Annotating Temporal Relations to Determine the Onset of Psychosis Symptoms

Natalia Viani, PhD

IoPPN, King’s College London
Introduction: clinical use-case

• For patients with schizophrenia, longer durations of untreated psychosis (DUP) are associated with worse intervention outcomes
• In electronic health records (EHRs), information on symptom and treatment onset is often documented in the form of free text

Date: 2010-03-01
The patient reported she has been hearing voices since the age of 14…
First assessed on 07/10/2009, started treatment in November…

MeDESTO project: Measuring Duration of Untreated Psychosis by Extraction of Symptom and Treatment Onset from mental health records using language technology.

Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.
Introduction: information extraction

Entities in the texts
- Time expressions
- Events, i.e. symptoms
- Temporal relations

Our goal: Identification of temporal links (TLINKs) related to symptom onset

Image courtesy of Dr. Sumithra Velupillai, King’s College London
Background

Natural Language Processing (NLP) in the clinical domain: few corpora developed for temporal information extraction

2012 i2b2
- Intensive care unit
- 310 discharge summaries
- 8 TLINK types between entity pairs (before, overlap, ...)

2012 i2b2 NLP Challenge

THYME corpus
- Oncology
- 1,254 records
- TLINKs to document creation time and narrative containers

2015, 2016, 2017 Clinical TempEval

Aims of the study

1. We propose a methodology for selecting the most relevant documents for the considered use-case.

2. We develop a manual annotation process to temporally anchor all the relevant symptoms, thus enabling the extraction of symptom onset and other information of interest.

3. We propose a preliminary NLP system to assess the utility of the created corpus.
Dataset: corpus extraction

Mental health records from CRIS: early intervention services

1) Extract all documents related to early intervention services
36,594 documents
4,166 patients

2) Filter documents
   • length > 50th percentile
   • avg_line_length > 25th percentile
16,318 documents
3,819 patients

3) Randomly select 20 patients and annotate documents for onset information
70 documents
20 patients (1-8 docs each)

Dataset: filtering steps

• 70 documents double-annotated for symptom onset information

  “Difficulties were noted for the first time when the patient was 7 years old, as he was displaying aggressive behaviour”

• Documents analyzed in terms of clinical and temporal content, automatically identifying symptoms (keyword list) and time expressions

• Final set of documents to be used for annotation

• Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. LREC 2012.
Temporal relation annotation

Documents pre-annotated with 26 symptoms and time expressions

- Each symptom linked to a time expression in the text (if possible)
- Each symptom assigned a polarity value (positive, negative)
- Three annotators, each document double annotated

Date of Birth: 1990-05-03
Visit Date: 2016-05-04

“The patient started hearing voices at the age of 16. On mental state examination, today, auditory hallucinations could not be elicited.”

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Polarity</th>
<th>Time Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>hearing voices</td>
<td>positive</td>
<td>At the age of 16</td>
</tr>
<tr>
<td>auditory hallucinations</td>
<td>negative</td>
<td>2016-05-04</td>
</tr>
</tbody>
</table>
Automated information extraction

Dataset split into training, development, test sets

Temporal relation module

- Rule-based system: section names ("clinical history", "mse"), anchor dates (admission, discharge, clinic date)
- 10 rules

  E.g., if section = “mental state examination on admission” → link to ADM_DATE

Polarity module

- Rule-based ConText algorithm, 11 modifiers ("no", "denies", …)

Results: Corpus selection

Documents with many clinical/temporal elements are more likely to contain information on symptom onset.

Initial corpus filtered with additional criteria: 
\[ \text{Symptom\_count} > 0 \text{ and } \text{Timex\_count} > 5 \]  
\[ \rightarrow \]  
9,779 documents  
3,433 patients
Results: Corpus annotation

- **239** randomly selected patients (24 batches)
- **645** documents (2.7 per patient)

**2,590 symptoms**: hallucinations, delusions, delusional, paranoia, thought disorder, …
Results: TLINK analysis per patient

For each patient: difference (\textit{diff}) between max and min “positive” symptom dates (206 patients)

41 patients with \textit{diff} > 1 year

71 patients with \textit{diff} = 0 days
Results: NLP development

Accuracy of NLP modules

<table>
<thead>
<tr>
<th>Item</th>
<th>Model</th>
<th>Train</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLINK</td>
<td>baseline</td>
<td>0.47</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Rule-based</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>Polarity</td>
<td>baseline</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>ConText</td>
<td>0.93</td>
<td>0.95</td>
</tr>
</tbody>
</table>

More work on TLINK module needed

Polarity is well captured by rules
Discussion (1)

• First temporally-annotated corpus that was developed for a specific clinical use-case besides clinical timeline reconstruction

• Dataset selection is a crucial step

Novelty of TLINK annotation task

• Temporal links between entities that are not close to each other

• Only one type of temporal link
Discussion (2)

• Long-term goal: extracting information on a patient-level
• Analysis of 41 patients with a diff value > 1 year

<table>
<thead>
<tr>
<th>#</th>
<th>Clear onsets date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Clear onset date</td>
<td>He has been suffering from psychosis since he was 10 years old when he started experiencing hallucinations</td>
</tr>
<tr>
<td>15</td>
<td>Close to onset date</td>
<td>Other text would be more specific</td>
</tr>
<tr>
<td>9</td>
<td>Not an onset date</td>
<td>E.g., Error in written date</td>
</tr>
</tbody>
</table>
Ongoing work and future directions (1)

- **Corpus selection**: Extend annotation work on a new set of documents (first referrals)
- **Keyword extension**: Search for textual variants of symptoms

**Future directions**

- **NLP development**: Refine rule-based approach and explore supervised extraction methods → Apply on large patient cohort
Ongoing work and future directions (2)

Automatic keywords extension

• Word embedding models: capture similarity among words (unsupervised)
• First experiments: full early intervention services dataset (~36K documents)
• Future studies: look at additional CRIS documents and other clinical corpora

<table>
<thead>
<tr>
<th>word (bigram)</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>auditory_hallucinations</td>
<td>0.905801773</td>
</tr>
<tr>
<td>hallucination</td>
<td>0.878103852</td>
</tr>
<tr>
<td>visual_hallucinations</td>
<td>0.847281337</td>
</tr>
<tr>
<td>auditory_hallucination</td>
<td>0.845688939</td>
</tr>
<tr>
<td>perceptual_abnormalities</td>
<td>0.830208898</td>
</tr>
<tr>
<td>perceptual_disturbances</td>
<td>0.829540431</td>
</tr>
<tr>
<td>abnormal_perceptions</td>
<td>0.799558282</td>
</tr>
<tr>
<td>hallucinatory_phenomena</td>
<td>0.790262461</td>
</tr>
<tr>
<td>hallucinatory_experiences</td>
<td>0.789078951</td>
</tr>
<tr>
<td>auditory</td>
<td>0.785478473</td>
</tr>
<tr>
<td>psychotic_phenomena</td>
<td>0.783942819</td>
</tr>
<tr>
<td>passivity_phenomena</td>
<td>0.779936433</td>
</tr>
<tr>
<td>command_hallucinations</td>
<td>0.768690705</td>
</tr>
<tr>
<td>somatic_hallucinations</td>
<td>0.758699775</td>
</tr>
<tr>
<td>tactile_hallucinations</td>
<td>0.757349372</td>
</tr>
</tbody>
</table>
Acknowledgments

Sumithra Velupillai
Robert Stewart
Rashmi Patel
Joyce Kam
Lucia Yin
Somain Verma
Thank you!

Questions?