Text Mining in Clinical Domain: Dealing with Noise

Hoang Nguyen
Data61 - CSIRO
13 Garden Street
Eveleigh, NSW 2015, Australia
hoang.nguyen@data61.csiro.au

Jon Patrick
University of Sydney
1 Cleveland Street
Sydney, NSW 2006, Australia
jonpat@it.usyd.edu.au

ABSTRACT

Text mining in clinical domain is usually more difficult than general domains (e.g. newswire reports and scientific literature) because of the high level of noise in both the corpus and training data for machine learning (ML). A large number of unknown word, non-word and poor grammatical sentences made up the noise in the clinical corpus. Unknown words are usually complex medical vocabularies, misspellings, acronyms and abbreviations where unknown non-words are generally the clinical patterns including scores and measures. This noise produces obstacles in the initial lexical processing step as well as subsequent semantic analysis. Furthermore, the labelled data used to build ML models is very costly to obtain because it requires intensive clinical knowledge from the annotators. And even created by experts, the training examples usually contain errors and inconsistencies due to the variations in human annotators' attentiveness. Clinical domain also suffers from the nature of the imbalanced data distribution problem. These kinds of noise are very popular and potentially affect the overall information extraction performance but they were not carefully investigated in most presented health informatics systems.

This paper introduces a general clinical data mining architecture which is potential of addressing all of these challenges using: automatic proof-reading process, trainable finite state pattern recogniser, iterative model development and active learning. The reportability classifier based on this architecture achieved 98.25% sensitivity and 96.14% specificity on an Australian cancer registry's held-out test set and up to 92% of training data provided for supervised ML was saved by active learning.

Keywords

Clinical, active learning, text classification, named-entity recognition, natural languages processing

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939720

1. INTRODUCTION

Processing clinical texts is quite challenging. Firstly, clinical records comprise idiosyncratic spellings, abbreviations, acronyms, poor grammatical structure, and up to 30 percent of non-word tokens. Besides resolving misspellings, knowing the correct expansions of abbreviations and acronyms is critical to understanding the document for both automatic natural language processing as well as human comprehension and interpretation [27]. Secondly, an important part of narrative reports that needs to be captured is clinical scores and measures as doctors infer a patient's status by analyzing these complex patterns. For example, BP 140/65 (84) is an example of Blood Pressure; HR 72 is pattern of Heart Rate. These lexical obstacles in the clinical corpus should be addressed at an early step of processing to avoid inherited chain of errors.

Our research focuses on supervised ML approaches because designing rules and patterns is time consuming and complicated process for human experts to complete. The basic advantage of ML is that the concept characteristics and classification rules can be automatically learnt through training examples. Therefore, the high quality of a semantically classified and annotated corpus is mandatory to contribute to the success of supervised learning algorithms. However, the training examples are usually not free and obtaining these labels is a time consuming and high labour cost process. Narrative reports usually take longer time to annotate as they require expertise knowledge in the field. Even if tagging is provided by human experts, there is a high level of inconsistency in the labelled data because some instances are implicitly difficult for annotators and also they become distracted or fatigued over time, introducing variability in the quality of their annotations [29]. Besides the noise caused during the annotation process, the train set of clinical ML also suffers from the problem of imbalanced data distribution: one class (usually negative class) may have many more instances that dominate all other classes (e.g. non-cancer cases vs cancer cases in radiology reports). This can cause a bias in the training process and may require a special active learning method for the sample data to improve the overall ML performance.

To the best of our knowledge, none of the published health information system is capable of addressing all of the above challenges. In the present paper, we propose the novel data mining system that employed natural language processing techniques, supervised ML and active learning approaches to overcome the difficulties when working with clinical text.

Our system focuses on clinical text classification and named-entity recognition and includes three main components:

- The specialised pre-processing system to deal with noise in the corpus.
- The iterative model development process to reduce both the noise and cost of manual annotations.
- Active learning infrastructure to optimise the training production of supervised learning. The main purpose of the optimisation process is to achieve the best performance with a smaller number but higher quality training examples, hence minimising labelling costs. The problem of imbalanced data distribution is also improved by applying special active learning strategies (Section 3.3).

The rest of the paper is structured as follows. Section 2 gives an overview of the previous clinical and radiological information extraction systems. Our system architecture is presented in Section 3. The proposed system demonstrated the high performance in classifying real-world and large scale radiology reports provided by Australian Cancer Registries. Experiments and results related to the Cancer Council's project are reported in Section 4. We conclude in Section 5 with directions for future work.

2. RELATED WORK

A variety of methods and systems have been implemented in the clinical domain to extract information from free text. The popular learning tasks include document classification, named-entity recognition (NER) and classification of relationship between the entities (medical concepts).

Friedman et al. (1994) [9] developed the medical language extraction and encoding (MedLEE) system, which used a domain-specific vocabulary and semantic grammar to process radiology reports. It was initially used to participate in an automated decision-support system, and to allow natural language queries. MedLEE was then adapted to automatically identify the concepts in clinical documents, map the concepts to semantic categories and semantic structures [10]. The final semantic representation of each concept contained information on status, location and certainty of each concept instance. Haug et al. (1995) [12] introduced symbolic text processor (SymText), a natural language understanding system for chest x-ray reports. SymText processes each sentence in a document independently with syntactic and probabilistic semantic analysis. Bayesian networks are used in SymText to determine the probability that a disease is present in the patient.

An early combined classifier approach in biomedical NER proposed a two-state model in which boundary recognition and term classification are separated into two phases [17]. In each classification phase, different feature sets were selected independently, which is more efficient for each task. A comparative study between two classical ML methods, Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) for clinical NER shows that the CRFs outperformed SVMs in clinical NER [19].

When extracting information from narrative text documents, the context or assertion of the concepts extracted play a critical role. The NegEx algorithm of [3] implements dictionaries of pre-UMLS and post-UMLS phrases that are indicative of negation to identify positive and negative assertions. NegEx uses a rule-based method and heuristics to limit the scope of indicative phrases. The i2b2 challenge's assertion classification is an extension of a previous system designed by [35], in the new specification an uncertainty assertion is divided into values of hypothetical, conditional and possible. A combination of ML and rule-based approaches is utilised in the system of [35]. One of these approaches extends the rule-based NegEx algorithm to capture alter-association in addition to positive, negative and uncertain assertions; the other employs an SVM to present a ML solution to assertion classification.

For the relationship classification task, there are many definitions of relationships between concepts in which each system classifies different relationship types. In general, relevant features are extracted from the text and are usually selected on the basis of the experimental results and intuition, or by statistical techniques [11]. First, by experience and intuition, we designed feature sets that were expected to have a strong correlation with the target classification. Forward selection was applied by sequentially adding each feature set to the model and evaluating its performance. The feature set is retained if a better result is achieved otherwise it is discarded before the next cycle is repeated.

ML systems have demonstrated high accuracy in information extraction and classification from radiology reports. It has also been used for automatic structuring of important medical information from radiology reports [15, 30]. Thomas et al. used Boolean logic built from 512 consecutive ankle radiography reports to create a text search algorithm and then applied to a different set of 750 radiology reports with a sensitivity of 87.8% and specificity of 91.3% [31]. The LEXIMER automated engine classified 1059 unstructured reports of radiography examinations based on the presence of important findings and suggestions for further actions with 94.9% and 97.7% sensitivity and specificity respectively [6]. In other research specialised on lung cancer reports, McCowan el at. used SVM learning techniques to investigate the classification of cancer stages [21, 20]. This system achieved an accuracy of 74% for tumour (T) staging and 87% for node (N) staging on the complete 179-case trial data set. In recent work published by Cheng et al., they first accessed whether the text contains sufficient information for a classification process then the tumour status and progression was determined by utilizing the SVMs models that reached 80.6% sensitivity and 91.6% specificity [4]. However, the sizes of the corpora used in previous research were relatively small compared to the number of reports processed by a registry each year.

3. SYSTEM ARCHITECTURE

Figure 1 presents the system architecture which comprises of three phases: pre-processing, iterative model development and active learning. The electronic medical records (EMRs) are retrieved from the database then passed into the pre-processing phase for proof-reading, lexical verification and medical concepts identification. At the end of this step, the records are cleaned and annotated with the medical concepts hence they are ready to be used to develop training data for ML models. At each development cycle, the model's performance is evaluated using n-fold cross-validation method then more data will be added to the train set with the support of active learning algorithms to query the most infor-

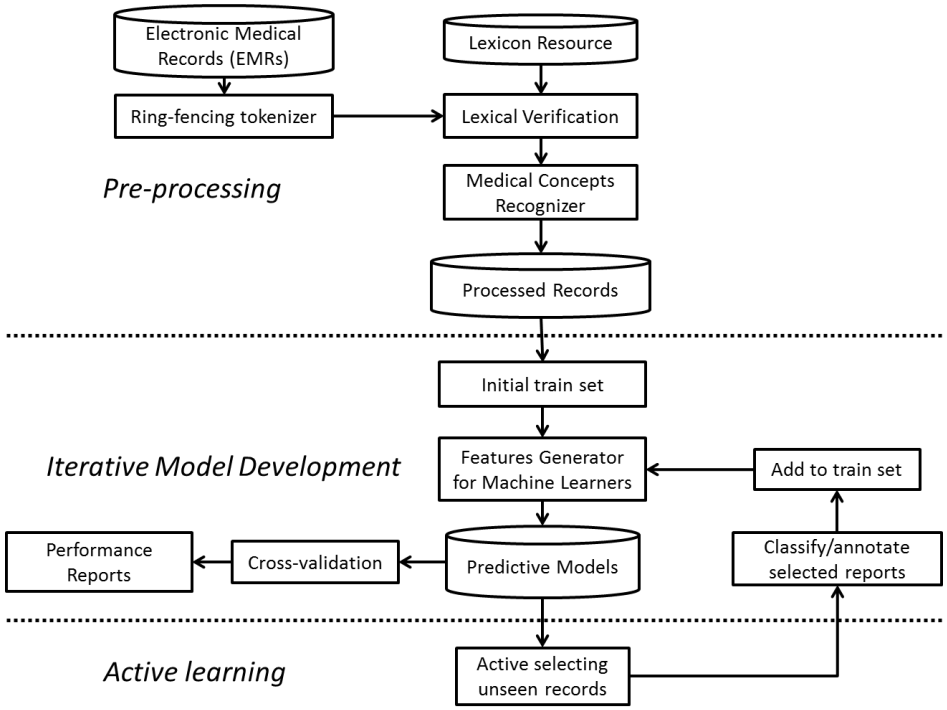

Figure 1: System architecture.

mative instances. This process is repeated until the desired performance is met.

3.1 Proof-reading (pre-processing system)

Proof-reading is a process whereby a clinical text is validated to identify unknown tokens/words and their valid forms. There are two principal tasks to be achieved, these are normalisation and standardisation. The normalisation process changes the texts in a way so that a human reader would consider it as normal, such as correcting spelling, expanding abbreviations and acronyms. The standardisation process converts the text into certain formats that an expert community has defined as standard; a good example is converting scores and measures into a standard layout.

3.1.1 Standardisation

The ring-fencing tokeniser is capable of capturing popular basic and complex patterns in clinical text. It is a cascaded Finite State Recognizer (FSR) which uses training examples to recognize token patterns constituting a score or measurement that requires standardisation [26]. There are a large number of different scores and measurements in clinical notes. Some other types of measurements and scores in the training patterns are illustrated in Table 1.

When using regular expressions (REs) to describe patterns as more rules are developed to capture missed items, the rules became so complicated that it makes them difficult to update as any change has the risk of losing previously recognized patterns or introducing new false positives. Another problem is that the rule updating task requires an exhaustive knowledge of REs and a considerable amount of time modifying the rules. Consequently, the automated learning process to capture patterns using REs is particularly difficult. On the other hand, a trainable FSR can be built directly from training examples of data with high accuracy and efficient computational time.

Type	Pattern
BP	BP 140/65(84)
ABG	ABG's: 0355 7.41/41/103/26/2
Lipids	Lipids 10% at 20mls/hr
Measurement	7mg/hr
SaO2	O2 sats 91%

Table 1: Examples of clinical measures and scores can be recognized by FSR.

To capture composite patterns in clinical text, the cascaded FSR is generated by several levels of generalisation. The basic patterns is recognised first by simple FSR and then passed to complex FSR to capture the final complicated patterns. The example of applying a cascaded FSR to capture complex patterns in the text is showed in Figure 2.

3.1.2 Normalisation

After standardisation, each token is passed through the lexical verification process and then inserted into the Lexicon Management System (LMS) which supports automated and manual resolution of unknown tokens. The LMS is a system developed to store the accumulated lexical knowledge and contains categorizations of spelling errors, abbreviations, acronyms and a variety of non-word tokens. It also has a web interface that supports rapid manual correction of unknown words with a high accuracy clinical spelling suggestor plus the addition of grammatical information and the categorization of such words into gazetteers [25]. The method of the clinical spelling suggestor is based on combining heuristic-based suggestion generation and ranking algo-

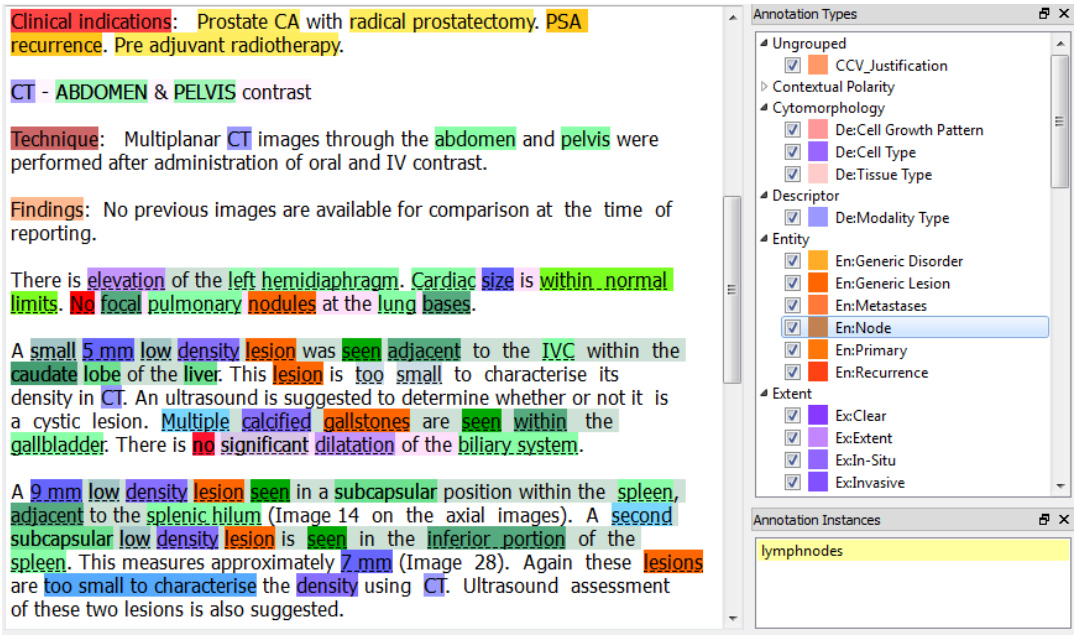

Figure 3: Example of annotated clinical report with in-house design of annotation schema for cancer radiology reports classification case study.

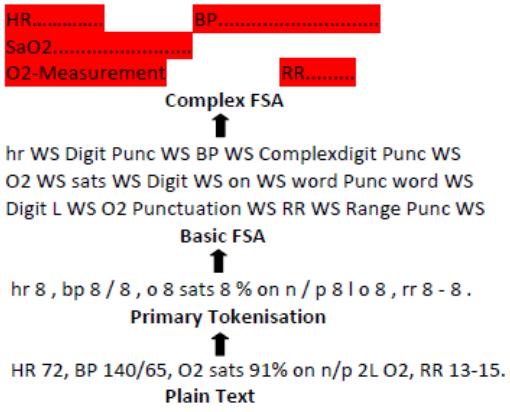

Figure 2: Example of cascaded FSR.

rithms based on word frequencies and trigram probabilities. The lexical verification process contains an additional step to resolve misplaced whitespace (e.g. 'looka fter' should be 'look after') and punctuation (e.g. 'natio;n' should be 'nation' or 'nation.The' should be 'nation. The').

3.1.3 Clinical Concepts Recognition

In EMRs, the same clinical concepts are usually expressed in different ways causing the models interpreted them as different features. Hence, there is a need for normalization of the medical concept using universal terminologies such as $UMLS^{®}$, $SNOMED-CT^{®}$ or their subsets. Besides general medical terminologies, the in-house design tag set was utilised in many studies because it is better controlled and more relevant to the predictive tasks, or to identify specific clinical entities in the text. Figure 3 illustrates the example annotation schema which was specially designed for classification of cancer radiology reports.

A detailed and well-designed tagging system can contribute significantly to the classification and extraction results. For example, the sentence 'There is no convincing metastatic bone lesion' in the conclusion will be tagged as:

There is $\underbrace{\text{no}}_{\text{LPN}}$ $\underbrace{\text{convincing}}_{\text{Modality}}$ $\underbrace{\text{metastatic}}_{\text{Metastases}}$ $\underbrace{\text{bone}}_{\text{Site}}$ $\underbrace{\text{lesion}}_{\text{Lesion}}$.

The occurrence of popular cancer terms (e.g. metastatic, lesion) in a sentence in the conclusion section is not enough to conclude that the cancer is reportable. The complete investigation has to consider whether the cancer term is negated or modalised on the basis of linguistic tags such as Lexical Polarity Negative (LPN) and Modality in the classification process.

3.2 Iterative Model Development

In the annotation process, free-text reports are annotated for examples of the information to be extracted and then algorithms are developed that use the examples to compute a more general model of the desired content. The small initial data set is selected to train the first model. The model is evaluated and the algorithm is revised in a feedback process to produce a more accurate result. This is continued over a series of experiments until an optimal model is identified.

In our model of iterative development, the annotators use a Visual Annotator (VA) tool as shown in Figure 3 which contains computational models of the tag set so as to support the manual annotation and classification processing. Hence, they no longer need to annotate each report de novo but employ knowledge from previously annotated instances and knowledge from all annotators as learnt by the computational model. This not only reduces the workload and annotation time per report but also reduces the error rate

and inconsistencies of human annotation which were generated by different levels of expertise.

After each cross-validation cycle, the new model is delivered to the VA to perform manual correction of the current gold-standard with the support of an annotation validation tool. This model is further supported by active learning algorithms to query the most informative instances to enrich the train set. The active learning selection process can query instances as a group in batch-mode which are suitable for parallel processing environment.

SVMs combined with CRFs are the main ML methods proposed for text classification and named-entity recognition for clinical textual data [14, 16]. For a large-scale classification problem with millions of instances and features such as in the radiology reports classification task, a linear kernel is usually a promising learning technique. Experiments were therefore performed with optimized linear kernel as the base classifier rather than SVMs with non-linear kernels [8]. The recommended sets for feature selection experiments include but not limited to:

- Bag of words (BOW): the feature value of the binary term weight is 0 or 1 corresponding to the existence of that feature in the text. The frequency term weight with normalised vector can also be applied.

- Proof reading: corrections and expansions, when used as features, will support the model in learning correct forms of misspelt words ('medicla' and 'medcial' refer to the same word 'medical') and variations of abbreviations ('amnt' and 'amt' are both 'amount'), and multiple acronyms of the same term ('ABG', 'ABGs' are both 'arterial blood gases').

- Ring-fencing: the basic patterns (date, time, number, etc) and standard patterns (blood pressure, heart rate, cancer stage, etc) are used as features to indicate whether a token belongs to any kind of scores or measures.

- Lemma, part of speech, chunk from the GENIA tagger: The GENIA tagger analyses English sentences and outputs the base forms, part-of-speech tags, chunk tags. The tagger is specifically tuned for biomedical text and is a useful pre-processing tool to extract information from biomedical documents [33].

- Medical terminology and gazetteer: checks whether a term belongs to a specialized clinical concept (UMLS, SNOMED-CT) or gazetteer (e.g. cancer terminology).

- Bag of tags (BOT): this feature indicates the clinical entities tagged by the computational annotation model. Entity is the generalisation of the frequent key words used to report a medical concept.

- Context feature: adds features to indicate whether a word belongs to a specific context (e.g. clinical indication, conclusion).

- Negation and modality feature: the occurrence of negation and modality tags help to identify whether a phrase is modalised or negated.

3.3 Active Learning

Active learning (AL) is a subfield of ML where the learner is allowed to query the most informative instances to retrain the model instead of making a random selection. Based on this approach, with the same number of sample selections, the performance of active learners dominates random learners in most cases [29]. This approach requires significantly fewer sample reports while maintaining comparable performance to traditional supervised learning with all training data or even bettering it.

In the present work, the main focus is on pool-based sampling which was introduced by Lewis and Catlett [18]. In this scenario, the learner has access to a pool of unlabelled instances and can request the labels for some number of them. Among many AL algorithms have been introduced in the literature, the four algorithms investigated and suggested for clinical text classification in this paper are Simple, Self-Confident (Self-Conf), Kernel Farthest-First (KFF), and Balanced Exploration and Exploitation (Balance-EE). These algorithms appear to be among the best performers based on empirical studies. Furthermore, they are reasonably well motivated and achieved high performance on real-world data sets [1, 24].

3.3.1 Simple AL

The Simple algorithm is based on the kernel machines and was independently proposed by three different research groups [28, 32, 2]. The name Simple (simple margin) used uncertainty estimation as its selection strategy [32]. In SVMs kernel space, the highest uncertain instance, which is defined as the most informative instance, is the one that lies closest to the decision hyperplane. For each unlabelled instance x, the shortest distance between the feature vector $\Phi(x)$ and the hyperplane w_i in the feature space is easily computed by $|w_i \cdot \Phi(x)|$. Hence, the querying function of Simple uses the current classifier to choose an unlabeled instance which is closest to the decision boundary.

3.3.2 Self Confident

The Self-Conf algorithm chooses the next example to be labeled so that, when it is added to the training data, the future generalization error probability is minimized [1]. Since true future error rates are unknown, the learner attempts to estimate them using a 'self-confidence' heuristic, which uses its current classifier for probability measurements. The future error rate is estimated by a log-loss function, which uses the entropy of the posterior class distribution on a sample of the unlabeled instances. Each instance from the unseen pool is examined by adding it to the training set with a sample of its possible labels and estimating the resulting future error rate as described in equation 1; the instance with the smallest expected log is then chosen.

Let $P(y|x)$ be an unknown conditional distribution over inputs x, and output $y \in \{y_1, y_2, ..., y_n\}$. At each query, a trained probabilistic (soft) classifier given by $\hat{P}(y|x)$ is already built from current training data. For each instance x from the unseen pool U, the algorithm trains a new classifier P' over $L'(x, y) = L \bigcup \{(x, y)\}$ and the expected log-loss is defined as:

$$E(\hat{P'}_{L'}(x, y)) = \frac{1}{|U|} \sum_{y' \in Y, x' \in U} \hat{P'}(y'|x') \log \hat{P'}(y'|x') \quad (1)$$

The original SELF-CONF employed Naive Bayes probability estimates. In the experiments described in this chapter, [1] implemented SELF-CONF using soft (confidence rated) SVMs. Probabilistic estimates are obtained in a standard way using the logistic regression transform. However, in each selection round, the expected log-loss is re-calculated for all instances in the unseen pool based on their possible labels and then the model is re-trained.

3.3.3 Kernel Farthest-First

The KFF algorithm uses a simple AL heuristic based on the 'farthest-first' traversal sequence in kernel space [13]. In this algorithm, the most informative instance is the farthest instance in the unseen pool from the current training set, where the distance from a point to a set is defined as the Euclidean distance to the closest point in the set. The assumption behind the KFF heuristic is that the farthest instance is considered to be the most dissimilar to the current training data and needs to be learned first.

Given a set L of labelled examples, KFF chooses the next farthest example x from L in the feature space induced by the SVM's kernel K to be labelled:

$$argmax_{x \in U}(min_{y \in L}||\Phi_K(x) - \Phi_K(y)||), \quad (2)$$

where

$$||\Phi_K(x) - \Phi_K(y)|| = \sqrt{K(x,x) + K(y,y) - 2K(x,y)} \quad (3)$$

is the Euclidean distance from $\Phi_K(x)$ to $\Phi_K(y)$, which are the projections of x and y in the feature space induced by the kernel K.

The advantage of KFF over Simple and Self-Conf algorithms is that it does not use the model to evaluate the unseen pool during the querying process. Hence, there is no need to retrain the model after each AL trial, and it can be applied to any learning algorithms.

3.3.4 Balanced Exploration and Exploitation

The Simple active learner is good at 'exploitation' by selecting the examples near the boundary, but it does not carry out 'exploration' by searching for large regions in the instance space that it might incorrectly predict. The Balance-EE method, which is based on a combination of Simple and KFF, to address the problem of balancing between exploitation of labeling instances that are near the current decision boundary (Simple) and exploration by searching for instances that are far from the already labeled points (KFF) [24]. At each trial, Balance-EE randomly decides whether exploration (Simple) or exploitation (KFF) will be used. If the choice is exploration (Simple), the algorithm evaluates the efficiency of the exploration to adjust its probability of exploring again.

Let h and h' be the hypothesis before and after the new example from Simple is added. The change induced from h to h' $d(h,h') \in [-1,+1]$ is evaluated. If $d(h,h')$ is positive, the exploration was efficient and p will be kept high and vice versa.

Let $S = x_1, x_2, ..., x_n = L \cup U$ be the set of labelled and unlabelled instances. For each of the hypotheses $h(\cdot), h'(\cdot)$, vectors of the predictions of h and h' on S are defined as $H = (h(x_1), h(x_2), ..., h(x_n))$ and $H' = (h'(x_1), h'(x_2), ..., h'(x_n))$. Then $d(h,h')$ is defined as:

$$d(h,h') = 3 - 4\frac{\langle H, H' \rangle}{||H||||H'||}, \quad (4)$$

The probability p for exploration will be updated as:

$$p' = \max(\min(p\lambda \exp(d(h,h')), 1-\epsilon), \epsilon), \quad (5)$$

where ϵ defines the upper and lower-bounds for the value of p, and λ is a learning rate for updating p.

3.3.5 Imbalanced Data Problem and AL

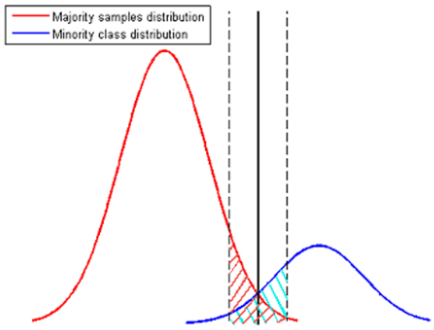

Figure 4: Data within the margin is less imbalanced than the entire data.

Ertekin et al. demonstrated that Simple method has the capability of overcoming the imbalanced data problem by providing the learner with much better balanced classes. Figure 4 presents an example of the imbalanced data distribution (image source [7]). Uncertainty AL tends to select the instances that are close to the decision boundary (within the margin), this selection strategy more likely ends up with a better balanced class distribution than that of the entire data set. This problem can also be improved by not querying the redundant samples from the dominated class or querying the positive and negative examples repeatedly from the labelled pool.

4. EXPERIMENTS AND RESULTS

Patrick and Nguyen (2011) conducted the evaluation for the role of pre-processing phase in improving the ML performance [25]. System performance was evaluated before and after an automatic proof reading process by comparing the computed SNOMED-CT codes to the coding created originally by the clinical staff. The automatic coding of the texts increased the coded content by 15% after the automatic correction process and the number of unique codes increased by approximately 5%.

The early design of our system without the integration of AL was initially ranked among the top three teams in the i2b2 2010 international challenge on clinical information extraction [25, 34]. Our follow-up challenge experiments with AL methods recorded the equivalent performance to the winning team with 15% less training data used [22, 5].

In this section, we focus on the evaluation of the general system which was fully developed for a project with Victorian Cancer Council in Australia[1]. This project has the

[1]http://www.cancervic.org.au/research/registry-statistics/capture-stage-recurrence

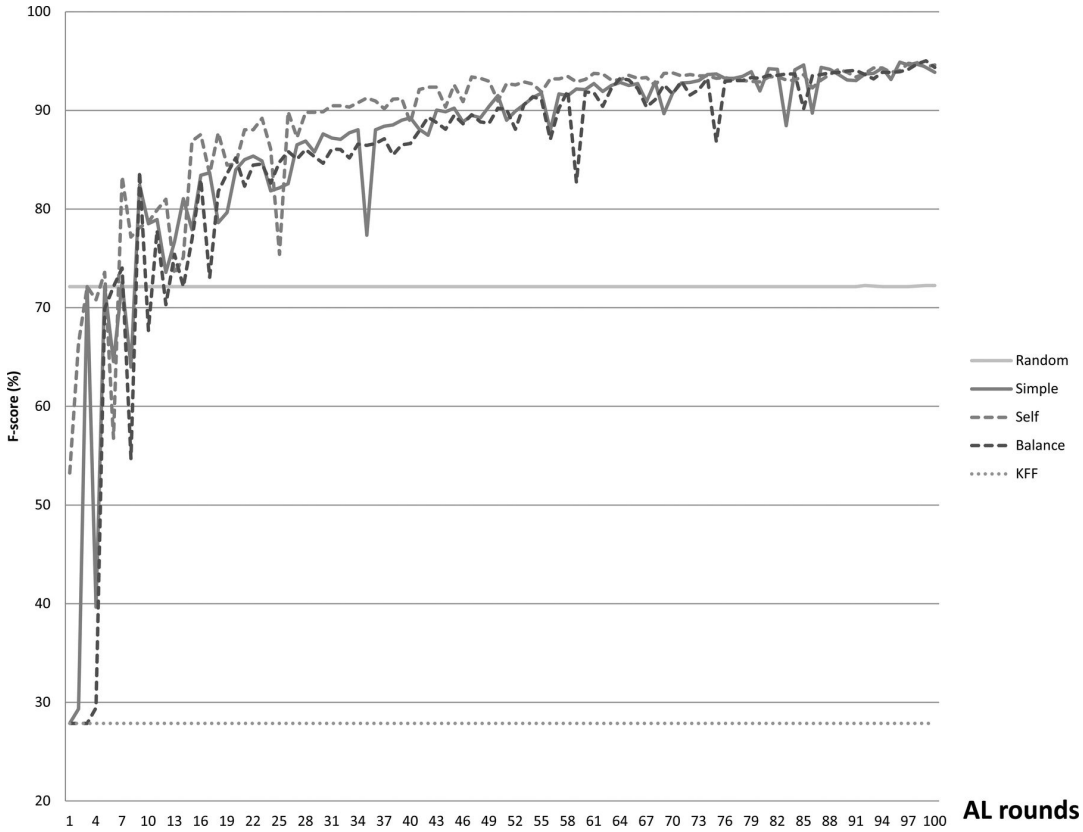

Figure 5: Evaluation of active and random sampling on test data.

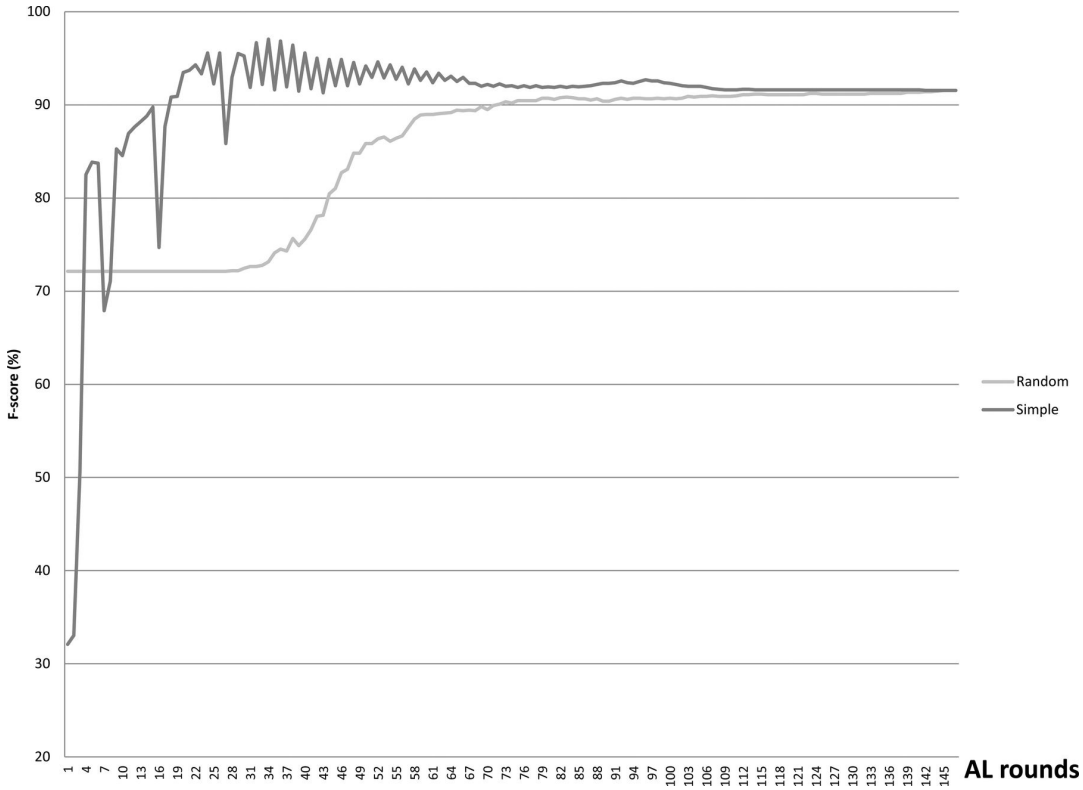

Figure 6: Full learning curves for Simple active learning and random learning with a batch size of 100.

Tag	TPs	FPs	FNs	P	R	F
De	230615	17980	16682	0.928	0.933	0.930
En	71340	6214	7640	0.920	0.903	0.912
Ra	1417	114	182	0.926	0.886	0.905
Li	141257	8464	8948	0.944	0.940	0.942
St	20370	204	391	0.990	0.981	0.986
All	464999	32976	33843	**0.934**	**0.932**	**0.933**

Table 2: Fivefold cross-validation tagging performance.

special requirement of separating the cancer and non-cancer reports by using a deliberate bias to ensure virtually all cancer reports are recognised. It has the consequence of producing as high an accuracy for sensitivity as possible with the challenge of maintaining reasonable specificity for the classifier. Ideally, the cancer registry does not want to miss any cancer cases, however they have accepted a sensitivity greater than 98% and specificity greater than 96%.

4.1 Data Description

The cancer cases covered in this study included all reports provided in a year's data collection by the imaging services in Australia. The pilot sites were Lake Imaging, Ballarat, Peter MacCallum Cancer Centre, Melbourne, and Westmead Hospital, Sydney in conjunction with the Cancer Institute of NSW. The process of creating the classifiers relied on using a manually trained corpus drawn from each site. Initially, a sample of 16 472 reports was drawn from Lake Imaging and assigned to cancer (4784 reports) or non-cancer (11 688 reports) classes by the cancer registry and then incrementally delivered to our system.

4.2 Named-Entity Recognition

To support the classification process, the training reports were tagged to identify structure and cancer-related information based on an in house design annotation schema. More than 3000 cancer reports were annotated with approximately 500 000 tag instances. The overall F-score for fivefold cross-validation of the Named-Entity Recognizer (NER) is 93%. The CRF++ tool was used in our NER experiments[2].

Our designed tag sets for cancer information extraction are well controlled and do not contain superfluous information, which can mislead the classification process. Table 2 presents the 5-fold cross validation performance of the computational tagger. In this table, the tags are divided into five subsets:

- Descriptor (De): morphology, topography, cytomorphology and modality type tags.
- Entity (En): objects of interest within a report. They are usually the subject of the report, which is cancer in this case.
- Linguistic (Li): includes lexical polarity, normality and modifier tags. Linguistic tags are not directly related to cancer content, but they are crucial for the confirmation of reportability.
- Radiologist's coding (Ra): includes cancer stage, TNM (tumor-nodal-metastases) values which are recorded directly in the text.
- Structure (St): includes heading tags. The structure tags are not directly related to the cancer content but support the use of context as features in the classification process. These headings are also used to structure the report body when populating the output in an XML format.

[2]http://crfpp.googlecode.com/svn/trunk/doc/index.html

4.3 Classification and Active Learning

Figure 5 shows the accuracy of the four AL algorithms and random sampling in classifying the radiology reports. The AL experiment was executed in batch mode with 10 reports/round for 100 rounds. Random sampling gave a consistent performance of 72.13% throughout the learning process; this accuracy is equal to the rate of non-reportable instances in the test set (1188/1647). Except for the last few cycles which can only capture one or two positive instances, the random classifier predicted every instance in the test set as non-reportable. This is because of imbalanced data distribution problem, e.g. the number of non-reportables in the training set is 2.6 times greater than the number of reportables, so they exceeded the reportable instances in the early cycles of random selection. The worst performance can be seen with the KFF algorithm, with 27.87% over time; this accuracy is equal to the rate of reportables in the test set (459/1647). Different from the random sampling, the KFF algorithm mostly selected the positive class (766 positives out of 1002 selected examples). Hence, the KFF classifier categorized all instances in the test set as reportable.

The three other AL algorithms show comparable results, with over 94.5% accuracy at the peak points. Balance-EE is a combination of Simple and KFF with a choice of algorithm in each trial. In this experiment, KFF was selected by Balance-EE for the first six trials for 60 examples, and then Simple was applied for the subsequent instances. As a result, except for several 'drop points', Balance-EE had a similar learning curve to Simple because most examples were selected using the 'exploitation' (Simple) strategy. The Self-Conf algorithm showed consistently higher accuracy than Simple and Balance-EE for the initial AL queries, and it quickly reached the top performance with only 60% queries used.

From these analyses, the Simple algorithm was chosen as the AL strategy for generating the priority list for manual reportability classification of radiology reports. As seen in Figure 5, the Simple method had comparable results to Self-Conf and Balanced-EE, but its implementation was simpler and more efficient. For 100 trials, Balanced-EE was slightly slower than Simple, while Self-Conf was five times slower than Simple.

The full learning curves for Simple active sampling and random sampling with a batch size of 100 reports per round for 145 rounds are presented in Figure 6. The batch size was increased from 10 to 100 to speed up the process in order to generate an overview of the comprehensive learning curves. However, the performance of the active learner with the same training size was reduced as compared to batch of 10 results because the model was updated 10 times less frequently. Figure 6 shows that the performance of the random sampler increased only when it had 3000 reports. At that point, the Simple active learner had already reached its top performance, which was over 23% higher than the random learner. There was not much difference in the performances of the two methods since 10 000 reports had been selected.

A known problem with many AL algorithms, especially in the early steps of learning, is that they are prone to generate a biased training set rather than be representative of the true underlying data distribution. As can be seen in Figure 6, there are a few points where the performance of the active learner dropped dramatically - for example, the performance fell below 80% when 1600 samples were used. This is due to limitations in the initial model that will perform AL. Many algorithms just randomly select a few instances to train the first model, which is usually not a good starting point for the real data distribution [22].

4.4 Held-out Test Set Result

The evaluation of the reportability classifier presented here was executed independently at the Cancer Registry. They used sensitivity and specificity as evaluation metrics, while precision, recall, and F1-score were calculated in our experiments. For the binary classification problem, 'sensitivity' is equal to 'recall' of the positive class (reportable), and 'specificity' is the 'recall' of the negative class (non-reportable). The Registry maintains the held-out test set to evaluate the system independently until the required sensitivity and specificity are achieved. None of the held-out test set was used for any part of the system development - for example, they were not used to build the gazetteers or the ML models. This held-out test set comprised 400 reportables and 2100 non-reportables, which is a similar distribution to the released data.

The approved version of the classifier achieved the sensitivity of 98.25% and specificity of 96.14% (Table 3). The final version is implemented based on two ML algorithms, they are CRFs and SVMs. In addition, the special cancer gazetteers collected by the linguists who have experience at interpreting the reports are used to support the ML models. The addition of gazetteer features slightly increased the sensitivity of the model (1.5%) while decreased the specificity level (2.2%). The significant improvement of nearly 5.3% in the targeted sensitivity score was experienced when the classifier was supported with the tagging features (BOT) generated by the computational annotation models. Furthermore, the specificity was still maintained at over 96% which was fulfilled the requirement of the project.

Sensitivity	Specificity	BOW	Gaz	BOT
91.46%	98.76%	Yes		
92.96%	96.53%	Yes	Yes	
98.25%	**96.14%**	**Yes**	**Yes**	**Yes**

Table 3: Reportability classifier's performance on evaluation (held-out) set for Lake Imaging

5. CONCLUSIONS

This paper presents a general system for text mining in clinical domain with a focus on dealing with multiple frequent kinds of noise. This system is then become part of an industrial-strength processing pipeline built to extract content from radiology reports for use in the Victorian Cancer Registry. The most important practical application of the reportability classifier is that it can dramatically reduce human effort in identifying relevant reports from the large imaging pool for further investigation of cancer. The classifier is built on a large real-world dataset and can achieve high performance in filtering relevant reports.

In future work, we will investigate the specialised parser to deal with the problem of poor grammatical sentences in the clinical corpus. In the patient notes, very few sentences could be successfully explored by full constituent parse tree due to the frequent ungrammatical notes written by the doctors. However, the partial trees are usually utilised when the complete parse could not be generated. Exploring the subtrees between two target concepts can help to better classify the relationship between them.

In the present system, the models stopped learning when it reached the pre-defined performance. However, it is more ideal if the model only finishes the querying process when the best performance is archived. We plan to improve the current system by introducing the stopping criteria during the AL process. Several stopping criteria have been introduced and are based on measures of stability or self-confidence within the learner [36, 23].

6. ACKNOWLEDGEMENTS

This work was supported by the Victorian Cancer Registry, Cancer Australia (2012-2013) and Data61 - CSIRO (2015). The authors would like to thank Helen Farrugia and Georgina Marr of the Cancer Council of Victoria, who provided the funding for this project and the registry expertise, and Dr Alex Pitman of Lake Imaging for contributing his radiological expertise.

7. REFERENCES

[1] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, 2004.

[2] C. Campbell, N. Cristianini, A. Smola, et al. Query learning with large margin classifiers. In *Machine Learning-International Workshop then Conference*, pages 111–118, 2000.

[3] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[4] L. Cheng, J. Zheng, G. Savova, and B. Erickson. Discerning tumor status from unstructured mri reports-completeness of information in existing reports and utility of automated natural language processing. *Journal of Digital Imaging*, 23:119–132, 2010.

[5] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562, 2011.

[6] K. Dreyer, M. Kalra, M. Maher, A. Hurier, B. a. Asfaw, T. Schultz, E. Halpern, and J. Thrall. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology*, 234(2):323–9, Mar. 2005.

[7] S. Ertekin, J. Huang, L. Bottou, and L. Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth*

ACM conference on Conference on information and knowledge management, pages 127–136. ACM, 2007.

[8] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[9] C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[10] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[11] B. Haddow. Using automated feature optimisation to create an adaptable relation extraction system. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 19–27, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[12] P. Haug, S. Koehler, L. Lau, P. Wang, R. Rocha, and S. Huff. Experience with a mixed semantic/syntactic parser. page 284, 1995.

[13] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

[14] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142, 1998.

[15] D. Johnson, R. Taira, A. Cardenas, and D. Aberle. Extracting information from free text radiology reports. *International Journal on Digital Libraries*, 1(3):297–308, Dec. 1997.

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[17] K. Lee, Y. Hwang, and H. Rim. Two-phase biomedical ne recognition based on svms. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13*, BioMed '03, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[18] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[19] D. Li, K. Kipper-Schuler, and G. Savova. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 94–95, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[20] I. McCowan and D. Moore. Collection of Cancer Stage Data by Classifying Free-text Medical Reports. *Journal of American Medical Informatics Association*, pages 736–745, 2007.

[21] I. McCowan, D. Moore, and M. Fry. Classification of cancer stage from free-text histology reports. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1:5153–6, Jan. 2006.

[22] D. Nguyen and J. Patrick. Reverse active learning for optimising information extraction training production. In *AI 2012: Advances in Artificial Intelligence*, pages 445–456. Springer, 2012.

[23] F. Olsson and K. Tomanek. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 138–146, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[24] T. Osugi, D. Kun, and S. Scott. Balancing exploration and exploitation: A new algorithm for active machine learning. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 330–337, 2005.

[25] J. Patrick and D. Nguyen. Automated proof reading of clinical notes. In *25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 303–312. aclweb, 2011.

[26] J. Patrick and M. Sabbagh. An active learning process for extraction and standardisation of medical measurements by a trainable fsa. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 151–162. Springer Berlin / Heidelberg, 2011.

[27] J. Patrick, M. Sabbagh, S. Jain, and H. Zheng. Spelling correction in clinical notes with emphasis on first suggestion accuracy. *2nd Workshop on Building and Evaluating Re-sources for Biomedical Text Mining*, pages 2–8, 2010.

[28] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Machine Learning-International Workshop then Conference*, pages 839–846. Citeseer, 2000.

[29] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[30] R. Taira. Automatic Structuring of Radiology Free-Text Reports. *Radiographics*, 98105:237–245, 2001.

[31] B. Thomas, H. Ouellette, E. Halpern, and D. Rosenthal. Automated computer-assisted categorization of radiology reports. *American Journal of Roentgenology*, 184(2):687–690, 2005.

[32] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.

[33] Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392, 2005.

[34] O. Uzuner, B. South, S. Shen, and S. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[35] O. Uzuner, X. Zhang, and T. Sibanda. Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association*, 16(1):109–115, 2009.

[36] A. Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.