WME 3.0:  
An Enhanced and Validated Lexicon of Medical Concepts  

Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay
Department of
IEEE Intelligent Systems on
Affective Computing and Sentiment Analysis

http://sentic.net/acsa
Presentation Outline:

- Introduction
- Challenges
- Motivations
- Previous versions of WME
  - WME 1.0
  - WME 2.0
- WME 3.0 Building
  - Enhancement of medical concepts
  - Category Assignment for medical concepts
- Evaluation of WME 3.0
  - Validation of overall lexicon
  - Validation of Individual Features
- Conclusions
- References
Introduction:

➢ Information extraction system is essential in healthcare due to the following issues,
  – Structured Corpus Preparation
  – An automated annotation system development
  – Ontology design for medical concepts
  – Medical concepts and its related features extraction
  – Understand the knowledge-based information for the medical corpus

➢ A domain-specific lexicon is important to build an automated information extraction system
Introduction:

➢ A lexicon provides the following information to recognize the contextual knowledge from the medical corpus
  – Medical concepts (e.g. Abdomen, Mass)
  – Linguistic Features
    • Category
    • Parts-Of-Speech (POS)
    • Gloss (Descriptive Definition)
  – Conceptual Features
    • Affinity score (relation between a pair of concepts)
    • Gravity score (relation between concept and its gloss)
  – Sense-based Features
    • Polarity score
    • Sentiment
    • Similar Sentiment Words (SSW)
Challenges:

➢ Unavailability of structured corpora
➢ It is challenging to find an annotated dataset, which combinedly labels the fundamental categories of medical concepts such as Diseases, Symptoms, and Drugs
➢ Isolation of general concepts and medical concepts are difficult
➢ Disambiguation of polarities of the medical concepts
➢ Hard to recognize similar types of diseases or symptoms of a particular disease or symptom
Motivation:

- Development of Structured Corpus
- Medical concepts and related information extraction
- Enrichment of our previously developed medical lexicon (WME 2.0):
  - Enhance more number of medical concepts
  - Recognize the existing features of WME 2.0 for the additional medical concepts
  - Additional category assignment for medical concepts
Motivations:

- Medical Concepts
  - Concept
    - Title: amnesia
    - Properties
      - POS: noun
      - Category: disease
      - Gloss: Loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.
Motivation:

<?xml version="3.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
    .........
  - <Concept>
    - <Title>amnesia</Title>
    - <Properties>
      <POS>noun</POS>
      <Category>disease</Category>
      <Gloss>Loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.</Gloss>
      - <SSW and Affinity score>
        blackout (0.674)
        memory loss (0.534)
        stupor (0.429)
        fugue (0.345)
      - <SSW and Affinity score>
        <Polarity score> - 0.375</Polarity score>
        <Gravity score>0.170</Gravity score>
        <Sentiment>negative</Sentiment>
  - <Properties>
  - <Concept>
    .........
  - <Concept>
  </Medical Concepts>

Annotated Corpus

Hundreds of foods and plant-based fiber products are available to relieve constipation naturally.

sounding as if the nose were pinched.

abnormal dryness of the conjunctiva and cornea of the eyes; may be due to a systemic deficiency of vitamin A.

Giant cell interstitial pneumonia (GIP) is a rare form of pulmonary fibrosis.
Previous versions of WME
Seed list and Useful resources for the previous versions of WME:

- SemEval 2015, Task-6 Trial and Training Datasets
- Pre-processed English Medical Dictionary
- Conventional WordNet
- SentiWordNet
- SenticNet
- Bing Liu subjective list
- Taboada’s adjective list
WME 1.0 Lexicon:

- Total number of medical concepts: 6415
- Parts-Of-Speech viz. Noun, Verb, Adjective etc.
- Gloss: Descriptive definition
- Polarity score: ranges from -1 to +1
- Sentiment: positive or negative
Sample output of WME 1.0 Lexicon:

```xml
<?xml version="1.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
    ........
  </Concept>
  - <Concept>
    <Title>abdominal_cavity</Title>
    - <Properties>
      <POS>noun</POS>
      <Gloss>The cavity containing the major viscera; in mammals it is separated from the thorax by the diaphragm.</Gloss>
      <Polarity score> - 0.500</Polarity score>
      <Sentiment>negative</Sentiment>
    </Properties>
  </Concept>
  - <Concept>
    ........
  </Concept>
</Medical Concepts>
```

WME 2.0


WME 2.0 Lexicon:

- Total number of medical concepts: 6415
- Parts-Of-Speech viz. Noun, Verb, Adjective etc.
- Gloss: Descriptive definition
- Polarity score: ranges from -1 to +1
- Sentiment: positive or negative
- SSW: Similar Sentiment Words
- Affinity score: ranges from 0 to +1
- Gravity score: ranges from -1 to +1
Sentiment based Relational Features:

➢ Sentiment based relations help to identify the hidden links between medical concepts

➢ It also assists in recognizing the proper link between various concepts and their different source of glosses

➢ The proposed sentiment based relational features are Affinity score and Gravity score
Affinity Score:

➢ Affinity refers to the linking between pair of medical concepts by determining sentiment from their common Similar Sentiment Words (SSW)

➢ Affinity score is obtained by a probabilistic count of similar sentiment-based concepts as shown in Equation 1 and 2,

First, we define the overlapping SSW of each concept pair as

\[ \text{Affinity}_c = MC_1 \cap MC_2 \]  \hspace{1cm} (1)

where \( MC_1 \) and \( MC_2 \) represents SSW sets of two different medical concepts and the \( \text{Affinity}_c \) implies the number of common SSW of \( MC_1 \) and \( MC_2 \).

Finally, Affinity score for a concept (\( MC_1 \)) is

\[ \text{Affinity Score}_c = \frac{\text{Affinity}_c}{MC_1 + MC_2} \]  \hspace{1cm} (2)
Affinity Score:

(a) 4 common SSW *breathing, respiration, ventilation,* and *external respiration* are determined for the concept pair of *abdominal breathing* and *hypopnea* out of total 8 SSW, resulting in affinity score Affinity Score \( c = 0.500 \).

(b) 3 common SSW *breathing, ventilation,* and *external respiration* out of total 12 SSW with respect to the pair *abdominal breathing* and *wheeze* and it provides Affinity Score \( c = 0.250 \) that indicates finite but weaker relations of the pair in (a).
Gravity Score:

➢ Gravity presents the relevance of sentiment appeared between a medical concept and its glosses.

➢ Gravity score confirms whether a gloss (or contextual information) pertaining a medical concept appropriates or not.

In order to achieve gravity score, we first consider the polarity score of each word in glosses and is denoted as Polaritygc.

Secondly, we calculate the gravity score Gravitygc for each medical concept which aims to determine the sentiment relations with the attached gloss by considering the polarity of medical concepts Polarityc.

\[ \text{Gravitygc} = \text{Polarityc} \times \text{Polaritygc} \]  \hspace{1cm} (3)

The final gravity score is then simply as,

\[ \text{Gravity Scorec} = \sum \text{Gravitygc}, i=1 \text{ to } N \]  \hspace{1cm} (4)
Gravity Score:

➢ The medical concept *sickness* with the polarity score $\text{Polarity}_c = -0.619$ is given with its gloss along with the medical concepts of WME 2.0 and represented in blue color whereas all others are shown as red.

➢ Every word is provided with its gross polarity $\text{Polarity}_{gc}$ as mentioned in the parenthesis.

➢ The score is assigned by the corresponding lexicons, e.g., the scores of medical concepts are collected from WME 2.0 and the others from either SenticNet or SentiWordNet.
Sample output of WME 2.0 Lexicon:

```
<?xml version="2.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
    ............
  </Concept>
- <Concept>
  <Title>amnesia</Title>
  <Properties>
    <POS>noun</POS>
    <Gloss>Loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.</Gloss>
    - <SSW and Affinity score>
      blackout (0.674)
      memory loss (0.534)
      stupor (0.429)
      fugue (0.345)
    </SSW and Affinity score>
    <Polarity score> - 0.375 </Polarity score>
    <Gravity score> 0.170 </Gravity score>
    <Sentiment> negative </Sentiment>
  </Properties>
  </Concept>
- <Concept>
  ............
  </Concept>
</Medical Concepts>
```

WME 3.0 Building
WME 3.0 Lexicon:

- Medical ontology based resource preparation
- Total number of medical concepts: 10186
- Parts-Of-Speech viz. Noun, Verb, Adjective etc.
- Gloss: Descriptive definition
- Polarity score: ranges from -1 to +1
- Sentiment: positive or negative
- Category: Diseases, Symptoms, Drugs, Human Anatomy and MMT (Miscellaneous Medical Terms)
- SSW: Similar Sentiment Words
- Affinity score: ranges from 0 to +1
- Gravity score: ranges from -1 to +1
Additional Medical Concepts Identification:

➢ Two resources namely WordNet and MedicineNet assist in enhancing 3771 number of medical concepts with WME 2.0

➢ Recognize features viz. POS, gloss, SSW, polarity score, sentiment, affinity score and gravity score for these additional concepts using following resources and machine learning approaches
  ➢ SentiWordNet
  ➢ SenticNet
  ➢ Bing Liu subjective list
  ➢ Taboada’s adjective list
  ➢ A preprocessed medical dictionary
Category Assignment to Medical Concepts:

- Assigned categories are diseases, symptoms, drugs, human anatomy and Miscellaneous Medical Terms (MMT)

- WME 3.0 assigned features of medical concepts help to assign the category of medical concepts in the presence of two well-known classifiers

- The classifiers are Naive Bayes and Logistic Regression
Sample output of WME 3.0 Lexicon:

```xml
<?xml version="3.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
    
  - <Concept>
    <Title>amnesia</Title>
    - <Properties>
      <POS>noun</POS>
      <Category>disease</Category>
      <Gloss>Loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.</Gloss>
      - <SSW and Affinity score>
        blackout (0.674)
        memory loss (0.534)
        stupor (0.429)
        fugue (0.345)
      </SSW and Affinity score>
      <Polarity score>- 0.375</Polarity score>
      <Gravity score>0.170</Gravity score>
      <Sentiment>negative</Sentiment>
    </Properties>
  </Concept>
  - <Concept>
    
  - <Concept>
</Medical Concepts>
```
A comparative study between all three versions of WME:

<table>
<thead>
<tr>
<th></th>
<th>WME 1.0</th>
<th>WME 2.0</th>
<th>WME 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Concepts</td>
<td>6415</td>
<td>6415</td>
<td>10186</td>
</tr>
<tr>
<td><strong>n-grams</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uni-gram</td>
<td>2956</td>
<td>2956</td>
<td>3722</td>
</tr>
<tr>
<td>Bi-gram</td>
<td>2837</td>
<td>2837</td>
<td>3866</td>
</tr>
<tr>
<td>Tri-gram</td>
<td>622</td>
<td>622</td>
<td>1762</td>
</tr>
<tr>
<td><strong>POS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td>4248</td>
<td>4248</td>
<td>7677</td>
</tr>
<tr>
<td>Verbs</td>
<td>2056</td>
<td>2056</td>
<td>2352</td>
</tr>
<tr>
<td>Adjectives</td>
<td>111</td>
<td>111</td>
<td>157</td>
</tr>
<tr>
<td><strong>Sentiment and Polarity score</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive (&gt; = 1)</td>
<td>2800</td>
<td>2800</td>
<td>3227</td>
</tr>
<tr>
<td>Negative (&lt; 1)</td>
<td>3615</td>
<td>3615</td>
<td>6959</td>
</tr>
<tr>
<td>Affinity score</td>
<td>WME 1.0</td>
<td>WME 2.0</td>
<td>WME 3.0</td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>0 to 0.5</td>
<td>-</td>
<td>4325</td>
<td>7177</td>
</tr>
<tr>
<td>0.5 to 1</td>
<td>-</td>
<td>2090</td>
<td>3009</td>
</tr>
<tr>
<td>Gravity score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than zero</td>
<td>-</td>
<td>2320</td>
<td>3783</td>
</tr>
<tr>
<td>Equal to zero</td>
<td>-</td>
<td>732</td>
<td>1961</td>
</tr>
<tr>
<td>Greater than zero</td>
<td>-</td>
<td>3363</td>
<td>4442</td>
</tr>
<tr>
<td>Category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diseases</td>
<td>-</td>
<td>-</td>
<td>3243</td>
</tr>
<tr>
<td>Drugs</td>
<td>-</td>
<td>-</td>
<td>3390</td>
</tr>
<tr>
<td>Symptoms</td>
<td>-</td>
<td>-</td>
<td>1409</td>
</tr>
<tr>
<td>Human anatomy</td>
<td>-</td>
<td>-</td>
<td>227</td>
</tr>
<tr>
<td>MMT</td>
<td>-</td>
<td>-</td>
<td>1917</td>
</tr>
</tbody>
</table>
Evaluation
Evaluation Process:

➢ We have conducted an agreement analysis with the help of two manual annotators to evaluate WME 3.0 lexicon
➢ The manual annotators are medical practitioners
➢ The agreement score ($k$ score) has been calculated using the following equation,

$$
K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
$$

where, $k$ is the cohen's kappa agreement score
$Pr(a)$ is the observed proportion of full agreement between two annotators.
$Pr(e)$ is the proportion expected by a chance which indicates a kind of random agreement between the annotators.
Validation of overall WME 3.0 lexicon:

<table>
<thead>
<tr>
<th>No. of Concepts: 10186</th>
<th>Annotator-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Annotator-2</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

Cohen's Kappa score \((k) = 0.79\)
Validation of individual features of WME 3.0:

<table>
<thead>
<tr>
<th>Annotator-2</th>
<th>Category</th>
<th>Yes</th>
<th>No</th>
<th>K score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>8778</td>
<td>93</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>161</td>
<td>1154</td>
<td></td>
</tr>
<tr>
<td>POS</td>
<td>Yes</td>
<td>9229</td>
<td>52</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>92</td>
<td>813</td>
<td></td>
</tr>
<tr>
<td>Gloss</td>
<td>Yes</td>
<td>8805</td>
<td>97</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>172</td>
<td>1112</td>
<td></td>
</tr>
<tr>
<td>No. of Concepts: 10186</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annotator-2</td>
<td>No. of Concepts: 10186</td>
<td>Annotator-1</td>
<td>K score</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------------------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>SSW</td>
<td>Yes</td>
<td>8767</td>
<td>137</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>256</td>
<td>1026</td>
<td></td>
</tr>
<tr>
<td>Sentiment</td>
<td>Yes</td>
<td>8727</td>
<td>67</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>124</td>
<td>1268</td>
<td></td>
</tr>
</tbody>
</table>
Validation of individual categories of WME 3.0:

<table>
<thead>
<tr>
<th>Annotator-2</th>
<th>No. of Concepts: 10186</th>
<th>Annotator-1</th>
<th>K score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Disease (3243)</td>
<td>Yes</td>
<td>2794</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>51</td>
<td>367</td>
</tr>
<tr>
<td>Symptom (1409)</td>
<td>Yes</td>
<td>1214</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>26</td>
<td>155</td>
</tr>
<tr>
<td>Drug (3390)</td>
<td>Yes</td>
<td>2922</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>53</td>
<td>381</td>
</tr>
<tr>
<td>Annotator-2</td>
<td>No. of Concepts: 10186</td>
<td>Annotator-1</td>
<td>K score</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Human anatomy (227)</td>
<td></td>
<td>Yes</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>3</td>
</tr>
<tr>
<td>MMT (1917)</td>
<td></td>
<td>Yes</td>
<td>1652</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>28</td>
</tr>
</tbody>
</table>
Conclusions
Conclusions:

➢ An enriched version of a medical lexicon viz. WME 3.0

➢ Category assignment for Medical concepts

➢ Various syntactic and semantic features extraction for medical concepts

➢ WME 3.0 may help to reduce the gap between medical experts and non-experts

➢ Assistance to prepare an annotated corpus
References
References:

References:


Thank You....