Temporal Information Extraction from Clinical Narratives

Natalia Viani

Institute of Psychiatry, Psychology and Neuroscience
King's College London

About me

- **2012-2014**: MSc in Bioengineering, “health technologies” area (University of Pavia)
- **2014-2017**: PhD student in Bioengineering and Bioinformatics (University of Pavia, laboratory of BioMedical Informatics “Mario Stefanelli”)
- **Since Jan 2018**: postdoc at the Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King’s College London

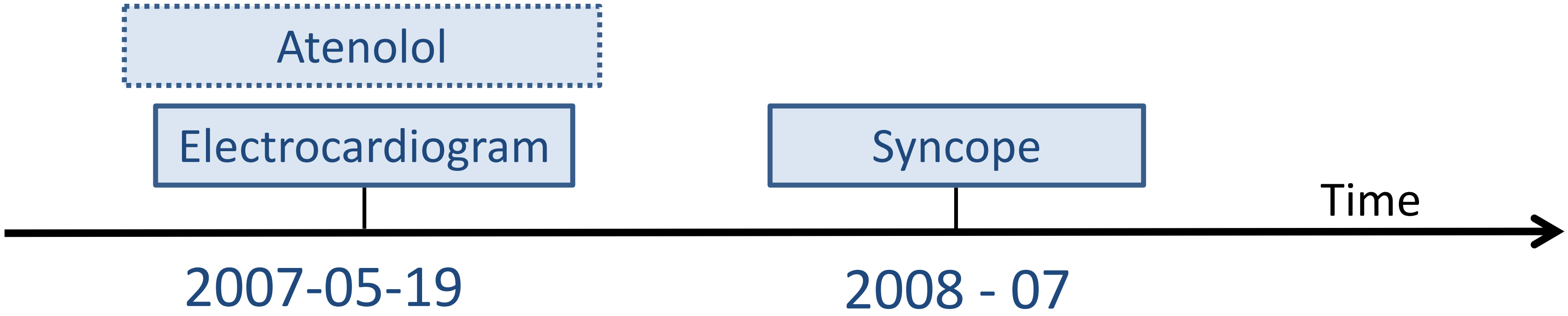
Research interests: clinical and temporal natural language processing

Clinical text: unstructured information

On 19.05.2007 electrocardiogram and Atenolol dosage increased to 50 mg twice a day.

In July 2008 syncope during physical exercise.

Denies any other symptom.



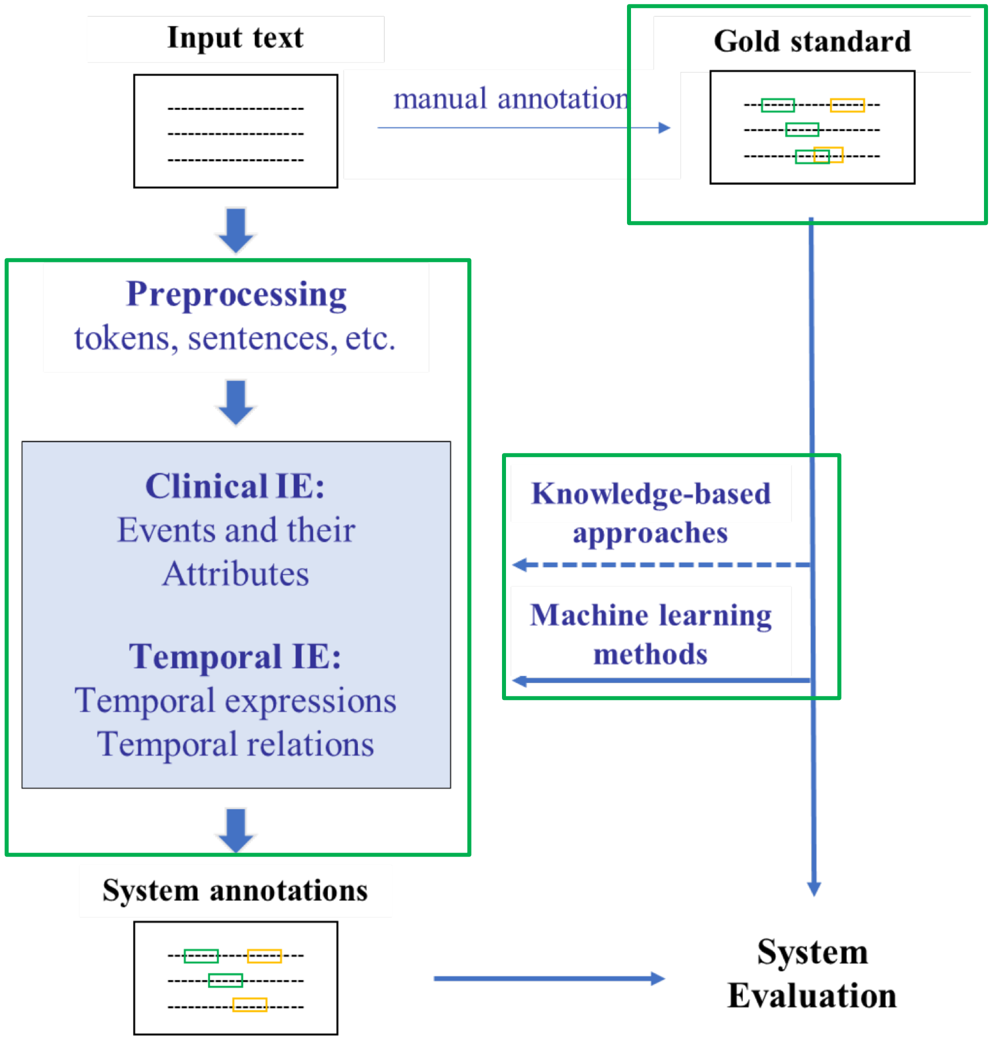
Natural language processing

*<< **Natural language processing (NLP)** is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content >>*

Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015 Jul 17;349(6245):261–6.

Information extraction: basic steps

NLP pipeline



Information extraction: methods

Knowledge-based approaches

- External dictionaries and terminologies
- Rules and regular expressions, e.g. *“\d+ %unit_of_measurement”*

→ Require manual rule engineering

Machine learning approaches

- Common classifiers (e.g., SVM, CRF)
- Deep learning approaches
- Need for annotated data

→ Need for many annotated data

Information extraction in a non-English language: Italian

Introduction and Background

Clinical IE for the Italian language

Challenges

- Lack of freely available annotated medical corpora
- Limited coverage of available clinical dictionaries
- Lack of medical-specific taggers

Temporal IE

- Annotation efforts mostly in the general domain
- Only one medical corpus (semi-automatically) annotated

Experiments in Identification of Italian Temporal Expressions

Giuseppe Attardi
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy

Luca Baronti
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy

… No temporally annotated medical corpora!

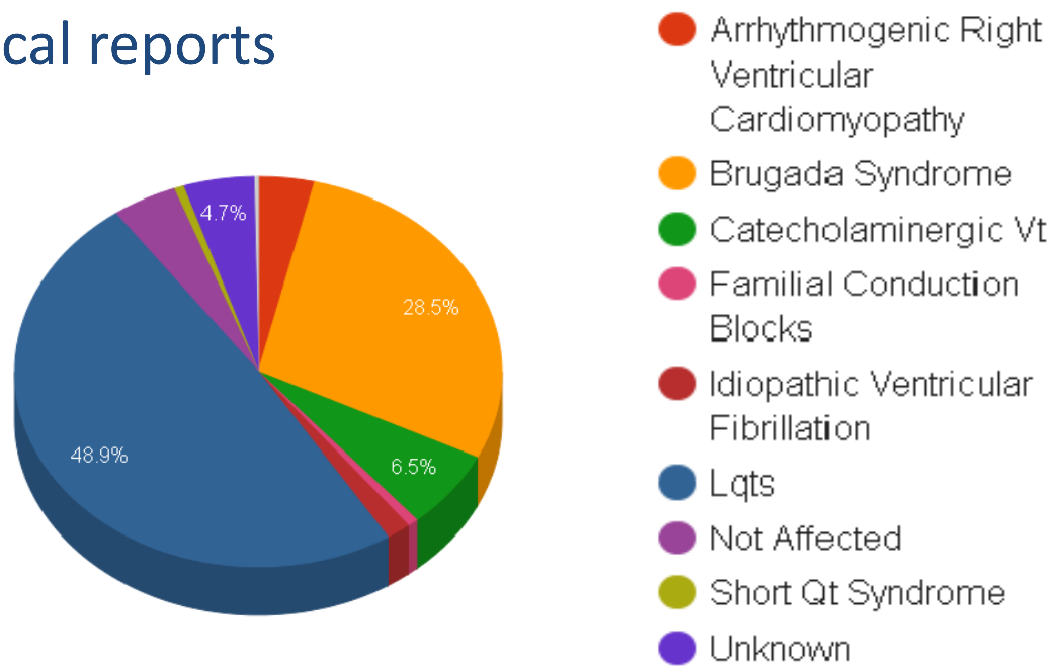
RIS: Ricerca e innovazione nella sanità. 2014. POR RIS of the Regione Toscana. Home page: http://progetto-ris.it/

Information extraction in a non-English language: Italian

Materials and Methods

Main CARDIO dataset

- Istituti Clinici Scientifici Maugeri Hospital (Pavia), Molecular Cardiology Unit
- Genetic variations in the field of inherited arrhythmogenic diseases
- 5432 medical reports



Information to be extracted (1)

Clinical events

- problems (“Brugada Syndrome”)
- tests (“ECG”)
- treatments (“Flecainide”)
- occurrences (“medical visit”)

Gli accertamenti eseguiti, in particolare, l'esito del test alla flecainide eseguito nel 2003, hanno portato a porre diagnosi di Sindrome di Brugada.

ECG: Ritmo sinusale. FC 57 bpm; PR 156 msec; QRS 106 msec; asse QRS 40°; QT 430 msec; QTc 425 msec.

accertamenti

test alla flecainide

Sindrome di Brugada

ECG

Information to be extracted (2)

Event attributes

- Test: results, findings
- Drug: dose, frequency
- ….

Gli accertamenti eseguiti, in particolare, l'esito del test alla flecainide eseguito nel 2003, hanno portato a porre diagnosi di Sindrome di Brugada.

ECG: Ritmo sinusale. FC 57 bpm; PR 156 msec; QRS 106 msec; asse QRS 40°; QT 430 msec; QTc 425 msec.

ECG

Ritmo: sinusale
Frequenza cardiaca: 57 bpm
PR: 156 msec
QRS: 106 msec
Asse QRS: 40°
QT: 430 msec
QTc: 425 msec

Information to be extracted (3)

Temporal expressions (TIMEXes)

- dates (“16/09/2010”)
- times (“2pm”)
- durations (“two months”)
- sets (“twice a day”)

Gli accertamenti eseguiti, in particolare, l'esito del test alla flecainide eseguito nel 2003, hanno portato a porre diagnosi di Sindrome di Brugada.

2003

ECG: Ritmo sinusale. FC 57 bpm; PR 156 msec; QRS 106 msec; asse QRS 40°; QT 430 msec; QTc 425 msec.

Manual annotation: 75 documents

Annotation process: created specific annotation guidelines

EVENTs

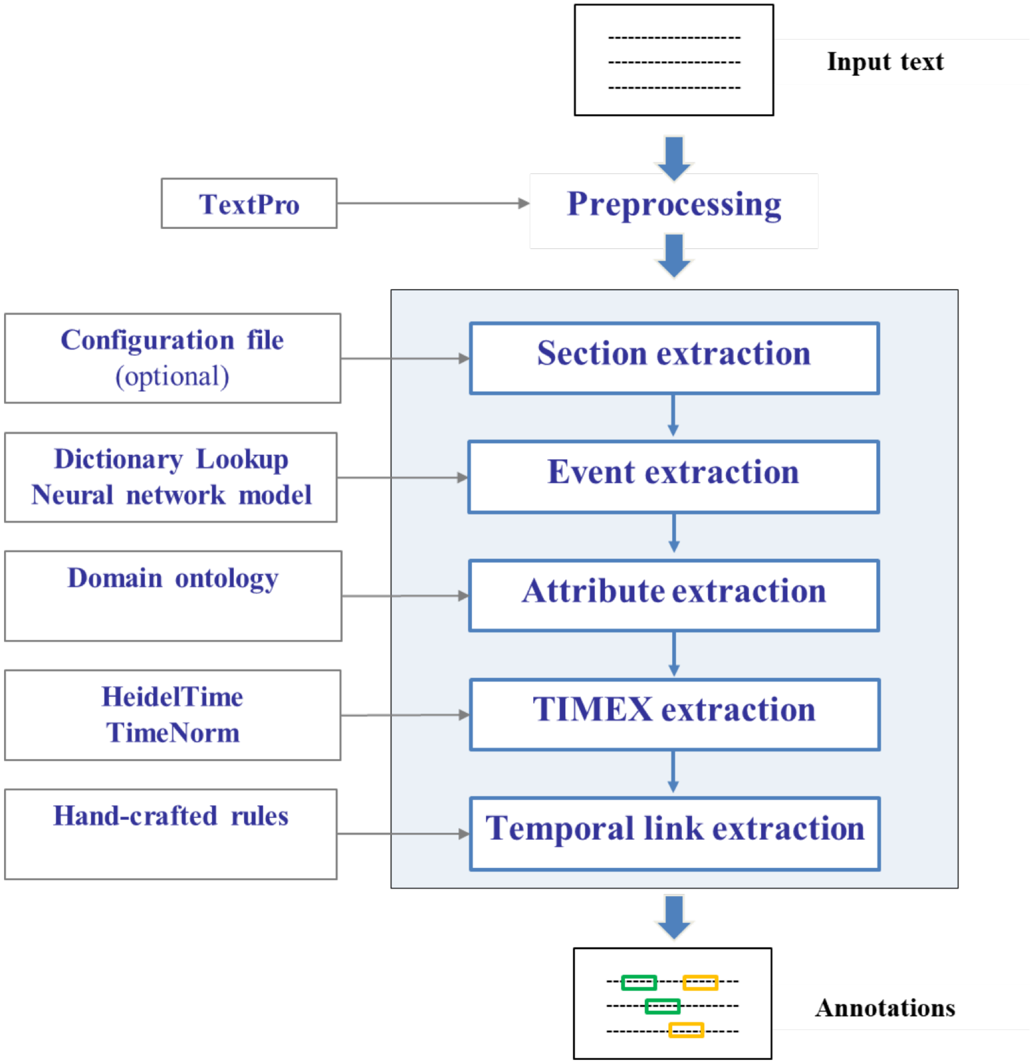
- Semantic type (problem, …)
- DocTimeRel (overlap, before, …)
- Polarity (positive, negative)
- Modality (hypothetical, …)
- Experiencer (patient, other)

TIMEXes

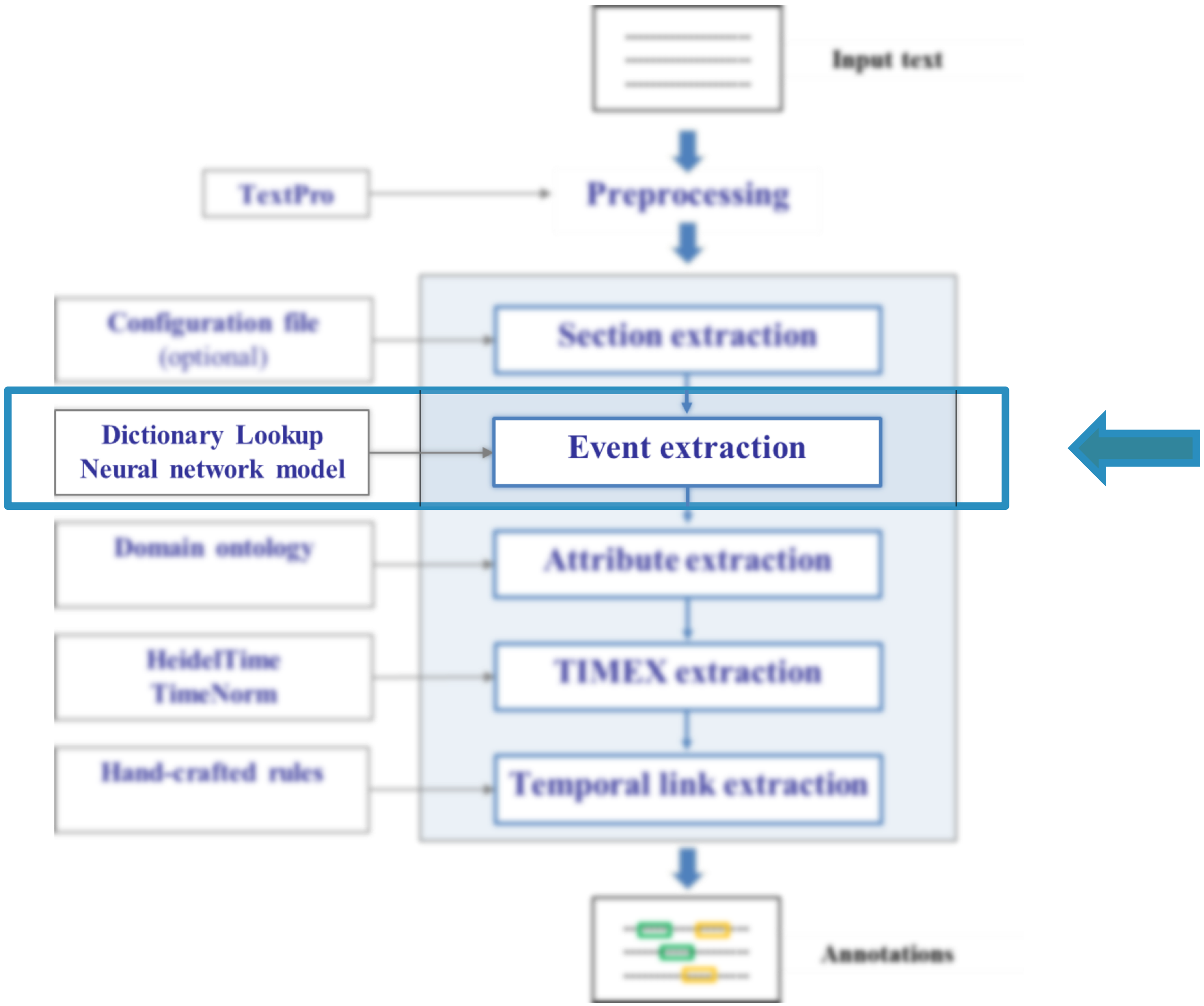
- Type (date, time, duration, set)
- Value
- Mod
- Quant (optional)
- Freq (optional)

• Chen W, Styler W. Anafora: A Web-based General Purpose Annotation Tool. Proceedings of the North American Association for Computational Linguistics Conference. 2013 Jun 9-13.

Information extraction pipeline



Event extraction



Dictionary lookup

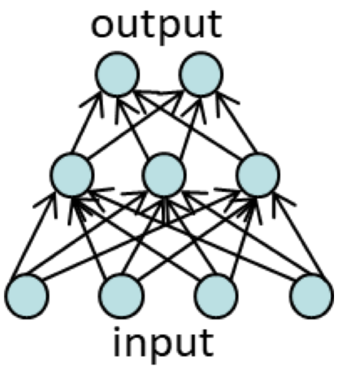
- problems ("Sindrome di Brugada", "episodi sincopali")
- tests ("ECG", "Test da Sforzo")
- treatments ("Flecainide", "Amiodarone")
- occurrences ("ricovero", "visita di controllo")

Lookup: search for dictionary entries in the text

- Dictionaries: UMLS, FederFarma, and two hand-crafted lexicons, acronyms
- TextPro for plural forms
- cTAKES UMLS Dictionary Lookup Fast Annotator

•Pianta E, Girardi C, Zanoli R. The TextPro tool suite. Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference, Marrakech, Morocco. 2008 May 28-30.
•Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc JAMIA. 2010;17(5):507–13.

Neural networks and entity recognition



Neural network models

- Automatically extract features for supervised learning
- Applied to NLP tasks with promising results

Token	Tag
The	O
ECG	B-test
test	I-test
revealed	O
features	O
consistent	O
with	O
Brugada	B-problem
Syndrome	I-problem

Sequence labeling problem

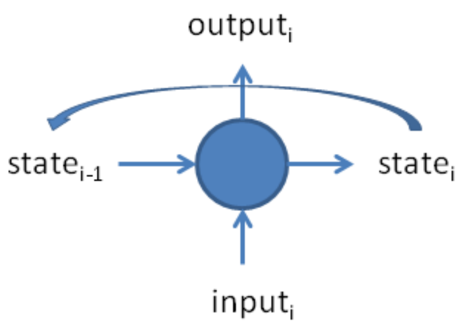
B (Beginning), I (Inside), O (Outside) tagging

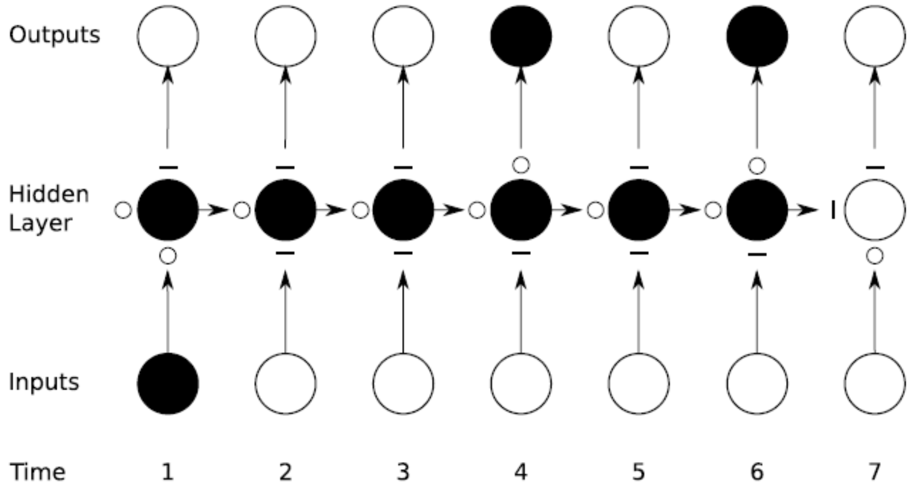
- Input: sequence of tokens
- Output: sequence of labels

Recurrent neural network models

Recurrent neural networks (RNNs)

- Flexible use of context information
- Map from an entire history of previous inputs to each output



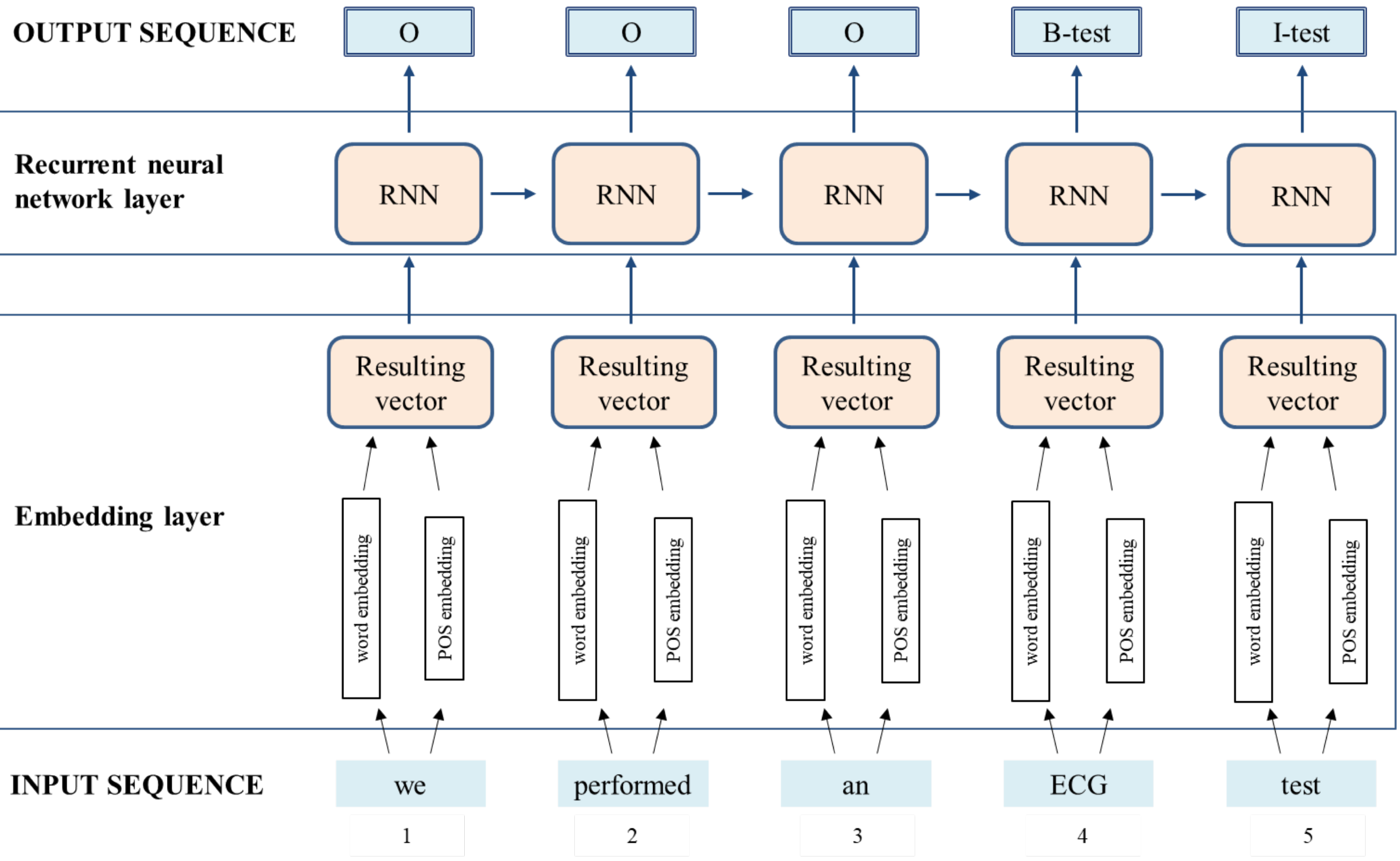


Long Short-Term Memory (LSTM): learn long-term dependencies

Gated Recurrent Unit (GRU): simpler variation

- Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks. In: Studies in Computational Intelligence. Springer (2012).

Developed model for Event recognition



Identification of Event properties

DocTimeRel: SVM classifier

token	POS tag	section	verb temp tense	token -1	token +1	token -2	token +2

DocTimeRel value (overlap, before, before/overlap, after)

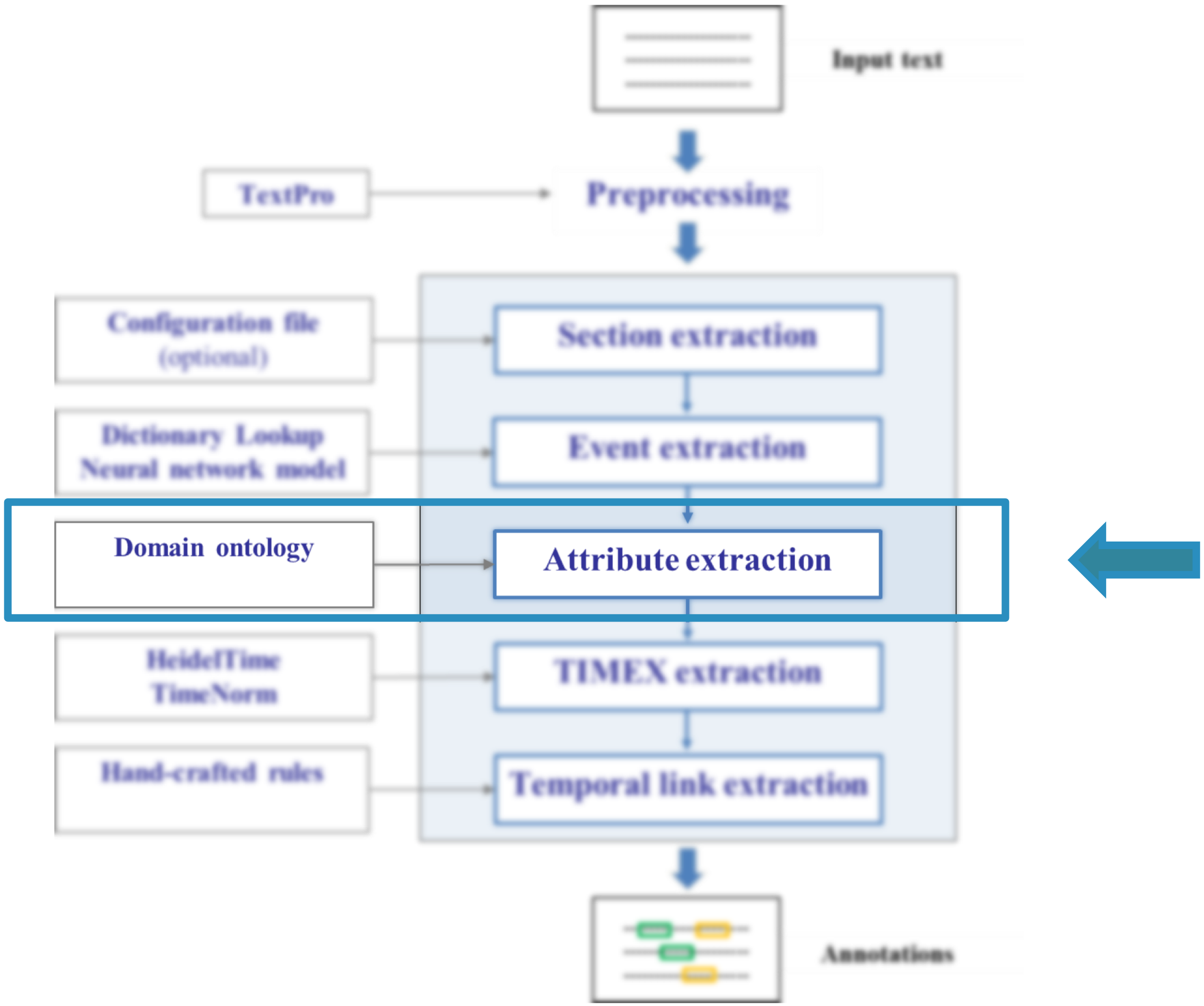
Polarity, Contextual modality, and Experiencer: ConText

The patient did **not** experience any symptom

Polarity: NEGATIVE

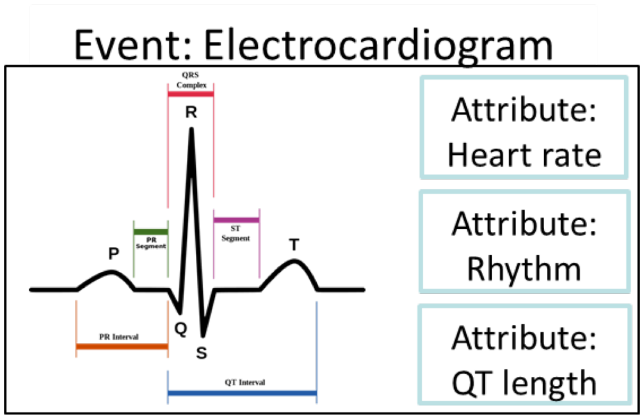
• Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009 Oct;42(5):839-51

Attribute extraction

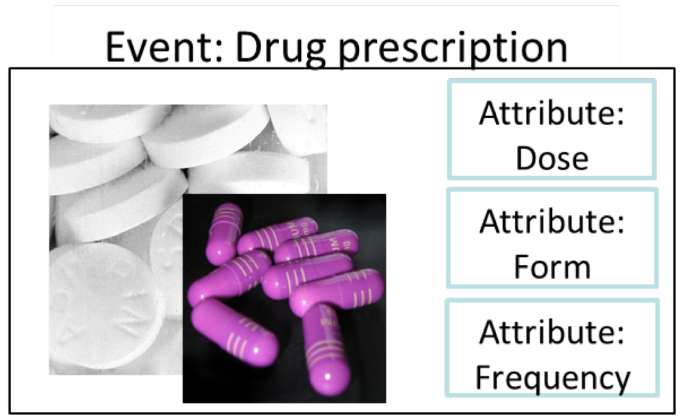


Ontology-driven approach (1)

In clinical reports, it is frequent to find occurrences of events that are related to a set of attributes



Source: Wikipedia



Source: Wikipedia

Ontologies: advantages

- can be easily updated to add/modify concepts
- can be enriched with new information
- regular expressions can be translated to other languages

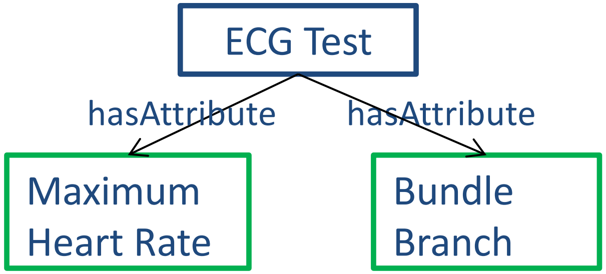
Ontology-driven approach (2)

Ontology-driven approach

1. Consider the medical problem to identify events and related attributes
2. Define a domain-related ontology containing events and attributes
   E.g. Cardiology domain
3. Automatically create an ontology-based configuration file

Electrocardiogram

- Rhythm
- Maximum Heart Rate
- QT length
- Bundle Branch



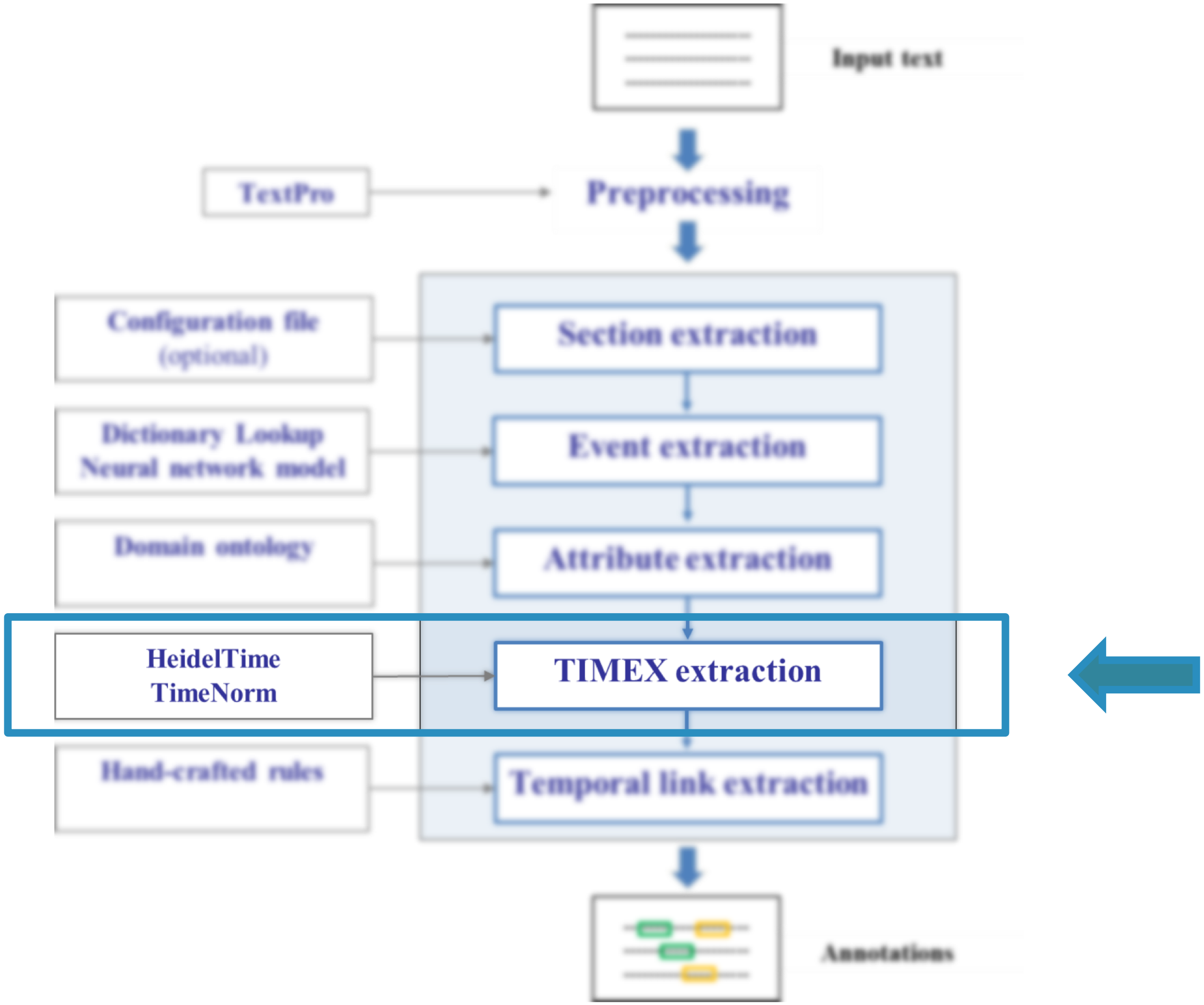

```xml
<event>
    <type>test</type>
    <name>ECGTest</name>
    <regex>(ECG|[Ee]lettrocardiogramma)</regex>
    <attributes>
        <attribute>
            <name>BundleBlock</name>
            <regex>(BBD|[Bb]locco di branca destra|[Ee]miblocco
            <type>string</type>
            <pattern>(incompleto|completo)?</pattern>
        </attribute>
        <attribute>
            <name>HeartRate</name>
            <regex>(FC|[Ff]requenza [Cc]ardiaca)</regex>
            <type>integer</type>
            <value_min>40</value_min>
            <value_max>200</value_max>
            <um>bpm</um>
        </attribute>
```

Attribute annotation process

- Extracted events are linked to their attributes through the configuration file
- Attribute names and values (regular expressions) are looked for in suitable lookup windows

Event semantic type	Contextual information	Lookup window
Test	No sections available	One paragraph*
Test	Included in matching section	One section
Test	Not included in any section	One paragraph*
Test	Included in non-matching section	One sentence
Treatment	NA	One sentence*

Temporal expression extraction



HeidelTime and TimeNorm adaptation

HeidelTime

- Rule-based tool
- TIMEX extraction and normalization
- Available also in Italian

TimeNorm

- Tool based on synchronous context-free grammars
- TIMEX normalization
- Available also in Italian

Adaptation to the clinical domain

- HeidelTime rules and TimeNorm grammar entries updated
- HeidelTime annotator modified to better deal with implicit TIMEXes (e.g. “the day after”)

- Strötgen J, Gertz M. HeidelTime: High Qualitiy Rule-based Extraction and Normalization of Temporal Expressions. Proceedings of the 5th International Workshop on Semantic Evaluation. 2010:321-324.
- Bethard S. A Synchronous Context Free Grammar for Time Normalization. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2013:821-826.

Main adaptations

- Extension of general domain rules: dates in the format DD.MM.YYYY, sets such as “every six months”, …

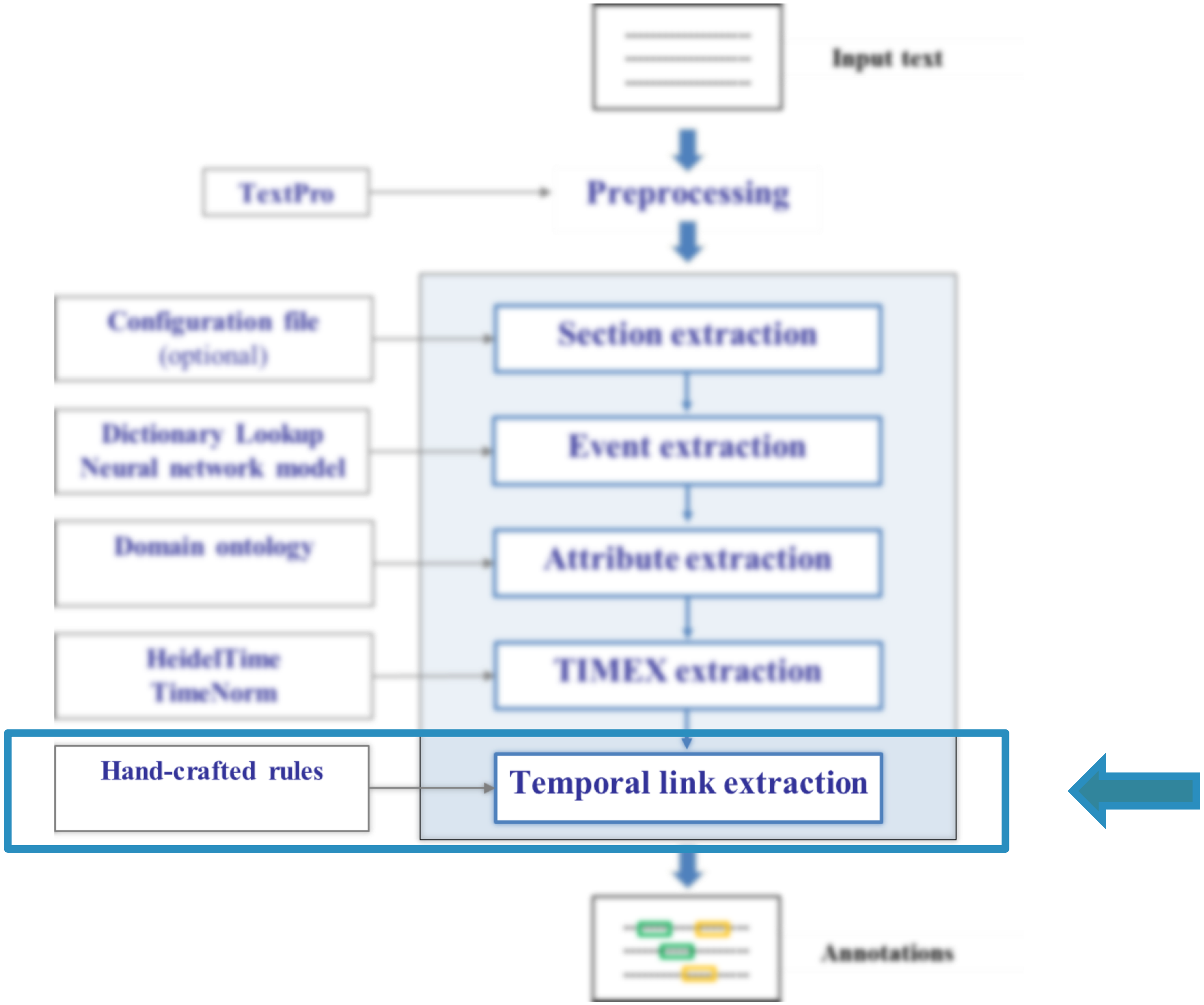
- Creation of domain-specific rules

IT *Atenololo: 1 cp x 2/die*
EN Atenolol: 1 tablet twice a day
→ Type = SET
Value = P1D, Freq = 2X

IT *Atenololo: 1 cp Ore 16*
EN Atenolol: 1 tablet at 4 pm
→ Type = TIME
Value = XXXX-XX-XXT16:00

• Strötgen J, Armiti A, Van Canh T, Zell J, Gertz M. Time for more languages: Temporal tagging of arabic, italian, spanish, and vietnamese. ACM Transactions on Asian Language Information Processing. 2014.;3(1):1–21.
• Mirza P, Minard A. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-Evalita 2014. Proceedings of the 4th International Workshop EVALITA-2014. 2014:44–49.

Temporal link extraction

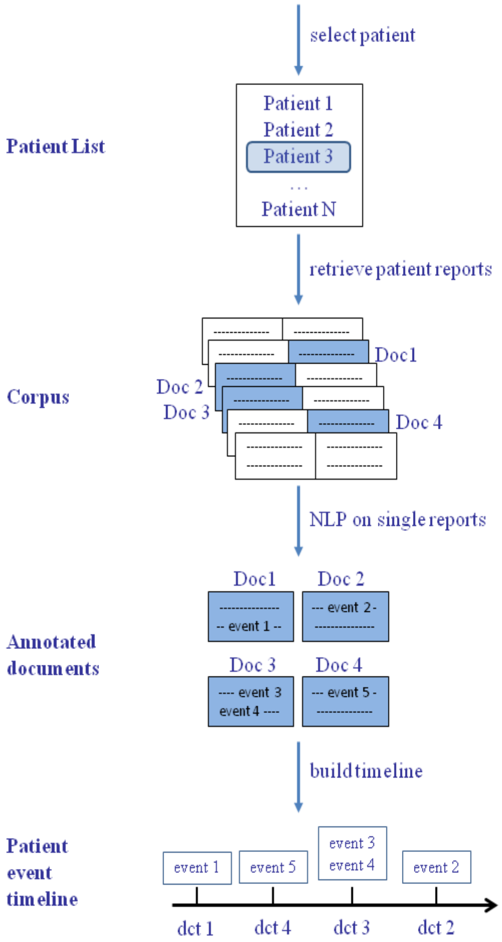


Rule-based approach

- Links between Event-TIMEX pairs included in the same sentence
- Five possible links: BEFORE, BEGINS_ON, ENDS_ON, CONTAINS, OVERLAP
- Manual creation of rules, based on 12 features:

1. Event	7. TIMEX
2. Event section	8. TIMEX type
3. Event DocTimeRel	9. TIMEX value
4. Event semantic type	10. Temporal preposition
5. Event polarity	11. Verb temporal tense
6. Event-TIMEX distance	12. Temporal verbs

Clinical timeline construction



The patient of interest is selected

The medical reports referred to the selected patient are retrieved

The NLP pipeline processes the retrieved documents

The events extracted from all patient documents are visualized on a timeline

Viani N, Tibollo V, Napolitano C, Priori SG, Bellazzi R, Sacchi L. Clinical Timelines Development from Textual Medical Reports in Italian. Proceedings of RTSI 2017, 3° International Forum on Research and Technologies for Society and Industry. 2017.

Information extraction in a non-English language: Italian

Results and Discussion

Statistics on the annotated corpus

	Training set	Test set	Corpus
Documents	60	15	75
Tokens	44115	13148	57263
Sentences	3347	941	4288
Events	3159	992	4151
TIMEXes	814	288	1102

Most events are Problems (42%), with an Overlap relation to the DCT (44%)

Most time expressions are Dates (61%)

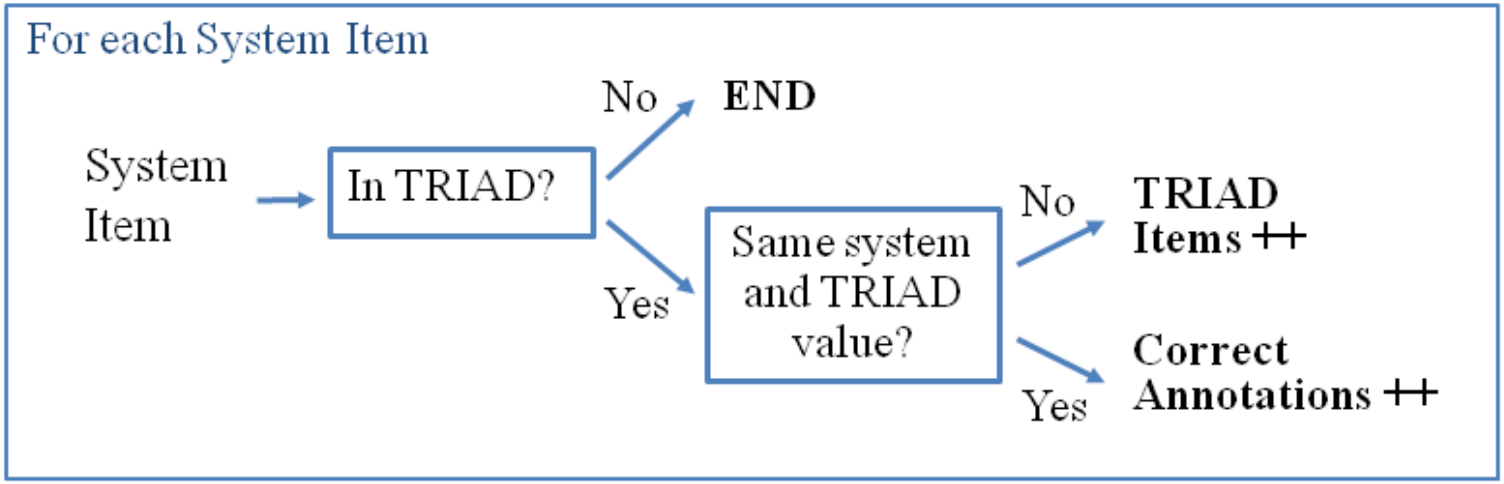
Event extraction: results

Annotated test set (15 documents)

Extraction method	TP	FP	FN	P	R	F1
Dictionary lookup	548	118	444	82.3%	55.2%	66.1%
CRF classifier	795	189	197	80.8%	80.1%	80.5%
SVM classifier	748	103	244	87.9%	75.4%	81.2%
GRU classifier	844	111	148	88.4%	85.1%	86.7%
GRU classifier with POS input	863	107	129	89.0%	87.0%	88.0%
Dictionary lookup + GRU classifier with POS input	**895**	**114**	**97**	**88.7%**	**90.2%**	**89.5%**

Attribute extraction: evaluation

- **TRIAD**: database for clinical and genetic variations in the field of inherited arrhythmogenic diseases
- Data on diagnoses, genetic mutations, cardiac events, performed tests, prescribed treatments and device implants



Attribute extraction: ontology

The developed ontology contains 11 events and 61 attributes

- Attribute
  - ProblemAttribute
  - TestAttribute
    - CardiacMorphologyAttribute
    - ECGSignalAttribute
      - Blocks
      - HeartRate
      - HolterPause
      - LatePotentials
      - Rhythm
      - SignalInterval
      - VentriculoAtrialConduction
    - ExerciseAttribute
    - PhysicalAttribute
  - TreatmentAttribute
- Event
  - Problem
    - Diagnosis
  - Test
    - GenericTest
    - InstrumentalTest
      - ECGTest
      - EchocardiogramTest
      - EffortStressTest
      - HolterMonitorTest
      - LatePotentialsTest
      - OtherImagingTest
      - ProgElectricalStimulatio
    - LaboratoryTest
  - Treatment
    - DrugTreatment

EVENT: ECGTest

- hasRegularExpression: “ECG|[Ee]lettrocardiogramma”
- hasAttribute: BundleBlock
- hasAttribute: HeartRate
- hasAttribute: QT
- hasAttribute: Rhythm

Numeric Attribute: AverageHeartRate

- hasRegularExpression: “FC|[Ff]requenza cardiaca”
- hasUnitOfMeasurement: “bpm”
- hasNumericValue: Integer >= 40 and Integer <= 200

String Attribute: Rhythm

- hasRegularExpression: “[Rr]itmo”
- hasStringValue: “(sinusale)?[ ]*bradicardico|(sinusale)?[ ]*tachicardico|sinusale”

Attribute extraction: results

SV	Set	Event	System item	TRIAD item	Correct annotations	Accuracy
1	**Dev** 4429 reports	Main Diagnosis	4202	4077	3607	**88.5%**
		ECG	26669	22546	21352	**94.7%**
		Holter ECG	26767	21538	19058	**88.5%**
		Effort Stress Test	9683	3978	2367	**59.5%**
		Prescribed Drug	8720	2436	2186	**89.7%**
2	**Test** 1003 reports	Main Diagnosis	927	913	845	**92.6%**
		ECG	7452	5070	4885	**96.4%**
		Holter ECG	7173	5127	4757	**92.8%**
		Effort Stress Test	2543	1118	1064	**95.2%**
		Prescribed Drug	1999	538	435	**80.9%**

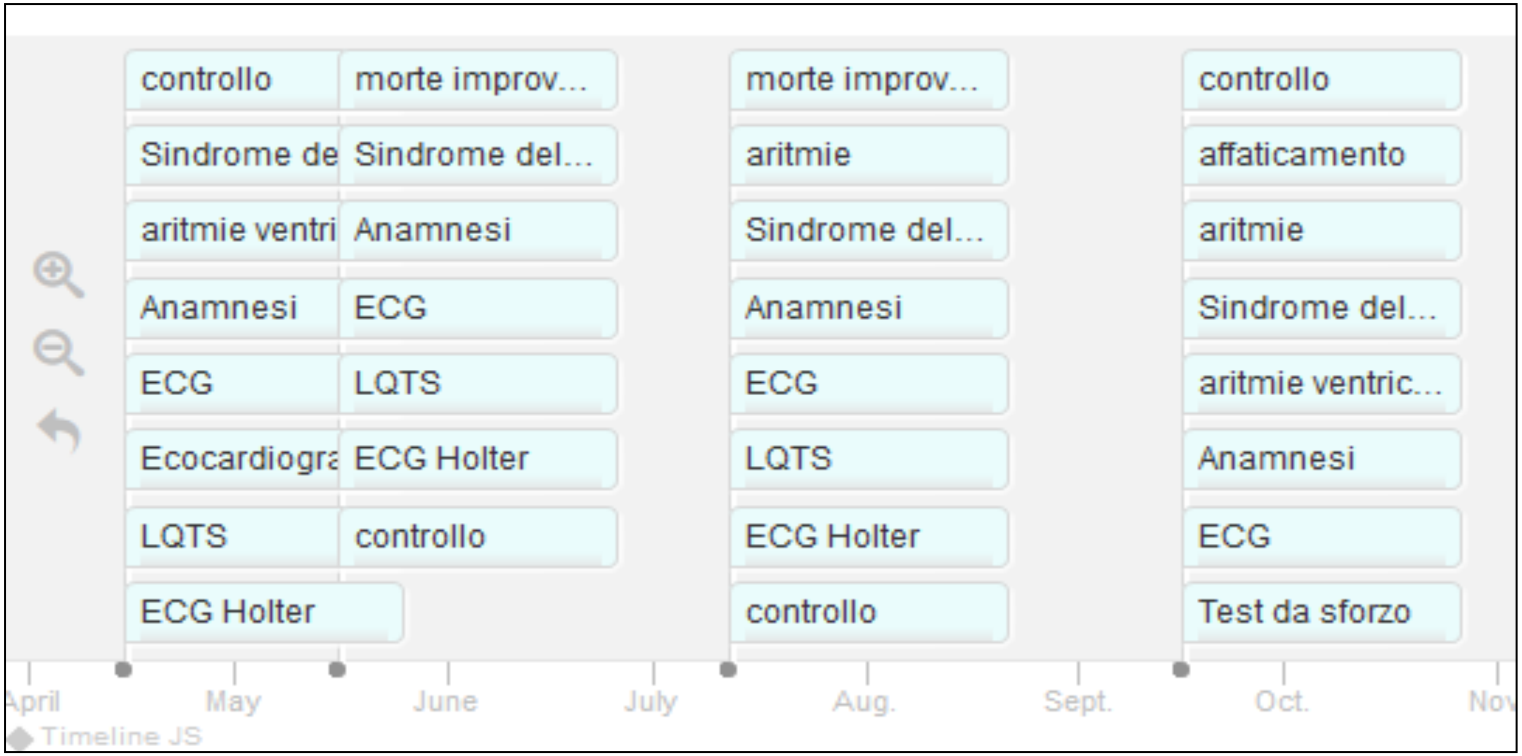
Time expression extraction: results

Annotated test set (15 documents)

System	Set	TP	FP	FN	F1
HeidelTime original	Training	425	196	389	59.2%
HeidelTime updated	Training	760	47	54	93.8%
HeidelTime updated	Test	273	13	15	**95.1%**

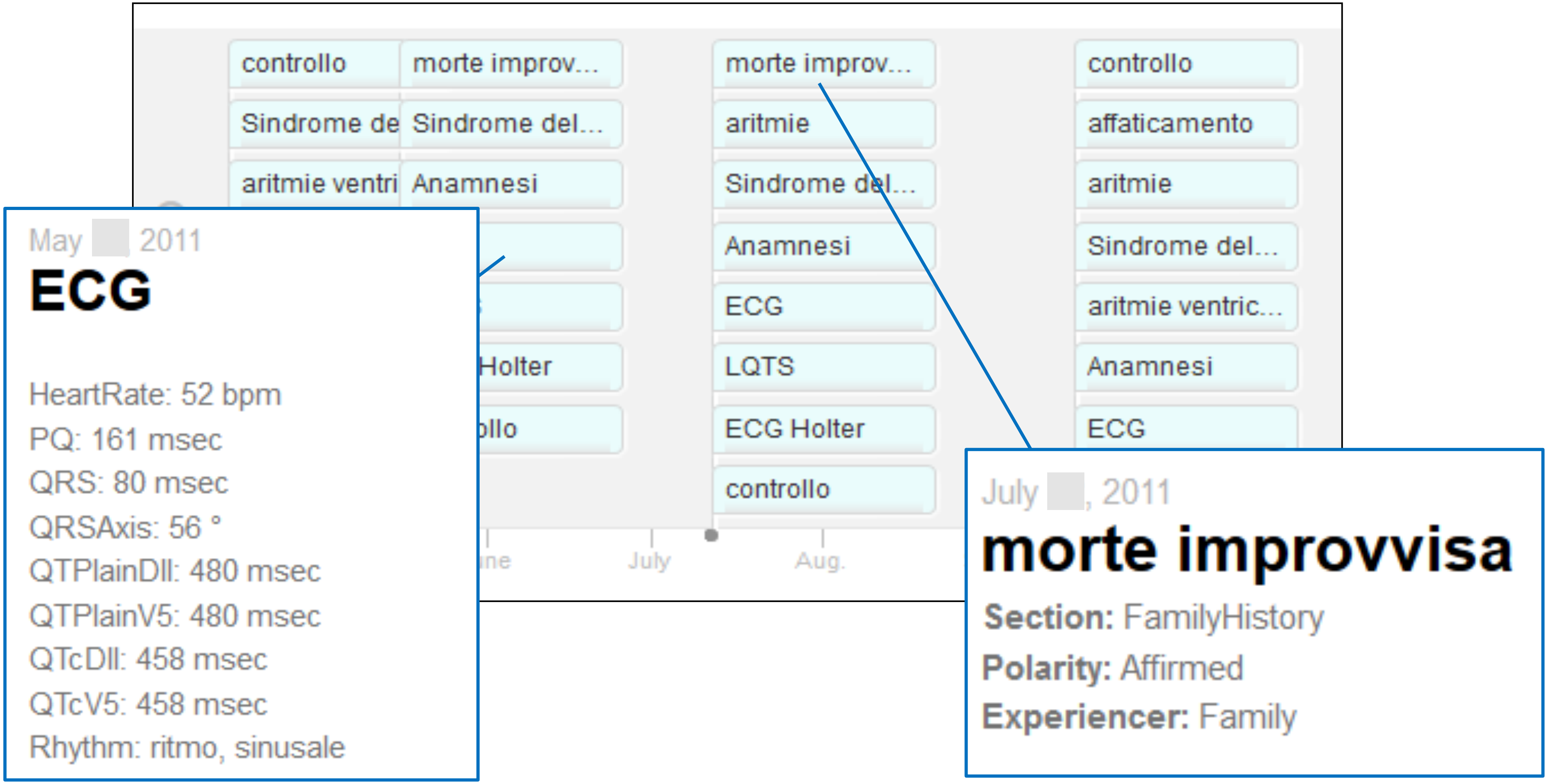
System	Set	TP	Property	Accuracy
HT original	Training	425	value	91.5%
HT updated	Test	273	value	**93.8%**
TN original	Training	760	value	56.7%
TN updated	Test	273	value	**89.0%**

Reconstructed patient timeline



Viani N, Tibollo V, Napolitano C, Priori SG, Bellazzi R, Sacchi L. Clinical Timelines Development from Textual Medical Reports in Italian. Proceedings of RTSI 2017, 3° International Forum on Research and Technologies for Society and Industry. 2017.

Reconstructed patient timeline



Viani N, Tibollo V, Napolitano C, Priori SG, Bellazzi R, Sacchi L. Clinical Timelines Development from Textual Medical Reports in Italian. Proceedings of RTSI 2017, 3° International Forum on Research and Technologies for Society and Industry. 2017.

Information extraction in a non-English language: Italian

Extensions and Integrations

Extension to a different domain (1)

- 221 anatomic pathology reports
- Hospital Papa Giovanni XXIII in Bergamo, Italy
- 20 reports: ontology design set
- 34 reports: test set

SECTION:NOTIZIE_CLINICHE
Vedi I17-xxx (core biopsy), T17-xxx (indagine FISH) e I17-xxx (linfonodo sentinella).

← Clinical information

SECTION:MATERIALE_INVIATO
1. Quadrante supero-interno della mammella destra.
2. Margine profondo.
3. Margine superiore.
4. Margine inferiore.

← Sent specimen

SECTION:TESTO_MACRO
1- Frammento di parenchima mammario di6x6x2 cm con losanga di cute di 5x0,7 cm, pervenuto già sezionato in corrispondenza di una neoplasia di 0,9 cm di asse maggiore.
2- Frammento di parenchima mammario di 4x3,5x1 cm, orientabile.
3- Frammento di parenchima mammario di 7x2,5x1 cm, orientabile.
4- Frammento di parenchima mammario di 8x2,5x2 cm, orientabile.

← Specimen description

SECTION:TESTO_DIAGNOSI
1- Carcinoma duttale infiltrante a medio grado di differenziazione. [...]
2,3- Parenchima mammario esente da neoplasia.
4- Focolaio di carcinoma lobulare in situ di tipo classico (diametro istologico pari a 3 mm) distante 3 mm dal margine di resezione; si associa iperplasia lobulare atipica. [...]
Stadiazione istopatologica sec. TNM VIII edizione: pT1b G2
Linfonodo sentinella esente da metastasi, esaminato con metodica molecolare O.S.N.A. (I17-xxx).

← Diagnosis

Viani N, Chiudinelli L, Tasca C, Zambelli A, Bucalo C, Ghirardi A, Barbarini N, Sfreddo E, Sacchi L, Tondini C, Bellazzi R. Automatic Processing of Anatomic Pathology Reports in the Italian Language to Enhance the Reuse of Clinical Data. Accepted at MIE 2018, 29th Medical Informatics Europe conference.

Extension to a different domain (2)

- Events (with Attributes): specimen, diagnosis, histopathological stage, prognostic factor
- **New IE task**: specimen-diagnosis Event-Event links

- Event
  - HistopathologicalStage
  - Diagnosis
    - Lesion
      - BenignLesion
      - NeoplasmDiagnosis
  - PrognosticFactor
  - Specimen
    - Biopsy
    - SurgicalResection
      - Lump
      - Organ
      - OrganPortion
        - Areola
        - Lozenge
        - Margin
        - Nipple
        - Parenchyma
        - Quadrant

- Attribute
  - HistopathologicalStageAttribute
  - DiagnosisAttribute
  - LocalizedOrgan
  - Nodule
  - OrganPortion
  - PrognosticFactorAttribute
  - SpecimenAttribute
    - LymphNodeLevel
    - LymphNodeNumber
    - OSNAMethod
    - ReferenceToExam
    - SpecimenNumber
    - SpecimenSize

44 events and 16 attributes

Viani N, Chiudinelli L, Tasca C, Zambelli A, Bucalo C, Ghirardi A, Barbarini N, Sfreddo E, Sacchi L, Tondini C, Bellazzi R. Automatic Processing of Anatomic Pathology Reports in the Italian Language to Enhance the Reuse of Clinical Data. Accepted at MIE 2018, 29th Medical Informatics Europe conference.

Extension to a different domain (3)

- Validation with Expert
- 476 system items
- Three types of errors: missing items, FN, FP

Items	Raw count	Distinct count
Missing items	57	38
FN	15	11
FP	26	21

Missing items:

Information to be added to the ontology

Precision: 94.5%

Recall: 96.8%

F1 score: 95.6%

Information extraction in the mental health domain

My experience at BRC

Natalia Viani, July 4th 2018

Mental health domain

Challenges for NLP

- large proportion of free-text
- heterogeneity in self-reported experiences, circumstances, treatment and outcomes
- symptomology and health progression often described without relying on structured fields

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 (1)

Aim: Identification of time expressions (TIMEXes) and symptom onset in mental health records for patients with a diagnosis of schizophrenia.

Relevance: For this disease, analysing symptom and treatment onset is essential to measure the duration of untreated psychosis (DUP).

The patient's partner reports that the patient was diagnosed with ***schizophrenia*** *in* ***1990****....*

Past *medication trials that the patient reports include* ***haloperidol*** *and* ***lithium*** *(started in* ***1991****, on and off* ***for 2 years****), neither of which particularly helpful....*

Introduction (2)

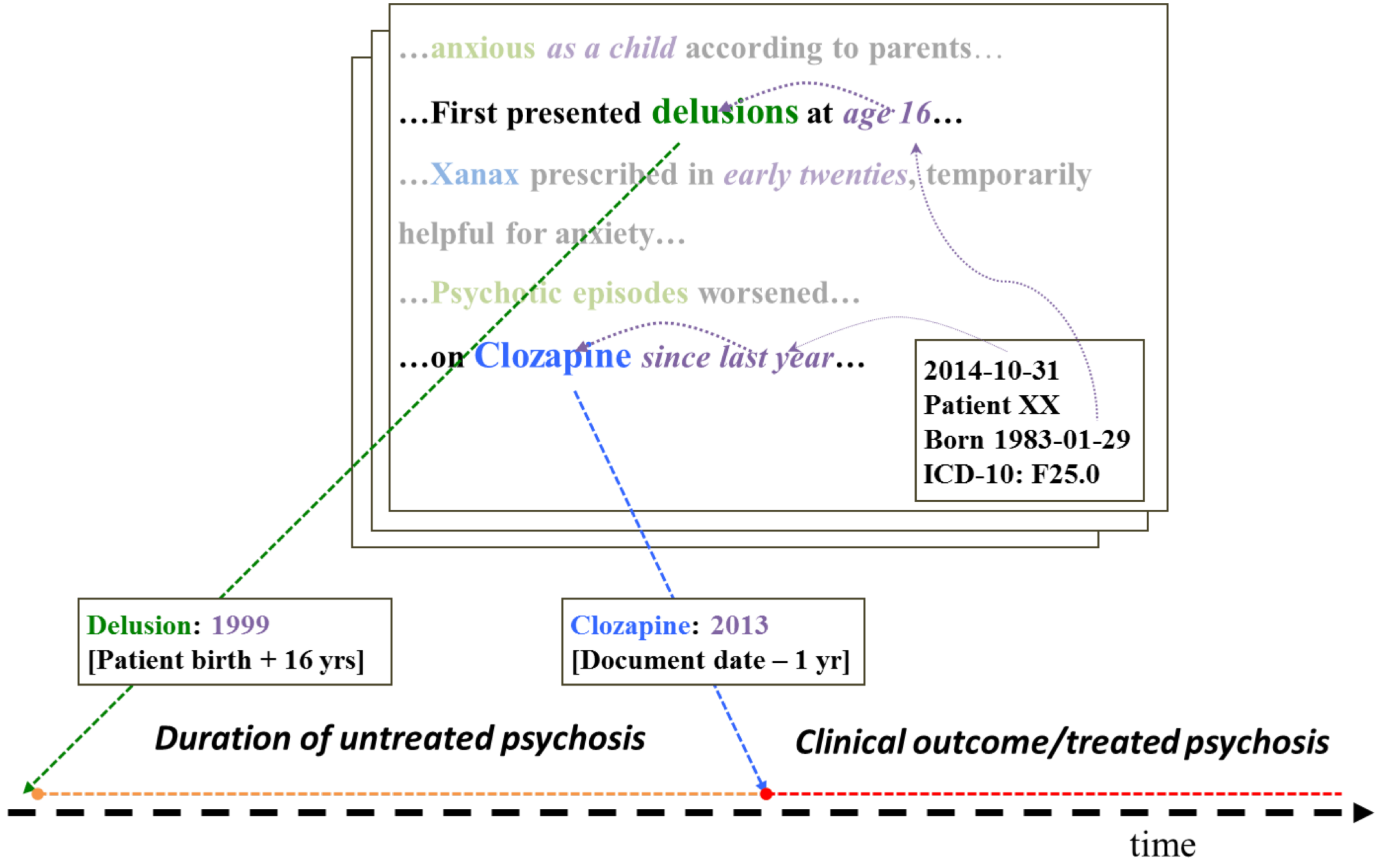


Image courtesy of Dr. Sumithra Velupillai, King's College

Background

Temporal link extraction from clinical narratives in English

2012 i2b2

- intensive care unit
- 310 discharge summaries
- events, temporal expressions, and 8 types of temporal relations (e.g., before, overlap)

>>> 2012 i2b2 NLP Challenge for Clinical Records

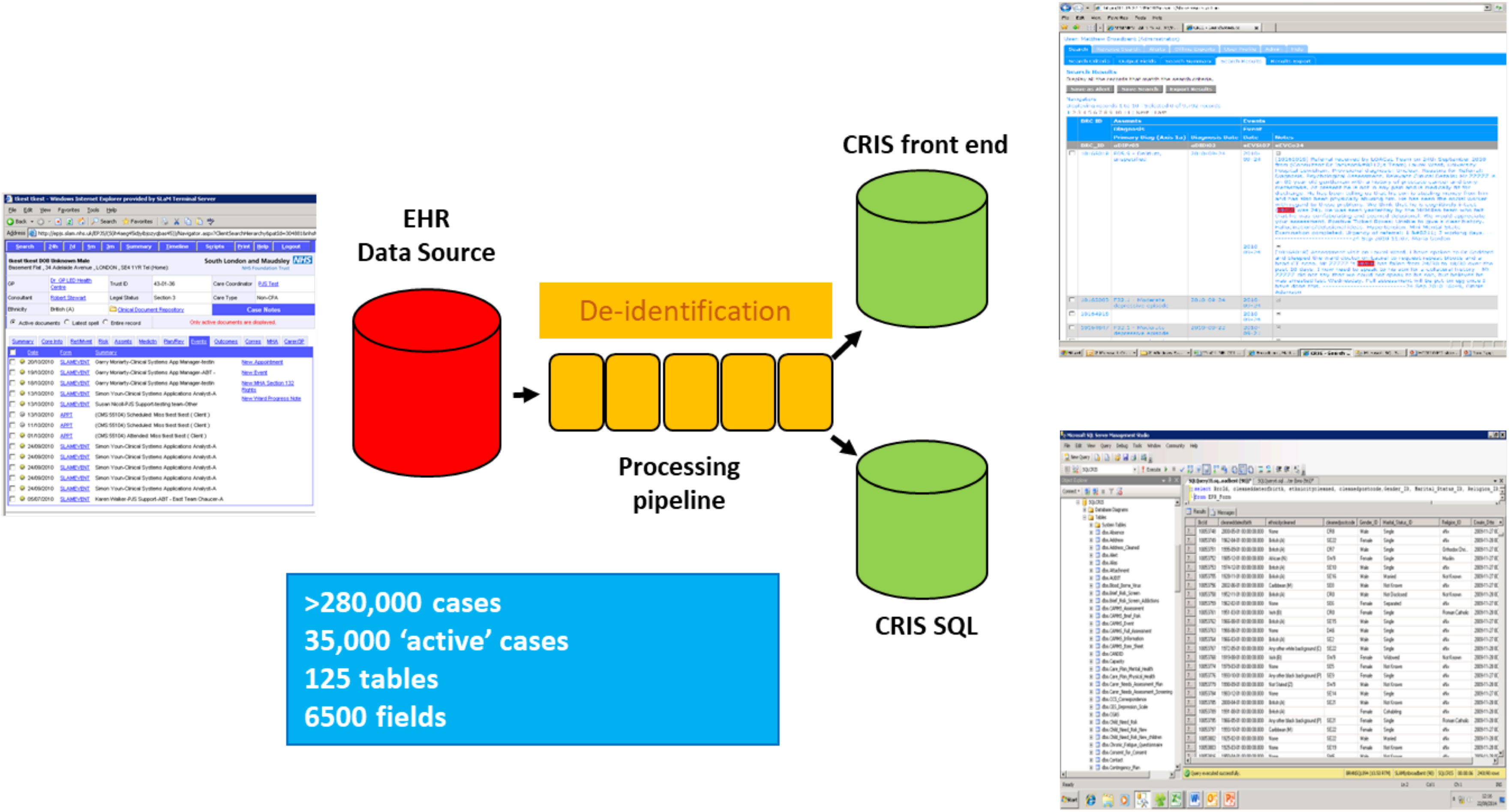
THYME corpus

- breast cancer, colon cancer
- 1,254 records
- events, temporal expressions, and 2 types of temporal links: DocTimeRel, and relations to narrative containers

>>> 2015, 2016, and 2017 Clinical TempEval (440, 591, and 1186 docs)

- 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, Palmer M, Pradhan S, de Groen PC, et al. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics. 2014;2:143.

CRIS – core functionality

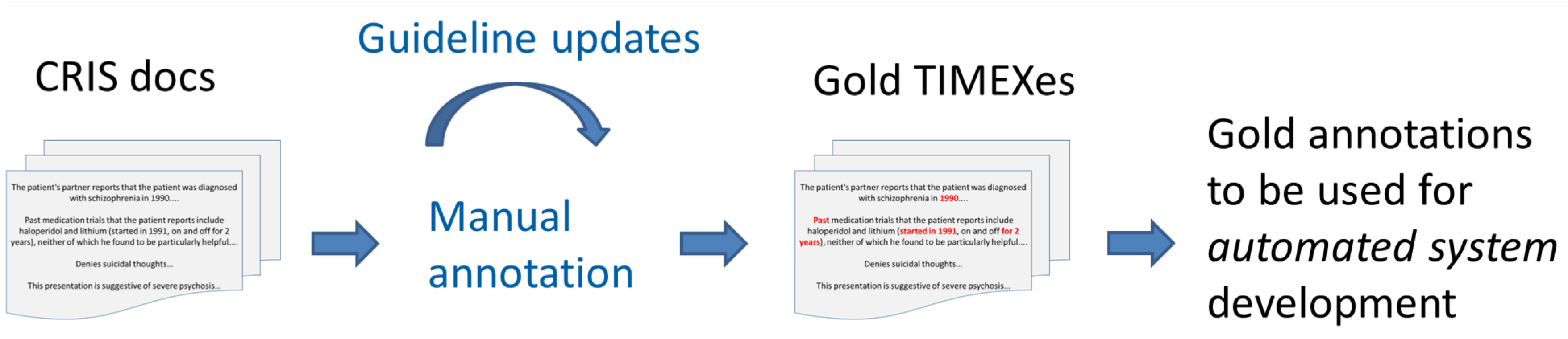


>280,000 cases
35,000 'active' cases
125 tables
6500 fields

Image courtesy of Prof. Rob Stewart, King's College

Temporal expression annotation

Data: Mental health records from the Clinical Record Interactive Search (CRIS) database were manually annotated for TIMEXes.



Guidelines development

- annotation guidelines developed based on previous work.
- discussion stage for guideline updates.

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.

First annotation process

Document selection

- Documents written within three months from first referral
- Longest document per each patient

Three annotators independently annotated **20 documents.**

New TIMEX type referred to the patient's age: **Age-related**

- *"she first experienced hallucinations at the age of 18..."*
- *"he has been hearing voices since his teens..."*

MeDESTO corpus

Extension of annotations

- **52 documents** annotated for time expressions
- 65.6 annotations per document

# TIMEXes	3413 (65.6/doc)
Date	1903 (55.8%)
Duration	563 (16.5%)
Time	366 (10.7%)
Frequency	276 (8.1%)
Age-related	305 (8.9%)

Automated system development

Annotated corpus used to adapt two rule-based TIMEX extraction systems:

- SUTime
- HeidelTime

Main adaptations:

- Added age-related TIMEXes and domain-specific expressions (e.g., OD for once daily)
- Post-processing for determining the age-related type

- Strötgen J, Gertz M. HeidelTime: High Qualitiy Rule-based Extraction and Normalization of Temporal Expressions. Proceedings of the 5th International Workshop on Semantic Evaluation. 2010:321-324.
- Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In Lrec, volume 2012, pages 3735–3740.

Event annotation

MeDESTO corpus manually annotated with events

- **symptoms**
- **signs**
- **diagnoses**
- medications
- life events or social circumstances
- healthcare services
- patient behaviour
- other health problems

Guidelines development

- discussion stage for guideline updates
- input by domain experts

Onset information annotation

Problem: find documents that are likely to contain the onset information.

- Extract all documents related to <u>early intervention services</u> (services that support people who are experiencing the symptoms of psychosis for the first time)

- Filter documents according to:
  - Length
  - Average line length

First document selection

1) extract all documents related to early intervention services

18281 documents
3840 patients

2) Filter documents
- length > 50th percentile
- avg_line_length > 25th percentile

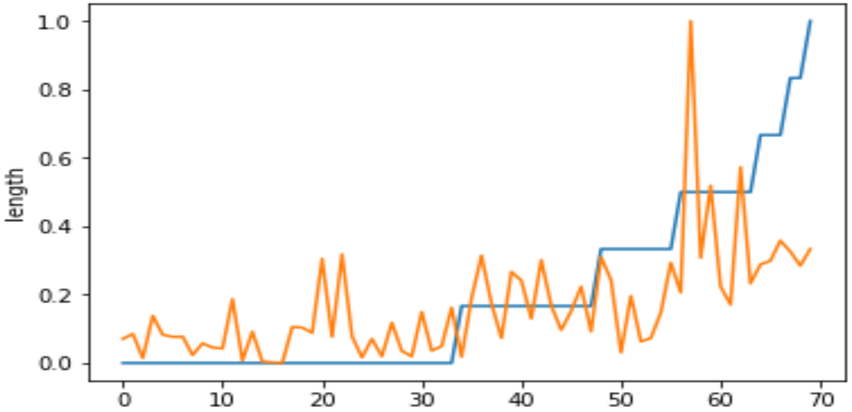
8496 documents
3198 patients

3) Randomly select 20 patients and save all their documents

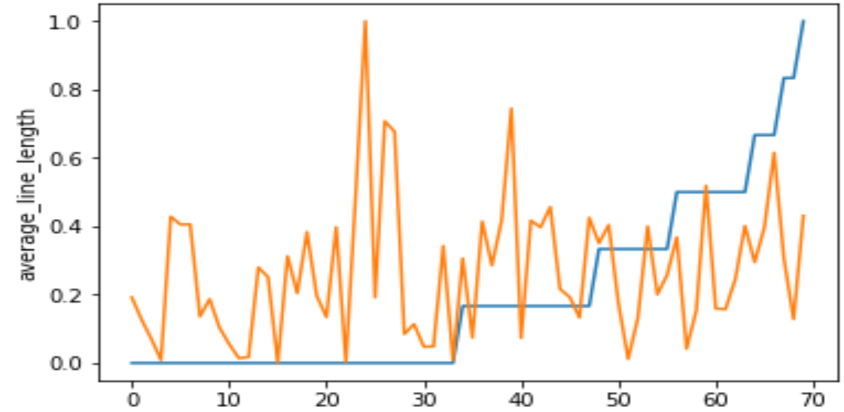
70 documents
20 patients (1-8 docs each)

Annotation results

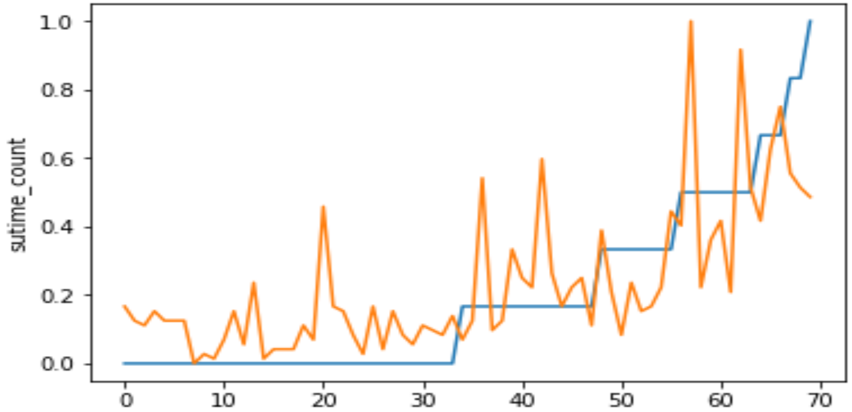
Length



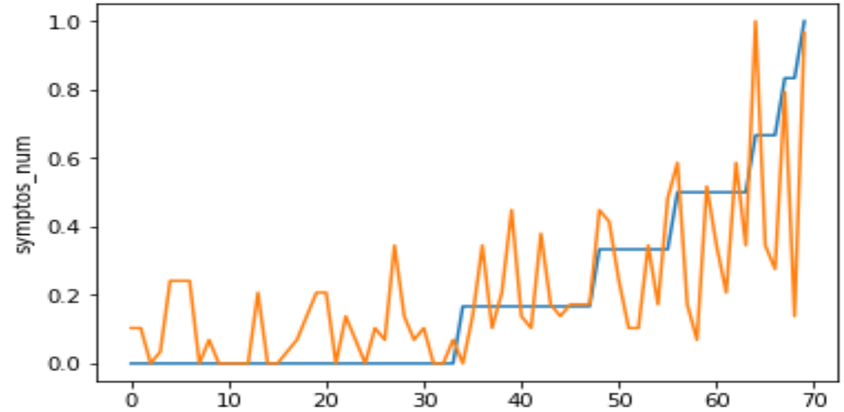
Avg line length



Num timexes (SUTime processing)

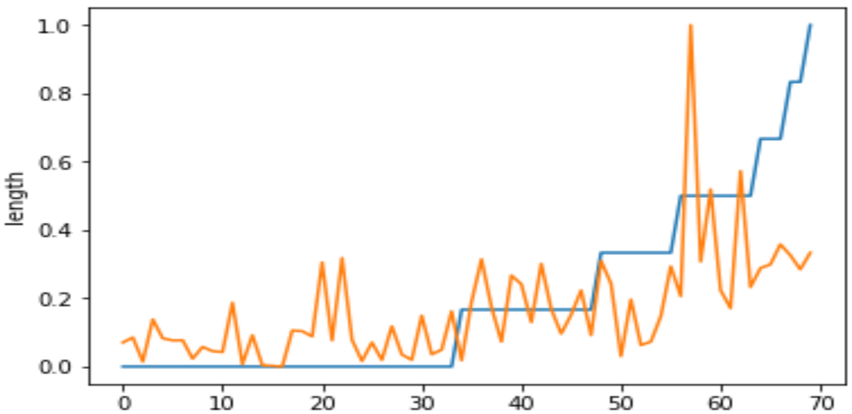


Num symptoms (list of 598 keywords)

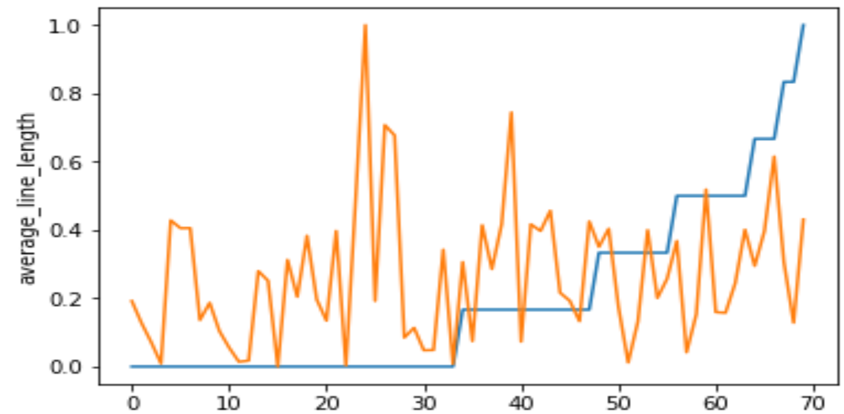


Annotation results

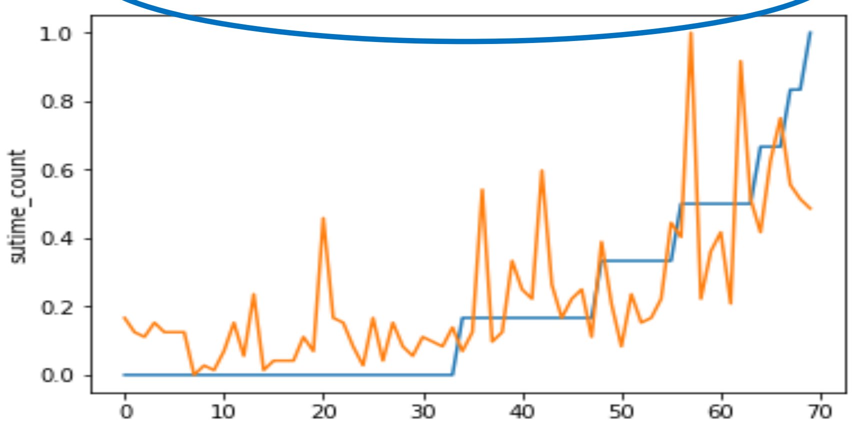
Length



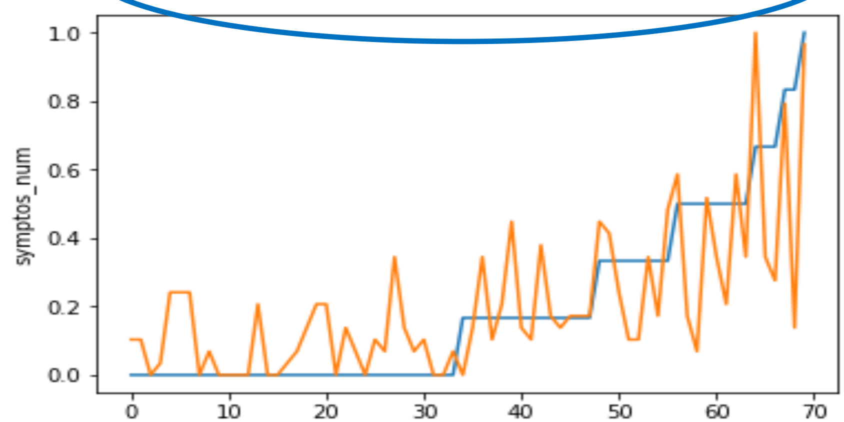
Avg line length



Num timexes (SUTime processing)



Num symptoms (list of 598 keywords)



Second document selection

1) extract all documents related to early intervention services

36594 documents
4166 patients

2) Filter documents
- length > 50th percentile
- avg_line_length > 25th percentile

16318 documents
3819 patients

3) New filters:
- Num symptoms > 3
- Num timexes > 5

8842 documents
3308 patients

4) Randomly select 20 patients and save all their documents

54 documents
20 patients (1-6 docs each)

Ongoing work

- Definition of additional symptom keywords
- Time expression normalization
- Temporal link annotation (just started)

Date: 2018-05-04

She reported she has been hearing voices since last year…"

hearing voices → last year (2017)

Conclusions

- Extracting information from clinical text is essential to make unstructured data available for further research
- Developing NLP applications for a specific clinical use-case is challenging
  - domain-specific language
  - lack of annotated resources

Future directions → system adaptation

→ multilingual approaches

Acknowledgments

Silvia Priori
Carlo Napolitano
Valentina Tibollo

Carlo Tondini
Alberto Zambelli

Lucia Sacchi
Riccardo Bellazzi
Silvana Quaglini

Centre for Health Technologies
University of Pavia

Cristiana Larizza

Boston Children's Hospital
Until every child is well

Guergana Savova
Timothy Miller

CLINICAL RECORD INTERACTIVE SEARCH

NHS National Institute for Health Research

King's College London

Sumithra Velupillai
Robert Stewart
Rashmi Patel
Rina Dutta
Ayunni Alawi

Joyce Kam
Lucia Yin
Somain Verma

Thank you!

Questions?