Information Extraction from Medication Prescriptions Within Drug Administration Data

Andrew MacKinlay and Karin Verspoor

NICTA Victoria Research Laboratories
University of Melbourne, VIC 3010 Australia
{andrew.mackinlay,karin.verspoor}@nicta.com.au

Abstract. We compare methods to convert a free text electronically-recorded medication prescription into a structured representation, including the dosage amount, frequency and route. We evaluate a dependency graph-based system on this data, achieving accuracies of 50–90% over various fields, and 68.3% over all fields simultaneously. This achieves superior performance to a majority baseline, and to repurposing the cTAKES drug mention annotator for this task.

1 Introduction

Clinical facilities record administration of prescription drugs to their patients, for tracking patient care and billing. Such records may include information on a drug that is administered to a particular patient, in a specific dosage, at a particular time and date. The records will also typically include the original prescription information for the patient. These data points can be used together in post hoc quality control or real-time monitoring to ensure that patients are given appropriate dosages of drugs.

The prescription text in Example 1 indicates that the dosage which the patient should have of a particular drug is between one and two tablets, that this should occur at nighttime, and that it is not compulsory. This places constraints on the allowed dosages – in particular, more than two tablets in one 24 hour period should be considered as too much. In a health-care facility with electronic records of drug administration, such prescriptions can be cross-referenced to actual recorded administered dosages, and any discrepancies could be flagged.

Example 1. Take ONE to TWO tablets at night when required

Here, we describe the application of an information extraction system to a corpus of electronically-recorded prescriptions. In previous work [2], we described a dependency-based method that transforms a free text prescription into a structured representation capturing the core constituents of the prescription: the drug dosage, frequency of administration, and the dosage route. We now evaluate that system on a newly annotated held-out test set to assess the generalisability of the approach. We also provide an analysis of the performance of the cTAKES drug mention annotator [3], developed for the somewhat different context of processing clinical narratives, on this gold standard data.

2 Prescription Data

Nature of the free-text prescription data. We work with a corpus of 500 English-language medication prescriptions for patients in long term care facilities, as would appear on a label printed by a pharmacy. These prescription instructions are stored in a database, in a text field which is separate from the drug field. The prescriptions are free text, although the format is fairly uniform, as the range of information encoded is quite narrow – primarily dosage frequency and amount (with occasional special instructions). In addition, many are filled semi-automatically using pharmacy point-of-sale software. These regularities mean that the data should be amenable to automatic processing. However, the prescriptions are intended for human readability rather than machine readability; we investigate two different techniques to interpret them automatically.

Target prescription schema. Previously, [2] we introduced an annotation schema targeted at this dosage monitoring scenario, including fields for the dosage quantity given to a patient: AmtMin (minimum dosage amount), AmtMax (maximum dosage amount), AmtUnits, (units of amount) and Opt (optionality). We also annotate dosage frequency. Most commonly, the dosages are specified as some requisite number of times within a time window, such as a day or a week. This is captured in the field PerWdw, with subfields PerWdwDays, the size of the dosage window in days, and DosesPerWdw, the number of dosages per window. Alternatively, the dosage may be specified by defining the target separation time between dosage instances, particularly when there are multiple doses per day. This is denoted as SepDoses, with subfields SepMinHrs and SepMaxHrs, the minimum and maximum separation between doses in hours. There is a clear relationship between PerWdw and SepDoses (for example the PerWdw phrase 'Once per day' would have a very similar dosage pattern to the SepDoses phrase 'every 24 hours') but the schema aims to capture the textually-specified dosage constraints as precisely as possible.

Building a gold standard set. MacKinlay *et al* produced a 40-item development set with the fields described above. This development set was used to guide the development and assess the performance of an information extraction system for the prescription data. To more rigorously evaluate the effectiveness of the system, we have created a new 60-item test set using the same schema.

3 Related Work

3.1 The i2b2 Medication Challenge

The i2b2 Medication Challenge (IMC) shared task [5] was concerned with extracting drug dosage information from medical discharge summaries. Participants were required to classify elements in the running text as denoting one of seven possible classes relating to drug delivery information. These included **medication** (such as drug name), **dosage** (referring to the amount), **mode** (drug route, e.g. *orally*) and **frequencies** (how often the drug is administered).

Our task is similar to the IMC task, but it also differs in important ways. The nature of the text is different – rather than the extended multiple-paragraph prose of clinical narratives where only a few tokens relate to medication dosages, we have terse entries of one or two sentences, almost completely concerned with dosage. In this way, the task is easier, as there is no precursor step required to identify the relevant text. However, our task requires abstractive IE over the data, i.e. extraction of specific values for specific categories of information into a structured representation rather than annotating strings, to enable comparison with structured data fields in the gold standard. Thus dosage amount, for example, is not simply the text *TWO to THREE*, but must be specified as minimum and maximum values (the real numbers 2 and 3, for AMTMIN and AMTMAX). Simple rules could often synthesise these values from annotated text, but this requires an extra processing step with a risk of introducing errors.

This distinction is clear when comparing our annotation schema with the IMC schema. Primarily, our schema lacks an explicit **medication** element, since the prescription text in our data does not include a drug name, but our schema is also more fine-grained than that of IMC. The single **dosage** class of the IMC schema corresponds to the full set of dosage-related information, but for dosage validation, it is necessary to explicitly identify the components of the dose, as defined by our schema. The IMC **frequencies** class encapsulates the PERWDW and SEPDOSES categories we introduce. Here again the schema makes a distinction, not available in the IMC schema, among frequencies that are tied to a specific time window, and those that are expressed as a time separation. Our schema does not consider the **c**lasses of **mode**/route or **reason**. In our data set, **reason** was never stated. The class **mode** was only occasionally explicitly indicated in our data; it was implicitly oral in most cases.

3.2 Medication IE using Dependencies

Our previous system [2] used a hand-crafted set of rules operating over the dependency parses derived from the relevant prescription text. For example, a node in the graph linked by *OBJ* (direct object) to the main verb of a sentence is presumed to denote dosage amount, and the links to that node are explored further to determine the DOSAGE and AMTUNITS; similar rules also yield dosage frequency information. This system is used here without modification.

3.3 Medication IE using cTAKES

There are a number of systems for information extraction over clinical text designed to work with medication descriptions. Most significant here is cTAKES, a modular collection of NLP components tailored towards processing of clinical text. Savova et al [4] augmented cTAKES with a customised drug mention annotator, which looks for drug mentions within clinical narratives to gather historical statistics on medications used to treat a particular class of disease. There are clearly important differences between the target domain of clinical narratives for which the cTAKES drug mention annotator designed and the prescription text

	Dev. Set				Test Set			
Fields	**#**	**Dep**	**cTk**	**MB**	**#**	**Dep**	**cTk**	**MB**
AmtMin	40	92.5	82.5	42.5	60	90.0	75.0	48.3
AmtMax	40	95.0	90.0	40.0	60	90.0	80.0	43.3
AmtUnits	40	92.5	90.0	52.5	60	81.7	81.7	58.3
AmtMin, AmtMax	40	92.5	82.5	37.5	60	90.0	75.0	43.3
AllDoseAmt	40	87.5	77.5	22.5	60	80.0	70.0	28.3
Opt	40	90.0	72.5	72.5	60	88.3	86.7	86.7
PerWdwDays	34	79.4	73.5	97.1	52	90.4	75.0	98.1
DosesPerWdw	34	76.5	67.6	55.9	52	90.4	73.1	63.5
PerWdw	34	76.5	67.6	52.9	52	90.4	73.1	61.5
SepMinHrs	5	60.0	0.0	0.0	2	50.0	0.0	0.0
SepMaxHrs	5	60.0	0.0	0.0	2	50.0	0.0	0.0
SepDoses	5	60.0	0.0	0.0	2	50.0	0.0	0.0
AllDoseAmt, Opt, PerWdw	34	70.6	44.1	20.6	52	71.2	51.9	23.1
AllDoseAmt, Opt, SepDoses	5	60.0	0.0	0.0	2	50.0	0.0	0.0
ALL	40	67.5	37.5	17.5	60	68.3	55.0	20.0

Table 1. Accuracy over development set and test set by field; AllDoseAmt denotes AmtMin, AmtMax and AmtUnits; PerWdw aggregates PerWdwDays and DosesPerWdw; SepDoses aggregates SepMinHrs and SepMaxHrs. The '#' column indicates the number of applicable items which have all relevant fields annotated. 'Dep' = the dependency-based IE system; 'cTk' = cTAKES; 'MB' = majority baseline.

which is our focus here. Nonetheless, we found that with some effort, we could apply cTAKES to the standalone prescription text, and adapt the outputs to make them compatible with our data format (We also experimented with MedEx [6, 1] but found that it would have required a large amount of post-processing to map from their prescription schema, so we report no results from this tool).

If Example 2 occurred in a narrative, *Aspirin* would be tagged by cTAKES as a drug mention, with associated metadata values MedicationStrength (*325*, units of *mg*), MedicationFrequency (*1*, units of *day*), MedicationForm (*tablet*), and MedicationDosage (*one* – this denotes the number of units).

Example 2. Aspirin 325 mg tablet once a day.

The configuration we used was a cTAKES drug annotator which is very similar to the annotator described by Savova *et al* [4]. cTAKES would not detect drug mentions without an explicit drug name, so we prepended the string 'aspirin' to all prescriptions in our corpus. With this modification, all prescriptions were tagged as containing drug mentions. In most cases, some metadata was extracted. We defined a mapping from the cTAKES metadata categories to our annotation schema for evaluation purposes.

4 Results and Discussion

We show results from applying the dependency graph-based approach of MacKinlay *et al* [2] to the development and test sets in Table 1. Information relating to dosage amount is extracted quite accurately over the development set, with at

least 92.5% accuracy over the individual fields, and 87.5% correct when all fields relating to dosage amount are considered together. The accuracy for PerWdw dosage frequencies is also respectable, with both fields exactly correct in 76.5% of instances. The rarer SepDoses fields are not identified as accurately, although we cannot draw strong conclusions from only 5 instances. Overall, we are able to populate all fields with exactly correct data on 67.5% of prescriptions. Our results are largely consistent on the new test data, with overall performance of 68.3%. There are some small improvements, e.g. for PerWdw, while for AmtUnits there is a slight decrease. cTAKES also showed a decrease over this field; there may be a few particularly challenging cases in the test set.

We include results for a "majority-class" baseline which returns the most frequent values for each field. We only postulate values for the most commonly observed combination of fields, since specifying values for e.g. SepMinHrs and PerWdwDays at the same time does not make sense. In all cases apart from PerWdwDays, we achieve strong improvements over the majority baseline. In a dosage validation scenario, PerWdwDays is most useful in combination with DosesPerWdw, when it can meaningfully be used to validate dosages. When these fields are combined, our rule-based method shows superior performance.

Our system consistently outperforms the cTAKES system on this data. This could be due to the adaptations that it required, beyond the fundamental difference in its target application scenario. Ad hoc prepending of a drug name can produce ungrammatical sentences and could conceivably have been detrimental to the cTAKES performance although it did not appear to be a significant problem. Our heuristic mapping of cTAKES annotation types to our schema may also have introduced errors. The performance of cTAKES is somewhat lower overall on the development set than on the test set, other than for dosage information (see line AllDoseAmt). The primary reason for this is that cTAKES is not designed to handle SepDoses dosage information, and there are more such cases in the development set than in the test set. The better performance of all systems on the PerWdw frequency information in the test set suggests that the test set is slightly easier than the development set for this data type.

While there is some room for improvement in these results, the accuracies obtained using our information extraction system should be sufficient to be used in a system for automated dosage monitoring. The system does make certain assumptions about the structure of the prescription based on the application context, so it would require some modification to generalise to IMC-style narrative text or indeed to prescriptions from other sources. Code for running the experiments described here is available as the package StructuRX, at http://nicta.com.au/business/health/biomedical_informatics/software/structurx.

5 Conclusions and Future Work

We have presented a scenario for monitoring prescription drug administration that requires mapping a free-text electronically-recorded prescription to a structured format. Such information extraction from free text fields enables cross-

referencing of the dosages required by the prescription for a given patient with the actual dosages given to that patient.

We have presented the results of applying an information extraction system over a corpus of free-text prescriptions. The system extracts structured information relevant for medication administration, capturing the dosage amount and frequency. Over a new test set, we were able to determine all fields related to dosage amount correctly for 80% of instances, and fields relating to dosage frequency for 50–71.2% of instances, depending on the method of specifying frequency. Overall, our system classified 68.3% of instances completely correctly according to our gold standard test set. We showed that these results over the test set were stable in comparison with the development data set.

We also compared the performance of our system to the publicly available cTAKES toolkit, using the components for annotating drug mentions [4]. We found that with a few adaptations we were able to apply that toolkit to this data. While our system does outperform cTAKES on this data, that tool did show reasonable performance given that it was developed to address a somewhat different task. The results of both of these methods are already highly usable, although there is doubtless room for improvements. We intend to explore the utility of these approaches to support quality control of drug administration in clinical settings in future work.

Acknowledgements

We wish to thank Jia-Yee Lee, Lawrence Cavedon and David Martinez for their advice. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

1. Doan, S., Bastarache, L., Klimkowski, S., Denny, J.C., Xu, H.: Integrating existing natural language processing tools for medication extraction from discharge summaries. JAMIA 17(5), 528–531 (2010)
2. MacKinlay, A., Verspoor, K.: Extracting structured information from free-text medication prescriptions using dependencies. In: DTMBIO'12. Maui, Hawaii, USA (2012)
3. Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. JAMIA 17(5), 507–513 (2010)
4. Savova, G., Olson, J., Murphy, S., Cafourek, V., Couch, F., Goetz, M., Ingle, J., Suman, V., Chute, C., Weinshilboum, R.: Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. JAMIA (2011)
5. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. JAMIA 17(5), 514–518 (2010)
6. Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., Denny, J.C.: Medex: a medication information extraction system for clinical narratives. JAMIA 17(1), 19–24 (2010)