Preventing Adverse Drug Events by Extracting Information from Drug Fact Sheets

Stefania Rubrichi*[1], Alex Spengler[2] , Patrick Gallinari[2] , Silvana Quaglini[1]

[1]Laboratory for Biomedical Informatics, Department of Computers and Systems Science, University of Pavia, Pavia, Italy
[2]Laboratoire d'Informatique de Paris 6, Université Pierre et Marie Curie, Paris, France

Email: Stefania Rubrichi*- stefania.rubrichi@unipv.it; Alex Spengler - alex.spengler@lip6.fr; Patrick Gallinari - patrick.gallinari@lip6.fr; Silvana Quaglini - silvana.quaglini@unipv.it;

*Corresponding author

Abstract

Background: The increasing volume and growing complexity of drugs lead to an increased risk of prescription errors and adverse events. A correct drug choice must be modulated to acknowledge both patients' status and drug-specific information. This information is reported in free-text on drug fact sheets. It is often overwhelming and difficult to access. There is thus a rising need for generating comprehensive and structured data that help prevent such events by improving access to fact sheet information. This work presents a machine learning based system for the automatic prediction of drug-related entities (active ingredient, interaction effects, etc.) in textual drug fact sheets, focusing on drug interactions.

Results: Our approach learns to classify this information in the structured prediction framework, comparing conditional random fields and support vector machines. Both classifiers are trained and evaluated using a corpus of 100 drug fact sheets. They have been hand-annotated with fourteen semantic labels that have been derived from a previously developed domain ontology. Our experimental results show that the two models exhibit similar overall performance. They achieve an average F_1-measure of about 93 per cent, which is promising. The performance results of both models on the individual labels are also comparably good.

Conclusions: We have shown that it is possible to perform the task of information extraction from drug fact sheets using supervised machine learning techniques. Although we have focused on drug interactions, the encouraging results and the adaptability of the approach we adopted means that our system has general significance for the extraction of detailed information on drugs (drug targets, contraindications, side effects, etc.).

Background

The medication management process is highly complex and involves a large number of choices from different health care professionals. Medication errors occur frequently among patients, at any point in the medication administration process. The Institute of Medicine (IOM) reports that more than 1.5 million adverse drug events (ADEs) are preventable each year in the US alone [1]. Examples of errors include a patient receiving the wrong medication, a medication to which they have a known allergy, or a patient receiving an incorrect dose of medicine. This phenomenon is aggravated by aging patients' multi-pathologies and the ever-growing number and complexity of drugs (e.g. drugs combining more than one active ingredient deserve more attention for interactions and contraindications). Physicians need to take into account many drug-specific and patient-specific characteristics and studies show that these factors are often overlooked or recognized too late.

In recent years, considerable efforts have been made to reduce medication errors and to detect and prevent ADEs. Computerized systems that incorporate specific applications in electronic medical records or in clinical information systems support medication ordering, dispensing and administration functions. These systems are referred to as Computerized Provider Order Entry – CPOE [2]. To enhance their performance, such systems have to include rich domain knowledge. Thus, they will be able to support clinical decision, assisting the physician, for example, by screening orders for allergies and drug-drug or drug-laboratory tests interactions, generating alerts tailored to patients' characteristics. Not much of such knowledge is available in semi-structured form, and even less in normalized, structured form. In particular, drug-related information is reported in free-text on fact sheet. For effective use, this information locked in natural language must first be transformed into structured data.

In this work, we consider the problem of automatic extraction of drug information conveyed in the Summary of Product Characteristic (SPC), focusing on a specific section concerned with drug-related interactions. Our contributions are:

1. We formulate the problem in a machine learning framework, in which we seek to assign the correct semantic label, such as *InteractionEffect* or *ActiveDrugIngredient*, to each word on a drug fact sheet. To this end, we employ two state-of-the-art classifiers: linear-chain conditional random fields (CRFs) and structured support vector machines (SVMs). These classifiers discriminate the semantic labels trough the automatic adaptation of hundreds of engineered text features, taking into account both local (on a word level) and global (sentence or fact sheet level) information.
2. We introduce a corpus of 100 interaction sections in Italian language that have been annotated with fourteen semantic labels, with respect to a previously implemented ontology.
3. We apply both CRF and SVM to our data set and evaluate their overall and individual label performances. Both classifiers achieve an average F_1-measures of about 93%—a promising result with regard to real-world applications.

Methods

Named Entity Recognition in SPCs

SPCs represent a source of information for health professionals on how to use medicines safely and effectively. It forms an intrinsic and integral part of the marketing authorisation. In order to obtain an authorization to place a medicinal product on the market, a SPC shall be included in the application made to the competent authority. Its content is regulated by Article 11 of Directive 2001/83/EC. A SPC sets out the agreed position (results of physico-chemical, biological or microbiological tests, toxicological and pharmacological tests, clinical trials etc.) on the medicinal product as collected during the course of the assessment process.

SPCs of specialty medicines for human use are organized into 12 sections: name, therapeutic categories, active ingredient, excipients, indications, contraindication/side effects, undesired effects, posology, storage precautions, warnings, interactions, use in case of pregnancy and nursing. Access to this comprehensive information provides a wide range of coded data, which are then available for new or improved clinical applications, facilitating and improving the prescription process. It is therefore an important step in preventing medical errors.

We propose a first approach to extracting drug-related interaction information reported as free-text in SPCs, following a named entity recognition (NER) approach. NER is an important step of an integral information extraction task and aims at identifying words or phrases in natural language text that belong to certain classes of interest, and labeling them according to their type. As an example, consider the following sentence (translated from Italian):

> $\langle$**Enoxaparin**$\rangle_{ActiveDrugIngredient}$ **dosed as a** $\langle$**1.0 mg/kg**$\rangle_{Posology}$ $\langle$**subcutaneous injection**$\rangle_{IntakeRoute}$ **for** $\langle$**four doses**$\rangle_{Posology}$ $\langle$**did not alter the pharmacokinetics**$\rangle_{InteractionEffect}$ **of** $\langle$**eptifibatide**$\rangle_{ActiveDrugIngredient}$.

In NER, each token is sought to be associated with a label that indicates its appropriate domain-specific category.

Typically, the first step in most NER tasks is to identify the named entities (labels) that are relevant to the concepts, relations and events described in the text. A system for NER is thus based on specific knowledge on the domain. Thus, as part of an understanding of the factual information process, we previously developed a domain ontology defining the entity classes, relations and attributes [3]. Based on the ontology, an extensive knowledge base of concepts is maintained.

One of the most successful methods for performing such labeling and segmentation tasks is that of employing supervised machine learning techniques. These methods automatically tune their own parameters to maximize their performance on a set of example texts that have been annotated by hand. The machine then generalizes from these examples.

We developed a framework for simultaneously recognizing occurrences of multiple entity classes using linear-chain CRFs [4] and structured SVMs [5,6]. Both supervised machine learning approaches predict words' labels using a large number of descriptive characteristics (features) of the input by assigning real-valued weight to these features. They can be seen as a way to "capture" the hidden patterns in both labels and features, and "learn" what would be the likely output considering these patterns. Due to paucity of space, however, we limit the treatment of these subjects to a presentation of the employed features and refer the reader to the original publications for details on the statistical properties.

Features

The feature construction process aims at capturing the salient characteristics of each token in order to help the system predict its semantic label. Since statistical models such as the CRF and the SVM crucially depend on a wise choice of these features, their defintion has critical impact on the overall performance of the system.

Defining features means to construct a set of generally binary-valued feature functions $f(x, t)$ for a sentence x and a word position t. For example,

$$f_{\texttt{enoxaparin}}(x, t) = \begin{cases} 1: & \text{if the word at position } t \\ & \text{in } x \text{ is } \texttt{enoxaparin} \\ 0: & \text{otherwise} \end{cases}$$

is a binary word feature which returns 1 whenever the word at position t in sentence x is `enoxaparin` and 0 otherwise.

We implemented and employed a large variety of informative features that can be derived from the fact sheets – both locally on a token or word level and globally on a sentence or section level.

Before the actual feature assignment process, we split each input sentence into tokens. We use a simple, but robust tokenization method which considers white-space, colon and parenthesis as token boundaries. We then remove all punctuation with the exception of hyphens occuring between alphanumeric strings in a second preprocessing step. Due to some length constraints our original database is not properly hyphenated. To remedy this we consulted an Italian language lexicon [7]. We now describe the features we used in our experiments.

Word and Neighbouring Word Features

Each word in the stream of tokens has been converted into a binary feature. Moreover, we equally created features for the words preceding or following the current position t in a sentence x, modeling local context. Consider the excerpt from page 2, for instance. Apart from $f_{\texttt{enoxaparin}}(x, t)$ which is 1 only for $t = 1$, we also have a feature $f_{\texttt{enoxaparin, -1}}(x, t)$ which is 1 whenever the preceding word is `enoxaparin`, here for $t = 2$. In our experiments, we report results for context sizes 0 (no context), $-3/3$ and $-7/7$.

Orthographical Features

Besides word features, we added orthographical features that indicate whether a token consists of digits. This is useful for identifying *Posology* entities.

Word Substring Features

Some substrings can provide good clues for classifying named entities. In particular, we identified a set of words which occur frequently with the same label; for example Italian words which start with "effet-" (effect) are usually *Interaction Effects*, those starting with "mg-" (mg) have usually been tagged as *Posology*, and so on.

Punctuation Features

Also notable are features which characterize interesting punctuation in sentences. After browsing our corpora we found that colons and brackets may be helpful. Given a medication, colons are usually preceded by the interacting substance and followed by the explanation of the specific interaction effects. Round brackets show extra information. For each token, the punctuation features test if it is preceded or followed by a colon or a parenthesis. All punctuation features have been used in conjunction with a context window of sentence length.

Active Ingredient Dictionary Feature

Finally, we added an examplary feature carrying domain-specific knowledge. Farmadati Italia [8] database provides a complete archive of active ingredients. The dictionary feature is based on these database entries and tells us whether a token is an active ingredient or not.

Table 1: Overall experimental results (in %) of CRF and SVM.

Model	Micro-averaged			Macro-averaged			Overall
	Precision	Recall	F_1-measure	Precision	Recall	F_1-measure	Accuracy
CRF	93.89	93.88	93.72	95.80	77.31	81.50	93.88
SVM	93.72	93.76	93.61	95.30	76.15	80.59	93.75

Experiments

Data Collection

The goal of this work lies in extracting information from SPCs, with a focus on drug-related interactions. We created a corpus which consists of 100 manually annotated interaction sections of specialty medicines for human use. They have been extracted from SPCs selected uniformly at random from the Farmadati Italia database. We used the BDF (Bancadati Federfarma) software [9] for an exploratory data analysis and for exporting the SPCs to a text file.

Ontology-based Annotation Process

Semantic annotation is used to establish links between the tokens in the SPCs and their semantic descriptions or concept classes. For a reliable annotation, the semantic descriptions must be well defined and easy to understand by the domain expert who annotates the text. We therefore annotated the text with respect to a previously developed ontology-based model of drug information as conveyed in the SPCs [3], which specifies the classes of concept (i.e. concepts representing drug characteristics), the relationships that bind them and other distinctions that are relevant for modeling the application area. The annotation process was performed by a biomedical engineer with domain knowledge. A review of the data has been used to validate and, when necessary, correct the annotations.

Leveraging the established ontology, we mapped its elements to the SPCs' text content. We accurately inspected all the corpus lines distinguishing the different senses with respect to the ontology; we then annotated each word in the extracted SPC interaction sections with the corresponding class in the ontology. Active ingredients have also been automatically extracted using the Farmadati database.

Experimetal Setup

We randomly split the 100 interaction sections into two sets; one for training which consists of 60 sections and one for testing which contains 40 sections. In total, there are 840 input sentences for training and 413 input sentences for testing.

We measure and evaluate the performance of our models based on precision (P), recall (R) and F_1-measure (F_1) [10]. We report results for the two classifiers in terms of overall and individual label performance. When dealing with multi-label classification and imbalanced labels, the performance on the individual labels can essentially be aggregated into overall performance results in two complementary ways: either we compute their arithmetic mean, giving equal weight to each of the labels (macro-averaged); or we compute the mean by weighting each label by the number of times they occur in the data set (micro-averaged).

Results and Discussion

Table 1 presents a summary of key performance figures for both CRF and SVM. Overally, our experiments show that the two classifiers, with carefully designed features, can identify information related to drug interactions with very high accuracy (about 93%). There is no clear superiority of one model over the other. Although the data might contain noise inherent to manual annotation, the learning algorithms reach high performance. This high recognition performance can be attributed to the latent structural regularities of natural language text and the regularity of the appearance of groups of named entities in the investigated paragraphs. Approaching the problem of information extraction from SPCs in the described machine learning approach is promising.

Table 2 shows the performance of CRF and SVM on the individual labels, employing all available features and a context of size -7/7. The labels *OtherSubstance* and *DiagnosticTest* are most difficult to extract, which is probably due to the tiny number of examples available. Rare labels as *AgeClass*, *IntakeRoute* exhibit better performance, they may in fact profit from a precise definition, contributing to high performance.

Moreover, we investigate the influence of the word neighbouring features with regard to overall performance. Using local word context appears to

Table 2: Performance results (in %) of the two classifiers on individual labels.

Label	N_{train}	N_{test}	CRF			SVM		
			Precision	Recall	F_1-measure	Precision	Recall	F_1-measure
ActiveDrugIngredient	1405	685	97.94	96.93	97.43	98.36	96.50	97.42
AgeClass	16	8	1	1	1	1	1	1
ClinicalCondition	61	41	1	82.93	90.67	1	80.49	89.19
DiagnosticTest	95	33	1	60.61	75.47	1	48.48	65.31
Drug	1	1	1	1	1	1	1	1
DrugClass	1397	764	90.18	87.70	88.92	90.61	87.17	88.86
IntakeRoute	40	21	94.12	76.19	84.21	93.75	71.43	81.08
InteractionEffect	1927	936	91.83	87.71	89.73	89.56	88.89	89.22
None	12711	6290	94.33	97.36	95.82	94.46	97.07	95.75
OtherSubstance	91	86	1	44.19	61.29	93.18	47.67	63.07
PharmaceuticalForm	1	0	-	-	-	-	-	-
PhysiologicalCondition	3	0	-	-	-	-	-	-
Posology	412	219	88.44	90.87	89.64	92.09	90.41	92.09
RecoveringAction	856	495	92.79	80.61	86.27	89.80	81.82	85.62

be useful for determining the semantic labels. The larger the context window size, the better and the more precise the results. Table 3 illustrates the performance of both classifiers for varying context window sizes. We followed an additive strategy: starting with no word neighbouring features (i.e. a window size of 0), we increased the window size piecemeal, measuring the performance of the resulting classifiers at each step. The initial classifiers didn't use a word neighbouring feature set. The addition of neighbouring words in the window -3/3 as features improves the F_1-measure by about $3-4\%$ (micro-averaged), and $7-9\%$ (macro-averaged). Incrementing the context window to a size of -7/7 gives rise to an improvement on all metrics, boosting micro-averaged F_1 by about 6% and the macro-averaged F_1 by about 7%. Non-zero context window sizes hence provide an important benefit with respect to the overall classification performance. An analysis of the different performance increments for contexts of size -3/3 and -7/7 will be left to future work.

Conclusions

We have presented a framework for simultaneously recognizing occurrences of multiple entity classes in textual drug fact sheets, using supervised machine learning techniques. We compared the performance of two state-of-the-art discriminative classifiers with carefully engineered features. Our empirical evaluation shows that the two classifiers exhibit similar overall performance, achieving high overall accuracy. Although we have focused on drug interactions, the encouraging results and the adaptability of adopted approach show that our system is significant for the extraction of detailed information about drugs (drug targets, contraindications, side effects).

Table 3: Variation in performance (in %) for different word feature context sizes.

Metric		Context Size					
		CRF			SVM		
		0	-3/3	-7/7	0	-3/3	-7/7
Micro-averaged	Precision	84.12	87.64	93.89	83.70	87.93	93.72
	Recall	84.60	87.83	93.88	84.34	88.22	93.76
	F_1-measure	84.14	87.31	93.72	83.45	87.96	93.61
Macro-averaged	Precision	82.80	91.45	95.80	83.45	83.12	95.30
	Recall	60.16	68.62	77.31	61.92	72.30	76.15
	F_1-measure	65.70	74.67	81.50	66.34	73.10	80.59

References

1. Institute of Medicine (Ed): *Preventing Medication Errors*. Washington: The National Academics Press 2007.

2. Sittig D, Stead W: **Computer-based physician order entry: the state of the art**. *Journal of the American Medical Informatics Association* 1994.

3. Rubrichi S, Leonardi G, Quaglini S: **A Drug Ontology as a basis for safe therapeutic decision**. In *Proceedings of the Second National Conference of Bioengineering*, Patron 2010.

4. Lafferty JD, McCallum A, Pereira FCN: **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *Proceedings of the International Conference on Machine Learning, Volume 18* 2001:282–289.

5. Tsochantaridis I, Joachims T, Hofmann T, Altun Y: **Large Margin Methods for Structured and Interdependent Output Variables**. *Journal of Machine Learning Research* 2005, **6**:1453–1484.

6. Bordes A, Usunier N, Bottou L: **Sequence Labelling SVMs Trained in One Pass**. In *ECML PKDD 2008*, Springer 2008:146–161.

7. Zanchetta E, Baroni M: **Morph-it!: a Free Corpus-Based Morphological Resource for the Italian Language**. In *Proceedings of the Corpus Linguistics 2005 conference* 2005.

8. **http://www.farmadati.it/**.

9. **http://www.farmadati.it/navigate.aspx?id=17**.

10. Van Rijsbergen CJ: *Information Retrieval*. Department of Computer Science, University of Glasgow 1979.