Topic Models Incorporating Statistical Word Senses

For CICling 2014

Outlines

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

Introduction(1/5)

- LDA
  - relies on the co-occurrences of surface words to capture their semantic relations.

- In reality, a surface word is likely to be highly associated to more than one topic and presents different word senses in different topics.

Introduction(2/5)

- *Robot*
  - S#1:machine robot
  - S#2:film robot
- In LDA
  - Two topics: *electronics technology* , *film*.
  - LDA considers the surface word 'robot‘ to be identical in both contexts and leverages on its co-occurrences with other words in the context to differentiate those two topics..
  - With word sense information
    - a document with context of word sense S#1 is expected to earn more probability mass for topic T#1 and less probability mass for topic T#2,

Introduction(3/5)

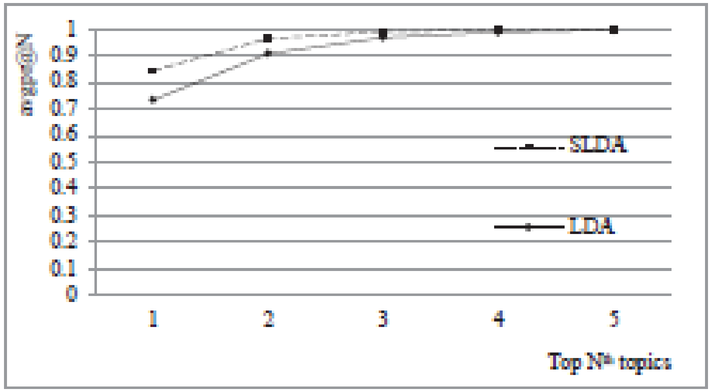


Fig. 1. Averaged per word (sense) topic distribution on the top-5 topics where the cumulative curve presents *avpgr* over the top-k topics.

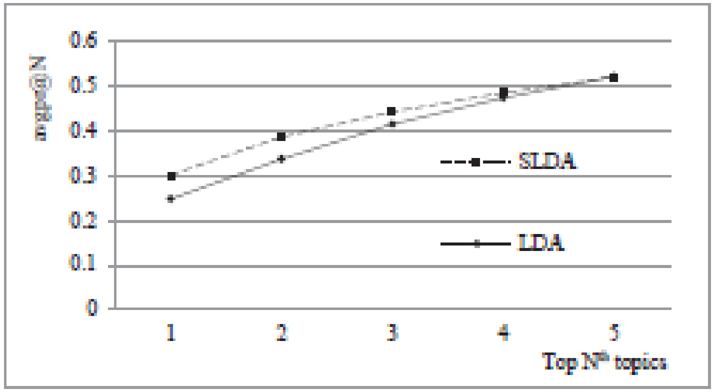


Fig. 2. Averaged per document topic distribution on the top-5 topics. where the cumulative curve presents the *avpgr* over the top-k topics.

Introduction(4/5)

- Incorporate the word sense information in the LDA generative story and construct a joint model to infer word senses for words and topics for documents simultaneously.

- Our model is completely unsupervised and is able to work with external resources minimized.

Introduction(5/5)

- HDP for word sense induction
- Two models are proposed in this paper:
  - Standalone SLDA(SA-SLDA)
    - word sense induction and document representation as standalone modules;
  - Collaborative SLDA(CO-SLDA)
    - takes the topics of senses from SLDA as the pseudo feedback for Word Sense Induction (WSI) and iteratively infers both topics and word senses.

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 senses
    - 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

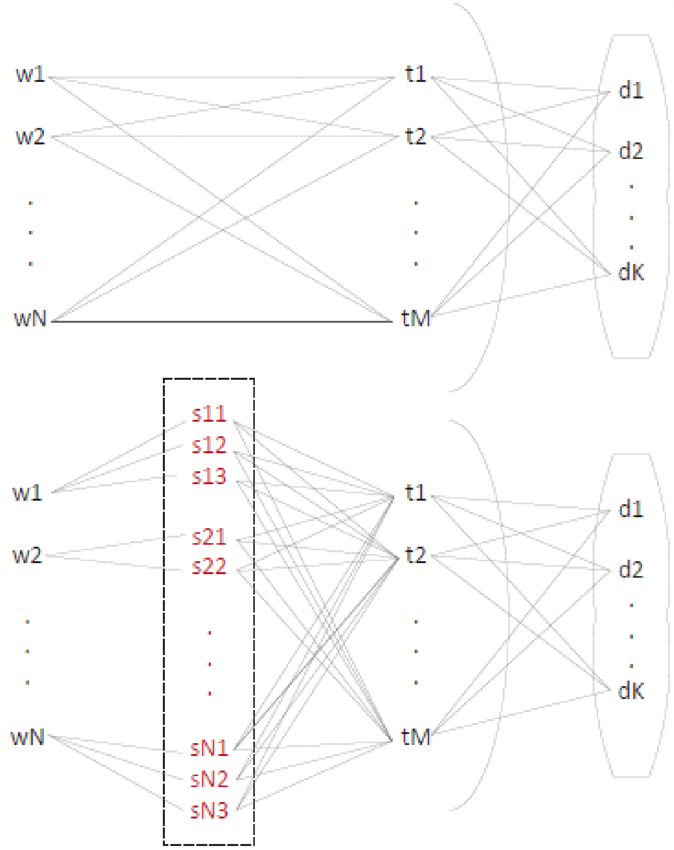


Fig. 3 Illustration of the classic LDA model (above) and the word sense extended LDA models (below). The values in the dot rectangle are assigned to the latent variable (i.e., word sense).

SA-SLDA

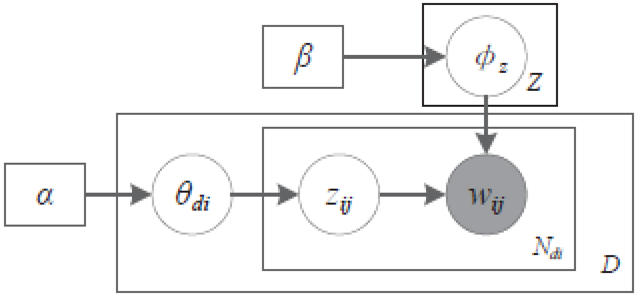


Fig. 4. Illustration of the standard LDA model.

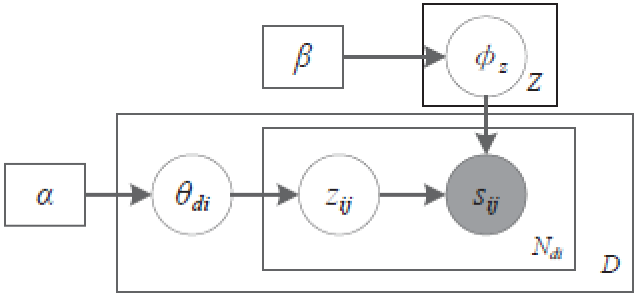


Fig. 5. Illustration of the SA-SLDA model.

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-SLDA(1/2)

- Can the topics of words make a positive impact on the indication of senses ?
  - Word *robot* in topic *film* has a higher probability to contain sense *robot#1*.

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

CO-SLDA(2/2)

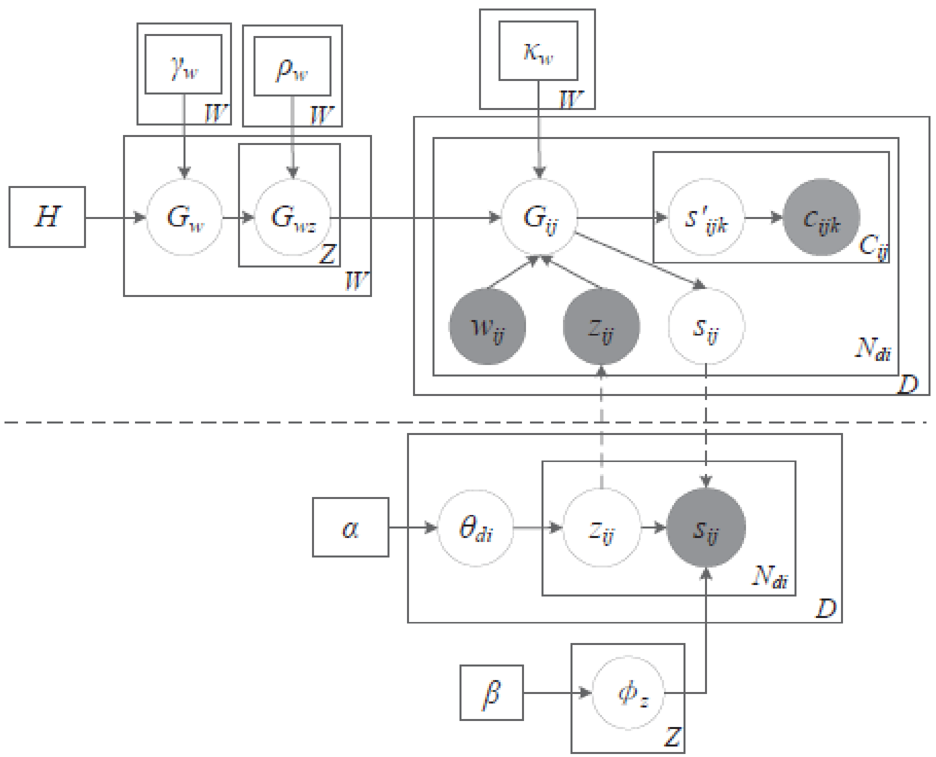


Fig. 6. Illustration of the CO-SLDA model.

Evaluation-Document clustering

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

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

Experiment 1.1: Different Word Sense Induction Approaches

Table 1. Results of SA-SLDA with different WSI approaches (i.e., LDA and HDP).

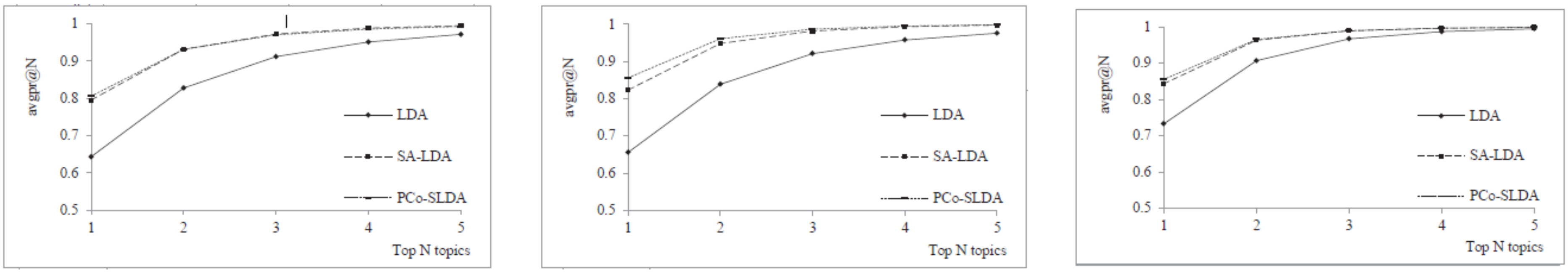
Method	TDT41	TDT42	Reuters20
SA-SLDA(LDA)	0.787	0.842	0.490
SA-SLDA(HDP)	0.792	0.870	0.512

Experiment 1.2: Different Extended LDA Models

Table 2. Results of the proposed models and baselines.

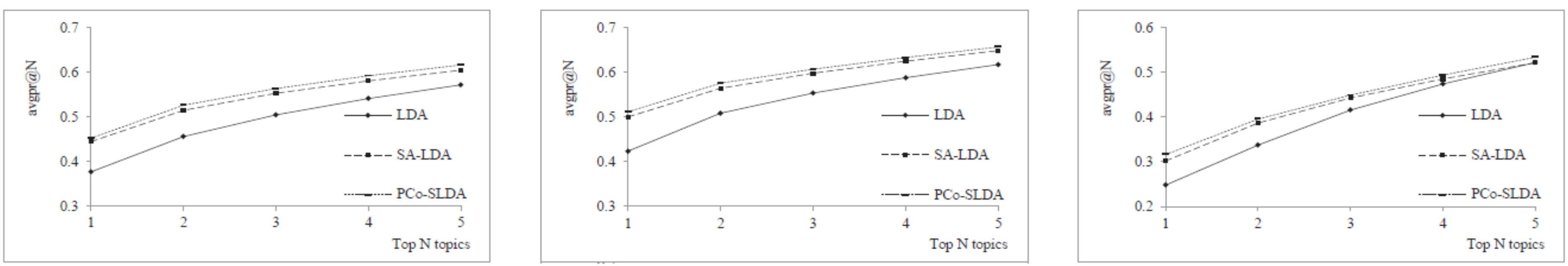
Method	TDT41	TDT42	Reuters20
K-Means	0.727	0.843	0.501
LDA	0.744	0.867	0.496
SA-SLDA	0.792	0.870	0.512
CO-SLDA	0.825	0.874	0.597

Evaluation-Distribution Analysis



(a) TDT41 (b) TDT42 (c) Reuters20

Fig. 7. Averaged per word (sense) topic distribution on the top-5 topics .



(a) TDT41 (b) TDT42 (c) Reuters20

Fig. 8. Averaged per document topic distribution on the top-5 topics.

Conclusion

- In this paper, we propose to represent topics with distributions over word senses.
  - SA-SLDA, CO-SLDA
- Distributions analysis shows a sharper distribution of topics in SLDAs which suggests that the proposed models provide more confidence on the posterior estimation.
- Empirical results verify that the word senses induced from corpora can facilitate the LDA model in document clustering.
- Specifically, we find the joint inference model (CO-SLDA) outperforms the standalone model (SA-SLDA) as the estimation of sense and topic can be collaboratively improved.

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