Estimation of errors in text and data processing

Angela Slavova, Borislav Valkov, Krasimir Tonchev, Nina Daskalova, Margarita Nikolova, Mira Bivas, Plamen Mateev, Roumyana Yordanova, Stela Zhelezova

1. Problem description

The company Adiss Lab Lts. obtained 1 000 000 medical reports that are either in free form text, or in XML format. One of the main goals of their development is to integrate an algorithm for information extraction (IE) in their platform. Since the algorithm will work with medical data, its performance should be as reliable as possible. Due to confidentiality, we were not given access to the data or to the algorithm, but only to a simple example of a report's content. That is why we considered them mostly as a black box. The verification of the algorithm's output for a report is done by a medical doctor (MD) for a certain fee. Validating the correctness of all data would be overwhelming and very expensive. Hence, the problem, as presented by the company, is to provide a method (algorithm) which determines the minimum amount of reports that will validate the correctness of the IE algorithm and a procedure for selecting these reports.

2. Considered approaches

In order to solve the problem we have considered two types of approaches:

- *algorithm centric* – active learning and semi-supervised learning;
- *data centric* – sampling methods (e.g. bootstrapping) and power calculation.

Since we do not have the data, we have settled on an algorithm centric approach, specifically the active learning.

3. Active learning

3.1. General algorithm

The primary goal of active learning (AL) [1] is exploiting unlabelled data. Nowadays such data is *plentiful and cheap* – e.g documents off the web, speech samples, images and video. However labelling it is *expensive*, just as in our case. Here is the active learning typical algorithm:

1. Start with a pool of unlabeled data
2. Pick a few points at random and get their labels
3. Repeat:
   - Train a classifier using the labels seen so far
   - Label randomly selected unlabeled points that are closest to the decision boundary (or most uncertain)

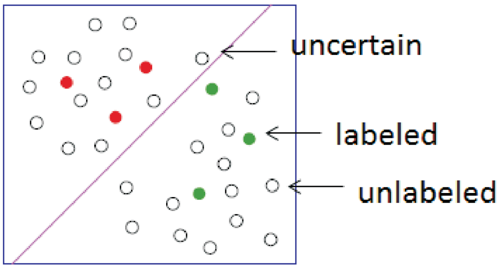


Figure 1: General AL algorithm

As it is seen in the algorithm, the active learning requires a *classifier with uncertainty output*. Another important specific is that *sampling is biased* i.e. the labelled samples are not representative for the underline distribution.

3.2. With clustering

An improvement of the active learning algorithm in our case would be to exploit the natural structure of the data by using *clustering*.
In general, clustering is incorporated in active learning [2] with the following steps:

1. Find a clustering of the data
2. Sample a few randomly-chosen points in each cluster and label them
3. If the granularity is right, assign each cluster its majority label; if not refine the clustering

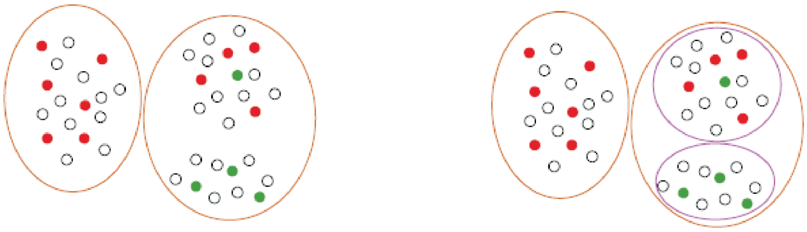


Figure 2: Finding the right granularity of the clustering

4. Proposed solution of the problem

Our proposition is *active learning with clustering* to be used in order to select the minimal amount of reports required to validate the correctness of the information extraction. We propose an algorithm for doing this. Additionally we suggest the rules for clustering, measure of uncertainty for algorithm decision and measure of accuracy of the extraction.

4.1. Algorithm

The algorithm consists of the following steps:

1. Select a classifier suitable for active learning

2. Select a sampling scheme

3. Fix K – the number of samples labelled at each iteration

4. Select the stopping threshold $t > 0$

5. Extract the meta-data from each document (city, hospital, etc.)

6. Cluster the reports according to the meta-data

7. Select samples from each cluster using the sampling scheme, so that their total count equals to K

8. Label the selected samples and store them in a buffer

9. Do

    (a) Train the classifier with the data in the buffer

    (b) Apply the classifier to the remaining unlabelled data

    (c) Measure the accuracy of the current iteration (A_i) using all labelled samples

(d) Select the uncertain samples (close to the decision threshold)

(e) Select samples from the uncertain ones in each cluster using the sampling scheme, so that their total count equals to K

(f) Label the selected samples and store them in the buffer

While $|A_i - A_{i-1}| > t$

4.2. Algorithm details

Let us elaborate on the parts of the algorithm specific for the problem.

Clustering procedure. The clustering is based on the assumption that the MDs in same areas and hospitals have the same convention for writing reports. This leads to similarity in notations between the reports from the same area/ hospital/MD. We should note that this clustering may not be hierarchical.

Measure of uncertainty. Since we are dealing with medical data, we assume that the dictionary is fixed and known and can be reduced depending on the medical tests. Therefore, we treat the different phrases as elements of a vector. For each report the information extraction algorithm outputs for each phrase from the dictionary whether it is fully recognised, or its value is uncertain, or it is not tested.

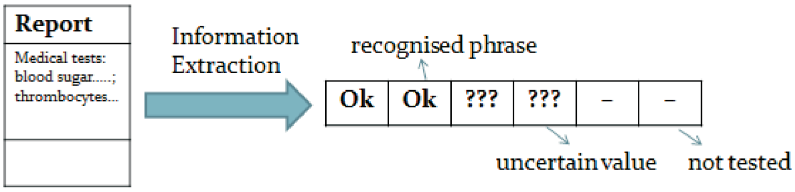


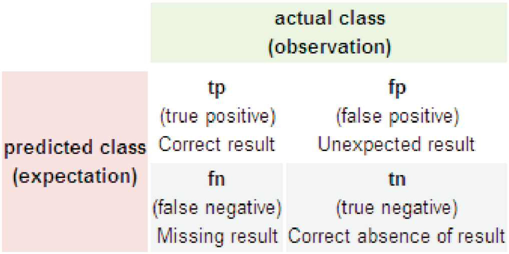
Figure 3: Output of the information extraction

We suggest the quotient of the count of the uncertain phrases - k divided by the count of all phrases - n to be used as a measure of uncertainty:

$$u = \frac{k}{n}, \quad u \in [0, 1]$$

Measure of accuracy. We suggest a few ways for measuring the accuracy of the algorithm.

The first one is the wide-known measures *precision* and *recall* [3] to be used. In order to define them, we introduce some terms. The terms true positives, true negatives, false positives, and false negatives compare the results of the information extraction with the MD judgement. The terms positive and negative refer to the information extraction prediction, and the terms true and false refer to whether that prediction corresponds to the MD judgment. This is illustrated by the table below:



	actual class (observation)	
predicted class (expectation)	tp (true positive) Correct result	fp (false positive) Unexpected result
	fn (false negative) Missing result	tn (true negative) Correct absence of result

Precision and recall are then defined as:

$$Precision = \frac{tp}{tp + fp}$$
$$Recall = \frac{tp}{tp + fn}$$

Another suitable for the problem measure is one that combines precision and recall. This can be the harmonic mean of precision and recall, the traditional *F-measure* or balanced F-score [4]:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The *Matthews correlation coefficient*(MCC) [4] can also be used for measuring the accuracy in our case. It is a measure of the quality of binary (two-class) classifications and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classification:

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

5. Suggestions for implementing our solution

Our solution can be easily implemented using open-source tools for active learning. We suggest the DUALIST (`https://code.google.com/p/dualist/`) interactive machine learning system to be used. Its main advantages are that it

quickly building classifiers for text processing tasks and also has web-based user interface in Java.
For further information on active learning, we recommend the *Active Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning* book by Burr Settles (`http://active-learning.net/`).

References

[1] Dasgupta S. and Langford J., *A tutorial on active learning*, ICML 2009.

[2] Dasgupta S. and Hsu D., *Hierarchical sampling for active learning*, ICML 2008.

[3] Van Rijsbergen C. J., *Information Retrieval* (2nd ed.), Butterworth, 1979.

[4] Baldi P., Brunak S., Chauvin Y., Andersen C. A. F., Nielsen H., *Assessing the accuracy of prediction algorithms for classification: an overview*, Bioinformatics 2000, 16, p. 412-424