Estimation of errors in text and data processing

Adiss Lab Lts.

Zhivko Angelov

Problem description

Much important information today is stored in free text format. This includes patient-related records which contain essential clinical facts related to the patient status and treatment. The recent achievements in automatic text analysis enable extraction of specific entities and/or events with high accuracy. In this way it becomes possible to plan large-scale initiatives in recognition of particular entities and storing them in data bases for further processing by data analytics software. This process is illustrated in Fig. 1: separate text descriptions are analysed, in order to structure their content, and the resulting data base is subject of further mining for statistical purposes.

Typically, the success of Information Extraction (IE, a kind of partial automatic text analysis) is measured by:

- ***Precision***: the number of correctly extracted entity descriptions, divided by the number of all recognised entity descriptions in the test set;
- ***Recall***: the number of correctly extracted entity descriptions, divided by the number of all available entity descriptions in the test set (some of them may remain unrecognised by the particular IE module).

Thus the Precision measures the success and the Recall – the

recognition ability and "sensiti-vity or coverage" of the algorithms. The F-score (harmonic mean of Precision and Recall) combines the two measures and is defined as $F = 2 \times Precision \times Recall / (Precision + Recall)$.

It order to support the IE tasks, various types of language resources are incrementally developed since decades. This includes corpora used as training or test data sets as well as “gold standard” annotated corpora that enable comparisons of different software systems.

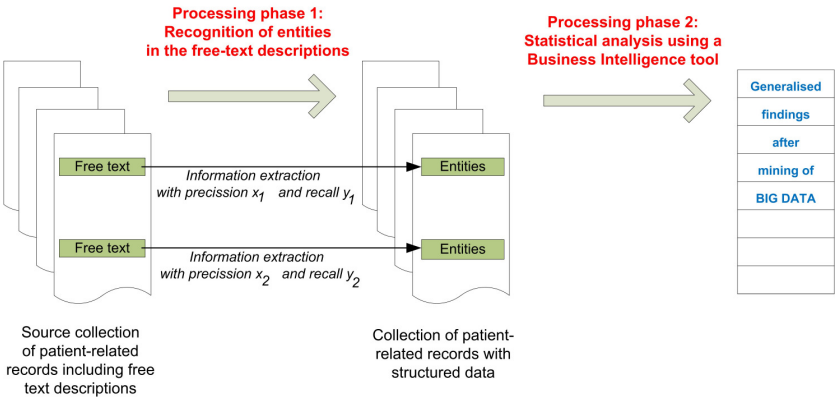

Figure 1. Phases in text analytics: extraction of specific entities and further data mining

QUESTIONS:

1. Million Patient records exist, typed in by thousands medical experts. In this way automatic assessment of the IE correctness by mapping the extracted entities to certain “gold standard” is impossible. Let

- the source collection of documents contains *N* records,
- one kind of focal event is DRUG THERAPY where an IE module extracts from the text drug codes, dosage, frequency and admission route, for 1500 drugs,
- another kind of extracted events are LAB TESTS where another IE module extracts from the text the Lab test type and the particular value, for 800 kinds of lab tests.

How many records from the collection with the structured data have to be inspected manually, in order to assess the precisions x_1 and x_2 and the recall y_1 and y_2?

2. Could you please suggest some measure Z for the "overall correctness" of the collection with the structured data, which integrates the results of several (at least two) independent IE modules?

3. What should be the minimal correctness Z for a collection with N records, to claim, that the data mining in phase 2 delivers credible observations?