Structured Information Extraction from Pharmaceutical Records

Michaela Bamburová and Zuzana Nevěřilová

Natural Language Processing Centre
Faculty of Informatics
Botanická 68a, Brno, Czech Republic

Abstract. The paper presents an iterative approach to understanding semi-structured or unstructured tabular data with pharmaceutical records. The task is to split records with entities such as drug name, dosage strength, dosage form, and package size into the appropriate columns. The data is provided by many suppliers, and so it is very diverse in terms of structure. Some of the records are easy to parse using regular expressions; others are difficult and need advanced methods. We used regular expressions for the easy-to-parse data and conditional random fields for the more complex records. We iteratively extend the training data set using the above methods together with manual corrections. Currently, the F1 score for correct classification into 5 classes is 95%.

Keywords: structured information extraction, table understanding, entity recognition

1 Introduction

National authorities for drug control and administration publish the drug data including prices in order to provide pharmaceutical companies, patient organizations, and other interested parties the most recent data. Price Monitor, a product of COGVIO[1] company, is a data processing and analysis tool for planning, payer negotiations, and price management of drugs with always up-to-date information. It integrates more than 100 global public data sources of drug-related data and offers various insights and analysis for market access and pricing teams. For comparing product prices, it is necessary not only to know what products are similar by dosage strength but also by dosage form, package size, distributor, or country. This information is present in the provided data; however, not always in an easy-to-parse form.

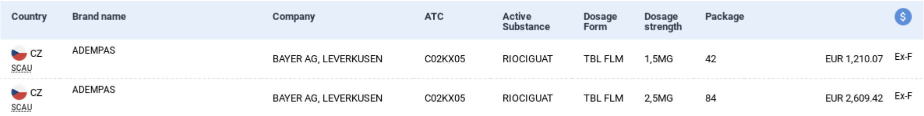
Figure 1 illustrates a typical case. One drug is provided in several dosage strengths, and the comparison of prices has to take these aspects into account.

Some suppliers provide already structured data. This data is easy to parse into the appropriate columns such as *name*, *dosage strength*, *dosage form*, *package size*, *company*, *price*, *ATC*[2] and many others. On the other hand, many suppliers

[1] `https://www.cogvio.com/`

[2] Anatomical Therapeutic Chemical is a classification system for active substances.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 55–62, 2019. © Tribun EU 2019

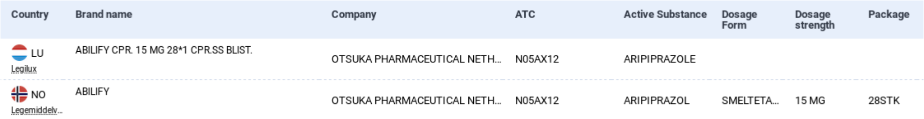


Country	Brand name	Company	ATC	Active Substance	Dosage Form	Dosage strength	Package		$
CZ SCAU	ADEMPAS	BAYER AG, LEVERKUSEN	C02KX05	RIOCIGUAT	TBL FLM	1,5MG	42	EUR 1,210.07	Ex-F
CZ SCAU	ADEMPAS	BAYER AG, LEVERKUSEN	C02KX05	RIOCIGUAT	TBL FLM	2,5MG	84	EUR 2,609.42	Ex-F

Fig. 1: Two examples of the same drug but in different dosage strengths and package sizes.

provide all the data in one column, and it is often difficult to identify and parse the data precisely, which is crucial for further analysis and reports.

Figure 2 illustrates one already structured drug data and one with part of the data in the same column.



Country	Brand name	Company	ATC	Active Substance	Dosage Form	Dosage strength	Package
LU Legilux	ABILIFY CPR. 15 MG 28*1 CPR.SS BLIST.	OTSUKA PHARMACEUTICAL NETH...	N05AX12	ARIPIPRAZOLE			
NO Legemiddelv...	ABILIFY	OTSUKA PHARMACEUTICAL NETH...	N05AX12	ARIPIPRAZOL	SMELTETA...	15 MG	28STK

Fig. 2: Examples of semi-structured and structured drug data.

In this paper, we describe the methods we use for parsing the data from individual data sets. The result of the work is the same data but fully structured. The parsing has to be performed repeatedly as the suppliers provide new drug data sets with updated prices and other drug information. The frequency of updates depends on national authorities and can vary from one day to one month.

1.1 Paper Outline

Section 2 mentions similar tasks in various domains, Section 3 describes the available data. In Section 4, we provide detailed information of the two basic methods we have used. Section 5 contains results of cross-validated evaluation. In Section 4, we plan further work on the topic.

2 Related Work

Tabular data are a very common form of information transfer. However, *tabular* not always means fully *structured*, i.e. usable by computer programs. A large survey on table understanding [2] describes various steps of table understanding in terms of dimension, nesting, generalization, and further processing (e.g., recognition of scanned tables) out of the scope of this work. Our case is one of the easiest: understanding of 1D nesting.

Understanding tables is a common task in the processing of Web documents since web tags for tables are also used for layout. [3] describes various challenges of table understanding in the context of the Web.

A major approach to table understanding is rule-based, as described e.g., in [4]. However, table cell parsing has a lot in common with sequence parsing and named entity recognition. We use two approaches, rule-based (regular expressions) and machine learning based.

3 Data Characteristics

In November 2019, Price Monitor database contained 115 million drug records that are daily growing. The data comes from more than 100 data sources and more than 45 countries in different languages. Among all the provided drug data, we are only interested in those that are usually put together in an unstructured form that makes it complicated to parse. Those data are:

- DRUG NAME (e.g., *Humira*, *Maxitrol*, *Nalgesin*)
- DOSAGE STRENGTH – indicates the amount of active substance in each dosage (e.g., *3 mg/ml*, *1.5 mg*, *100 ic/ml*)
- PACKAGE SIZE – indicates size of the package for certain drug (e.g., *28*, *2 × 330 ml*, *500 ml*)
- DOSAGE FORM – form of a drug product which contains active substance and other ingredients (e.g., solution for injection, tablet, concentrate for infusion, capsules, cream)

Because the suppliers provide this data in various forms, we divided the subset of data sets into three categories:

- YELLOW – all the data we are interested in are split into desired columns. The data are either split in the provided data sets so that no further processing is needed, or the data are split by simple processing in the form of regular expressions.
- BLUE – some of the data are split into the right columns, but some are not. For example, a name and a dosage strength are in separate columns, while a dosage form and a package size are in the same column.
- GRAY – all the data are in one column and often with different positions of values. For example, a brand name is not always at the beginning of a column, or a package size is not always followed by a dosage form in the same data set.

The data in the yellow category are entirely structured, which also means they are very uniform and contain precisely the estimated information. The yellow category contains 16 different data sets with 103 thousand records, which makes it the smallest one. The blue category contains 20 different data sets with 1.1 million records. The rest of the data sets are in the gray category. Because of the unstructured character of the data sets in the last two categories, columns contain other unnecessary information we could not easily eliminate.

3.1 Languages

The suppliers provide their data sets mostly in English, but some of them are only in their official language and therefore we need to tackle with the various cross-lingual variants (e.g., *tablet*, *tablety*, *tablett*, *tavoletta*) as well as non-Unicode characters. We also need to deal with various abbreviated words (e.g., *ml*, *mg*, *inj sol*, *tabl*, *tbl*, *filmtabl*).

4 Methods

4.1 Regular Expressions

We used regular expressions as the first method for parsing the data. However, we could use this method only for a small number of data sets and after a detailed analysis of data. We have to take into account positions of values and find patterns in their representation. Some data sets always respect their predefined format that allowed us, for instance, to rely on a brand name being at the beginning of line always followed by the same separator; or the same type of value always being in the parentheses. This method helps us mainly with enlarging the yellow category that we later used as an initial training set for another approach.

4.2 Conditional Random Fields

Conditional Random Fields (CRFs, [5]), a discriminative classifier, is a widely used statistical method for sequenced prediction. With the possibility to take context into account, it appears to be a promising method for parsing drug information.

Firstly, we had to create an appropriate training set. The first training set consists only from data sets from the yellow category, which contains already structured data, and therefore it was easy to label them. The training set was structured as a set of rows where every row represents one drug with its drug information. Every row was split into words and every word had its label. Because we are interested only about certain drug information, we used labels as DRUG NAME, DOSAGE STRENGTH, DOSAGE FORM and PACKAGE SIZE. For other unrelated drug information and punctuation marks, such as brackets, we used label OTHER. The data in training sets for the CRF method are also usually labeled with part-of-speech (POS) tags. Since our drug data are not in the form of sentences, we decided to omit this method because it would not have any additional value.

Secondly, we specified feature functions that are the key components of CRFs because they affect probabilities for sequence labeling. We started with functions such as word identity, word suffix and prefix, and whether the word is a number or punctuation. Later we experimented with more specific functions for our data, and we also specified the feature functions for neighboring words.

We set several experiments using `sklearn_crfsuite.CRF`[3] with various feature functions and training parameters described in detail in Section 4.3. For analyzing the feature weights of the model, we used `eli5`[4] Python library that provides transition features table and also state features table.

4.3 Experiments

We started with training data extracted only from the yellow category. The first experiments have shown weak predictions on data sets in languages that were not covered in the training set, especially in predictions of DOSAGE FORM (e.g., *tablet* and *tavoletta*). On the other hand, predictions of DRUG NAME or DOSAGE STRENGTH have shown above-average results since drug names usually have the same, or very similar, name in different languages, and DOSAGE STRENGTH usually follows a similar format (e.g., X ml/ X mg, X mg, X ml, X g).

Another finding was that because of the small amount of data in the training set, the predictions were biased. As we can see in Table 1, the model remembered whole words and assigned them high prediction weights. For instance, feature weights for Swedish words *peritonealdialysvätska* and *infusionsvätska* were relatively large, taking into account the fact that the values appeared only in one specific data set. As a result, the model could not perform well on new data on which it was not trained on.

To reduce over-fitting we used two different approaches, one is to enlarge the training set with more diverse data, the second is regularization.

Since we could not easily parse and label data sets from the blue category by regular expressions because of their inconsistent format, we labeled them with the classifier trained on the initial training set and then fix incorrect labels by hand. This approach allowed us to iteratively enrich the training set with data from the blue category and improved predicted results on a wider range of data with different drugs and in different languages. After several iterations, when the training set was large and diverse enough, we also labeled some data sets from the gray category to enrich the training set even more.

Another option to prevent over-fitting is to tune training parameters, especially regularization parameters L1 and L2. Regularization [1] is a smoothing technique that adds some penalty as the model complexity increases and the model consists of a large number of features. Because L1 regularization leads to feature selection and produces a sparse model by eliminating less important features [1], we tried to train a model using this technique. The cross-validation results have shown a significant change of weight from +6.780 to +4.390 for word *peritonealdialysvätska*; the model stopped to rely on particular words and started to use context more, which led to better generalization.

We performed further improvement in prediction by adding a feature that takes the prefix of a word into account. Many words in the domain start with the same prefix and differ in the endings in different languages. For example,

[3] `https://github.com/TeamHG-Memex/sklearn-crfsuite`

[4] `https://pypi.org/project/eli5/`

Table 1: Example of feature weights for dosage form after the first experiment

y='DOSAGE FORM'	top features
Weight	Feature
+10.693	word.lower():tablet
+7.471	-1:word.lower():surepal
+7.469	word.lower():tabletės
+7.270	word.lower():gelis
+6.921	-1:word.lower():trockensub
+6.780	+1:word.lower():peritonealdialysvätska
+6.513	+1:word.lower():infusionsvätska
+6.176	word[-3:]:eet
+6.061	-1:word.lower():stk
+5.872	word[-3:]:tfl
+5.808	word.lower():tabletė
+5.797	+1:word.lower():injektionsvätska
+5.555	word.lower():por
+5.555	word[-3:]:por
+5.501	word[-3:]:tti
+5.407	word.lower():capsule
+5.243	word.lower():krem
+5.221	word[-3:]:kum
	...
-9.010	word.isdigit()

the word *tablet*: in Czech it is *tableta*, in Swedish *tablett*, in German *tablette*, in shortened form can be *tabl*. After adding feature `'word[:-3]': word[:-3]` into the set of features, the prediction of DOSAGE FORM improved significantly for words with the same prefix.

Another feature – `features['BOS']`, which stands for *Beginning Of Sentence* – helps to predict the name of the drug as it is in most cases the first word in the sequence.

Since the drug data contains a lot of short words, such as *10 ml/mg*, *10 ml*, *1 vial of injection*, we decided to add features not only for 1, but also for 2 words before and after the current one to cover the context better.

We also noticed a small improvement by adding features such as `is_unit()` and `is_punctuation()` for words.

After those improvements, we achieved satisfactory results for the next iterations of the data labeling, and we could use the data from the blue category for making the training set more extensive and more diverse.

5 Results

The final training set consists of 1,687,187 drugs from all yellow, 4 blue, and 2 gray categories. The table shows the final 5-fold cross-validated results.

Table 2: 5-fold cross-validated results

	precision	recall	f1-score	support
DOSAGE FORM	0.95	0.94	0.94	388,489
DRUG NAME	0.94	0.92	0.93	257,298
OTHER	0.93	0.91	0.92	98,360
PACKAGE SIZE	0.94	0.94	0.94	391,810
DOSAGE STRENGTH	0.96	0.98	0.97	551,230
accuracy			0.95	1,687,187
macro avg	0.94	0.94	0.94	1,687,187
weighted avg	0.95	0.95	0.95	1,687,187

5.1 Error Analysis

After feature and training optimization and enlarging the training set with more various drug data, there are still some errors that often occur.

For instance, the model was trained that higher number values, such as *100, 200* are more likely a part of DOSAGE STRENGTH, whereas smaller number values, such as *10, 24,* or *50,* are more likely PACKAGE SIZE. However, when a bigger package of a drug appears (*100 tablets*), the model incorrectly predicts the value as DOSAGE STRENGTH.

Another error that occurs is related to the order of values to be predicted. The most common order is a sequence of brand name, dosage strength, package size, and dosage form.

Tables 4a and 4b illustrate an incorrect prediction when the values are provided in a less common order. We can see that prediction of DOSAGE STRENGTH value is missing and values related to DOSAGE STRENGTH and PACKAGE SIZE are predicted as OTHER.

As we did not use for training all the provided data from all categories, there are still errors in predicting values on unknown words and in languages that are not covered in the training set.

6 Conclusion and Future Work

In this work, we created a parser for records with pharmaceutical data. The purpose is to split the text into appropriate predefined columns: DRUG NAME, DOSAGE STRENGTH, DOSAGE FORM, PACKAGE SIZE, and OTHER. For roughly

Table 3: Example of correct and incorrect PACKAGE SIZE prediction.

DRUG NAME	DOSAGE STRENGTH	PACKAGE SIZE	DOSAGE FORM
viron	200 mg	70	kapsul
viron	200 mg 168		kapsul

Table 4a: Example of input data

	PACKAGE SIZE
medroxyprogesterone acetate 150 mg/ml inj,susp	1ml

Table 4b: Incorrectly labeled data

DRUG NAME	DOSAGE FORM	OTHER	PACKAGE SIZE
medroxyprogesterone acetate 150 /	inj, sus	mg ml	1 ml

one-quarter of the data, the splitting is not needed since the data already has the desired structure. We used these records as the training data and iteratively added new training examples by using regular expressions and conditional random fields. The present F1-measure in the 5-folded cross-validation evaluation is 0.95.

One future direction is to experiment more with the iterative approach, add new feature functions and discover inconsistencies in the training data.

Another future direction is to experiment with recurrent neural networks (RNNs) since in similar tasks such as named entity recognition, RNNs used together with CRFs provide the state-of-the-art results.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781.

References

1. Chaturvedi, R., Arora, D., Singh, P.: Conditional random fields and regularization for efficient label prediction. ARPN Journal of Engineering and Applied Sciences **13**(20), 8332–8336 (Oct 2018)
2. Embley, D.W., Hurst, M., Lopresti, D., Nagy, G.: Table-processing paradigms: a research survey. International Journal of Document Analysis and Recognition (IJDAR) **8**(2), 66–86 (Jun 2006). https://doi.org/10.1007/s10032-006-0017-x, `https://doi.org/10.1007/s10032-006-0017-x`
3. Hurst, M.: Layout and language: Challenges for table understanding on the web. In: Proceedings of the International Workshop on Web Document Analysis. pp. 27–30 (2001)
4. Shigarov, A.: Rule-based table analysis and interpretation. In: Dregvaite, G., Damasevicius, R. (eds.) Information and Software Technologies. pp. 175–186. Springer International Publishing, Cham (2015)
5. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning **4**(4), 267–373 (2012)