Statistical Word Sense Improves Document Clustering

Guoyu Tang

Outlines

- Introduction
- Related work
- Document representation based on word sense
- Evaluation
- Conclusion and future work

Introduction (1/4)

- Document Clustering:
  - Automatically organize a large collection of documents into groups of similar documents
- How to represent document?
  - Vector Space Model(Salton et al., 1975)
- Two linguistic phenomena:
  - Synonymy
    - computer and PC
  - Polysemy
    - apple: ⑴ A pomaceous fruit; ⑵ A computer company founded by Steve Jobs.

Introduction (2/4)

VSM
- Synonymy
- Polysemy

Explicit Sematic Representation
- Need large, general purpose lexical resources
- Tend to over-represent rare word senses while missing corpus-specific senses.

Latent Semantic Representation
- According to previous research Lu et al., (2011) it cannot provide fine granularity discrimination.

Introduction (3/4)

- Our solution: represent document with statistical word sense.
  - Word senses are constructed in two steps:
    - Local word senses are induced from the development dataset by the LDA models (Brody and Lapata, 2009).
    - Local word senses are combined as global word senses by clustering technology
  - Global word senses are used to represent every document after word sense disambiguation on the document.

Introduction (4/4)

- The proposed model aims to well address the synonymy and polysemy issues in document representation.
  - Synonymy: Different words of the same meaning are identified as the same sense.
  - Polysemy: One word in different contexts can be identified as different sense in different contexts.
- Compared with previous researches,
  - Compared with the explicit sematic methods:
    - Word sense can be induced from the raw development dataset
    - It can be easily extended to process documents in other languages
  - Compared with the latent sematic methods:
    - It can achieve finer granularity discrimination in document representation

Related work(1/2)

- Document representation models
  - Classic model
    - VSM(Vector Space Model)
      - Problems: Synonymy Polysemy
  - Improvement:
    - Explicit Sematic Representation(Hotho et al., 2003; Gabrilovich and Markovitch, 2007; Huang and Kuo, 2010)
      - Lexical resources: WordNet and wikipedia
      - Represent documents in the concept space
    - Latent Semantic Representation
      - Probabilistic latent semantic analysis (Puzicha and Hofmann, 1999)
      - Latent Dirichlet Allocation (Blei et al., 2003)

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 (Brody and Lapata ,2009)
      - Use an extended LDA model to induce word senses
      - 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.

Document representation based on word sense

- How to represent word sense?
- How to obtain word sense?
- How to use word sense in document clustering?

How to represent word sense?(1/2)

- Local word sense:
  - A probability distribution over context words

Example #1: word sense arm#1 for word arm
arm: 0.159
sleeve: 0.069
sew: 0.019

Example #2: word sense arm#2 for word arm
arm: 0.116
weapon: 0.039
war: 0.026

How to represent word sense?(2/2)

- Global word sense:
  - A group of similar local word senses

Example #3: sense cluster c#1
{arm#1, sleeve#1}
arm#1={arm: 0.159, sleeve: 0.069, sew: 0.019}
sleeve#1={sleeve:0.179,arm:0.059,sew: 0.029}

How to obtain word sense?(1/2)

- Local word sense:
  - Bayesian word sense introduction model (Brody and Lapata, 2009)

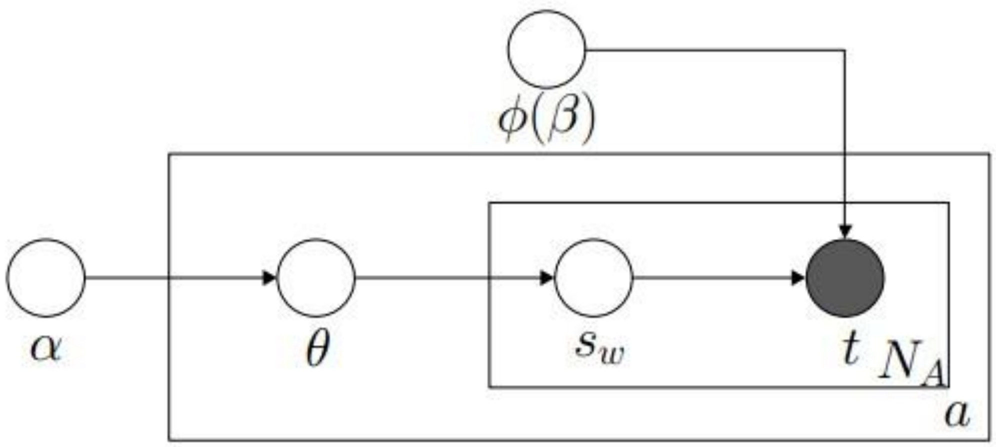


FIGURE 1 – Bayesian sense induction model (Brody and Lapata, 2009).

How to obtain word sense?(2/2)

- Apply clustering algorithm to obtain global word sense.
  - In the clustering algorithms, we take context words of local word senses as features and probabilities of the context words as the weights of features.
  - Bisecting K-Means
    - An extension of K-means, which is proved better than standard K-Means and hierarchical agglomerative clustering (Steinbach et al., 2000). It begins with a large cluster consisting of every element to be clustered and iteratively picks the largest cluster in the set, split it into two.

How to use word sense in document clustering?(1/3)

- Bayesian word sense disambiguation
  - Example:
    - ① There's a tear in the arm of my jacket.
      - P(arm#1| S1)=0.998005.
    - ②The nation must arm its soldiers for battle.
      - P(arm#2| S2)= 0.944096.

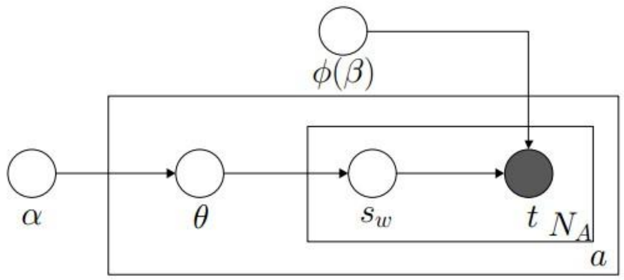


FIGURE 1 – Bayesian sense induction model (Brody and Lapata, 2009).

How to use word sense in document clustering?(2/3)

- Represent document in global word sense space

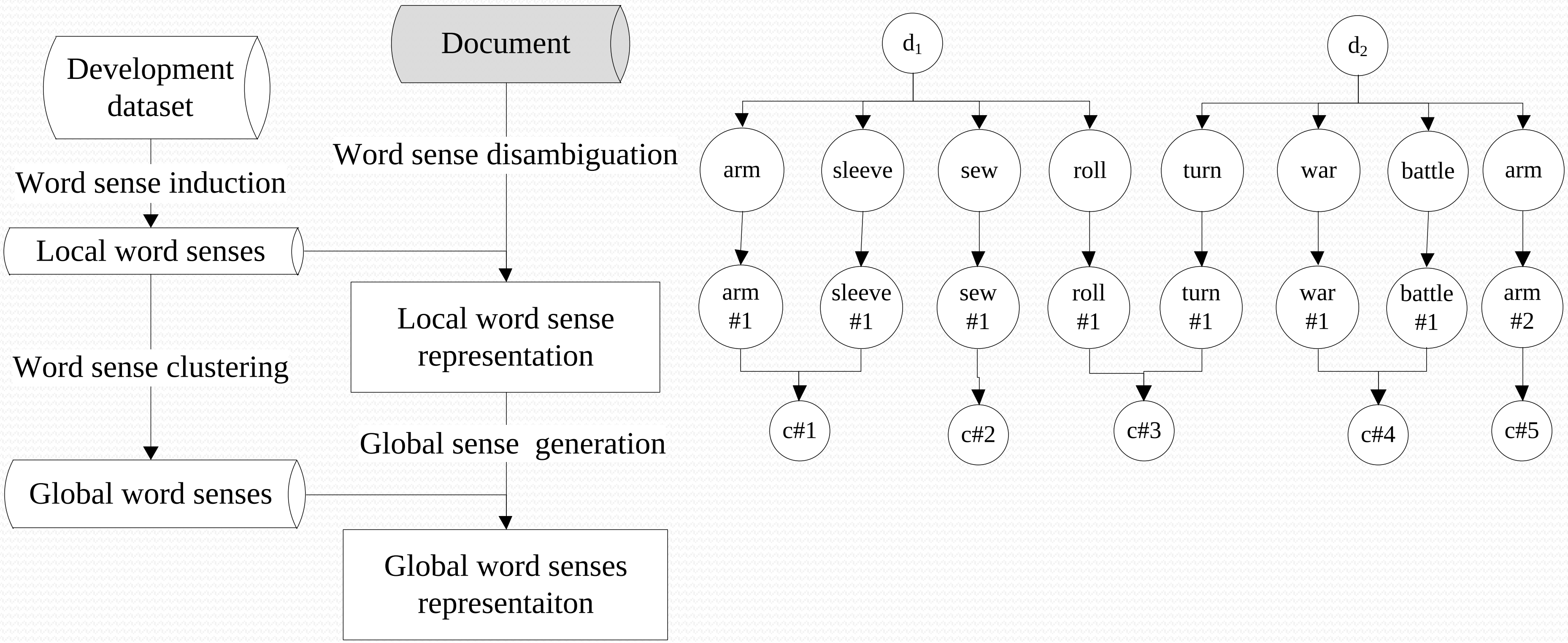
$$n(c,d) = \sum_{w_k \in d} n(c|w_k, d) \quad n(c|w,d) = \sum_{s_w \in c} p(s_w|d)$$

- Sense based TF-IDF

$$tf(c,d) = n(c,d), idf(c) = \sum_{n(c,d)>1} 1$$

- Clustering Methods:
  - Cosine similarity
  - Hierarchical Agglomerative Clustering

How to use word sense in document clustering?(2/3)



Evaluation

- Setup
  - Development Dataset: Giga Word（ 2.1 million English documents and 3.5 million Chinese documents ）
  - Test Dataset: TDT4 and CLTC in both English and Chinese language
  - Evaluation criteria
    - Precision
    - Recall
    - F-Measure

Dataset	**English**	**Chinese**
TDT41	38/1270	37/657
TDT42	33/617	32/560
CLTC1	20/200	20/200
CLTC2	20/600	20/600

Experiment

- Methods:
  - VSM (Vector Space Model)
  - LDA(Latent Dirichlet Allocation)
  - LSSM (Local Sense Space Model)
  - GSSM (Global Sense Space Model)
- Result

Methods	CLTC1	CLTC2	TDT41	TDT42
VSM	0.886	0.898	0.894	0.924
LDA	0.832	0.891	0.789	0.854
LSSM	0.888	0.893	0.922	0.964
GSSM	0.905	0.918	0.926	0.964

Methods	CLTC1	CLTC2	TDT41	TDT42
VSM	0.886	0.898	0.894	0.924
LDA	0.832	0.891	0.789	0.854
LSSM	0.888	0.893	0.922	0.964
GSSM	0.905	0.918	0.926	0.964

Conclusion and future work

- Our research on addressing synonymy and polysemy issues in document representation shows that document representation can be further improved with word sense.
- In this work, a new document represent model is proposed to make full use of global word sense.
  - The proposed model aims to well address the synonymy and polysemy issues
  - Experiments on four datasets of two language cases show that our proposed SCM model advances both baseline systems and LDA models in document clustering task in both language cases.
- In the future work, we will continue to evaluate the performance of our model with datasets of smaller samples, e.g., SMS messages and tweets.

Reference(1/2)

- G. Salton, A. Wong, and C. S. Yang(1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, v. 18(11): 613–620. 1975.
- A. Hotho, S. Staab,G. Stumme (2003). WordNet improves text document clustering. Proc.of SIGIR2003 semantic web workshop.ACM, New York, pp. 541-544.
- Evgeniy Gabrilovich and Shaul Markovitch.(2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India, January 2007
- D.M. Blei, A. Y. Ng, and M.I. Jordan (2003). Latent dirichlet allocation. J. Machine Learning Research (3):993-1022.
- 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.
- Yue Lu , Qiaozhu Mei , Chengxiang Zhai (2011), Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, v.14 n.2, p.178-203, April 2011
- S. Brody, M. Lapata (2009). Bayesian word sense induction. Proc. of EACL'2009: 103-111.
- H. Huang, Y. Kuo (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.
- J. Pessiot, Y. Kim, M. Amini, P. Gallinari (2010). Improving document clustering in a learned concet space. Information Processing and Management. vol. 46:180-192
- I. S. Dhillon (2001). Co-clustering documents and words using bipartite spectral graph partitioning. Proc. SIGKDD'2001:269–274.
- S. K. M. Wong, W. Ziarko, P. C. N. Wong (1985). Generalized vector model in information retrieval. Proc. of the 8th ACM SIGIR:18-25
- A.K. Farahat, M.S.Kamel (2010). Statistical semantic for enhancing document clustering. Knowledge and Information Systems.

Reference(2/2)

- R. Navigli (2009). Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages.
- C. Stokoe, M. P. Oakes, and J. Tait (2003). Word sense disambiguation in information retrieval revisited. In Proceedings of SIGIR '03:159-166.
- D. Tufi; and S. Koeva (2007). Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation. In Proceedings of the 7th international workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory (WILF '07), Francesco Masulli, Sushmita Mitra, and Gabriella Pasi (Eds.). Springer-Verlag, Berlin, Heidelberg, 447-455.
- M. Denkowski (2009). A Survey of Techniques for Unsupervised Word Sense Induction. Technical Report. Language Technologies Institute, Carnegie Mellon University
- E. Agirre and A. Soroa (2007). Semeval-2007 task02: Evaluating word sense induction and discrimination systems. SemEval2007.
- H. Schutze and J. Pedersen (1995). Information Retrieval based on word senses. Proc. of SDAIR'95: 161–175.
- R. Navigli and G. Crisafulli (2010). Inducing word senses to improve web search result clustering. Proc. of EMNLP '10:116-126.
- I. S. Dhillon and D. S. Modha (2001). Concept decompositions for large sparse text data using clustering. Mach. Learn., 42(1-2):143~175, 2001.
- Y. Zhao, G. Karypis, and U. Fayyad (2005). Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 10(2):141~168, 2005.
- C. Ordonez and E. Omiecinski (2002). Frem: fast and robust em clustering for large data sets. In CIKM '02, pp.590~599, New York, NY, USA, 2002. ACM Press.
- T. L. Griffiths and M.. Steyvers (2004). Finding scientific topics. Proc. Nat. Acad. Sci. 101: 5228--5235.
- M. Steinbach, G. Karypis, V. Kumar(2000). A comparison of document clustering techniques. KDD Workshop on Text Mining.
- Junbo Kong and David Graff (2005). TDT4 multilingual broadcast news speech corpus.
- G. Tang, Y. Xia, M. Zhang, H. Li, F. Zheng (2011) CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. Proc. of IJCNLP'2010: 580-588.
- H. Schmid (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK
- E. M. Voorhees (1986) Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Information Processing and Management. v.22(6):465-476. 1986.
- Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei(2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476):1566–1581.

Thank you ！

Q&A