BioChain: Lexical Chaining Methods for Biomedical Text Summarization

Lawrence Reeve, Hyoil Han, and Ari D. Brooks

College of Information Science and Technology, College of Medicine, Drexel University

Abstract

Concept chaining is a technique for identifying semantically-related terms in text. We propose concept chaining to link semantically-related concepts within biomedical text together. The resulting concept chains are then used to identify candidate sentences useful for extraction. The extracted sentences are used to produce a summary of the biomedical text. The concept chaining process is adapted from existing lexical chaining approaches, which focus on chaining semantically-related terms, rather than semantically-related concepts.

Approach

- Apply methods/concepts from lexical chaining:
  - Cluster (chain) words together based on semantic-relatedness
  - Words are chained together based on word 'senses' (concepts)
- Lexical Chaining...
  - Identifies lexical coherence
  - Captures core themes of a text (aboutness)
  - Is an intermediate format
- Example: (Doran et al., 2004)
  - "The house contains an attic. The home is a cabin."

Need for summarization

1) There is no 'ideal' summary
   - Depends on reader's information need
2) Abstract may be missing ideas from the full-text
3) May want alternative summaries
4) Use in question-answering systems to provide personalized information
5) Evaluate summary selection methods for use in multi-document summarization

Evaluation

- Avg precision=0.90, recall=0.92
- Avg 8 of strong chains in full text is 3
- Represents 2% of all possible semantic types
- Avg unique UMLS concepts in abstract is 8
- Avg 80% coverage of concepts in filter

Conclusions

- Use lexical chaining approaches with existing UMLS resources to identify the 'aboutness' of a text using concepts vs terms
- Extract sentences containing stronger concepts within a strong (semantic type) chain
- Result is an indicative summary of what text is about
- Evaluation shows concept chaining (BioChain) is strong between human summary and full text

Publications:


Figure 1 Concept Chaining Process

<table>
<thead>
<tr>
<th>Concept Chaining Implemented Using UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMLS (Unified Medical Language System)</td>
</tr>
<tr>
<td>Developed by National Library of Medicine</td>
</tr>
<tr>
<td>Resources used in BioChain:</td>
</tr>
<tr>
<td>Metathesaurus: Maps terms into concepts</td>
</tr>
<tr>
<td>Semantic Network: Organizes related concepts</td>
</tr>
<tr>
<td>MetaMap Transfer Application: text-to-concept mapping tool</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Source Text Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract or full text from PubMed</td>
</tr>
<tr>
<td>Need to identify noun phrases within each sentence:</td>
</tr>
<tr>
<td>Concepts are derived from noun phrases using vocabulary in Metathesaurus</td>
</tr>
<tr>
<td>Conversion from PDF-formatted files:</td>
</tr>
<tr>
<td>Columns, Captions, Bibliography, Reference numbers, Images of documents, Text tables</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept Chaining Implemented Using UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMLS (Unified Medical Language System)</td>
</tr>
<tr>
<td>Developed by National Library of Medicine</td>
</tr>
<tr>
<td>Resources used in BioChain:</td>
</tr>
<tr>
<td>Metathesaurus: Maps terms into concepts</td>
</tr>
<tr>
<td>Semantic Network: Organizes related concepts</td>
</tr>
<tr>
<td>MetaMap Transfer Application: text-to-concept mapping tool</td>
</tr>
</tbody>
</table>