Agile Text Mining for the i2b2 2014 Risk Factors Challenge (Track 2)

Linguamatics and Northwestern University

James Cormack

i2b2 workshop - AMIA

November 14th 2014

i2b2 Task 2: Definition

- 790 annotated training documents
- 269 held out for development
- 514 test documents

- Annotate risk factors at document-level with the temporal relation to the document (before, during, after):
  - Positive mentions of CAD, Obesity, Hyperlipidemia, Hypertension
  - Medications related to these conditions
  - High blood pressure, glucose, cholesterol, A1C, BMI
  - CAD events (such as heart attack), symptoms related to CAD and positive CAD tests (e.g. a positive stress test)
  - Family history of premature CAD
  - Smoking history

NORTHWESTERN UNIVERSITY

Linguamatics

i2b2 Strategy

Unannotated XML

↓

NLP-based Rule Engine

↓

Post-Processing

Filtering to relevant values according to guidelines

↓

Temporal Relation Extraction

Annotate temporal relation to document:

- Rules
- Machine learning

↓

Output Annotated XML documents

NORTHWESTERN UNIVERSITY

Linguamatics

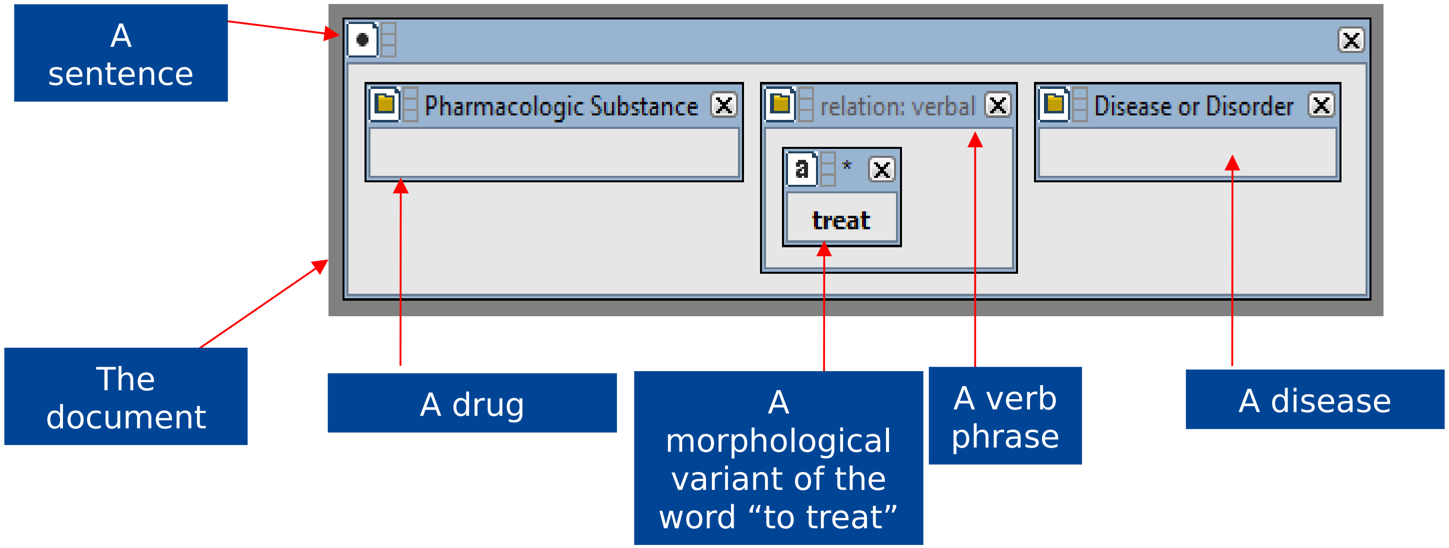
NLP-based Rule Engine

- We used I2E
  - commercial text mining platform
  - used in 17/20 top pharma companies
  - combines search technology and NLP to provide “agile” text mining
- Rules created in a graphical interface using an index of:
  - Tokens
  - POS tags
  - Shallow syntactic chunks
  - Semantic entities
  - Sentences
  - Document regions
- Results in seconds allowing interactive refinement

NORTHWESTERN UNIVERSITY

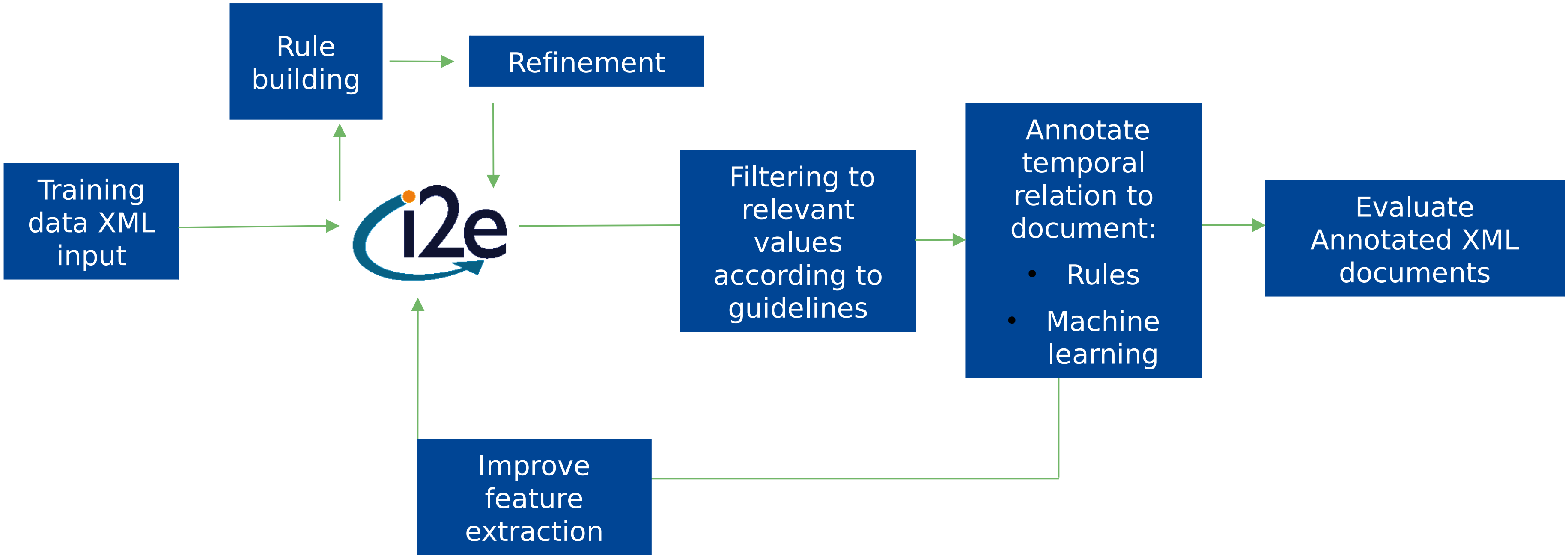
Linguamatics

An Example



- Linguistic items are ‘containers’ for other linguistic items
- Documents contain sentences, sentences contain words, phrases, syntactic chunks and semantic concepts.

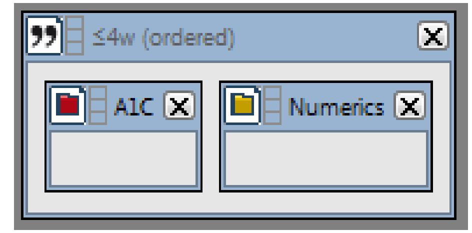
i2b2 Strategy - Training



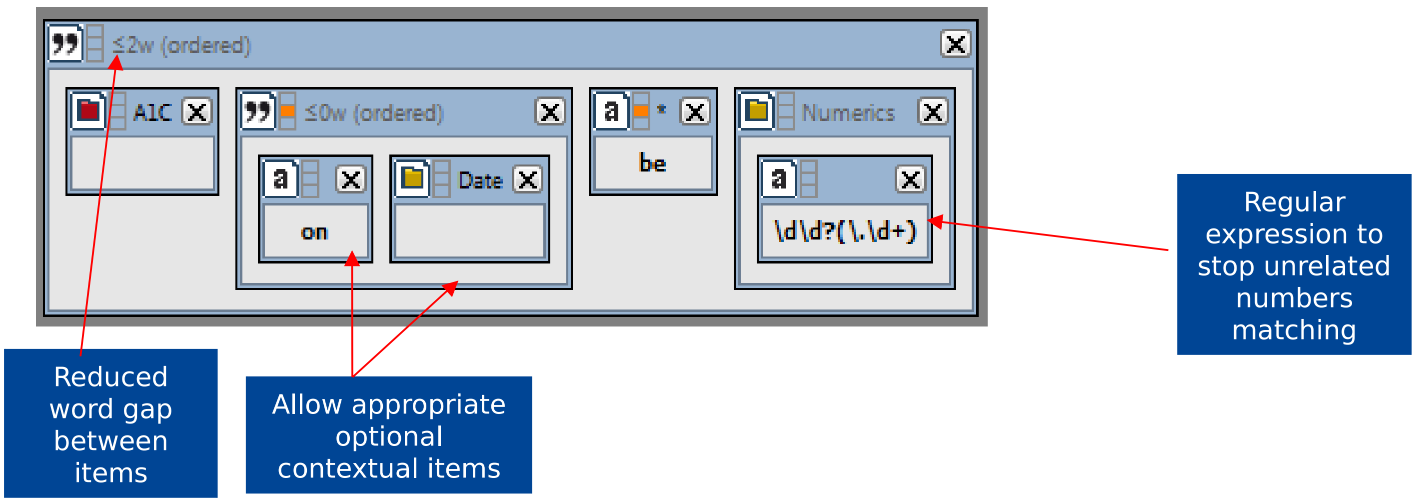
NORTHWESTERN UNIVERSITY
Linguamatics

Rule Engineering

- Start with high recall, low precision patterns:



- Start tightening up constraints to increase precision:

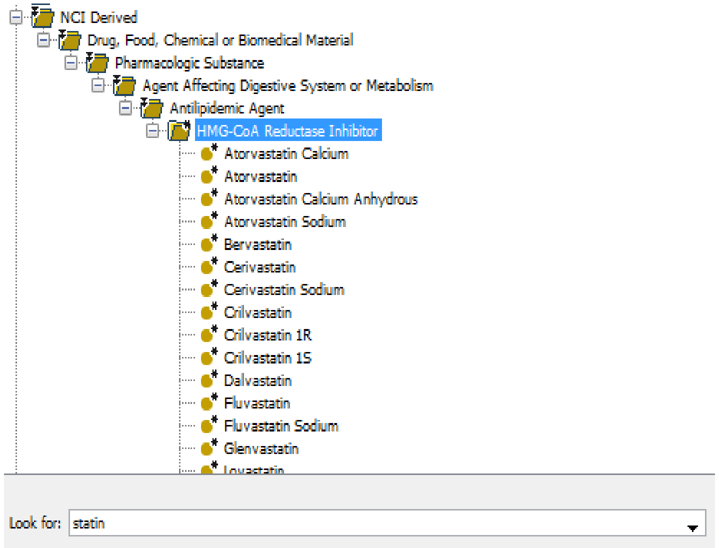


NORTHWESTERN UNIVERSITY

Linguamatics

Identifying Concepts

- Started with existing controlled terminologies
  - NCI Thesuarus
  - MeSH
  - RxNorm



- Terminologies modified while building rules:
  - Found similar words generated using distributional similarity
    - Word2vec over PubMed central/Wikipedia
    - Byblo + I2E patterns over MEDLINE
  - Words generated by generic rule patterns/regular expressions
    - Used I2E rules to find terms in similar contexts
  - Supplemented by synonyms (or common misspellings) present in the training data but not found with the above.
    - Used I2E rules over annotated text spans in the gold standard

NORTHWESTERN UNIVERSITY

Linguamatics

Smoking Categorization

- Trained separately on 2007 smoking challenge dataset and then tuned on this year's data.
- Classification is done at the sentence level
  - Conflicting sentences had to be resolved in post-processing
- This dataset had more ambiguous annotations in the training data: 'History of smoking' could be current, 'ever' or past.
- Annotators seemed reluctant to use the 'ever' category.
- More form-based/parameter value records made this challenging

    Smear on 14/02/10 negative

    smoking status 14/02/10 has not quit

NORTHWESTERN UNIVERSITY

Linguamatics

Lines and Tables

- Indexing the positions of the beginning and ends of a line
- Using the positional information of the word to extract result

Date/Time	CHOL	TRIG	HDL	**LDLCAL**
05/08/2066	272 (*)	301 (*)	46	**166** (*)

- Relies on similar table format throughout reports
- We had also hoped that the line breaks would be useful as sentence break indicators
- This seemed useful as table row indicators, but gave worse results in general

Post-processing

- Logical constraints
  - Do the values satisfy the guidelines?
    - Correct range for the risk factor?
    - Correct number of values? e.g. 2 high glucose values
  - Resolving duplicate candidates for annotation e.g. Past smoker/Current smoker.
    - "Notable for tobacco use...he has since quit smoking"

NORTHWESTERN UNIVERSITY

Linguamatics

Temporal Relation Extraction (1/3)

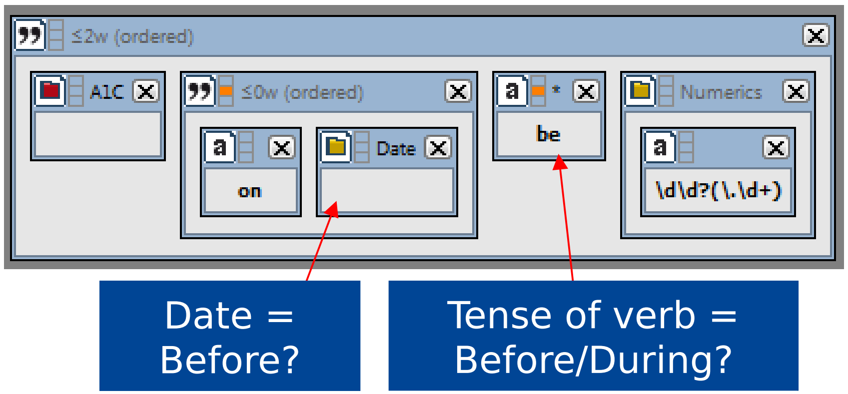
- Much of the temporal processing is implied by the risk factor/test type described rather than explicit in the text
- Need some way of representing the information the doctors have when they write patient notes
  - “Medications are presumed to be continuing unless I write otherwise”
- Rather than using specifically defined knowledge bases, this can be leveraged from the data
  - Provided that there are enough annotations for every risk factor…

NORTHWESTERN UNIVERSITY

Linguamatics

Temporal Relation Extraction (2/3)

- When building rule add any linguistic items that express time to the output representation.



A1C	Numerics	Time Evidence	Time Evidence		Doc		Hit
A1C	4.9	on 6-04-87	was	1	159-02	1	This is borne out by her last hemoglobin A1C on 6-04-87 which was 4.9.
	7.8	on 12/27/66	was	1	400-04	1	(A1C on 12/27/66 was 7.8.

- Normalise them to 'before', 'during', 'after' type expressions if possible
- Let the post-processing decide the temporal annotation on the basis of the evidence extracted

Temporal Relation Extraction (3/3)

- Fed these as high level features into a classifier which contained rules and machine learning components
- Data-derived rules – some risk factors had a very skewed distribution of possible values for temporal attributes
  - Almost all cardiac events were in the past
  - Medications and diseases were annotated as before, during and after unless there was an associated match with a word list of before/during/after expressions
- Logical rules:
  - Cardiac events can't be reported in the future
- Machine learned classifiers
  - Used for lab tests as they were the most varied in terms of time expressions in the training data

NORTHWESTERN UNIVERSITY

Linguamatics

Lab Test Temporal Relation Classifier

- Lab values temporal attributes classified using simple features
  - Test Type
  - Date associated with the test (considered past)
  - Matches with a ‘before’ and ‘during’ word list associated with test
  - Tense of the associated verb (this was not always useful though)
- Features were generally too sparse in the training data for huge improvements over the baseline (guessing the most common).

Classifier	Accuracy (on held out set)
Baseline	65%
Naïve Bayes	77%
CART Decision Tree	83%

NORTHWESTERN UNIVERSITY

Linguamatics

Results

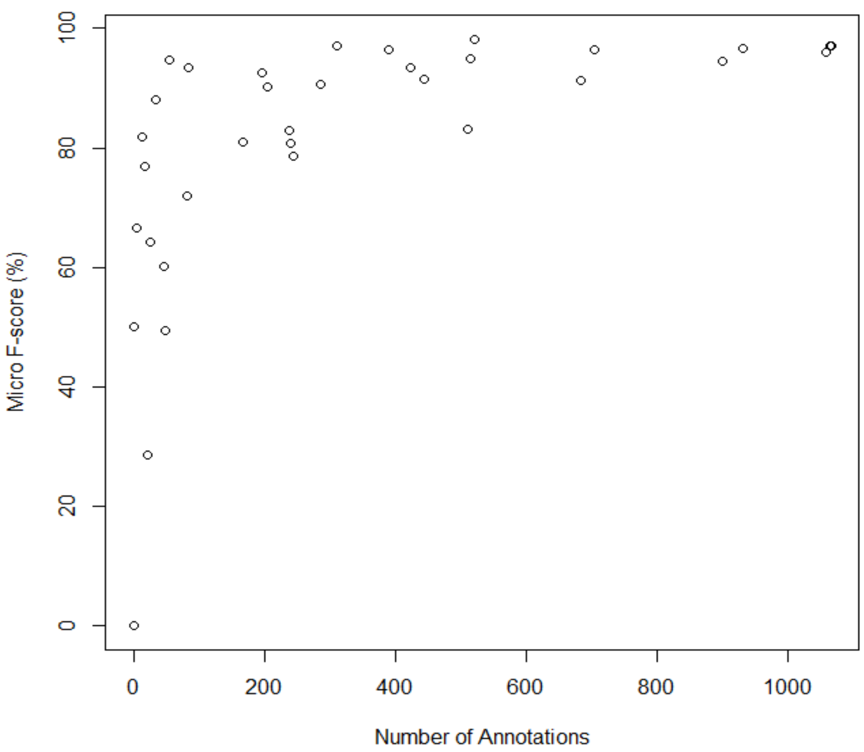
Data	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
Training	91.3	94.4	92.8	91.2	94.6	92.9
Development	88.2	92.7	90.4	88.7	92.9	90.7
Test	89.9	93.6	**91.7**	89.8	93.8	**91.7**

- For the test set, querying with I2E took 8.2 seconds and post-processing took 10 seconds (on a 3.1GHz Intel® Core™ i5-2400)

NORTHWESTERN UNIVERSITY
Linguamatics

Effect of skewed training data

- Even though this approach was mostly rule based, the number of annotations for a given risk factor had a big impact on the accuracy for that feature
- Rules still require enough examples



NORTHWESTERN UNIVERSITY

Linguamatics

Challenges

- For less frequent risk factors, it was unclear whether to use the guidelines or the annotations, where they appeared to conflict
- Loss of structure in the documents
  - Table processing produced false positives particularly for glucose
  - Results looked good to us, but actually lowered our score
  - Form based parameter-value caused difficulties in sentence break detection
- A very similar system for smoking for the i2b2 2007 dataset gave 90% Micro F-score on the test set [unpublished work], whereas the performance of smoking categorization on this set was 84%

NORTHWESTERN UNIVERSITY

Linguamatics

Conclusions

- An agile approach to text mining is a good fit for this task
  - A well performing system was developed in a few weeks, including contribution from Northwestern University who were new to I2E
- High-level features from rule based systems can provide discriminatory features for temporal relation classification
- More time could be spent trying to give structure to the document where the formatting of the document has some syntactic properties
- More unannotated data would have been useful to raise the scores for less frequent risk factors

NORTHWESTERN UNIVERSITY

Linguamatics

Acknowledgements

- Team from Linguamatics: James Cormack, David Milward
- Team from Northwestern University: Chinmoy Nath, Kalpana Raja, Siddhartha Jonnalagadda
- Thanks to i2b2 for challenge organization and data

NORTHWESTERN UNIVERSITY

Linguamatics