Chapter 6
Evaluation Metrics and Evaluation

The area of evaluation of information retrieval and natural language processing systems is complex. It will only be touched on in this chapter. First the scientific base for evaluation of all information retrieval systems, called the Cranfield paradigm will be described. Then different evaluation concepts such as precision, recall, F-score, development, training and evaluation sets and k-fold cross validation will be described. Statistical significance testing will be presented. This chapter will also discuss manual annotation and inter-annotator agreement, annotation tools such as BRAT and the gold standard. An example of a shared task on retrieving information from electronic patient records will be presented.

6.1 Qualitative and Quantitative Evaluation

There are two types of evaluation, *qualitative evaluation* and *quantitative evaluation*. In this book quantitative evaluation is mostly used and described. Qualitative evaluation means asking a user or user groups whether the result from an information retrieval system gives a satisfying answer or not. Qualitative evaluation focuses mostly on one or more users' experiences of a system. Quantitative evaluation means having a mechanical way to quantify the results, in numbers, from an information retrieval system. The Cranfield paradigm will be described, which was the first attempt to make a quantitative evaluation of an information retrieval system.

6.2 The Cranfield Paradigm

The evaluation methods used here are mainly quantitative and are based on the Cranfield tests that also called the *Cranfield Evaluation paradigm or the Cranfield paradigm*, carried out by Cleverdon (1967).

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_6

Cyril Cleverdon was a librarian at the College of Aeronautics in Cranfield (later the Cranfield Institute of Technology and Cranfield University), UK. Cleverdon conducted a series of controlled experiments in document retrieval. He used a search device, a set of queries, and test collections of documents and the correct answers. The correct answers consisted of documents answering the questions. These documents are also called relevant documents. The search device had indexed the document collections automatically before the search experiment started.

By using a controlled document collection, a set of controlled queries and knowing the relevant documents, he could run the experiments over and over again after changing different parameters and observe the outcome, measuring precision and recall. Cleverdon proved that single terms taken from the document collection achieved the best retrieval performance in contrast with what he had been trained to do and taught as librarian, which was to perform manual indexing using synonym words from a controlled list.

Voorhees (2001) elaborates on the Cranfield paradigm and argues that this is the only way to evaluate information retrieval systems, since manual objective evaluation is too costly and may be also too imprecise. The Cranfield paradigm has given rise to the Text REtrieval Conference (TREC) and the Cross-Language Evaluation Forum (CLEF), where large controlled document collections together with queries in specific topics are used for the evaluation of information retrieval.

6.3 Metrics

Evaluation, in this case quantitative evaluation, can have many different purposes. There may also be different limitations on the amount data used for training and for evaluation. In some cases, high recall is considered a priority over high precision, and in some cases it is the opposite.

When developing a system, a *development set* is used. A development set is either data for developing rules for an artefact or training material for a machine learning system. In machine learning the development set is often called the *training set* and is used for training the machine learning system. Part of the training set can be put aside for error analysis for the algorithm, and the machine learning algorithm can be adjusted according to the errors, so-called parameter tuning. This part of the training set is called development test set.

A test set is put aside to test the artefact, this test set is neither used for development nor for training and is sometimes called *held out data*, (Pustejovsky and Stubbs 2012).

If data is scarce a method called *k-fold cross validation* is used, this is carried out by dividing the whole dataset into k folds and the k-1 folds are used for training and the remaining one, the 1 fold for evaluation: the folds are switched until all folds are trained and tested on the remaining of the k-1 folds and finally an average is calculated. Usually 10-fold cross validation is used, (Kohavi 1995).

Table 6.1 Confusion matrix: predicted annotation is what the algorithm retrieves or annotates and gold annotation is what was marked up or annotated by a human

		Predicted annotation	
		Positive	Negative
Gold annotation	Positive	True positive (tp)	False negative (fn)
	Negative	False positive (fp)	True negative (tn)

Some false positives that are detected by the algorithm may be correct but wrongly classified by the human annotator (Pustejovsky and Stubbs 2012)

Two metrics used for measuring the performance of a retrieval system are *precision* and *recall*. *Precision* measures the number of correct instances retrieved divided by all retrieved instances, see Formula 6.1. *Recall* measures the number of correct instances retrieved divided by all correct instances, see Formula 6.2. Instances can be entities in a text, or a whole document in a document collection (corpus), that were retrieved. A confusion matrix, see Table 6.1 is often used for explaining the different entities.

Here follow the definitions of precision and recall, see Formulas 6.1 and 6.2 respectively.

$$Precision : P = \frac{tp}{tp + fp} \tag{6.1}$$

$$Recall : R = \frac{tp}{tp + fn} \tag{6.2}$$

The *F-score* is defined as the weighted average of both precision and recall depending on the weight function β, see Formula 6.3. The F_1-*score* means the harmonic mean between precision and recall, see Formula 6.4, when it is written *F-score* it usually means F_1-*score*. The F-score is also called the F-measure. The F_1-*score* can have different indices giving different weights to precision and recall.

$$F\text{-}score : F_\beta = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \tag{6.3}$$

With $\beta = 1$ the standard F-score is obtained, see Formula 6.4.

$$F\text{-}score : F_1 = F = 2 * \frac{P * R}{P + R} \tag{6.4}$$

Precision uses all retrieved documents for the calculation. If there are a large number of documents, there is a possibility to make the calculation simpler by using precision at a cut-off value, for example precision at top 5 or precision at top 10 written as P@5 or P@10 respectively. This measure is called precision at n, or with a general term precision at P@n.

For details on evaluation measurements see Van Rijsbergen (1979), Pustejovsky and Stubbs (2012) and Japkowicz and Shah (2011).

Two evaluation concepts used in medicine and health informatics are *specificity* and *sensitivity*.[1]

- Specificity measures the proportion of negatives that are correctly identified as negative, or not having the condition.
- Sensitivity (the same as recall) measures the proportion of negatives that are correctly identified (e.g., the percentage of healthy people who are correctly identified as not having the condition).

Yet another commonly used metric in medical and clinical systems is *positive predictive value (PPV)* corresponding to precision.

Accuracy is another measurement defined as the proportion of true instances retrieved, both positive and negative, among all instances retrieved. Accuracy is a weighted arithmetic mean of precision and inverse precision. Accuracy can also be high but precision low, meaning the system performs well but the results produced are slightly spread, compare this with hitting the bulls eye meaning both high accuracy and high precision, see Formula 6.5.

$$Accuracy : A = \frac{tp + tn}{tp + tn + fp + fn} \tag{6.5}$$

A *baseline* is usually a value for what a basic system would perform. The baseline system can be a system working in a random way or be a naïve system. The baseline can also be very smart and strong, but the importance of the baseline is to have something to compare with.

If there are other systems using the same data, then it is easy to compare results with these systems and a baseline is not so important.

6.4 Annotation

Annotation in this book means to manually add an annotation to a token or to a set of tokens in text. The reason for this is either to create training material for a machine learning system that uses supervised methods to train, or to create test material to evaluate both machine learning-based tools and rule-based tools. Annotation can also mean to manually classify documents according to predefined classes. *Labelling* is a synonym for annotation.

When performing an annotation task, a decision has to be made on what classes should be annotated and also how to annotate them. Should just the lexical item be

[1]Sensitivity and specificity, https://en.wikipedia.org/wiki/Sensitivity_and_specificity. Accessed 2018-01-11.

annotated or in the case of a personal name, both the first and surname? What about the initial or the title of the person? Likewise when annotating a finding, should the negation modifying the finding be annotated or just the finding? Therefore, guidelines for the annotation task are developed. Usually a couple of test runs of annotations are carried out to test the annotation classes, the settings and the guidelines. The annotators are then allowed to meet and discuss how they reasoned while annotating. After the discussion it is decided how to continue, usually by the chief annotator. New annotation classes are added or redefined and the guidelines are updated (Pustejovsky and Stubbs 2012).

6.5 Inter-Annotator Agreement (IAA)

When annotating data, preferably more than one annotator is used. For finding out the agreement between annotators and the difficulty of the annotation task *inter-annotator agreement (IAA)* is calculated. Inter-annotator agreement is sometimes also called inter-rater agreement. This is usually carried out by calculating the Precision, Recall, F-score and Cohen's kappa, between two annotators. If the IAA is very low, for example an F-score under 0.6, it is considered that the annotation task is difficult, but a low F-score can also be due to the annotators have not been instructed properly on what annotations they should do and the range and format of the annotation or they did not obtain any guidelines. Cohen's kappa measures whether if two annotators might annotate similarly due to chance and not because they agree. Another similar measurement is intra-annotator agreement where the same annotator does his or her task twice on the same text with some time interval to observe the difference between the two sets of annotations, (Pustejovsky and Stubbs 2012).

A gold standard is a manually created set of correct answers, annotations, which is used for evaluation of information retrieval tools, see Sect. 6.8 details.

When building a gold standard usually at least two annotators should agree on the gold standard. Theoretically Cohen's kappa can be used for this task to reveal if the agreement is by chance or not, but for using Cohen's kappa there should not be any variation of number of annotated tokens for each annotator, therefore, F-score is a better measurement between two annotators; however, a high F-score meaning high agreement indicates the task was easy for the annotators.

For document classification there is no variation for length and therefore Cohen's kappa is suitable for statistical significance testing for this task (Artstein and Poesio 2008; Hripcsak and Rothschild 2005; Japkowicz and Shah 2011).

6.6 Confidence and Statistical Significance Testing

The metrics precision and recall can be used to compare the performance of two different algorithms (or classifiers) on a dataset; however, these measurements do not measure if the results found occurred by chance. To confirm that the results are statistically significant *statistical significance testing* has to be carried out.

In Japkowicz and Shah (2011) there is a good overview of statistical significance testing. There any many different statistical methods available to find out if the different results are confident or significant. In this section three methods are going to be described: *McNemar's test*, the *sign test* and *Cohen's kappa* κ.

McNemar's test is commonly used in the case of comparing paired nominal data and specifically the erroneous data to observe if they occur by chance or are errors with some regularity. Produced errors should occur in a regular way, then we know that the results of the classifiers really are different and not by chance, and a conclusion can be drawn on which classifier performs best on this data set. For this to occur the null hypothesis should not be rejected.

McNemar's test is based on the chi-squared (χ^2) test, but while the standard χ^2 test is a test of independence of variables, McNemar's test is for consistency across two variables in a 2×2 contingency table (or misclassification matrix), see Table 6.2.

McNemar's test should only be used if the number of differently classified entities or errors is larger than 20, otherwise the sign test should be used. If there are less than 20 entities the null hypothesis is easily rejected.

To perform the McNemar's Chi-square test with continuity correction use Formula (6.6):

$$\chi^2 = \frac{(|b-c|-1)^2}{(b+c)} \tag{6.6}$$

where b and c can be found in the contingency table, see Table 6.2.

χ^2 should have a confidence level (P- or probability-value at the χ^2 distribution to not reject the null hypothesis commonly selected P-value is 0.95 with a significance level of 0.05.

The *sign test* is one of the simplest statistical tests. A sign test is a type of binomial test. It can be used to compare two algorithms on multiple domains or multiple data sets. The sign test calculates the number of times one algorithm

Table 6.2 Contingency table describing the output of two algorithms using the same test data giving errors b and c in two cases (Japkowicz and Shah 2011)

		Algorithm 1	
		0	1
Algorithm 2	0	a	b
	1	c	d

0 stands for misclassified and *1* for correctly classified

outperforms the other algorithm on different data sets. The more times one algorithm outperforms the other algorithm, the more confident the results are.

See Japkowicz and Shah (2011) for continued reading on statistical significance testing.

If considering big data, statistical significance testing is not really useful since there is so much data that the results always will be significant!

6.7 Annotation Tools

An *Annotation tool* is a tool for manually annotating of words or text pieces with grammatical categories, named entities, clinical entities etc. There are many different tools that can be used for the annotation work, for a nice review article on annotation tools see Neves and Leser (2012). One commonly used annotation tool is BRAT[2] (Stenetorp et al. 2012).

BRAT runs as a web server on the local computer, and a web browser is used as interface. It is possible to define annotation classes, attributes and relations, and use them for the annotation task. In the specific task shown in Fig. 6.1 a pre-annotation model was trained on annotated records from a medical emergency unit, and was used for the pre-annotation of the records that contain ADE. Observe the complicated *relation arches* for indications, adverse drug reactions and also for the *negated findings* (crossed at), for example "ingen" *andningspåverkan* ("no" respiratory distress). See Sect. 8.2.3 for an explanation of pre-annotation.

6.8 Gold Standard

A *gold standard* is a concept used in information retrieval and natural language processing. A gold standard is a set of correct annotations or correct answers to a query or the correct classification of documents.

A gold standard is created by letting the annotators who have carried out the annotations on the same set decide what the correct annotation is, and agree on this and hence produce a gold standard. There is always a plethora of annotation cases that need to be resolved between annotators, since some annotations are either missing or overlapping. If the annotators cannot agree they may end up with a "silver" standard. Often the chief annotator solve any arguments in the annotation process by making a final decision on the correct answer. There is also a possibility that one of the "best" annotation sets is selected and used as a gold standard.

A gold standard is usually used for evaluation of various developed systems. A gold standard can also be used for training and evaluation of the different systems.

[2]BRAT Rapid Annotation Tool http://brat.nlplab.org/index.html. Accessed 2018-01-11.

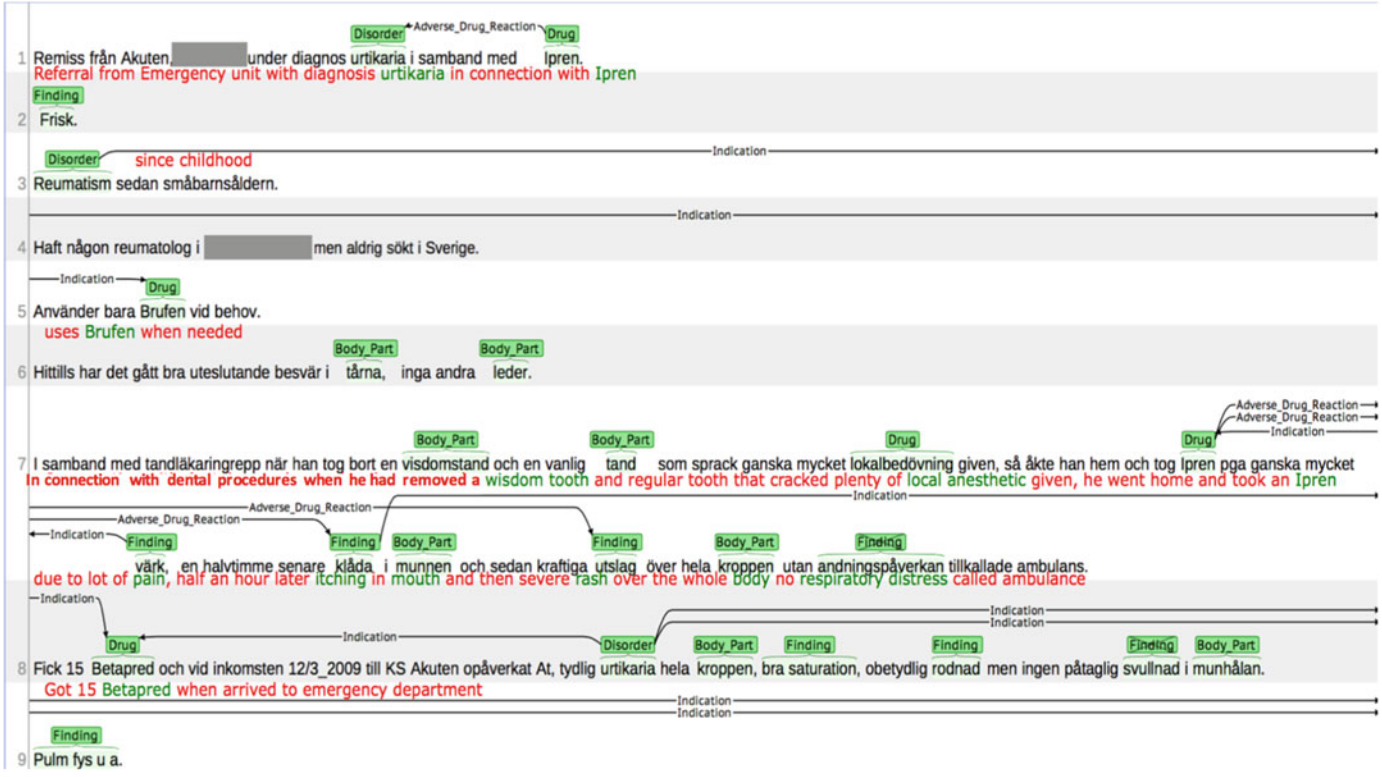

Fig. 6.1 Example of annotation using BRAT on a Swedish (anonymised) clinical text containing an adverse drug event. The example has been manually translated to English

If data is scarce and there is no possibility to divide the gold standard into one development set and a evaluation set, k-fold cross validation can be used instead (Pustejovsky and Stubbs 2012), see Sect. 6.3 on details about k-fold cross validation.

6.9 Summary of Evaluation Metrics and Annotation

This chapter presented evaluation metrics based on the Cranfield paradigm for information retrieval in document collections. Metrics such as precision, recall and F-score were introduced. The concepts development set, and evaluation set and k-fold cross validation were introduced. Manual annotation and inter-annotator agreement were described. The need for statistical significance testing was outlined and various test were presented. Different methods for manual annotations were discussed as well as tools for annotation of clinical text. The concept of the gold standard was introduced.