Drug Interaction Information Extraction from Text Using Conditional Random Fields

Stefania Rubrichi Silvana Quaglini

Laboratory for Biomedical Informatics “Mario Stefanelli”, Department of Computers and Systems Science, University of Pavia, Pavia, Italy.

NETTAB 2011
Pavia, October 12-14, 2011

Outline

Motivation

- Why is drug information needed?
  - Adverse drug events (ADEs) are a public health issue: aging patients multi-pathologies and growing complexity of drugs lead to an increased risk of medication errors and thus preventable ADEs.
  - Most of such errors occur during the prescription process and are commonly due to the lack of up-to-date knowledge about the drug and how it should be used [Leape et al 1995]

–> We propose a way of mining drug information from Summary of Product Characteristics (SPCs).

–> SPCs represent the official source of information on how to use drugs safely and effectively, the content is regulated by Article 11 of Directive 2001/83/EC.

Example of SPC

SUMMARY OF PRODUCT CHARACTERISTICS

1 NAME OF THE MEDICINAL PRODUCT

DIAMOX* 250mg Tablets
Acetazolamide 250mg Tablets

2 QUALITATIVE AND QUANTITATIVE COMPOSITION

Each tablet contains 250mg acetazolamide BP.
For excipients see 6.1.

3 PHARMACEUTICAL FORM

Tablet.

Round, convex, white tablets engraved with "FW 147" on one side and cored in quarters on the other.

4 CLINICAL PARTICULARS

4.1 Therapeutic indications

DIAMOX Tablets are for oral administration.
.
.
.

4.5 Interaction with other medicinal products and other forms of interaction
DIAMOX is a sulphonamide derivative. Sulphonamides may potentiate the effects of folic acid antagonists. Possible potentiation of the effects of folic acid antagonists, hypoglycaemics and oral anti-coagulants may occur. Concurrent administration of acetazolamide and aspirin may result in severe acidosis and increase central nervous system toxicity. Adjustment of dose may be required when DIAMOX is given with cardiac glycosides or hypertensive agents.

Objectives

–> Our goal: extract drug-related interaction information reported as free text in SPCs, following a statistic-based approach.

–> Main idea: formulate the content extraction problem as a classification problem in which we seek to assign the correct semantic label to each word of the text.

–> Our approach is based on a supervised learning technique.

–> We use a state-of-the-art classifier, linear chain conditional random fields (CRF), because of its known performance in text categorization.

Conditional Random Fields

Main idea:

Let X=$< x_1, x_2, \ldots x_n >$ random variable over data sequence to be labeled, such as a sequence of words in a text document.
Let Y=$< y_1, y_2, \ldots y_n >$ random variable over corresponding label sequence.
Let S=$< y_1, y_2, \ldots y_n >$ be a predefined set of labels.

The most appropriate labels sequence y^*:

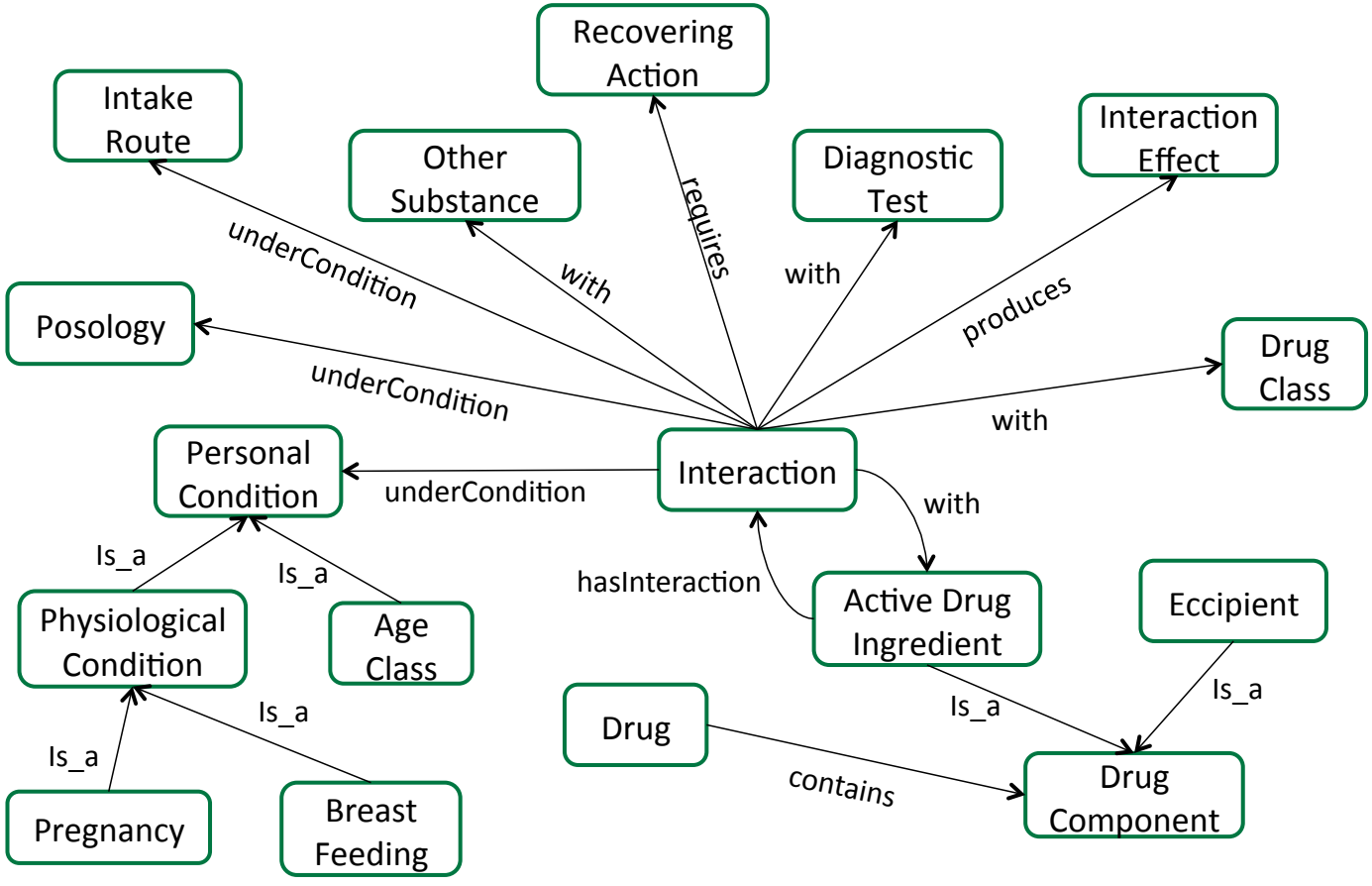
$$y^* = \arg\max_{y \in S} p(y|x)$$

Framework Outline

Our methodology is developed through five steps:

1. Semantic representation of drug information conveyed in the SPCs.
   –> *need for domain knowledge to identify the underlying semantic concept classes representing drug characteristics*.
2. Pre-processing step.
   –> *for preparing the dataset for the use by the extraction module*.
3. Hand annotation of the dataset according to the conceptual model.
   –> *for generating the gold standard*.
4. Feature definition and data conversion.
   –> *for generating the CRFs input data*.
5. Data processing through the CRFs.

1 Semantic representation: Medication Ontology



2 Pre-processing

Prediction is on a word-by-word basis, and decisions are made one sentence at a time.

–> *Split the text of SPC interaction section into sentences*

–> *Break the input sentences into tokens*

–> *Normalization step*:

- removing all punctuation except for colon and brackets
- adding white spaces between colon and brackets, and the previous word
- removing hyphens if they exist between strings
- replacing periods that occur between numbers (3.4) with commas (3,4)

3 Hand Annotation: Labeled Data

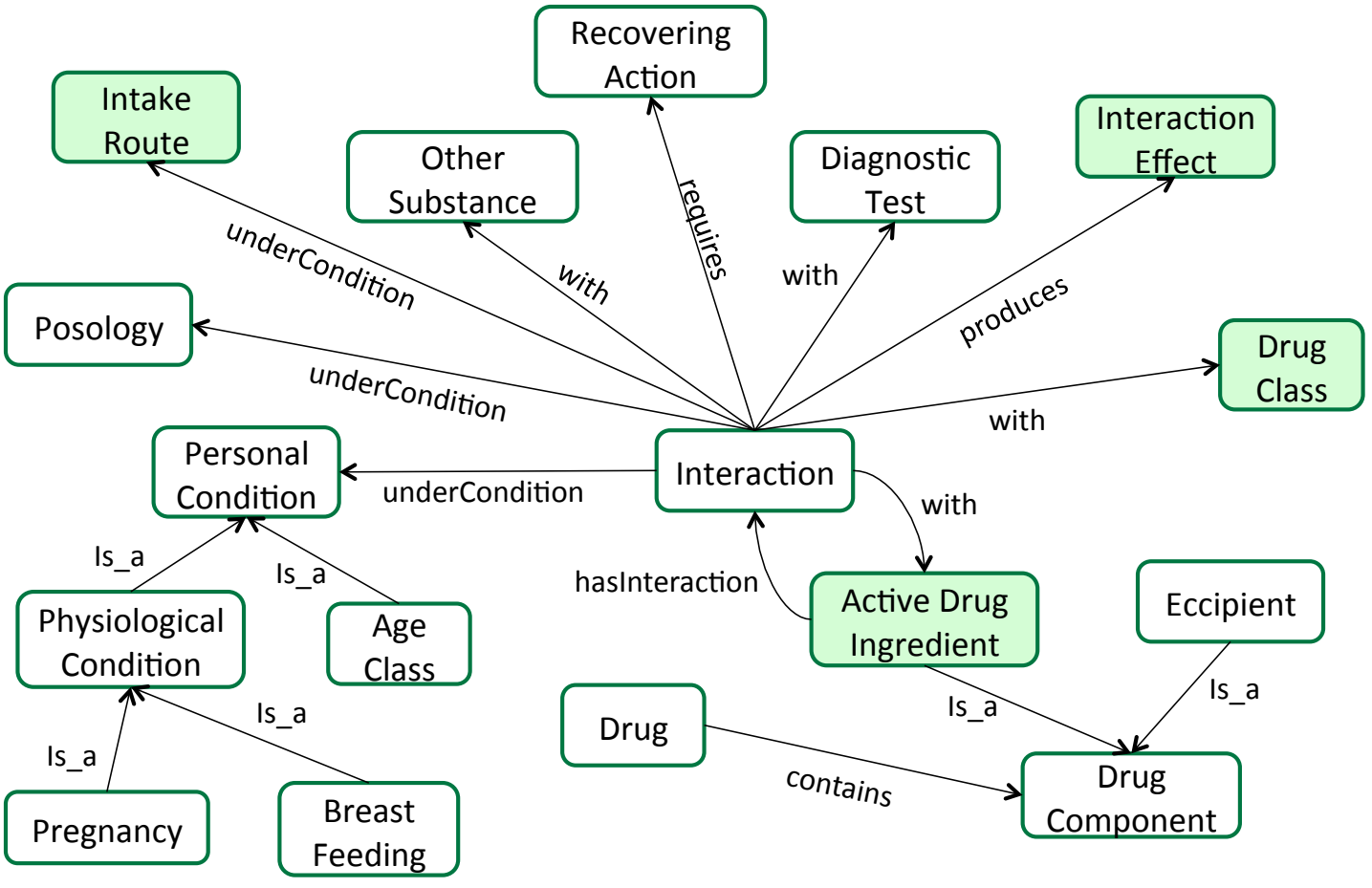
–> One hundred interaction sections in Italian language, found in the Farmadati Italia Database.
–> We annotated the corpus with 13 semantic labels according to the established ontology

Example

Salicylates may enhance the effect of oral hypoglycaemic agents, eptifibatide and sodium valproate.

⟨**Salicylates**⟩$_{DrugClass}$ ⟨**may enhance the effect**⟩$_{InteractionEffect}$ ⟨**of**⟩$_{None}$ ⟨**oral**⟩$_{IntakeRoute}$ ⟨**hypoglycaemic agents**⟩$_{DrugClass}$, ⟨**eptifibatide**⟩$_{ActiveDrugIngredient}$ ⟨**and**⟩$_{None}$ ⟨**sodium valproate**⟩$_{ActiveDrugIngredient}$.

Medication Ontology



4 Feature Definition

–> Feature definition is a critical stage regarding the success of CRFs.

–> CRFs label each token learning a correspondence between labels and features.

–> After a careful inspection of the corpus we identified a set of informative features that capture salient aspects of the data with respect to the tagging.

We compiled 5 types of features.

1. Orthographic Features;
2. Neighboring Word Features;
3. Prefix Features;
4. Punctuation Features;
5. Dictionary Features.

4 Feature Definition

–> Feature definition is a critical stage regarding the success of CRFs.

–> CRFs label each token learning a correspondence between labels and features.

–> After a careful inspection of the corpus we identified a set of informative features that capture salient aspects of the data with respect to the tagging.

We compiled 5 types of features.

5 Dictionary Features.

$$f_5(x,i) = \begin{cases} 1: & \text{if the observation at position } i \text{ is} \\ & \text{an Active Drug Ingredient} \\ 0: & \text{otherwise} \end{cases}$$

4 Data Conversion

–> Each token is represented by the set of active features.

Example

"...avoid drugs association:..."

The CRFs input corresponding to the token avoid will be:

$$f_{16}, f_6, f_{71}, f_{32}$$

$$f_{16}(x,i) = \begin{cases} 1: & \text{if the observation at position } i \text{ is } \texttt{avoid} \\ 0: & \text{otherwise} \end{cases}$$

$$f_6(x,i) = \begin{cases} 1: & \text{if the observation at position } i+1 \text{ is } \texttt{drugs} \\ 0: & \text{otherwise} \end{cases}$$

$$f_{71}(x,i) = \begin{cases} 1: & \text{if the observation at position } i+2 \text{ is } \texttt{association} \\ 0: & \text{otherwise} \end{cases}$$

$$f_{32}(x,i) = \begin{cases} 1: & \text{if there is a colon three positions after } i \\ 0: & \text{otherwise} \end{cases}$$

Results

Overall Results

Overall experimental results (in %) of CRFs.

Micro-average			Macro-average			Overall
Precision	Recall	F_1-measure	Precision	Recall	F_1-measure	accuracy
90.45	90.53	90.30	90.43	78.82	83.72	90.53

–> Micro-average: mean by weighting each label by the number of times it occurs in the data set.

–> Macro-average: arithmetic mean, giving equal weight to each of the labels.

–> In general, our experiments show that the classifier perform well, with a resulting overall accuracy of around 90%.

Results

Performance results on individual labels

Performance results (in %) of the classifier on individual labels.

Label	N_{train}	N_{test}	Precision	Recall	F_1-measure
ActiveDrugIngredient	1196	894	97.39	87.70	92.29
AgeClass	16	8	100	75.00	85.71
ClinicalCondition	77	25	100	100	100
DiagnosticTest	77	51	100	56.86	72.50
DrugClass	1527	634	87.23	70.03	77.69
IntakeRoute	40	21	80.00	76.19	78.05
InteractionEffect	1698	1165	85.75	78.54	81.99
None	11378	7623	91.04	96.39	93.64
OtherSubstance	119	58	76.47	67.24	71.56
PharmaceuticalForm	1	-	-	-	-
PhysiologicalCondition	3	-	-	-	-
Posology	256	375	94.02	88.00	90.91
RecoveringAction	787	564	82.85	71.1	76.53

Conclusion

–> Expressing the problem of content extraction in the described machine learning approach is therefore promising

–> The classifier achieves high overall accuracy.

–> The encouraging results and the ready adaptability show that our system has significance for the extraction of detailed information about drugs (drug targets, contraindications, side effects, etc.) more generally

Thank You!