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: (1) A pomaceous fruit; (2) A computer company founded by Steve Jobs.
Introduction (2/4)

VSM
- Synonymy
- Polysemy

Explicit Semantic 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.
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.
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 semantic 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 semantic 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 Semantic 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.
      - \( P(\text{arm}\#1 \mid S_1) = 0.998005 \).
    - ② The nation must arm its soldiers for battle.
      - \( P(\text{arm}\#2 \mid S_2) = 0.944096 \).

*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

\[ n(c, d) = \sum_{w_k \in d} n(c \mid w_k, d) \]
\[ n(c \mid w, d) = \sum_{s_w \in c} p(s_w \mid d) \]

- Sense based TF-IDF

\[ tf(c, d) = n(c, d),\quad idf(c) = \sum_{n(c,d)>1} 1 \]

- 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT41</td>
<td>38/1270</td>
<td>37/657</td>
</tr>
<tr>
<td>TDT42</td>
<td>33/617</td>
<td>32/560</td>
</tr>
<tr>
<td>CLTC1</td>
<td>20/200</td>
<td>20/200</td>
</tr>
<tr>
<td>CLTC2</td>
<td>20/600</td>
<td>20/600</td>
</tr>
</tbody>
</table>
Experiment

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

- **Result**

<table>
<thead>
<tr>
<th>Methods</th>
<th>CLTC1</th>
<th>CLTC2</th>
<th>TDT41</th>
<th>TDT42</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSM</td>
<td>0.886</td>
<td>0.898</td>
<td>0.894</td>
<td>0.924</td>
</tr>
<tr>
<td>LDA</td>
<td>0.832</td>
<td>0.891</td>
<td>0.789</td>
<td>0.854</td>
</tr>
<tr>
<td>LSSM</td>
<td>0.888</td>
<td>0.893</td>
<td>0.922</td>
<td>0.964</td>
</tr>
<tr>
<td>GSSM</td>
<td>0.905</td>
<td>0.918</td>
<td>0.926</td>
<td>0.964</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>CLTC1</th>
<th>CLTC2</th>
<th>TDT41</th>
<th>TDT42</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSM</td>
<td>0.886</td>
<td>0.898</td>
<td>0.894</td>
<td>0.924</td>
</tr>
<tr>
<td>LDA</td>
<td>0.832</td>
<td>0.891</td>
<td>0.789</td>
<td>0.854</td>
</tr>
<tr>
<td>LSSM</td>
<td>0.888</td>
<td>0.893</td>
<td>0.922</td>
<td>0.964</td>
</tr>
<tr>
<td>GSSM</td>
<td>0.905</td>
<td>0.918</td>
<td>0.926</td>
<td>0.964</td>
</tr>
</tbody>
</table>
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 representation 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)


Thank you!

Q&A