zhu, X. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL'2007*:1024-1033.
- Brody, S., Lapata, M. 2009. Bayesian word sense induction. In *EACL'2009*: 103-111.
- Chemudugunta, C., Smyth, P. and Steyvers, M. 2008. Combining concept hierarches and statistical topic models. In *CIKM'2008*: 1469-1470.
- Denkowski, M. 2009. A Survey of Techniques for Unsupervised Word Sense Induction. *Technical Report*.Language Technologies Institute, Carnegie Mellon University
- Dietz, L., Bickel., S., Scheffer, T., 2007. Unsuperservised prediction of citation influence. In *ICML'2007*: 233-240.
- Gabrilovich, E. and Markovitch, S.2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI'2007*, Hyderabad, India, January 2007
- Griffiths, T. L., Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, 101:5228-5235
- Guo, W. and Diab, M. 2010. Combining orthogonal monolingual and multilingual sources of evidence for all words wsd. In *ACL'2010*: 1542-1551.
- Ferguson, T.S.. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2): 209-330
- Hotho, A., Staab, S., Stummem, G.. 2003. WordNet improves text document clustering. In *SIGIR2003 semantic web workshop*. ACM, New York, pp. 541-544.
- Huang, H., Kuo, Y., 2010. Cross-Lingual Document Representation and Semantic Similarity Measure: A Fuzzy Set and Rough Set Based Approach. *Fuzzy Systems, IEEE Transactions,* vol.18, no.6, pp.1098-1111.
- Kong, J. and Graff, D. 2005. TDT4 multilingual broadcast news speech corpus. In LDC link: http://www.ldc.upenn.edu/Catalog/index.jsp

Reference(2/2)

- Navigli, R. and Crisafulli , G. 2010. Inducing word senses to improve web search result clustering. *Proc. of EMNLP* '10:116-126.
- Oakes, M. P., and Tait , J. 2003. Word sense disambiguation in information retrieval revisited. In *Proc. of SIGIR '03*:159-166.
- T. K. Landauer and S. T. Domais(1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*. 104(2):211-240.
- Lewis, D.. Reuters-21578 text categorization test collection. http://www.research.att.com/~lewis, 1997.
- Schmid, H.. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *EMNLP'1994*, Manchester, UK
- Schutze, H. and Pedersen , J. 1995. Information Retrieval based on word senses. In *SDAIR'95*: 161–175.
- Steinbach, M., Karypis, G., Kumar, V.. 2000. A comparison of document clustering techniques. In *KDD'2000* Workshop on Text Mining.
- Stokoe, C., Oakes, M. P., and Tait, J. 2003. Word sense disambiguation in information retrieval revisited. In *SIGIR '03*:159-166.
- Teh, Y. W., Jordan, M. I. , Beal, M. J., and Blei, D. M. 2004. Hierarchical dirichlet processes. In *NIPS*, 2004.
- Tufi, D., and Koeva, S.. 2007. Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation. In *WILF '07*: 447-455.
- Wang, X. , McCallum, A. , Wei, X. . 2007. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval, In *ICDM'2007:* 697-702, October 28-31, 2007
- Yao, X., Durme, B.V.. 2007. Nonparametric Bayesian Word Sense Induction. In *TextGraphs-6* Workshop:10-14, June 19-24, 2011.

Thank you ！

Q&A