Topic Analysis for Psychiatric Document Retrieval

Liang-Chih Yu
Department of Information Management,
Yuan Ze University, Taiwan, R.O.C.
Outline

- Introduction
  - Motivation
  - Topics in Psychiatric Documents
- Methodology
  - Topic Analysis
  - Retrieval Model
- Experimental Results
- Conclusion
Introduction

- People in their daily life may suffer from some negative life events, which may trigger depressive symptoms.
- In this circumstance, people often search for help via psychiatric web sites.
  - People can describe their problems via message boards, emails, or other services.
  - Health professionals will then make suggestions to the problems as soon as possible.

-produce thousands of psychiatric consultation documents (problem-response pairs)
Bibliotherapy

- A kind of cognitive behavior therapy (CBT)
- It can help people deal with their mental health problems through reading self-help literature or web resources
- Consist of three stages
  - Identification: identify a character, situation or problem in literature
  - Catharsis: release pent-up emotions by recognizing that others have similar problems to their own
  - Insight: realize that their problems can also be addressed by learning through others’ experiences
I broke up with my dear but cruel boyfriend recently. I was late getting to the restaurant where we were meeting and he had already been waiting for 30 minutes and was very angry. When we left, he started yelling, slapped me in the face and said he never wanted to see me again.

Since then, I have often felt like crying out of nowhere, and I feel pain every day. So, I think that continuing to live like this is meaningless.

After that, it took me a long time to fall asleep at night. In recent months, I often lose my temper for no reason.

Feeling this way is normal when going through these kinds of struggles, but over time your emotions should level out. Suicide doesn't solve anything; think about how it would affect your family. It's only when we learn to face our despair that we can learn the value of life, and also how to help other people. There are a few things you can try to help you get to sleep at night, like doing some light exercise in the evening, drinking warm milk, and listening to relaxing music; all of these can be conducive to sleep. If you still have trouble dealing with the pain, and you feel as if your mood is getting worse, it wouldn't hurt to get seek help from a healthcare professional, who can help you work through your emotions.
Topics

- Characteristics of a psychiatric consultation document
  - Document-like query
  - Topics
    - Negative life events
    - Depressive symptoms
    - Semantic relations between symptoms
      - Cause-Effect relation
      - Temporal relation
Goal

- Develop a **topic-based retrieval system** to assist people efficiently and effectively locating relevant documents
  - identify topics in queries and documents
  - retrieve relevant documents based on topics
- People can thereby immediately understand how to deal with their problems according to the suggestions
- We expect that the topic-based approach can improve retrieval precision
Outline

- Introduction
- Methodology
  - Topic Analysis
    - Symptom identification
    - Relation identification
    - Negative life event identification
  - Retrieval Model
    - Linear combination of symptom, relation and event similarities
- Experiments
- Conclusion
## Definition of Depressive Symptoms

### 17-item Hamilton Depression Rating Scale (HDRS)

<table>
<thead>
<tr>
<th>No.</th>
<th>Item</th>
<th>Example Rules</th>
<th>No.</th>
<th>Item</th>
<th>Example Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Depressed Mood</td>
<td>Feelings of sadness, hopeless, helpless, and so on.)</td>
<td>10</td>
<td>Anxiety</td>
<td>Worry about minor matter. Feel tense, feel fear, get mad</td>
</tr>
<tr>
<td>2</td>
<td>Feelings of guilt</td>
<td>Self reproach, feels he has let people down.</td>
<td>11</td>
<td>Anxiety – somatic</td>
<td>Such as flushing, sweating, tremor, and so on.</td>
</tr>
<tr>
<td>3</td>
<td>Suicide</td>
<td>Feels life is not worth living</td>
<td>12</td>
<td>Somatic Symptoms (gastrointestinal)</td>
<td>Loss of appetite.</td>
</tr>
<tr>
<td>4</td>
<td>Insomnia-Early</td>
<td>Complains of difficulty falling asleep</td>
<td>13</td>
<td>Somatic Symptoms</td>
<td>Such as backaches, headache, muscle aches, and so on.</td>
</tr>
<tr>
<td>5</td>
<td>Insomnia-Middle</td>
<td>Complains of being restless and disturbed during the night</td>
<td>14</td>
<td>Genital Symptoms</td>
<td>Such as loss of libido and menstrual disturbances.</td>
</tr>
<tr>
<td>6</td>
<td>Insomnia-Late</td>
<td>Waking in early hours of the morning</td>
<td>15</td>
<td>Hypochondriasis</td>
<td>Preoccupation with health.</td>
</tr>
<tr>
<td>7</td>
<td>Work and Activities</td>
<td>Loss of interest in activity; hobbies or work.</td>
<td>16</td>
<td>Loss of Weight</td>
<td>Loss of weigh.</td>
</tr>
<tr>
<td>8</td>
<td>Retardation</td>
<td>Slowness of thought and speech.</td>
<td>17</td>
<td>Insight</td>
<td>Acknowledges or denies being depressed</td>
</tr>
<tr>
<td>9</td>
<td>Agitation</td>
<td>Such as fidgetiness, playing with hand, biting of lips, etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Traditional Approach

- Bag-of-words representation

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Features (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>Sad, depress, hopeless</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Worry, tense, fear</td>
</tr>
</tbody>
</table>
Symptoms are usually embedded in sentences
- We propose the use of semantic dependencies to analyze sentence structure
- A semantic dependency represents word relationship in a sentence

[I]  [often]  worry about  [some]  [minor matters]

Format: (who worry about what)
Once the semantic dependencies have been extracted, symptoms can be identified by the probability distribution of dependencies.

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Features (semantic dependencies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>(I, feel, sad), (I feel like, cry)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>(I, worry, matter), (I, get mad, friend)</td>
</tr>
</tbody>
</table>
Symptom Mining

\[ W \sqsubseteq I \text{ often worry about some minor matters} \]

**Step 1**

Word Segmentation & POS tagging

\[ I \text{ (Nh) often (Dd) worry about (VK) some (DM) minor (VH) matters (Na)} \]

**Step 2**

Parser □ Analyze sentence structure

\[ D = \{ (I, \text{ worry about, experiencer}), (\text{often, worry about, time}), (\text{matters, worry about, goal}), (\text{some, matters, quantifier}), (\text{minor, matters, property}) \} \]

**Step 3**

Semantic Label Inference

\[ \hat{l}_k = \arg \max \prod_{d \in D} P(l_k | l_{k-1}, d) \quad \rightarrow \quad \hat{l}_k = \langle \text{Anxiety} \rangle \]
Relation Identification

- Two types of relations
  - Cause-effect relation
  - Temporal relation
- Discourse markers (連接詞) are the significant features for relation identification
  - “because”, “therefore”
  - “before”, “after”
- A total of 121 discourse markers are collected from Web
Relation Identification

I often felt like crying and felt pain every day. <Depressed>

So, I tried to kill myself several times. <Suicide>

After that, it took me a long time to fall asleep at night. <Insomnia>

In recent months, I often lose my temper for no reason. <Anxiety>

- Cause-Effect relation
- Temporal relation

A query (or document) is divided into different time intervals using the temporal discourse markers.
Definition of Negative Life Events

- Referring to recent studies on negative life events

<table>
<thead>
<tr>
<th>Types</th>
<th>Example patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>&lt;parents, divorce&gt;, &lt;son, injure&gt;, &lt;daughter, sick&gt;</td>
</tr>
<tr>
<td>Love</td>
<td>&lt;husband, argue&gt;, &lt;girlfriend, break up&gt;, &lt;wife, cheat&gt;, &lt;marriage, break&gt;</td>
</tr>
<tr>
<td>School</td>
<td>&lt;teacher, blame&gt;, &lt;exam, fail&gt;, &lt;classmate, fight&gt;</td>
</tr>
<tr>
<td>Work</td>
<td>&lt;salary, cut&gt;, &lt;job, lose&gt;, &lt;work, stop&gt;, &lt;boss, fire&gt;, &lt;performance, drop&gt;</td>
</tr>
<tr>
<td>Social</td>
<td>&lt;friend, die&gt;</td>
</tr>
</tbody>
</table>
Goal

- Induce more event patterns from Web
  - Pattern: $k$-words combination (2, 3, 4, …)
  - Format (2-words): <Noun, Verb>
  - Format (>2 words): <Noun, Verb, *, … >

<table>
<thead>
<tr>
<th>Types</th>
<th>Seed Pattern</th>
<th>Pattern Induction from Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>&lt;son, injure&gt;, &lt;husband, argue&gt;</td>
<td>&lt;husband, fight&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;husband, yell&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;wife, argue&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;spouse, fight&gt;</td>
</tr>
<tr>
<td>Love</td>
<td>&lt;marriage, break&gt;, &lt;wife, cheat&gt;</td>
<td>&lt;husband, fight, money&gt;</td>
</tr>
<tr>
<td>School</td>
<td>&lt;teacher, blame&gt;, &lt;exam, fail&gt;</td>
<td>&lt;wife, argue, money&gt;</td>
</tr>
<tr>
<td>Work</td>
<td>&lt;salary, cut&gt;, &lt;work, stop&gt;</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td>&lt;friend, die&gt;</td>
<td></td>
</tr>
</tbody>
</table>

How to determine word meanings and similar words
HAL Model (1/3)

- Hyperspace Analog to Language (HAL) model
  - Cognitive model
  - The notion is derived from observations of human behavior
- Human beings may determine the sense of an unknown word by referring to its context (neighboring words)

```
context

stress    colleague    client    company    work

boss

chief

flower
```
Similarity of words is computed by comparing their contexts.

Two words sharing more common contexts are more similar.

\[ Sim(w_0, w_n) = \text{Average}(Sim_{\text{Left}}(w_0, w_n), Sim_{\text{Right}}(w_0, w_n)) \]
HAL Model (3/3)

- Computation of similar words
  - ♦ husband $\rightarrow$ wife, spouse
  - ♦ argue $\rightarrow$ fight, yell
- Computation of similar patterns
  - ♦ <husband, argue> $\rightarrow$ <husband, fight>  
    - <wife, argue>  
    - <wife, argue, money>
An Evolutionary Inference Algorithm (EIA) is devised based on evolutionary computation (genetic algorithm)

- combine the HAL model

- Induce more similar patterns from the Web based on a set of seed patterns
EIA

- Initial population
- Fitness function (HAL similarity)
- Selection
- Crossover and Mutation
- Population Update
- Relevance feedback
Initial Population

- Seed Pattern
  - <husband, argue>

- Initial population (size = 100)
  - a set of candidates
    - <husband, guide>
    - <police, fight>
    - <wife, hear>
    - <car, yell>
    - <flower, talk>
    - ...
Fitness Function

- **Seed Pattern**
  - ♦ <husband, argue>

- **Fitness Function** (HAL Similarity)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;husband, guide&gt;</td>
<td>0.38</td>
</tr>
<tr>
<td>&lt;police, fight&gt;</td>
<td>0.36</td>
</tr>
<tr>
<td>&lt;wife, hear&gt;</td>
<td>0.31</td>
</tr>
<tr>
<td>&lt;car, yell&gt;</td>
<td>0.29</td>
</tr>
<tr>
<td>&lt;flower, talk&gt;</td>
<td>0.08</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Diagram:
- Seed Patterns → EIA → Discovered Pattern Set
- Initial population (length 2) → Generation=0 → Fitness function
- Initial population (length 3) → ...
- Initial population (length k) → ...

HAL Space Construction → HAL model
Psychiatry Web Corpora
Selection

- Seed Pattern
  - <husband, argue>

- Select parents
  - In proportion to fitness values (similarity scores)
  - Roulette wheel
Crossover and Mutation

- **Crossover**
  - <police, fight>
  - <wife, hear>
  - <wife, fight>
  - <police, hear>

- **Mutation**
  - <flower, talk>
  - <flower, shout>

---

**Diagram:**
- **Seed Patterns** → **EIA** → **Discovered Pattern Set**
- **Initial population (length 2)** → **Generation=0** → **Fitness function** → **Termination criteria** → **Selection**
- **Parents** → **Parents** → **Offspring** → **Offspring**

**Processes:**
- **Crossover**
- **Mutation**

**HAL Space Construction** → **Psychiatry Web Corpora** → **HAL model**
Population Update

- Replace parents with superior offspring
- Re-calculate fitness values

Parents

<table>
<thead>
<tr>
<th>Pair</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;husband, guide&gt;</td>
<td>0.38</td>
</tr>
<tr>
<td>&lt;police, fight&gt;</td>
<td>0.36</td>
</tr>
<tr>
<td>&lt;wife, hear&gt;</td>
<td>0.31</td>
</tr>
<tr>
<td>&lt;car, yell&gt;</td>
<td>0.29</td>
</tr>
<tr>
<td>&lt;flower, talk&gt;</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Offspring

<table>
<thead>
<tr>
<th>Pair</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;flower, shout&gt;</td>
<td>0.30</td>
</tr>
<tr>
<td>&lt;wife, fight&gt;</td>
<td>0.82</td>
</tr>
<tr>
<td>&lt;police, hear&gt;</td>
<td>0.07</td>
</tr>
<tr>
<td>&lt;flower, shout&gt;</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Relevance Feedback

- Seed Pattern
  - <husband, argue>
- New population

Relevant

- <wife, fight> 0.82

Non-relevant

- <police, fight> 0.36
- <wife, hear> 0.31
- <flower, shout> 0.30
- <car, yell> 0.29
Relevance Feedback

- Improve seed pattern by adjusting its context distribution according to relevant patterns

<husband, argue>

<wife, fight>

<police, fight>

<wife, hear>

<flower, shout>

<car, yell>

Seed pattern

Relevant

Non-relevant

Non-relevant

Non-relevant

Non-relevant

- Relevance feedback can guide the evolutionary process
  - induce more relevant patterns and move away from noisy patterns in future generations
Summary

- Topic Analysis
  - Semantic dependencies
    - Symptoms and relations
  - HAL + EIA
    - Negative Life events
Retrieval Framework

- Natural Language Query
- Topic Analysis
  - Negative Life Event Identification
  - Symptom Identification
  - Relation Identification

Consultation Documents

Retrieval Model

Ranking

Combine the similarity of each topic
Retrieval Model

- Combining similarities of events, symptoms and relations between queries and documents

\[
Sim(q,d) = \alpha Sim_{Evn}(q,d) + \beta Sim_{Sym}(q,d) + (1 - \alpha - \beta) Sim_{Rel}(q,d)
\]

- \(Sim_{Evn}(q,d)\): similarity of negative life events between \(q\) and \(d\)
- \(Sim_{Sym}(q,d)\): similarity of symptoms between \(q\) and \(d\)
- \(Sim_{Rel}(q,d)\): similarity of relation between \(q\) and \(d\)

\(\alpha, \beta\): combination factors
Experimental Results

- **Experiment setup**
  - **Baselines:** (no topic knowledge, just word matching)
    - Vector Space Model (VSM)
    - Okapi model (BM25, BM15, BM11)
  - VSM and Okapi models **compare each word** in queries and document

- **Comparative results**
  - Precision
  - Efficiency (Query processing time)
Data

- **Web sites**
  - John Tung Foundation (http://www.jtf.org.tw)
  - PsychPark (http://www.psychpark.org)

- A total of 3,650 documents were collected
  - Test query set: 20 documents (randomly selected)
  - Tuning set: 100 documents (randomly selected)
  - Retrieval set: the remaining 3,530 documents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Avg. Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Life Event</td>
<td>1.45</td>
</tr>
<tr>
<td>Depressive Symptom</td>
<td>4.40</td>
</tr>
<tr>
<td>Semantic Relation</td>
<td>3.35</td>
</tr>
</tbody>
</table>

Table 1. Number of topics in test query set
## Results (Precision)

<table>
<thead>
<tr>
<th></th>
<th>Prec(5)</th>
<th>Prec(10)</th>
<th>Prec(20)</th>
<th>Prec(50)</th>
<th>Prec(100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>4.4624</td>
<td>6.7023</td>
<td>7.1156</td>
<td>7.8129</td>
<td>8.6597</td>
</tr>
<tr>
<td>BM11</td>
<td>3.8877</td>
<td>4.9328</td>
<td>5.9589</td>
<td>6.9703</td>
<td>7.7057</td>
</tr>
<tr>
<td>VSM</td>
<td>2.3454</td>
<td>3.3195</td>
<td>4.4609</td>
<td>5.8179</td>
<td>6.6945</td>
</tr>
<tr>
<td>BM15</td>
<td>2.1362</td>
<td>2.6120</td>
<td>3.4487</td>
<td>4.5452</td>
<td>5.7020</td>
</tr>
</tbody>
</table>

* Topic vs BM 25 significantly different (t-test, p<0.05)

Topic model outperformed both Okapi and VSM
Results (Efficiency)

- **Query processing time**
  - The time for processing all test queries
  - Tested on a PC with Windows XP operating system, a 2.4GHz Pentium IV processor and 512MB RAM

- **Topic model required more processing time than did word-based models**

<table>
<thead>
<tr>
<th>Retrieval Model</th>
<th>Avg. Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>17.13</td>
</tr>
<tr>
<td>VSM</td>
<td>0.68</td>
</tr>
<tr>
<td>BM25</td>
<td>0.48</td>
</tr>
</tbody>
</table>

*Topic model involved more detailed analysis of queries*
Why Topic model outperformed word-based models?

- Symptoms and events can each be expressed by different words
  - Word-based models failed to retrieve relevant documents when different words were used in the input query
- Word-based models cannot capture relations between symptoms
Conclusion

- NLP techniques are useful for many web applications
- Information retrieval
  - Semantic dependencies
  - HAL model + Evolutionary + Relevance feedback
  - Improve retrieval results
Thank you!