Precise Medication Extraction using Agile Text Mining

Chaitanya Shivade[*], James Cormack[†], David Milward[†]
[*]The Ohio State University, Columbus, Ohio, USA
[†]Linguamatics Ltd, Cambridge, UK
shivade@cse.ohio-state.edu,
{james.cormack,david.milward}@linguamatics.com

Abstract

Agile text mining is widely used for commercial text mining in the pharmaceutical industry. It can be applied without building an annotated training corpus, so is well-suited to novel or one-off extraction tasks. In this work we wanted to see how efficiently it could be adapted for healthcare extraction tasks such as medication extraction. The aim was to identify medication names, associated dosage, route of administration, frequency, duration and reason, as specified in the 2009 i2b2 medication challenge.

Queries were constructed based on 696 discharge summaries available as training data. Performance was measured on a test dataset of 251 unseen documents. F1-scores were calculated by comparing system annotations against ground truth provided for the test data.

Despite the short amount of time spent in adapting the system to this task, it achieved high precision and reasonable recall (precision of 0.92, recall of 0.715). It would have ranked fourth in comparison to the original challenge participants on the basis of its F-score of 0.805 for phrase level horizontal evaluation. This shows that agile text mining is an effective approach towards information extraction that can yield highly accurate results.

1 Introduction

Medication information occupies a sizeable portion of clinical notes, especially discharge summaries. This includes medications on admission, during hospital course, and at discharge. This information is useful for clinical tasks such as inferring adverse drug reactions, clinical trial recruitment, etc. The i2b2 Natural Language Processing (NLP) challenges encourage the development of systems for clinical applications, using a shared task, publicly available clinical data, and comparison of performance with the other participating systems, subject to rigid evaluation metrics. The 2009 challenge (Uzuner, Solti, & Cadag, 2010) aimed to extract mentions of medication names, associated dosage, route of administration, frequency, duration and the reason for medication.

The project used the Linguamatics Interactive Information Extraction (I2E) platform. This combines NLP, terminologies and search technology to provide a unique “agile” text mining approach (Milward et al., 2005) that can yield highly precise results in a small amount of time. The approach involves semantic annotation and indexing of data followed by interactive design of queries that capture typical syntactic and semantic features of the desired information. While the system uses machine learning approaches within its core linguistic processing, the final set of queries are essentially syntactic/semantic rules identifying specific information in the text.

2 Section Identification

Although discharge summaries are considered to be unstructured data, there are typical characteristics associated with them. There is a specific flow of information within every discharge summary, starting with details of patient’s admission, followed by the hospital course and ending with discharge instructions. Other common sections include chief complaint, physical examination, etc. There were more than twenty headings to express discharge medications in the training data (“Medications on discharge,” “Discharge meds,” etc.). The training data was processed to identify section headings and multiple forms of the same heading were normalized to a single heading. The plain text

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 75–79,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

was converted into XML with tags representing section names.

3 Offset Information

To allow evaluation of results in the i2b2 format, the text was preprocessed to include line numbers and word numbers as further XML annotations**.**

4 Natural Language Processing

Indexing documents with I2E uses a standard NLP pipeline involving tokenization of the text, part-of-speech tagging, and linguistic chunking. The output of the pipeline provides useful linguistic information, particularly about the location of noun phrases and verb phrases, for use in entity extraction and querying.

5 Terminologies

The I2E platform uses hierarchical terminologies to extract entities from the text. These can include freely available terminologies such as MeSH, and the NCI thesaurus, as well as proprietary terminologies such as MedDRA. A series of regular expressions allow for the indexing of numeric terms (integers, fractions, decimal numbers) and measurement units (length, time, weight, etc.). In addition, custom terminologies can be created for specific tasks by combining or merging existing terminologies, or by using the system itself to help discover terminology from the data.

6 Querying

The I2E framework provides an interactive querying experience that is similar to a web search. While users can enter text queries just as one might in an internet search engine, the query interface also allows specification of linguistic and non-linguistic units as 'containers' for other units. For example, it is possible to search for a noun phrase within a sentence and to specify words, regular expressions and concepts from terminologies. Non-linguistic units can be customized to regulate the ordering of items within the container, the number of items that may occur between two items and whether they are constrained by linguistic boundaries, such as the sentence. The output of the query can also be customized so as to provide structured representation of the query results.

As an example, one of the typical ways a medication is prescribed follows the construct: "Aspirin 625 mg p.o. b.i.d." This means Aspirin with a dosage of 625 milligrams is to be consumed orally (p.o.), twice a day (b.i.d.). A query to capture this construct can be constructed as a non-linguistic phrase, starting with (a) a pharmacological substance (a concept from the appropriate branch of the NCI-thesaurus), followed by (b) a numerical term, (c) a unit for measuring weight, (d) a dosage abbreviation and finally, (e) an abbreviation for the frequency of medication.

A query containing items only for (a), (b) and (c) will give results for all phrases containing a pharmacological substance followed by its dosage (Aspirin 625 mg, Tylenol 350 mg, etc.). The graphical query interface is sufficiently flexible to allow many different orderings of these constructs and to negate false positive results.

User defined terminologies can be systematically constructed to allow consistent matching of lists of terms and to generate concise queries. For example, candidates for abbreviations corresponding to the route of administration were found by constructing a query with items for (a), (b), (c) and (e) and an empty word container for (d). This gave all phrases containing (a), (b), (c), and any word in the discharge summary that was followed by (e). The results of this query were candidates for route of administration. The efficiency of querying in I2E provides an opportunity to interactively refine parts of the final query and discover terms in the training data that might be missed by regular expressions and thesauri.

Queries can also be limited to specific sections of the document. The pre-processing step described above identified sections in discharge summaries of the i2b2 medications challenge corpus. The queries can thus be limited to only a few specific sections such as “Medications on Admission” and “Medications on Discharge” by embedding the query in a section container. The challenge specified not to include medications mentioned as allergies for a patient. Results obtained in the allergies section of discharge summaries were therefore ignored using this approach.

7 Post-processing

I2E’s default output is an HTML table with columns corresponding to different containers used in the query. Output can also be limited to predefined columns of interest. Multiple queries are often required to capture different pieces of information spread across the corpus. In the i2b2 challenge, there are multiple fields associated with every mention of medication. A single structured record corresponding to every mention of medication is expected as an output. Spasic et al. (2010) view the challenge as a template filling task where the participating system is expected to fill slots in a template. Thus, the output can be configured to be 6 columns representing each of the templates. Following their terminology, different semantic queries filled different slots of the same template. These slots were aggregated into a single template using post-processing.

Multiple issues had to be taken care of in this step. Different queries captured parts of the text corresponding to the same slot. For example, a query aimed at capturing a particular linguistic construct may extract frequency as “daily after dinner,” while another query may capture its substring “daily.” In this case, the former extraction, which is the longer string, received priority as per the challenge specifications. Another important problem encountered was that of multiple matches for the same field. For example, Insulin and Aspart were identified as separate pharmacological substances during the indexing process. However, “Insulin aspart” is considered as a single medication name as per the challenge specifications. Two separate templates are thus created. The results of the post processing collapse them into one. Certain terms from the terminologies did not match the definition of a medication, since terminology branches are often generic. For example, the Chemicals and Drugs branch of MeSH constitutes terms such as coffee. Therefore, a list of false positives for medication names corresponding to these matches was generated from the training data.

8 Experiments

The i2b2 website offers downloading of the NLP dataset for the 2009 challenge after signing a Data Usage Agreement. The training data consists of 696 discharge summaries. A subset of ten documents with gold standard annotations has been made available by the organizers. The test dataset consists of 251 documents which were annotated by the participants under a community annotation experiment conducted by the organizers (Uzuner, Solti, Xia, et al. 2010). These 251 documents and their corresponding gold standard annotations are also available. The performance was calculated using phrase level and token level metrics for horizontal and vertical evaluations as defined in (Uzuner, Solti, & Cadag, 2010). The phrase level horizontal evaluation measures the performance of a system across all six fields. This was used as a primary metric to rank the results in the challenge.

Terminology	P	R	F1
NCI	0.953	0.657	0.777
MeSH	0.923	0.563	0.699
NCI + MeSH	0.932	0.688	0.792
NCI + FDA	0.947	0.678	0.790
MeSH + FDA	0.921	0.571	0.705
NCI + MeSH + FDA	0.931	0.698	0.798
NCI + MeSH + FDA + RxNorm	0.92	0.715	0.805

Table 1: Comparison of Different Terminologies.

In order to assess the utility of different terminologies, the same set of queries were modified by replacing the concept from one with the corresponding concept in another. For example: Pharmacological substance from NCI was replaced with Chemicals and Drugs from MeSH. This offered an objective way to compare the coverage of MeSH and NCI with respect to medication names. Coverage of multiple terminologies can be leveraged by aggregating the results of queries resulting from different terminologies. NCI thesaurus, MeSH, a list of FDA drug labels, and RxNorm were used. In addition a custom terminology was prepared by capturing medication names in the training data that were missed by the terminologies. The best F-score was obtained when query results for all sources were aggregated. Addition of sources resulted in a drop in precision but increased recall. Table 1 summarizes these results, where columns P and R denote precision and recall respectively.

9 Results

Twenty teams representing 23 organizations and nine countries participated in the medication challenge. The other systems used a variety of rule-based, machine-learning and hybrid systems, with the most popular being rule-based systems (Uzuner et al., 2010). The best ranked system, detailed in Patrick & Li (2009), was an example of a hybrid system, using both rule-based and statistical classifiers.

No.	Group	P	R	F1
1	USyd	0.896	0.82	0.857
2	Vanderbilt	0.840	0.803	0.821
3	Manchester	0.864	0.766	0.812
*	I2E	0.920	0.715	0.805
4	NLM	0.784	0.823	0.803
5	BME - Humboldt	0.841	0.758	0.797
6	OpenU	0.850	0.748	0.796
7	UParis	0.799	0.761	780
8	LIMSI	0.827	0.725	0.773
9	UofUtah	0.832	0.715	0.769
10	U Wisconsin Madison	0.904	0.661	0.764

Table 2: Phrase level horizontal evaluation

Phrase level horizontal evaluation was used as a metric to rank the performance of participants in the challenge. Table 2 compares the performance of I2E with the top ten participants in the challenge using this metric. It achieves highly precise results as compared to other participants of the challenge. The vertical evaluation which measures the performance along individual fields showed that the system performed poorly on duration and reason, in common with other systems. As reported by the organizers of the challenge (Uzuner et al., 2010), capturing duration and reason is a hard task. They report that this is primarily due to the variation in length and content of these fields in the training and testing data.

10 Conclusion

Extracting information through interactive design of queries can achieve highly precise results in a short amount of time. Much of the time in this project was spent on pre-processing documents to allow the results to conform to the i2b2 format. The time taken on query development was of the order of a few weeks, including a couple of days training in the system at the start of the project. This process requires far less specialist knowledge of Artificial Intelligence than other solutions to this challenge and the easy to use interface means refinement is straightforward. Clearly, recall still needs to be improved: our best system would have been ranked 4th out of 21 systems in the phrase level horizontal evaluation. Examination of the training material suggests this is due to gaps in the drug coverage provided by the terminologies rather than gaps in the query patterns. We will therefore concentrate on extending drug coverage in our future work.

Acknowledgments

The authors would like to thank Tracy Gregory, Himanshu Agarwal and Matthijs Vakar for their help in this project.

References

Milward, D. et al., 2005. Ontology-based interactive information extraction from scientific abstracts.

Comparative and functional genomics, 6(1-2), pp.67–71.

Spasic, I. et al., 2010. Medication information extraction with linguistic pattern matching and semantic rules. *Journal of the American Medical Informatics Association* : *JAMIA*, 17(5), pp.532–5.

Uzuner, O., Solti, I., Xia, F., et al., 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17(5), pp.519–23.

Uzuner, O. Solti, I. & Cadag, E., 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 17(5), pp.514–8

Patrick J & Li M. A Cascade Approach to Extracting Medication Events. Proceedings of the Third i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2009.