Categorizing Medications from Unstructured Clinical Notes

Faisal Farooq, PhD, Shipeng Yu, PhD, Vikram Anand, MS, Balaji Krishnapuram, PhD
Siemens Medical Solutions, USA Inc., Malvern, PA 19355

Abstract
One of the important pieces of information in a patient’s clinical record is the information about their medications. Besides administering information, it also consists of the category of the medication i.e. whether the patient was taking these medications at home, were administered in the Emergency Department, during course of stay or on discharge etc. Unfortunately, much of this information is presently embedded in unstructured clinical notes e.g. in ER records, History & Physical documents etc. This information is required for adherence to quality and regulatory guidelines or for retrospective analysis e.g. CMS reporting. It is a manually intensive process to extract such information. This paper explains in detail a statistical NLP system developed to extract such information. We have trained a Maximum Entropy Markov model to categorize instances of medication names into previously defined categories. The system was tested on a variety of clinical notes from different institutions and we achieved an average accuracy of 91.3%.

Introduction
Medications form an important entity in the documentation of patient encounters. The information about medications consists of administering information (dosage, route etc.) as well as temporal information (taking at home, administered in ED, prescribed at discharge etc.). This information is indispensable for various purposes notably diagnoses, treatment, retrospective analysis for quality and clinical adherence, research, clinical trials and so on. For example, Center for Medicare and Medicaid Services (CMS) requires hospitals to report the adherence to guidelines for the various quality indicators (Heart Failure, Pneumonia, Acute Myocardial Infarction and Surgical Care). Amongst these are metrics like whether a pneumonia (PN) patient received an antibiotic within a defined time period of arrival or documentation of not doing so (e.g. if the patient was on an antibiotic at home just prior to arrival). Thus, it is important to know the category of any medication that is identified in the clinical record of a patient. Some of the common categories are Home Medications, Between Arrival and Admission (e.g. ED), During Stay and Discharge. However, much of this information is available in unstructured (free) text. Typically, this information is embedded deep into the text in various types of documents. In spite of the recommendations to push this information into structured data, much of it is still unstructured. This is often due to the fact that medications are reconciled (home, hospital stay etc.) at discharge time e.g. by stopping, continuing or changing. This reconciliation information is easy to enter using spoken text. It is a manually intensive process to read and extract such information from a record. This paper explains in detail the system developed to extract such information. After briefly discussing the related work, we present the data, our methodology and the results obtained.

| Category                      | Examples                                                                 |
|-------------------------------|--------------------------------------------------------------------------|
| Home                          | Had Z-Pak previously. No other home medications.                         |
| **HOME MEDICATIONS:**         |                                                                          |
| 1. Bacitracin                  |                                                                          |
| 2. Levofoxacin for two weeks   |                                                                          |
| **ED COURSE:**                 | Considered the differential for PE despite a therapeutic INR on coumadin. |
| 2/ Levofloxacin infusing from ED |                                                                         |
| This patient was given an aspirin initially upon arrival. |                                                          |
| **During Stay**                |                                                                          |
| 1. Ampul hung at this time    |                                                                          |
| 2/ Zosyn done infusing         |                                                                          |
| Times a day, Lovenox 40 mg subcutaneous daily. |                                                                 |
| **Discharge**                 |                                                                          |
| This nurse called into MyLocal Drugstore on Month 9th, 2011, Levaquin    |                                                                          |
| Discharge instructions given over with patient and stated understanding. prescriptions for Levaquin, Augmentin |                                                                 |
| Family requesting oral antibiotic for patient to take post discharge. This nurse spoke with Dr. Bice who wrote |                                                                 |
| **None**                      |                                                                          |
| Screening for methicillin resistant staphylococcus aureus was positive. |                                                                 |
| I will give the patient c/o Zithromax because of her pain during infusion and will give her Exelon for her |                                                                 |
| **ALLERGIES:**                |                                                                          |
| Penicillin and iodine.         |                                                                          |
| Patient agreeable to receiving primaxin via IV. |                                                                  |
| We gave the patient bactrim.   |                                                                          |

Table 1. Example of medication instances (bold) from each category in the dataset.

Related Work
Information extraction from clinical text has recently received a lot of attention. Researchers are applying machine learning and natural language processing for text mining in systems. Friedman et al. discuss the potential of using
NLP techniques in the medical domain, and also provide a comparative overview of the state-of-the-art NLP tools applied to biomedical text. Literature\textsuperscript{4,5,6} provides a survey of various approaches to information extraction from biomedical text including named entity tagging and extracting relationship between different entities and between different texts.

In general the systems that have been designed by various researchers can be classified into two broad categories i) Rule Based and ii) Learning Based. Rule based systems\textsuperscript{8,9,10} encode knowledge into grammars, if-else rules and regular expressions. They can be easy to implement, however, suffer from major drawbacks like hard maintainability and very low generalization. The rules engine is hard to maintain and number of rules can increase exponentially. On the other hand learning based systems\textsuperscript{11,12} take advantage of data and the inherent statistical knowledge. They do not require any hand crafting of rules and are easy to maintain. However, they require hand labeled data on which these systems can be trained which can be laborious to create. Given the nature of the problem where it is not feasible to encode diverse rules into a small subset, it is better, however to invest one-time effort in creation of such datasets (often called ground truth). Researchers have also focused on medication information extraction. Jagannathan et al\textsuperscript{7} assessed four commercial NLP engines for their ability to extract medication information. Uzuner et al\textsuperscript{13} describe various systems that were built as part of the i2b2 challenge. These systems, however, were dedicated to extract information such as medication name, dose, route, frequency etc. In this current work, we focus on the aspect of automatically classifying the medications into 5 categories: 1) Home, 2) Between Arrival and Admission, 3) During Stay, 4) Discharge and 5) None. Where as the first 4 categories are self explanatory, we defined the fourth category to account for the fact that sometimes, the medication names may be mentioned in some other context e.g. orders and allergies. Table 1 provides some real examples of medications in their corresponding contexts. We propose a maximum entropy based learning method which uses statistical and natural language features that are easy to extract and do not require high degree of domain expertise. The method is highly scalable and can be easily trained as more data is available.

**Data Collection and Annotation**

To account for variety of data, we collected data from four different institutions. Due to confidentiality agreements, we are unable to divulge the names of these institutions. However, it is worth mentioning that all of these hospitals are medium to large size institutions and are geographically very distant from each other within continental US. The documents were extracted for one quarter worth of patient visits and represented various types viz., ED Records, H&Ps, Consultation Reports and Discharge Summaries etc. To create the ground truth data, we developed an annotation tool to visualize and label data. This tool loaded one document at a time and highlighted single instances of medications identified. Since, our aim was to develop a module that could be used in the quality adherence space as required by CMS, we used a list of medication provided by CMS in its guidelines\textsuperscript{16}. Alternatively more comprehensive lists’ or automated methods can also be used for this step. The tool performed simple string matching (accounting for minor spelling variation) on the tokenized document in order to locate the instances of medications. Then these instances would be highlighted one at a time and the annotator (a clinical expert e.g. a nurse) could make the appropriate selection (using a radio button corresponding to each class) and click ‘Next’ to go to the second instance and repeat till the end of the document. This process was repeated for each document in the database till we had 560 documents from four different institutions (almost evenly distributed into the four document types mentioned above). After labeling, we had around 1200 instances of medications in these documents, thus averaging 2.14 meds per document with a variation of between zero to 20 medications in a single document.

**Feature Extraction**

Before extracting any features from the data, we passed it through a preprocessing module. This module performed tokenization, sentence boundary detection and stop word removal. It also performed class grouping e.g. all numbers (except dates) were replaced by $NUM, all measurements mg, cm, ml into $QUANT and all instances of medication names found by the system into $MED. The feature extraction was guided by some basic principles. We did not want features that would be extremely hard to generate or would be specific to institutions or document types. In addition, we wanted features that would generalize across languages and take the language structure into account. Lastly, we wanted features that a clinical expert would understand and often use in their judgment (sometimes unknowingly). Some of the features classes that we experimented with are described below.

*Context:* The phrasal context (or the words surrounding) the medication ($MED\textsuperscript{instance in this case}) often provide information about its group. For example, “She is allergic to $MED and $MED”, “Patient received antibiotics while in the emergency department. This consisted of both intravenous $MED and $MED”, “$MED IV complete that was hung in ER”. As we note in these cases, the tokens in italics often describe the context of the medication. This is
often referred to as n-gram features. Whereas classically n-grams mean left-context of the token in question, there is no reason to limit the scope and we extracted n-gram features from either sides of the $MED$ instance. Needless to mention, we performed stemming and canonical representation of the tokens while extracting these features.

**Tense and Negation:** In order to supplement the context (token) level features with richer knowledge, we also extracted tense (past, present, future) and negation information using the Stanford Parser\(^1\). Worth mentioning is the fact that examples like “We will give the patient $MED$” are labeled by the clinical experts as class 5 (None). The reason is that this is ‘intended’ rather than ‘performed’. In many situations, this can change and the medication may actually never be administered. Thus, this feature proves to be extremely beneficial.

**Part of Speech:** One of the most common ‘language’ features is the part of speech (POS) information which splits the tokens in a sentence into the part of the spoken text based on grammatical rules. For example a sentence, “We are going to put the patient on $MED$” would yield the following POS information

We/PRP are/VBP going/VBG to/TO put/VB the_DT patient/NN on/IN $MED$/NNP /.

We used Mayo Clinic’s cTakes\(^1\) POS tagger. In addition, we also extracted the phrase-level information such as Noun Phrase(NP), VP (Verb Phrase) etc. using the cTakes chunker. The former is often referred to as deep parsing and the latter as shallow parsing.

**Document metadata:** The document metadata such as document type and document date often provides useful insights about the content of the document. For example, in case a medication instance is described as being administered two days ago and the document is an ED Report, then most likely that instance of the medication refers to a medication being taken at home. Whereas documents like Discharge Summaries, Consultation Reports, and Progress Notes etc. represent a uniform distribution of all 5 classes, some document types are more representative of some classes than others. For example, an ED Report mostly represents Classes 1, 2 and 5 and an H&P mostly represents 1, 2, 3 and 5. Sometimes, the document type often acts as a tie breaker. As an illustration, without knowledge of document type, the instance in the statement “The patient was given $MED$”, it can equally represent Classes 1, 2 and 3. However, if this statement is found in an ED Record, then the tie breaks in favor of Class 2. Thus, using the document metadata would prove useful information. In order to account for unknown document types, we added a category Unknown to the types of documents.

**Heading:** A document is often divided into sections with each section starting with a heading. For example,

KNOWLEDGER ALLERGIES: Amiodarone Hydrochloride, Niacin

HOME MEDS: Tylenol, Levaquin, Insulin

Even though this is useful information, it is often not easy to detect headings. We used simple rules like uppercase, title case, semi-colon, paragraph etc. to detect certain tokens are headings and used it as a feature. A tribute to statistical systems is that all features are weighted by the learning algorithm and if we are not sure of the feature extracted, it is automatically assigned a low probability.

All of these groups of features were used to extract numerous features (often thousands to tens of thousands) and fed into a learning algorithm for classification. The next section describes the learning and classification algorithm we used in the system.

**Classification**

We used the maximum entropy markov models for classification. The motivating idea behind maximum entropy is that the most uniform model should be preferred when no information is present and whatever information is present should constrain that uniform model. Thus, the Maximum Entropy model prefers the most uniform model satisfying the constraint

$$P(c | F) = \frac{\sum_{c'} \lambda_i f_i(F,c)}{\sum_{c'} \sum_{c'} f_{i}(F,c')},$$

$\lambda_i$’s represent parameters that are estimated during the course of the Maximum Entropy training. The $f_i$’s are any real-valued functions describing features of the (feature, class) relationship, $F$ represents the features extracted, and $c$ represents a medication category (1-5). Since the Maximum Likelihood training overfits (Nigam et al.\(^1\)), we use a Maximum Aposteriori training with Gaussian priors over feature functions. The prior probability of the model is the product of Gaussians of each feature value $\lambda_i$ with variance $\sigma_i^2 = \frac{1}{\sqrt{2\pi\sigma_i^2}}$. The training has been implemented

50
using the Improved Iterative Scaling (IIS) algorithm which is of the Quasi-Newton family of numerical algorithms. We used the openNLP MaxEnt system for training and classification with the setting as mentioned above.

**Experiments and Results**
The data was split into training and test set. We did not have any separate validation (parameter tuning) sets. This was taken care of by the default configuration in the openNLP toolkit. We performed a wide variety of experiments to validate impact of features, generalization and scalability. They are described as follows:

*Feature Validation:* To evaluate each individual feature, we split the data evenly into training and test. We started off with the simplest context (ngram) feature described in the above section and trained our model following which we evaluated the model performance on the test data. After that, we started adding one feature group at a time and evaluate the differential gain. We added the feature groups in the following order: a) Context, b) +POS, c) +Tense, d) +Negation, e) +Metadata, f) +Heading. Figure 2 shows the gain in the accuracies by adding a feature group to the existing models. Note that even though the actual accuracy numbers would certainly depend on the order in which the features are added, the average gains would be similar.

*Leave One Out (Document Type):* In this case, we trained on all the data except the data corresponding to one document type. We tested on the held out data (corresponding to a document type never encountered during training). We repeated this such that each document type was held out at least once and then calculated the average of the accuracy scores. The average accuracy in this setup was 83.9%. This experiment gives us an idea of generalization within an institution in case unseen document types are encountered.

*Leave One Out (Institution):* Our next experiment was aimed at estimating the generalization across institutions. We held out all the data from one institution and trained on the other 3. We evaluated the trained model on the held out institution. We repeated the experiment till every institution was held out and averaged the accuracy scores. The average accuracy in this setup was 82.9%.

*Overall Results:* The final experiment we conducted was randomly splitting the data into a training set (75%) and a test set (25%) and performing a cross-fold validation. The average accuracy was 91.3 ± 2.11%.

**Discussion**
We achieved encouraging results using the system. In this section, we will also present some cases that we believe the system should be able to identify correctly, but is not doing so (1-2 in Table 2). Such cases show that there is some room for improvement. However, we will also show examples of some cases which are extremely hard (3-4 in Table 2). Example 3 is classified as Class 5 because it is an Order, which would make it correct. However, according to the experts since there is a second Signed entry by a nurse, this means the nurse actually administered the medication.

---

**Figure 1 Accuracy Gains by adding different feature groups**

| Example                                                                 | O | GT |
|------------------------------------------------------------------------|---|----|
| 1. $MED$ was not started here in the ER because the patient is given $MED$ for an unknown reason | 2 | 1  |
| 2. MEDICATIONS: $MED$, $MED$, $MED$ (which she stopped 3 days ago), $MED$ | 1 | 5  |
| 3. Order: $MED$ Signed: Dr. John Doe MM/DD/YYYY HH:MM:SS              | 5 | 2  |
Signed: Jane Doe, RN MM/DD/YYYY HH:MM:SS

| HOME MEDS: 1. $MED, 2. $MED |
|-----------------------------|
| She is a non-compliant patient and does not take her meds. |

### Table 2 Example failed cases

*(Med is in bold italics. O is system out and GT is groundtruth)*

| 4. | HOME MEDS: 1. $MED, 2. $MED |
|-----------------------------|
| She is a non-compliant patient and does not take her meds. |

### Conclusion and Future work

In this paper, we presented a system that classifies an instance of a medication into one of the pre-defined categories. We used statistical natural language processing and machine learning to address the problem and indeed achieved very encouraging results. The rationale for this system is CMS quality reporting where many quality indicator questions to be reported require this information. This is being performed manually in many hospitals. Even though this use case in itself is large enough to warrant addressing the challenge, the system can be used in many other cases also. Guideline adherence to best practices (e.g. aspirin within 24 hrs), support to med reconciliation tools, interactive discharge summaries etc. are other use cases for such a system. As part of the future work, we plan to improve the detection of headings (maybe make it stochastic) and also are looking into decomposing a document into logical sections which will prove useful features for the system.

### References

1. Guergana Savova, Karin Kipper-Schuler, James Buntrock, and Christopher Chute: UIMA-based clinical information extraction system. In: Language Resources and Evaluation: LREC: Towards enhanced interoperability for large HLT systems: UIMA for NLP. Marrakech, Morocco, 2008.
2. Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan and Stuart Nelson: RxNorm: Prescription for Electronic Drug Information Exchange, In: IT Professional 7(5), 17-23 2005, ryor TA, Gardner RM, Clayton RD, Warner HR. The HELP system. J Med Sys. 1983; 7:87-101.
3. C. Friedman, and G. Hripcsak, “Natural Language Processing and Its Future in Medicine: Can Computers Make Sense out of Natural Language Text”, in Academic Medicine, August 1999, Vol. 74(8), pp. 890-895.
4. Soderland, S., Aronow, D., Fisher, D., Aseltine, J., and Lehnert, W., Machine Learning of Text Analysis Rules for Clinical Records, CIIR Technical Report, University of Massachusetts Amherst, 1995.
5. Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A. 2006. Approaches to text mining for clinical medical records. In Proceedings of the 2006 ACM Symposium on Applied Computing (Dijon, France, April 23 - 27, 2006). SAC ’06. ACM, New York, NY, 235-239.
6. Angus Roberts, Robert Gaizauskas, Mark Hepple. Extracting Clinical Relationships from Patient Narratives, BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 10–18, Columbus, Ohio, USA, June 2008
7. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2008;78:284-91.
8. Hua Xu, Shane P Sterner, Son Doan, Kevin B Johnson, Lemuel R Waitman, Joshua C Denny, MedEx: a medication information extraction system for clinical narratives, JAMIA 2010 17: 19-24
9. Chhieng D, Day T, Gordon G, et al. Use of natural language programming to extract medication from unstructured electronic medical records. AMIA Annu Symp Proc 2007:908.
10. Levin MA, Krol M, Doshi AM, et al. Extraction and mapping of drug names from free text to a standardized nomenclature. AMIA Annu Symp Proc 2007:438-42.
11. Jon Patrick, Min Li, A Cascade Approach to Extract Medication Event, i2b2 NLP Workshop, AMIA 2009
12. Zuo Feng Li, Yonggang Cao, Qing Zhang, Shashank Agarwal, Bajaj P Ramesh, Daehyun Kim, Hong Yu, Extracting Medication Information from Patient Discharge Summaries, i2b2 NLP Workshop, AMIA 2009
13. Özlem Uzuner, Imre Solti, Eithon C dag: Extracting medication information from clinical text. JAMIA 17(5): 514-518 (2010)
14. Journal of American Medical Informatics Association (JAMIA), September 2011, Volume 18, Issue 5
15. Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.
16. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=122877152586
17. Kamal Nigam, John Lafferty and Andrew McCallum. Using Maximum Entropy for Text Classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67. 1999
18. Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430