Annotating Temporal Relations to Determine the Onset of Psychosis Symptoms

Natalia Viani, PhD

IoPPN, King’s College London

National Institute for Health Research

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

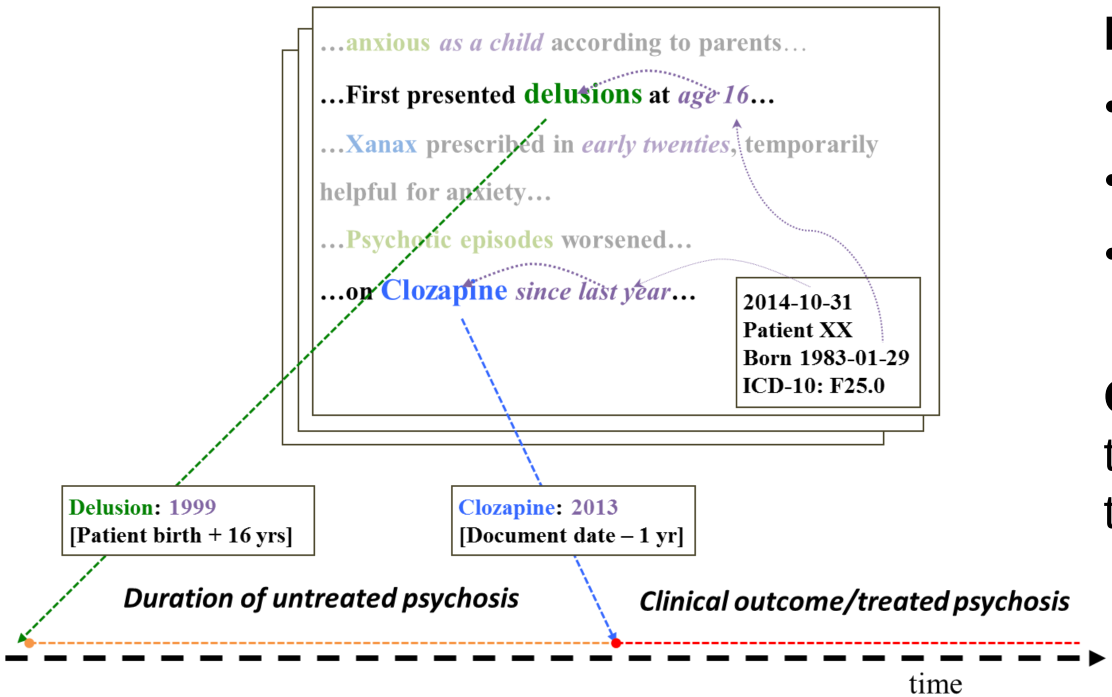


Image courtesy of Dr. Sumithra Velupillai, King's College London

Entities in the texts

- Time expressions
- Events, i.e. symptoms
- Temporal relations

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

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

- Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. Journal of biomedical informatics. 2013;46:S5–S12.
- Styler IV WF, Bethard S, Finan S, et al. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics. 2014;2:143.

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)

• Perera G, Broadbent M, Callard F, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):e008721.

National Institute for Health Research

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.
- Jackson R, Patel R, Velupillai S, et al. Knowledge discovery for Deep Pheno-typing serious mental illness from Electronic Mental Health records. F1000Res. 2018;7:210.
- Viani N, Yin L, Kam J, et al. Time Expressions in Mental Health Records for Symptom Onset Extraction. Proceedings of Louhi 2018.

NHS
National Institute for Health Research

Temporal relation annotation

Documents pre-annotated with 26 symptoms and time expressions

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."

hearing voices Polarity: positive	TLINK →	At the age of 16
auditory hallucinations Polarity: negative	TLINK →	2016-05-04

- 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

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

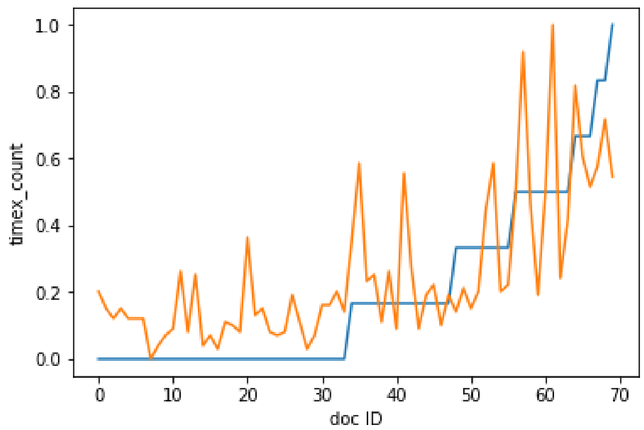
Polarity module

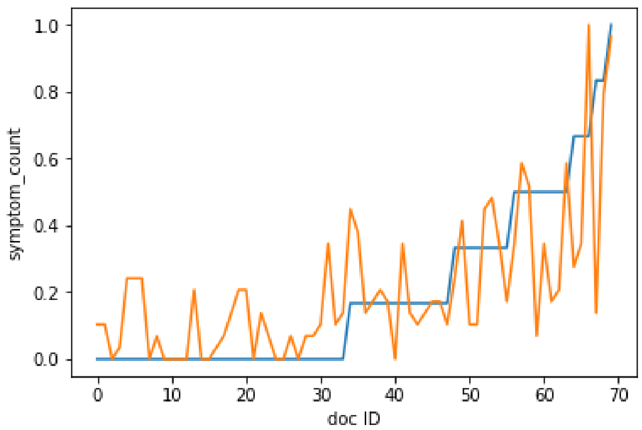
- Rule-based ConText algorithm, 11 modifiers (“no”, “denies”, …)

- Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform. 2011; 44 (5): 728–737.

Results: Corpus selection

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





Initial corpus filtered with additional criteria:
Symptom_count > *0* and *Timex_count* > *5*

→ 9,779 documents
3,433 patients

National Institute for Health Research

Results: Corpus annotation

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

Item	IAA (average)	IAA (range/batch)
TLINK	0.73	0.60 - 0.84
Polarity	0.95	0.81 - 1

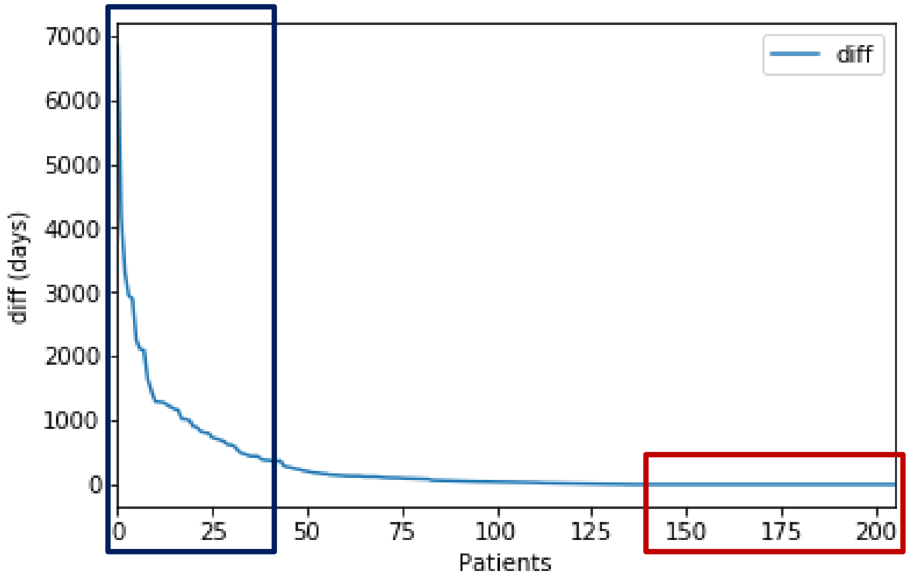
2,590 symptoms: *hallucinations*, *delusions*, *delusional*, *paranoia*, *thought disorder*, …

Item	Value	Total
TLINK	Yes	1,661 (64.1%)
	No	929 (35.9%)
Polarity	Pos	1,900 (73.4%)
	Neg	690 (26.6%)

Results: TLINK analysis per patient

For each patient: difference (*diff*) between max and min “positive” symptom dates (206 patients)

41 patients with
diff > 1 year



71 patients with
diff = 0 days

Results: NLP development

Accuracy of NLP modules

Item	Model	Train	Dev
TLINK	baseline	0.47	0.54
	Rule-based	0.67	0.58
Polarity	baseline	0.76	0.72
	ConText	0.93	0.95

More work on TLINK module needed

Polarity is well captured by rules

National Institute for Health Research

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

17	**Clear onset date**	*He has been suffering from psychosis since he was 10 years old when he started experiencing hallucinations*
15	Close to onset date	Other text would be more specific
9	Not an onset date	E.g., Error in written date

National Institute for Health Research

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

NHS
National Institute for Health Research

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

most_similar('*hallucinations*')

word (bigram)	similarity
auditory_hallucinations	0.905801773
hallucination	0.878103852
visual_hallucinations	0.847281337
auditory_hallucination	0.845688939
perceptual_abnormalities	0.830208898
perceptual_disturbances	0.829540431
abnormal_perceptions	0.799558282
hallucinatory_phenomena	0.790262461
hallucinatory_experiences	0.789078951
auditory	0.785478473
psychotic_phenomena	0.783942819
passivity_phenomena	0.779936433
command_hallucinations	0.768690705
somatic_hallucinations	0.758699775
tactile_hallucinations	0.757349372

National Institute for Health Research

Acknowledgments

National Institute for Health Research

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

National Institute for Health Research

Thank you!

Questions?