Grouping the Pharmacovigilance Terms with a Hybrid Approach

Marie Dupuch\textsuperscript{1}, Laëtitia Dupuch\textsuperscript{2}, Amandine Périnet\textsuperscript{3}, Thierry Hamon\textsuperscript{3}, Natalia Grabar\textsuperscript{1}

\textsuperscript{1}CNRS UMR 8163 STL, Université Lille 3, France
\textsuperscript{2}Université Toulouse III Paul Sabatier, France
\textsuperscript{3}LIM&BIO (EA3969) UFR SMBH, Université Paris 13, France
Plan

- Context
- Material and Methods
- Results and Discussion
- Limitations and Perspectives
Context
Pharmacovigilance

Pharmacovigilance:

- Collection, *analysis* and prevention of Adverse Drug Reactions (ADRs)

Grouping ADRs reports:

- Based on the MedDRA term coding
  - Medical Dictionary for Drug Regulatory Activities (Brown et al, 1999)
- Detection of the safety alerts (signal detection)
  - Statistical methods (ROR, IC, CHI2, PRR, YuleQ, EBGM...)
  - Groups of terms
  - Fine-grained terminology
  - *hepatitis infectious*, *hepatitis infectious mononucleosis*, *hepatitis viral*...
Context

MedDRA Terminology

- Coding Adverse Drug Reactions
- Signs and symptoms, diagnosis, medical and surgical procedures, etc.
- Hierarchically structured

Marie Dupuch et al.
Context

Objective: Group semantically similar terms

- **MedDRA**
  - HLT terms
  - SMQs (Standardized MedDRA Queries)
    - Manually built by experts (CIOMS, 2004)
    - MedDRA structure
    - Scientific literature
    ⇒ Long and meticulous work

- **Other resources (ontoEIM):**
  - Hierarchical subsumption (Alecu et al, 2008; Jaulent et al, 2009)
  - Semantic distance (Bousquet et al, 2005; Iavindrasana, 2006)
  - Subsets of MedDRA terms, no evaluation

- **Semantic methods to systematically group MedDRA terms**
  - Semantic similarity methods
  - Terminology structuring methods: synonymy, hyperonymy
  - Evaluation:
    - comparison with the SMQs
    - expert evaluation
Context

SMQs or Standardised MedDRA Queries

- Groups of MedDRA terms related to a diagnosis (84 SMQs)
  - *Acute renal failure, Hepatic disorders, Thrombocytopenia*

PTs within SMQs ∈ different SOCs

Number of SOCs per SMQ:
- Varies between 4 to 25
- The average is: 8.26 SOCs per SMQ
General schema of the method

Semantic distance and similarity approaches
- Computing of the semantic distance and similarity
  - TerminoEIM resource (Rada et al, 1989)
  - (Leacock & Chodorow, 1998)
  - (Zhong et al, 2002)
- Detection of synonymy relations
- Detection of hierarchical relations
- ontoEIM resource

Clustering of MedDRA terms
- Radius
- HAC (with the R project)

Pre-processing
- Raw list of MedDRA terms
- POS-tagging
- Syntactic analysis
- Ogmios platform
- Genia tagger
- YaTeA

Term structuring
- Lexical inclusion
- Faster
- Detection of synonymy relations
- Synoterm + resources
- Faster

Clustering within directed graphs
- Strongly connected components
- Connected components

Evaluation
- SMQs

Marie Dupuch et al.
General schema of the method

Semantic distance and similarity approaches
Computing of the semantic distance and similarity
(Leacock & Chodorow, 1998)
(Zhong et al, 2002)
Radius
HAC (with the R project)

Clustering of MedDRA terms

Terminology structuring approach
Pre-processing
POS-tagging
Syntactic analysis
Raw list of MedDRA terms
Ogmios platform
Genia tagger
YaTeA

Term structuring
Detection of hierarchical relations
Lexical inclusion
Faster
Detection of synonymy relations
Synoterm + resources
Faster

Clustering within directed graphs
Strongly connected components
Connected components

Evaluation
SMQs

Marie Dupuch et al.
General schema of the method

\[ \text{dist}_{\text{ontoEIM}}(A, B) = \sum_{i \in \{ADR, M, T\}} W_i \times \text{sp}(A_i, B_i) \]

\[ = \frac{\sum_{i \in \{ADR, M, T\}} W_i \times \text{sp}(A_i, B_i)}{\sum_{j \in \{ADR, M, T\}} W_j} = 3.5 \]
General schema of the method

Semantic distance and similarity approaches
- Computing of the semantic distance and similarity
  - ontoEIM resource (Zhong et al, 2002)
  - (Leacock & Chodorow, 1998)

Clustering of MedDRA terms
- Radius
- HAC (with the R project)

Pre-processing
- POS-tagging
- Syntactic analysis
- Syntactic analysis
- Ogmios platform
- Genia tagger
- YaTeA

Term structuring
- Detection of hierarchical relations
- Lexical inclusion
- Faster
- Detection of synonymy relations
- Faster
- Synoterm + resources
- Faster

Clustering within directed graphs
- Strongly connected components
- Connected components

Evaluation
- SMQs

Raw list of MedDRA terms

Terminology structuring approach

\{external ear lesion excision, lesion excision\}
\{malignant neoplasm of orbit, neoplasm of orbit\}
\{acute promyelocytic leukaemia, acute leukaemia\}
General schema of the method

Semantic distance and similarity approaches

Computing of the semantic distance and similarity

- OntoEIM resource
- Leacock & Chodorow (1998)
- Zhong et al. (2002)

Clustering of MedDRA terms

- Radius
- HAC (with the R project)

Terminology structuring approach

Pre-processing

- POS-tagging
- Syntactic analysis
- Ogmios platform
- Genia tagger
- YaTeA

Raw list of MedDRA terms

Term structuring

- Detection of hierarchical relations
- Lexical inclusion
- Faster
- Detection of synonymy relations
- Synoterm + resources
- Faster

Clustering within directed graphs

- Strongly connected components
- Connected components
- Faster

Merging of the clusters

Evaluation

SMQs

\{sepsis, infection\} \rightarrow \{wound sepsis, wound infection\}

\{arterial insufficiency, artery insufficiency\}

\{eye penetration, penetrating eye\}
General schema of the method

Semantic distance and similarity approaches
Computing of the semantic distance and similarity
- POS-tagging
- Syntactic analysis
- ontoEIM resource

Detection of hierarchical relations
(Rada et al., 1989)
(Leacock & Chodorow, 1998)
(Zhong et al., 2002)

Detection of synonymy relations

Clustering of MedDRA terms
- Clustering within directed graphs
- Faster
- Connected components
- SMQs

Pre-processing
- Raw list of MedDRA terms
- Ogmios platform
- Genia tagger
- YaTeA

Term structuring
- Detection of hierarchical relations
- Lexical inclusion
- Faster

Evaluation
- HAC (with the R project)
- Radius
- non-disjoint clusters

Terminology structuring approach
- Synoterm + resources
- Faster

Marie Dupuch et al.
Evaluation of the computed clusters

- Correctness of the acquired relations
- Quantitative evaluation (comparison with SMQs):
  - $P$ (specificity)
    - number of relevant grouped terms as a percentage of the total number of the grouped terms
  - $R$ (sensitivity)
    - number of relevant grouped terms as a percentage of the number of terms in the corresponding SMQ
  - $F$ (f-mesure)
    - the harmonic mean of $P$ and $R$ with $\beta = 1$
- Qualitative evaluation with an expert
## Results

### Terminology structuring methods

<table>
<thead>
<tr>
<th>Methods and relationships</th>
<th>#relations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hierarchical relations</strong></td>
<td></td>
</tr>
<tr>
<td>Lexical inclusions</td>
<td>3,366</td>
</tr>
<tr>
<td>Morpho-syntactic variants</td>
<td>743</td>
</tr>
<tr>
<td><strong>Medical synonyms</strong></td>
<td></td>
</tr>
<tr>
<td>3 biomedical terminologies</td>
<td>1,879</td>
</tr>
<tr>
<td>UMLS/Filtered UMLS</td>
<td>190</td>
</tr>
<tr>
<td>Morpho-syntactic variants</td>
<td>100</td>
</tr>
<tr>
<td><strong>Medical synonyms and WordNet</strong></td>
<td></td>
</tr>
<tr>
<td>3 biomedical terminologies</td>
<td>1,939</td>
</tr>
<tr>
<td>UMLS/Filtered UMLS</td>
<td>227</td>
</tr>
</tbody>
</table>
## Results

### Clustering of terms

<table>
<thead>
<tr>
<th>Approach</th>
<th>SMQs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#clus</td>
</tr>
<tr>
<td>Semantic distance</td>
<td>2,931</td>
</tr>
<tr>
<td>Structuring (hie)</td>
<td>748</td>
</tr>
<tr>
<td>Structuring (hie+syn)</td>
<td>748</td>
</tr>
<tr>
<td>Merging (hie)</td>
<td>2,998</td>
</tr>
<tr>
<td>Merging (hie+syn)</td>
<td>2,998</td>
</tr>
</tbody>
</table>
Results

Correctness of the acquired relations

- Manual analysis of the generated hierarchical relations
- 144 pairs (5%) with syntactic ambiguities:
  - *anticonvulsant drug level, anticonvulsant drug level*
  - *cranial nerve injury, cranial nerve injury*
  - *eye movement disorder, eye movement disorder*
- Semantic relations remain correct
  - the constraint involved through the syntactic analysis guarantees correct propositions
Results

Terminology structuring methods

Marie Dupuch et al.
Results
Semantic similarity methods
Results

Merging of the two methods
Results

Analysis with an expert

SMQ *Embolic and thrombotic events, arterial*

- **False positives:**
  - *Cerebral arteriosclerosis, Cervical myelopathy, Intercostal neuralgia, Acute disseminated encephalomyelitis...*

- **Too general terms:**
  - *Ischaemia, Arterial disorder, Artery thrombosis, Shock, Vascular insufficiency...*

- **Relevant terms missing in the SMQs:**
  - *Varicose veins vulval, Cerebral ataxia, Hepatic artery stenosis, Renal artery stenosis...*
## Results

### Analysis with an expert

<table>
<thead>
<tr>
<th>SMQs</th>
<th>Number of terms</th>
<th>Reference</th>
<th>After expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SMQ</td>
<td>clu</td>
<td>com</td>
</tr>
<tr>
<td>Angioedema$_{sd}$</td>
<td>52</td>
<td>32</td>
<td>13</td>
</tr>
<tr>
<td>Angioedema$_{st}$</td>
<td>52</td>
<td>31</td>
<td>19</td>
</tr>
<tr>
<td>Angioedema$_{mrg}$</td>
<td>52</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>Embolic and thrombotic...$_{sd}$</td>
<td>132</td>
<td>159</td>
<td>48</td>
</tr>
<tr>
<td>Embolic and thrombotic...$_{st}$</td>
<td>132</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Embolic and thrombotic...$_{mrg}$</td>
<td>132</td>
<td>130</td>
<td>49</td>
</tr>
<tr>
<td>Haemodynamic oedema...$_{sd}$</td>
<td>36</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>Haemodynamic oedema...$_{st}$</td>
<td>36</td>
<td>31</td>
<td>13</td>
</tr>
<tr>
<td>Haemodynamic oedema...$_{mrg}$</td>
<td>36</td>
<td>35</td>
<td>16</td>
</tr>
</tbody>
</table>
Discussion

- At least 2 hierarchical levels in the obtained groupings
  - Important contribution of lexical inclusions
- Hierarchical structure of MedDRA: very rough
  - Intermediate hierarchical levels can be created
- Generation of non-disjoint clusters
- Quality of the gold standard
  - Important terms may be missing in the SMQs
    (Pearson 2009; Mozzicato 2007)
  - Evaluation by an expert is required
- Correctness vs. relevance of the semantic relations
Limitations and Perspectives

- Semantic similarity: 51% of the PTs only
  - Improvement of the alignment of the MedDRA terms (Nadkarni et al, 2011; Mougin et al, 2011)
- Two semantic methods exploited
  - Assumption: complementary contribution of different methods
  - Corpora for the detection of other semantic relations (Resnik, 1999; Lin, 1998; Jian & Conrath, 1997)
- Only two types of semantic relations
  - Analysis of the types of semantic relations between the terms within the SMQs
- Exploitation of more sound clustering methods
  - Creation of hierarchies of clusters
- Variation of the results according to the SMQs
- Evaluation of this work for the creation of the SMQs
  - Systematic recruitment of the terms
  - Naming of the clusters
- Evaluation: SMQs vs. signal detection