Estimation of errors in text and data processing

European Study Group with Industry'95 (ESGI'95)
September 23 – 27, 2013, Sofia, Bulgaria

Workgroup 2:
Angela Slavova, Borislav Valkov, Krasimir Tonchev, Nina Daskalova, Margarita Nikolova, Mira Bivas, Plamen Mateev, Roumyana Yordanova, Stela Zhelezova

Problem description

- Given data
  - 1,000,000 medical reports – free form text or in XML format
  - algorithm for information extraction (IE) – black box
  - the algorithm correctness can be verified by a paid MD expertise
- Goal
  - determine the minimum amount of reports required to validate the correctness of the algorithm
  - provide a method for selecting these reports

Considered approaches

- Algorithm centric
  - active learning
  - semi-supervised learning
- Data centric
  - sampling method – e.g. bootstrapping
  - power calculation

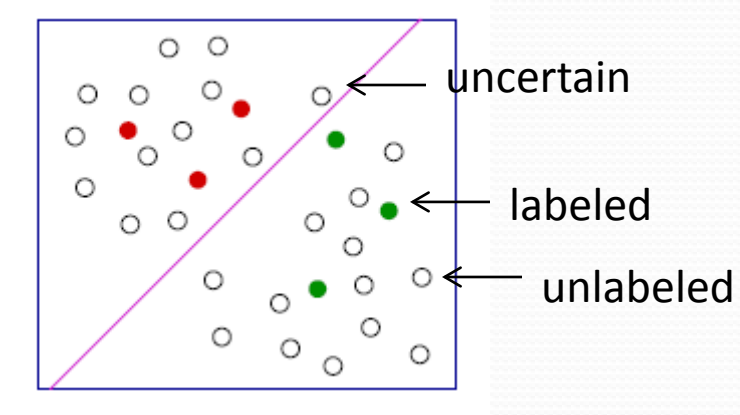
Active Learning (AL)

- Primary goal of AL – exploiting unlabelled data
- The unlabelled data is plentiful
  - documents off the web
  - speech samples
  - images and video
- Main characteristic of the data – **cheap**
- Main characteristic of labelling – **expensive**

- Matches our case exactly ☺

Active Learning (2)

- AL 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)



Active Learning (3)

- Requires classifier with uncertainty output
- Sampling is biased i.e. the labelled samples are not representative for the underline distribution

- We will exploit structure in data by using natural clustering

Active Learning (4)

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

Our proposition

- Use AL in order to select minimal amount of reports required to validate the correctness of the IE
- Additionally we suggest rules for:
  - clustering
  - measure of uncertainty for algorithm decision
  - measure of accuracy

Algorithm Initialisation

1. Select classifier suitable for AL
2. Select sampling scheme (SS)
3. Fix K – number of samples labelled at each iteration
4. Select 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 SS such that their total number equals to K
8. Label the selected samples and store them in a buffer(B)

Algorithm Iteration

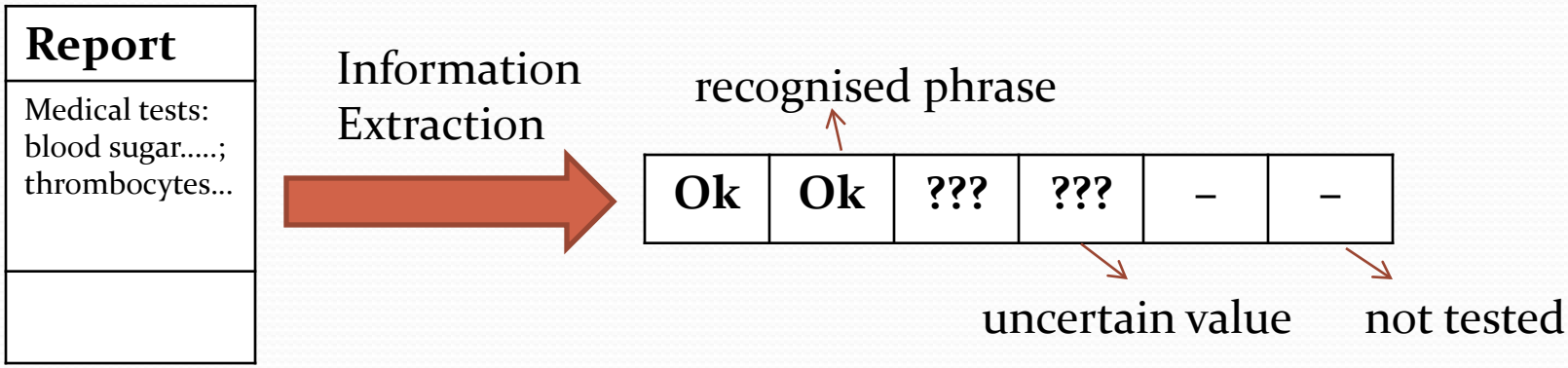
1. Do
   i. Train the classifier with data in B
   ii. Apply classifier to the remaining unlabelled data
   iii. Measure the accuracy (A_i) using all labelled samples
   iv. Select the uncertain samples(close to the decision threshold) (US)
   v. Select samples from US in each cluster using SS such that their total number equals to K
   vi. Label the selected samples and store them in a buffer(B)
2. While $|A_i - A_{i-1}| > t$

Algorithm details

- Clustering procedure
  - Assumption: 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
  - Important note: *this clustering may not be hierarchical*

Algorithm details (2)

- Measure of uncertainty
  - Assumption: since we are dealing with medical data, 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



$$u = \frac{k}{n}$$

k: count of uncertain phrases

n: count of all found phrases

$$u \in [0, 1]$$

Algorithm details (3)

- Measure of accuracy
  - precision and recall

	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

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

- Matthews correlation coefficient - the geometric mean of the regression coefficient of the problem and of its dual

$$\text{N} = TN + TP + FN + FP$$

$$\text{S} = \frac{TP + FN}{N}$$

$$\text{P} = \frac{TP + FP}{N}$$

$$\text{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$

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, 412–424*

Thank you!