

Using Semantic Relatedness to improve the Evaluation of Multi-label Classifiers

Master's Thesis

Christophe Deloo

Using Semantic Relatedness to improve the Evaluation of Multi-label Classifiers

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Christophe Deloo
born in Leiderdorp



Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
<http://www.wis.ewi.tudelft.nl/>



GridLine B.V.
Keizersgracht 520
Amsterdam, the Netherlands
<http://www.gridline.nl/>

Using Semantic Relatedness to improve the Evaluation of Multi-label Classifiers

Author: Christophe Deloo
Student id: 1369350
Email: c.p.p.deloo@gmail.com

Abstract

Different evaluation techniques can be used in order to assess the quality of a multi-label classification task. The most commonly used performance measures in classification are the classic information retrieval notions of precision and recall. These measures assume independence between the target labels meaning that the assignment of a label to an item can either be evaluated as correct or incorrect and not somewhere in between. The drawback of these binary evaluation measures is that they do not take into account the difference in semantics between the concepts depicted by the labels. For example, an item classified as plant instead of flower would be considered as wrong although the concepts are semantically related and a flower is a specialisation of a plant. A way of improving the evaluation of a classification would be to also consider the degree of semantic relatedness between concepts.

The work in this thesis focuses on improving the evaluation of a real-world multi-label classification task by using the notion of semantic relatedness. The labels of this classification task represent concepts originating from a hierarchically structured thesaurus. Intuitively, the concepts that are close to each other in the graph are more related than the ones that are more distant from each other. Several techniques that measure semantic relatedness based on this principle have been proposed in scientific literature. This work investigates the performance of such techniques in the evaluation of multi-label classifications by assessing them against human judgements of relatedness. It is shown that an evaluation incorporating semantic relatedness is likely to be more accurate than the traditional binary evaluations. The knowledge on classifier assessment acquired in this study is applied in practice in the form of an evaluation framework comprising several techniques for measuring and analysing the performance of classifiers.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P.M. Houben, Faculty EEMCS, TUDelft
University supervisor:	Dr. C. Hauff, Faculty EEMCS, TUDelft
Company supervisor:	Drs. J.N. Tiel Groenestege, GridLine B.V.
Committee member:	Prof. dr. ir. A.P. de Vries, Faculty EEMCS, TUDelft

Preface

Finally, after nine months of research I am proud to present you my Master's thesis. This thesis is submitted in partial fulfilment of the requirements for a Master's degree in Computer Science and represents my final work as a student of Delft University of Technology. The research conducted in this thesis has been carried out at GridLine, a small information technology company located in Amsterdam that is specialised in search and language technology.

The completion of my Master's thesis is the beginning of a new chapter in my life. It concludes my life as a student and marks the beginning of my professional career. This thesis does not only reflect my efforts, but also the instruction, guidance and support by many people. I want to take this opportunity to thank these people.

In the first place, I would like to extend my sincere gratitude to my Professor Geert-Jan Houben. He not only made me benefit from his professional counselling and constructive criticism, but also helped me in my search for a graduation project.

Furthermore, I would really like to thank my direct supervisors Claudia Hauff and Job Tiel Groenestege. They both provided me with invaluable guidance, assistance and feedback. I learned a lot from their expertise in the field of science and software development. I am very pleased that I still have the opportunity to work with them, even after the completion of my thesis.

My thanks also go to Jasper Oosterman, Dennis de Vries, Thomas Dohmen and my fellow students who were closely involved during the entire project, the persons who participated in my surveys and GridLine for giving me the opportunity to work on such a challenging project. Moreover I would like to thank Arjan de Vries for being a member of my graduation committee.

At last, special thanks go to my parents, brother, sisters, relatives and close friends. Thanks to their support and motivating words I was able to complete my academic education in beauty and begin this new chapter in my life.

Christophe Deloo
Delft, the Netherlands
February 17, 2013

Contents

Preface	iii
Contents	v
List of Figures	vii
List of Tables	ix
1 Introduction	1
1.1 Motivation	1
1.2 Objective	3
1.3 Outline	6
2 Evaluation of Classifiers	7
2.1 Binary classifier evaluation	7
2.2 Multi-label classifier evaluation	8
3 Measures of Semantic Relatedness	11
3.1 Graph-based measures	12
3.2 Corpus-based measures	14
3.3 Hybrid measures	15
3.4 Overview	17
4 Label-relatedness Mechanism	19
4.1 Method	19
4.2 Results	24
4.3 Discussion	26
5 Set-relatedness Mechanism	31
5.1 Method	32
5.2 Results	33
5.3 Discussion	34
6 Evaluation Framework	39

6.1	Features	39
6.2	Competencies	49
6.3	Components	52
7	Conclusions	57
7.1	Summary	57
7.2	Contributions	59
7.3	Future work	59
	Bibliography	63
A	Survey Results: Term Pair Relatedness	67
B	Example: Classified Document	69
C	Survey Results: Document Classification Quality	71
D	Glossary	73

List of Figures

1.1	Annotated document originating from the Dutch Parliament.	2
1.2	Fragment of the Parliament thesaurus.	3
1.3	Parameters of relatedness evaluation measure.	5
2.1	The ROC curve of three different classifiers.	9
4.1	Synonymous relations.	20
4.2	Hierarchical and related relations.	21
4.3	Overview of term pair ratings. It displays the number of times that each of the five ratings was assigned to a pair. The pairs are ordered in the same way as the pairs in the table of Appendix A.	29
4.4	Human judgements versus Leacock & Chodorow's similarity measure. . .	30
4.5	Human judgements versus Leacock & Chodorow's relatedness measure. .	30
5.1	First example of mechanism.	32
5.2	Second example of mechanism.	32
5.3	Expert versus relatedness evaluations.	36
5.4	Expert versus binary evaluations.	36
6.1	Data model.	39
6.2	Library page.	40
6.3	Study page.	42
6.4	Dataset page.	44
6.5	Graphical representation when exporting an evaluation as an histogram. .	47
6.6	Evaluation page.	48
6.7	Field validation.	49
6.8	Confirmation window.	50
6.9	Operation feedback.	50
6.10	Development process.	51
6.11	System routing.	52
6.12	System technologies.	53
6.13	System modules.	55

List of Tables

2.1	Confusion matrix for a binary classification.	7
3.1	Overview of measure characteristics.	17
4.1	Correlation of relatedness measures with human judgements of relatedness for the Parliament thesaurus.	25
4.2	Correlation of relatedness measures with human judgements of relatedness in all the different settings. Setting 1 is the Parliament graph in combination with the Parliament test-bed, setting 2 and 3 is the Wikipedia graph in combination with respectively the Parliament and Miller & Charles test-bed and setting 4 the WordNet graph with the Miller & Charles test-bed .	26
4.3	Numeric characteristics of the Parliament, Wikipedia and WordNet graph.	26
4.4	Correlations of similarity measures on WordNet for the Miller & Charles test-bed.	28
5.1	Expert, relatedness and binary evaluations per document.	34
5.2	Correlation of relatedness and binary evaluations with expert evaluations.	34
A.1	Overview of results from term pairs relatedness survey.	68
C.1	Overview of results from document classification quality survey.	71

Chapter 1

Introduction

This chapter introduces the problem addressed by this thesis. The main research objective and accompanying research questions are formulated. Finally, the structure of this document is outlined.

1.1 Motivation

The value of a digital collection does not only depend on the information it contains but also on the ease and effectiveness of retrieving this information. To enable effective retrieval it is therefore important that the content is richly annotated. However, the manual annotation of large collections is a time-consuming task. Annotating these collections by domain experts is expensive. This underlines the necessity of powerful techniques for the automatic annotation of digital content.

GridLine, a small information technology company specialised in search and language technology, have recently developed a system for the automatic annotation of a large document collection. The documents originate from the Dutch Parliament¹ and need to be annotated with terms, lexical expressions of concepts, from a hierarchically structured thesaurus². This thesaurus will be referred to as the Parliament thesaurus. The Parliament thesaurus contains nearly 8,000 Dutch terms oriented towards political themes such as welfare, healthcare, defence, culture, environment and so on. A typical example of such a document originating from the Dutch Parliament which has been manually annotated is given in Figure 1.1. This document is about matters regarding the defence policy of the Netherlands and has been assigned two terms by the experts: “kernwapenopslag” (nuclear weapons storage) and “wapenbeheersing” (arms control). The ultimate goal of GridLine is to produce a system that automatically annotates the documents with all relevant terms from the thesaurus.

In theory, the previously described task is equivalent to a multi-label classification problem. Multi-label classification is a variant of the classification problem where each document is assigned to one or more classes. The development of such a classification system is not straightforward. It is therefore essential to evaluate the quality of

¹The documents originally come from the Dutch House of Representatives, which is the lower house of the bicameral parliament of the Netherlands.

²In the context of information science a thesaurus can be defined as a type of controlled vocabulary in which relations between concepts are made explicit.

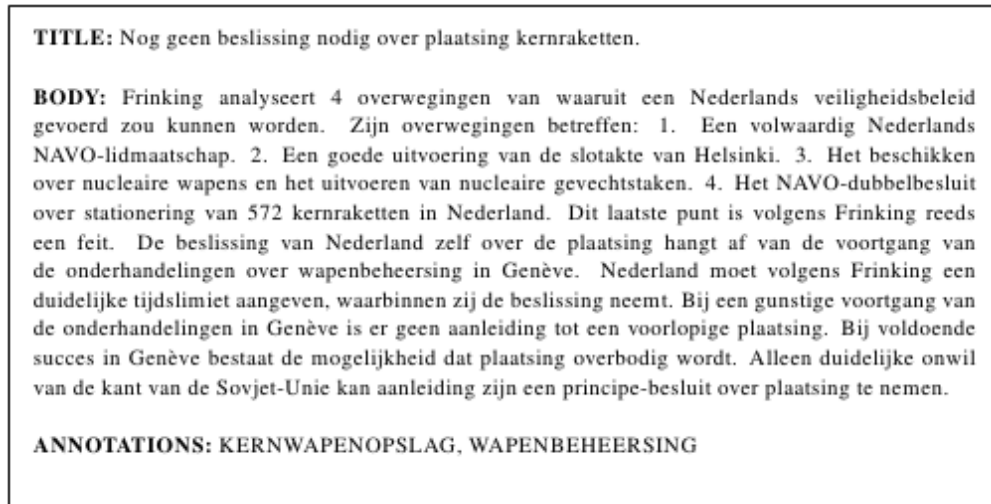


Figure 1.1: Annotated document originating from the Dutch Parliament.

annotation by the classifier. Such an evaluation can not only show if previous changes to the system led to a better annotation but may also give valuable insights that help to improve the classifier. GridLine evaluated their classification system by comparing the automatically produced annotations (classifier judgements) to a ground-truth, i.e. manual annotations of domain experts (expert judgements). The commonly used measures of precision and recall have been applied in this evaluation. These measures assume independence between the labels meaning that the assignment of a label to a document can either be evaluated as correct or incorrect and not somewhere in between³. This is why these measures are also known as binary evaluation measures.

The results of the evaluation performed by GridLine show that the classifier has difficulties in assigning labels in accordance to the ground-truth. A closer look at the annotations of each document individually reveals however that the classification is often still acceptable. Many documents have been assigned labels that bear some strong semantic relatedness with the manually assigned labels. The reason why the classifier has difficulties is mainly due to the large amount of related terms. The classifier learns the characteristic features for each term by training it on a large set of pre-labelled documents. In other words, the classifier learns when to associate a term to a particular context. One can thus imagine that it is difficult for a classifier to learn distinctive features for terms often appearing in very similar contexts.

The problem is that the commonly used performance measures of classification do not take into account the difference in semantics between the concepts depicted by the labels. For example, a document labelled as “renewable energy” instead of “solar energy” would be considered as wrong although the concepts are semantically related and “solar energy” is a specialisation of “renewable energy”. This relation is illustrated in Figure 1.2, which depicts a portion of the Parliament hierarchy⁴. A way of improving the evaluation of a classification would be therefore to also consider the

³Note that the words label and term are similar and therefore used interchangeably in this work.

⁴Note that these terms originate from the Dutch Parliament thesaurus and have been translated into English.

degree of semantic relatedness between terms.

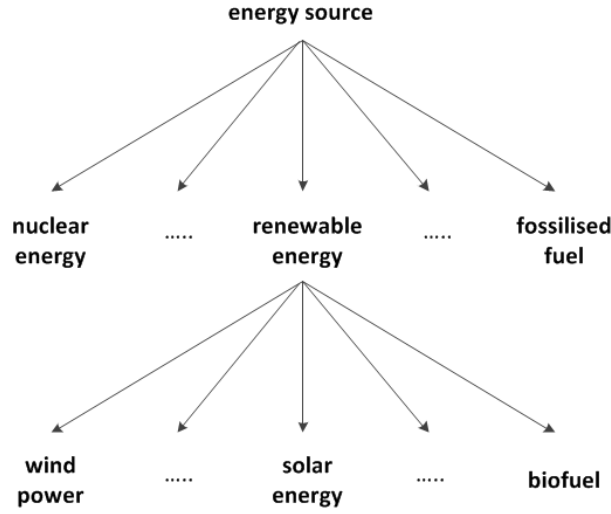


Figure 1.2: Fragment of the Parliament thesaurus.

This work focuses on improving the evaluation of a multi-label classification task by using the notion of semantic relatedness. The research will be specifically based on the previously described multi-label classification problem.

1.2 Objective

The primary objective of this thesis is to investigate the possibility of enriching the evaluation of a real-world multi-label classification task by using the notion of semantic relatedness. A multi-label classification is typically evaluated by comparing the classifier judgements to the expert judgements. In general, the more similar both judgements the better the classification. The main challenge therefore lies in quantifying the resemblance between the two judgements. This basically means comparing two sets of labels. However, before being able to compare two sets of labels one first needs to know a way how to determine the degree of semantic relatedness between two labels.

Semantic relatedness is the degree to which a given pair of lexically expressed concepts is related according to human judgements [21]. Several automatic techniques that attempt to imitate human judgements of relatedness have been proposed in scientific literature. These techniques are also known as semantic relatedness measures. They can be roughly divided into three categories, i.e. graph-based, corpus-based and hybrid techniques. The techniques, as their name suggests, rely on different resources. Graph-based techniques derive semantic relations from hierarchically structured graphs, corpus-based techniques from large corpora of text and hybrid techniques from both resources.

This work defines and validates a classification evaluation measure that incorporates the notion of semantic relatedness. Such an evaluation measure will also be

referred to as a relatedness evaluation measure ⁵. This to distinguish it from the traditional binary evaluation measures. The parameters of the relatedness evaluation measures are sketched in Figure 1.3. The relatedness evaluation measure basically contains three parameters. The two sets of labels (expert judgements and classifier judgements) and the external resource (graph and/or corpus). Note that the judgements of an actual document are used in this illustration, This is the document that has been shown earlier in Figure 1.1.

The relatedness evaluation measure basically determines the correctness of a classification based on two mechanisms, a low-level and a high-level mechanism. The low-level mechanism is responsible for determining the semantic relatedness between two labels, i.e. lexically expressed concepts. The high-level mechanism operates on top of the low-level mechanism. It defines a procedure for comparing two sets of labels, i.e. the labels from the classifier and expert judgements. The set comparison procedure quantifies the correctness of a classification based on measurements of relatedness provided by the low-level mechanism. The mechanisms will also be referred to as the label-relatedness and the set-relatedness mechanism.

The two mechanisms of the relatedness evaluation measure are defined and validated in two separate steps. The mechanisms will be validated in the context of our real-world multi-label classification task. Both steps are briefly described as follows:

1. **Label-relatedness mechanism** The low-level mechanism is defined according to the results of an experiment. In this experiment a selection of state-of-the-art semantic relatedness measures is evaluated against human judgements of relatedness. The goal of this experiment is to identify the most promising semantic relatedness measure, i.e. the measure that correlates most closely with human judgements of relatedness. The most promising relatedness measure will be used as the low-level mechanism of the relatedness evaluation measure. This experiment gives an answer to the following question:

How well do the different relatedness measures perform in accordance to human judgements of relatedness?

2. **Set-relatedness mechanism** In this step the remaining mechanism of the relatedness evaluation measure is defined after which the measure is examined in its entirety. The accuracy of the relatedness evaluation measure is evaluated against the judgements of classification experts. The classification experts are asked to rate a few document classifications on correctness, i.e. the better the classification, the higher the rating. These expert evaluations are not only used to evaluate the output of our relatedness evaluation measure, but also the output of one frequently used binary evaluation measure, the F-measure. This experiment gives an answer to the following questions:

⁵Note the difference between a relatedness measure and a relatedness evaluation measure. A relatedness measure quantifies the degree of relatedness between two concepts whereas a relatedness evaluation measure quantifies the correctness of a classification by taking semantic relatedness into account.

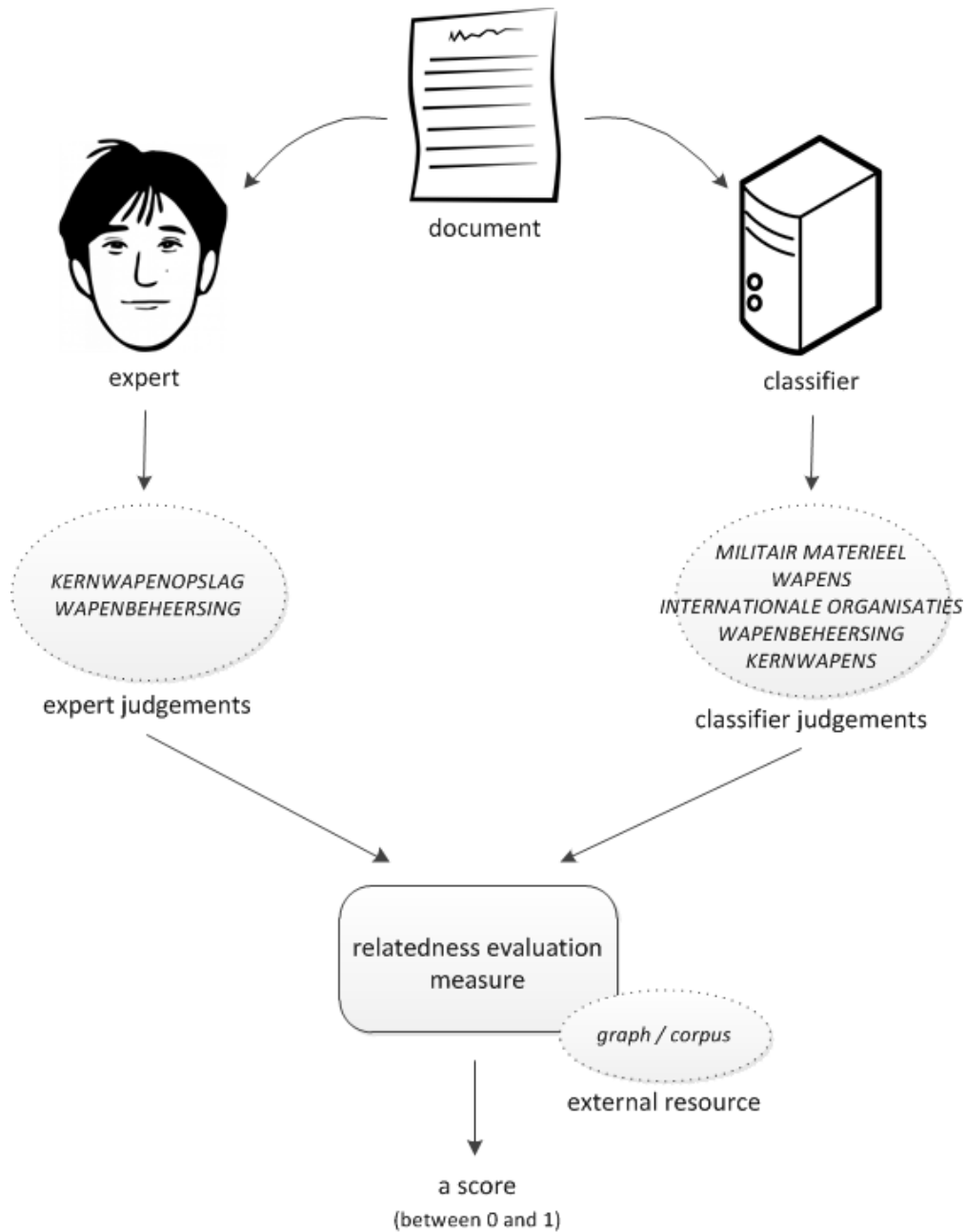


Figure 1.3: Parameters of relatedness evaluation measure.

How well does the relatedness evaluation measure perform according to the judgements of classification experts? Does it give more accurate evaluations than the commonly used F-measure?

The knowledge on classifier assessment acquired in this study is applied in practice in the form of an evaluation framework. The evaluation framework is a workbench which is developed for the purpose of automatic classifier assessment and is tailored to the needs of GridLine. It comprises several techniques for measuring and analysing

the performance of classifiers including both the common binary evaluation measures as well as the relatedness evaluation measure defined in this work. The insights that are provided by the workbench support GridLine in taking decisions regarding the development of their classification systems. How well does the classifier perform? Which classifier performs better? Does the classifier fails for certain specific cases?

1.3 Outline

This document is structured as follows. First of all, an overview of current multi-label evaluation practices and measures of semantic relatedness is given in respectively Chapter 2 and Chapter 3. The relatedness evaluation measure is defined and validated in two steps: the label-relatedness and the set-relatedness mechanism. Both mechanisms are defined and validated in respectively Chapter 4 and Chapter 5. Chapter 6 describes the evaluation framework including its features, competencies and components. Finally, conclusions and recommendations for future work are drawn in Chapter 7.

Chapter 2

Evaluation of Classifiers

Many of the classification problems that arise in practice are not easy to solve. For these problems it is often hard or nearly impossible to construct a perfect classification model that would correctly classify all possible cases. The goal is therefore to obtain a model approaching the perfect classification, in other words minimising the amount of classification errors. Different evaluation measures can be used in order to assess whether this goal is being met and to which extent. The most commonly used measures are the classic information retrieval notions of precision and recall [26]. The exact definitions of these notions depend on the type of classification problem. In this chapter an overview of current multi-label classifier evaluation methods is given after first introducing the commonly used evaluation practices for traditional binary classification problems. These multi-label classifier evaluation methods are essentially methods derived from binary classifier evaluation practices.

2.1 Binary classifier evaluation

In traditional binary classification problems each classified item can be either evaluated as true positive, true negative, false positive or false negative by comparing the predicted classification (classifier judgements) to the actual classification (expert judgements). These evaluations can be summarized in a so called confusion matrix as shown in Table 2.1. This matrix summarizes the true positives, true negatives, false positives and false negatives of all individual item evaluations.

actual \ predicted	yes	no
	yes	no
yes	#true-positives (TP)	#false-negatives (FN)
no	#false-positives (FP)	#true-negatives (TN)

Table 2.1: Confusion matrix for a binary classification.

The definitions of precision and recall are then defined as (2.1) and (2.2) respectively. From these definitions one can infer that precision is the fraction of cases classified as positive that were correctly classified whereas recall is the fraction of positive cases

that were correctly classified. Both measures fall in the range from 0 to 1, with 1 being the best possible score.

$$Pr = \frac{TP}{TP + FP} \quad (2.1)$$

$$Re = \frac{TP}{TP + FN} \quad (2.2)$$

Neither precision nor recall can effectively measure classification performance in isolation [26]. Therefore, classifier performance is often measured by the combination of precision and recall. One such measure is the F-measure. It is defined as (2.3). The F_1 -measure is a special case of the F-measure where β is equal to one and represents the harmonic mean of precision and recall. Two other commonly used F-measures are the F_2 -measure and $F_{0.5}$ -measure. The former attaches twice as much importance to recall as to precision whereas the latter weights them the other way around. The F-measure reaches its best value at 1 and worst at 0. Besides these measures many other performance measures based on the values of the confusion matrix exist. These include metrics such as the accuracy and error rate. An overview is given by Sebastiani [26].

$$F_\beta = (1 + \beta^2) \cdot \frac{Pr \cdot Re}{\beta^2 \cdot Pr + Re} \quad (2.3)$$

Another well-known measure in the evaluation of classifiers is the Area Under the Curve measure, also known as the AUC measure. The Area Under the Curve is a summary statistic often used to measure the ranking quality of a classifier [2]. It can be defined as the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. A random classifier has an area of 0.5, while an ideal one has an area of 1. The curve where the AUC metric refers to is the ROC curve. It is derived from signal detection theory where it is used to determine if an electronic receiver is able to satisfactorily distinguish between signal and noise [8]. The ROC curve for different classifiers are shown in Figure 2.1. The curve plots the true positive rate as a function of the false positive rate. It stands for Relative Operating Characteristic curve since it is a comparison of two operating characteristics (the true positive rate and false positive rate) as the criterion (classification threshold) changes. The points of the curve are obtained by sweeping the criterion value from the most positive classification value to the most negative. An ideal classifier is indicated by a ROC curve which follows the left hand and top edges of the square. A classifier that performs no better than chance is reflected by a ROC curve which follows the diagonal of the square, from the lower left corner to the top right corner. A classifier which performs worse than chance would produce a ROC curve which lays below the diagonal. Note however that in this case performance superior to chance could be obtained by inverting the predictions. The advantage of the AUC measure is thus that it is independent of the choice of classification threshold.

2.2 Multi-label classifier evaluation

A multi-label classification problem can be defined as variant of the traditional binary classification problem where each item is assigned to one or more classes. Multi-



Figure 2.1: The ROC curve of three different classifiers.

label classification should not be confused with multi-class classification, which is the problem of categorizing items into more than two classes. Multi-label classification requires different evaluation techniques than those used for binary classification problems [31]. However, one should note that these evaluation methods are essentially methods derived from binary classifier evaluation practices.

In multi-label classification problems multiple confusion matrices can be created, namely one for each class. These confusion matrices are constructed by first forming so-called bi-partitions. A bi-partition includes both the set of predicted and the set of actual items assigned to a class. The confusion matrix for a class is now simply created by comparing both sets of the bi-partition. The existence of multiple confusion matrices makes it possible to evaluate the classifier on two different levels. One can now not only produce values for precision and recall for a single class in isolation but also aggregate values considering all classes at once [26]. These aggregate values for precision and recall can be obtained in two ways called micro-averaging (2.4) and macro-averaging (2.5). The two methods may give different results depending on the generality of the classes i.e. the frequency of occurrence of a class. According to the definition, micro-averaging equally weights all items whereas macro-averaging equally weights all classes. The result of micro-averaging is therefore dominated by frequent classes, since they occur much more often than the rare ones. Whether one or the other should be used depends on the application requirements [26]. Just as for precision and recall a micro-average and macro-average of the F-measure can be computed [36]. In micro-averaging, the F-measure is computed by using the micro-averaged recall and precision, while in macro-averaging, the F-measure is computed by taking the average over all F-measures for each class. A similar class aggregation technique has also been proposed for the AUC measure [9].

$$Pr_{micro} = \frac{\sum_{\forall c \in C} TP_c}{\sum_{\forall c \in C} TP_c + FP_c} \quad Re_{micro} = \frac{\sum_{\forall c \in C} TP_c}{\sum_{\forall c \in C} TP_c + FN_c} \quad (2.4)$$

$$Pr_{macro} = \frac{\sum_{\forall c \in C} Pr_c}{|C|} \quad Re_{macro} = \frac{\sum_{\forall c \in C} Re_c}{|C|} \quad (2.5)$$

One thing that makes multi-label classifier evaluations different from both binary and multi-class classifier evaluations is that one can also construct bi-partitions for every item [7, 28, 32, 18]. In this case the sets of a bi-partition consist of classes assigned to an item instead of items assigned to a class. Measures such as precision and recall can now be computed in the same way as for class-based bi-partitions as described earlier. For further reference these two types of evaluations will be called class-based and item-based evaluations. It should be noted that in the case of item-based evaluations the macro-averaged F-measure, for example, is computed by taking the average over all F-measures for each item and therefore evenly weights all items. This in contrary to class-based evaluations.

As far as known the paper by Nowak et al. [17] is the only scientific publication using the notion of semantic relatedness in the evaluation of a multi-label classification problem. This paper quantifies the correctness of a classification by using measurements of semantic relatedness in comparing the set of classifier judgements to the set of expert judgements. The algorithm considers every label of both sets individually. Each label is matched with the label from the other set resulting in the lowest cost (i.e. the highest relatedness). These cost values are summed up into a total cost after being factored by the annotator agreement. The annotator agreement defines the subjectivity of a concept, in other words, the extent to which experts agree on classifying an item with a certain concept. The total cost is then normalized by dividing it by the number of labels appearing in both sets and reversed from a distance into a relatedness value by subtracting it from one. A simplified version of this algorithm is used for the high-level mechanism (set comparison procedure) of the relatedness evaluation measure that is going to be defined and validated in the upcoming chapters.

Chapter 3

Measures of Semantic Relatedness

Semantic relatedness refers to the degree to which a given pair of lexically expressed concepts is related according to human judgements [21]. The need to determine the degree of semantic relatedness between concepts is a problem that occurs in many applications of computational linguistics. Measures of semantic relatedness have been used in applications such as word sense disambiguation [20], spelling correction [3], ontology matching [5] and many others. Several measures of semantic relatedness using a variety of lexical resources have been proposed in scientific literature. In most cases semantic relations between concepts are either inferred from large corpora of text or lexical structures such as taxonomies or thesauri. In this chapter an overview of state-of-the-art semantic relatedness measures is presented. The measures are organised according to their type of knowledge resource, i.e. graph-based measures, corpus-based measures and hybrid measures.

It is important to note that semantic relatedness is a generalisation of semantic similarity [4]. Semantic relatedness considers any kind of semantic relation. It includes relations such as synonym, hyponymy/hypernymy (is-a), meronymy/holonymy (part-of), antonymy (inverse-of) or just any kind of functional relationship or frequent association. While the notion of semantic similarity is only limited to hyponymy/hypernymy relations. Many of the graph-based measures only consider relations expressing semantic similarity. Such measures are therefore further classified as semantic similarity measures. Besides this, one should also be aware of the fact that many of the semantic measures are expressed as semantic distance measures. Semantic distance can be defined as the inverse notion of semantic relatedness (similarity) and therefore quantifies the degree of unrelatedness (dissimilarity) between two lexically expressed concepts. Distance measures can, in most cases, easily be converted to relatedness measures. When the distance is a value between one and zero a relatedness measure is obtained by subtracting the distance from one or by taking the negative logarithm of the distance to base ten. Another way is to divide the value one by the distance. It should be noted that both the logarithmic and dividing function are undefined for distances equal to zero. Distance measures based on path lengths in graph-based measures are often converted by subtracting it from twice the maximum graph depth as this value represents an upper-bound for the path length.

The majority of the state-of-the-art relatedness measures presented in this chapter

have originally been developed for WordNet ¹. WordNet is a large lexical database for the English language that organises words in a graph-based structure according to their meanings. In WordNet each word is grouped into a set of synonyms called a synset. The synsets are in turn connected to other synsets via a number of semantic relations. These includes relations such as hyponymy/hypernymy, meronymy/holonymy and antonymy. Furthermore, each synset in WordNet is accompanied with a general definition. Such short definitions are also called glosses.

While WordNet represents a well structured taxonomy, questions were raised about the need for a larger coverage. This demand has resulted in some recent efforts focusing on the knowledge-rich source Wikipedia ². Semantic relatedness measures originally developed for WordNet have been thoroughly tested on Wikipedia as well as new measures which were specially tailored to take advantage of the structural components of Wikipedia. Strube et al. [29] were the first to compute measures of semantic relatedness using Wikipedia. They used the category-graph of Wikipedia to evaluate different prominent graph-based measures. The category-graph is an hierarchical structure used by Wikipedia to semantically organise its articles. A totally different type of graph-based approach was proposed by Milne et al. [34]. They take advantage of Wikipedia's hyperlink structure in defining relatedness. Wikipedia's extensive network of cross-references provides a huge amount of explicitly defined semantics. Despite the promising results reported by these studies no Wikipedia-based measures are introduced in this chapter. This not only because they are less prominent in literature, but also due to their specific nature. Most of the promising Wikipedia measures are tailored too closely to Wikipedia, in the sense that they rely heavily on specific structural components of Wikipedia. The strength of these measures can not come into full play on other domains.

3.1 Graph-based measures

Graph-based relatedness measures derive semantic relations from graph structures. The vertices in the graph represent lexically expressed concepts and are often hierarchically ordered. The vertices are interconnected by edges. These connections convey semantic relations. Graph-based measures rely on the intuition that concepts located near in the graph are more related than the ones that are located far. Semantic distance is quantified by computing the length between two vertices. The path length is typically defined by summing up the weights of every edge in the path. Different factors can play a role in determining the edge weight:

- **Edge type** Edges exhibit different strengths of semantic relatedness between concepts depending on the type of relation. For example, an edge representing a synonym relation is obviously semantically stronger than an is-a relation and therefore should be higher weighted.
- **Graph depth** The strength of semantic relations between concepts increases as one descends the hierarchy. This is based on the assumption that concepts located lower in the hierarchy are differentiated on finer details.

¹<http://wordnet.princeton.edu/>

²<http://www.wikipedia.org/>

- **Graph width** The density in a hierarchical semantic graph is generally not constant. Typically, regions with high density are of finer grain than regions of low density and one therefore argues that the strength of semantic relations is positively correlated with the density.

Rada [22]

The measure proposed by Rada et al. is conceptually one of the simplest semantic measure. It quantifies semantic distance based on the shortest path between two concepts in a graph by calculating the path length (4.1). Originally, this measure restricts itself to is-a relations and is therefore considered as a semantic similarity measure. However, one could easily introduce other types of semantic relations with varying weights in calculating the distance between two concepts.

$$rada(c_1, c_2) = \min_{p \in \mathcal{P}(c_1, c_2)} length(p) \quad (3.1)$$

Leacock & Chodorow [13]

Leacock & Chodorow proposed a similarity measure corresponding to a transformation of the Rada distance measure. It is normalized by dividing the path length by twice the maximum graph depth. This value ensures that the denominator is always at least as big as the numerator. The Leacock & Chodorow measure is defined as (4.3) where d equals the maximum graph depth.

$$leacock-chodorow(c_1, c_2) = -\log \frac{\min_{p \in \mathcal{P}(c_1, c_2)} length(p)}{2d} \quad (3.2)$$

Wu & Palmer [35]

The similarity measure proposed by Wu & Palmer is also based on path lengths. It is defined as (3.3). The concept specified as c_{mscs} represents the most specific common subsumer of concepts c_1 and c_2 . In the graph this concept is specified as the first vertex upward in the hierarchy that subsumes both vertices representing concepts c_1 and c_2 . The concept specified as c_{root} is the top concept of the hierarchy. The measure assumes that there is only one most specific common subsumer and root concept.

$$wu-palmer(c_1, c_2) = \frac{2 * \min_{p \in \mathcal{P}(c_{mscs}, c_{root})} length(p)}{\min_{p \in \mathcal{P}(c_1, c_2)} length(p) + 2 * \min_{p \in \mathcal{P}(c_{mscs}, c_{root})} length(p)} \quad (3.3)$$

Hirst & St.Onge [10]

Hirst & St.Onge proposed a distance measure tailored to the semantic network of WordNet. The idea behind this measure is that two concepts are semantically close if they are connected by a path that is not too long and that does not change direction too often regarding the hierarchical structure. The measure quantifies the semantic distance by finding the shortest path when only considering semantically correct paths.

Concepts are unrelated when no correct path exists. Semantically correct paths are defined as sequences of edge directions. Hirst & St.Onge define three kinds of directions: upward, downward and horizontal. An upward direction corresponds to a generalization and a downward direction to a specialization. Horizontal directions are less frequent than upward and downward directions and are usually very specific of meaning.

Sussna [30]

Sussna's distance measure is known as one of the first WordNet-based relatedness measures. It is calculated by summing all edge weights in the shortest path between two concepts. The edge between two adjacent vertices representing concept c_i and c_j is assigned a weight as defined in (3.4) where $\rightarrow_r$ is a relation of type r , $\rightarrow_{r'}$ is its inverse relation and d the depth of the deepest of the two concepts in the hierarchy. The two one-sided relationships are averaged to ensure a symmetrical distance between concepts. The weight of a one-sided relationship is defined as (3.5) where max_r and min_r are the maximum and minimum weights possible for a relation of type r and $n_r(c_i)$ is the number of relations of type r leaving vertex c_i . Sussna assigned weights to the different types of relations defined in WordNet: a weight of 0 to synonym relations, a weight range from 1 to 2 to hyponymy/hypernymy and meronymy/holonymy relations and a weight of 2.5 to antonym relations.

$$weight(c_i, c_j) = \frac{weight(c_i \rightarrow_r c_j) + weight(c_j \rightarrow_{r'} c_i)}{2d} \quad (3.4)$$

$$weight(c_i \rightarrow_r c_j) = max_r - \frac{max_r - min_r}{n_r(c_i)} \quad (3.5)$$

3.2 Corpus-based measures

Corpus-based relatedness measures derive semantic relations from large corpora of text. The degree of semantic relatedness between concepts is usually determined by comparing the textual surroundings of concepts in a corpus of text. The more overlap between the textual neighbourhoods of two concepts, the more related these concepts are.

Tversky [33]

Tversky proposed a measure that calculates a degree of relatedness by comparing the description sets of concepts. A description set typically contains lexical expressions of features describing the concept. The features may represent concrete properties such as size or color but also abstract properties such as quality or complexity. The more common the two description sets, the more related the terms. Tversky's measure is defined as (3.6) where C_1 and C_2 are the description sets of concept c_1 and c_2 respectively. This measure incorporates an additional parameter $\kappa \in [0, 1]$ defining the relative importance of the non-common characteristics.

$$tversky(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \kappa |C_1/C_2| + (1 - \kappa) |C_2/C_1|} \quad (3.6)$$

Patwardhan [19]

Patwardhan's relatedness measure is one of the few measures that does not rely on a semantic network. It combines the content of dictionary definitions with statistical information derived from large corpora of text. The relatedness between two concepts is determined by computing the cosine of their context vectors (3.7). A context vector can be defined as a second-order representation of co-occurrence information, a notion described by Schütze [25]. It is created by aggregating its word vectors. A word vector is basically a vector containing the frequencies of occurrence for each word appearing in the neighbourhood of a predefined word according to some corpus of text. Word vectors are constructed from the words occurring in the dictionary definition of the concept in question. Preferably, one should limit the construction of word vectors to words that are highly topical by filtering out stop-words such as "the", "for", "a", etc. This does not only reduce the amount of noise but also the dimensionality of the vectors.

$$patwardhan(c_1, c_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|} \quad (3.7)$$

The measure of Patwardhan is based on a famous technique proposed by Lesk used for word sense disambiguation [14]. Lesk disambiguates a word by comparing its textual surroundings in text to its dictionary definitions. Such short definitions are also called glosses. The selected sense is the sense whose gloss gives the maximum overlap of words according to its surroundings.

3.3 Hybrid measures

Hybrid measures derive semantic relations from both graph structures and corpora. Almost all prominent hybrid measures use the notion of information content introduced by Resnik [23]. The similarity between two concepts is then measured as the extent to which they share information in common. This common information is typically quantified as the information content of the most specific common subsumer. The value of the information content of a concept c is obtained by estimating the probability of occurrence of this concept in a large corpus of text. It is quantified as (3.8) where $P(c)$ is the probability of encountering an instance of concept c . According to the definition the concept informativeness decreases as the concept probability increases. The more abstract the concept, the lower its information content.

$$IC(c) = -\log P(c) \quad (3.8)$$

The concept probability is computed as relative frequency. It is defined as (3.9) where N is the total number of observed occurrences for all concepts. The frequency of a concept, $frequency(c)$, is counted as the number of times that the lexical expression of either this concept or a subordinate (hyponym) concept occurs in the corpus. For example, in Figure 1.2, the frequency of concept "renewable energy" would be counted

toward the frequency of “renewable energy”, “wind energy”, “solar energy”, “biofuel” and so on. This definition ensures that a top concept (in the case of a unique top vertex) will always receive a concept probability equal to one, resulting in an information content of zero.

$$P(c) = \frac{\text{frequency}(c)}{N} \quad (3.9)$$

More recently, Seco et al. [27] introduced an alternative approach for quantifying the information content of a concept which seem to outperform the original one. Their method solely rest on the hierarchical structure of a semantic network and is defined as (3.10) where $\text{hypo}(c)$ equals the number of hyponyms for a concept c and n_c the total number of concepts in the hierarchy. They assume that concepts with many hyponyms convey less information than concepts that are leaves. This assumption is based on the presumption that the more hyponyms a concepts has the less information it expresses, otherwise there would be no need to further differentiate it. Likewise, concepts that are leaf vertices are the most specified in the hierarchy so the information they express is maximal. One major advantage of this approach is that is does not rely on any corpora analysis and therefore reduces complexity and external resources.

$$IC(c) = 1 - \frac{\log(\text{hypo}(c) + 1)}{\log(n_c)} \quad (3.10)$$

Resnik [23]

Resnik introduced the notion of information content as a measure of semantic similarity for a pair of concepts. The similarity between two concepts is measured as the information content of their most specific common subsumer (3.11).

$$\text{resnik}(c_1, c_2) = IC(c_{mcs}) \quad (3.11)$$

Jiang & Conrath [11]

Jiang & Conrath use the notion of information content in computing a semantic distance. It is defined as (4.2). Jiang & Conrath also provide a more complex version of their measure including aspects such as graph depth, graph width and edge weight. Different conversions have been proposed for this measure.

$$\text{jiang-conrath}(c_1, c_2) = 2 * IC(c_{mcs}) - (IC(c_1) + IC(c_2)) \quad (3.12)$$

Lin [15]

The measure proposed by Lin uses exactly the same elements as in Jiang & Conrath’s measure but in a different configuration. It is defined as (3.13). It should be noted that this is a similarity measure and not a distance measure.

$$\text{lin}(c_1, c_2) = \frac{2 * IC(c_{mcs})}{IC(c_1) + IC(c_2)} \quad (3.13)$$

Banerjee & Pedersen [1]

Banerjee & Pederson's WordNet-based measure is like Patwardhan's measure, also based on the technique proposed by Lesk described earlier [1]. The measure is unique in that it extends the glosses of the concepts under consideration by also including the glosses of related concepts according to a given hierarchy. A semantic distance value is computed by summing the degree of overlap between the glosses belonging to different types of relations including the glosses of the concepts in question. The degree of overlap between two glosses is defined by a scoring mechanism. The score between two glosses is determined by summing up the individual scores for each word sequence that two glosses share in common and is defined as n^2 where n represents the length of a word sequence.

3.4 Overview

The characteristics of the measures presented in the previous sections are summarized in Table 3.1.

characteristics measures	resource	semantics	distance
Rada	graph	similarity	yes
Leacock & Chodorow	graph	similarity	no
Wu & Palmer	graph	similarity	no
Hirst & St.Onge	graph	relatedness	no
Sussna	graph	relatedness	yes
Tversky	corpus	relatedness	no
Patwardhan	corpus	relatedness	no
Resnik	graph & corpus	similarity	no
Jiang & Conrath	graph & corpus	similarity	yes
Lin	graph & corpus	similarity	no
Banerjee & Pedersen	graph & corpus	similarity	yes

Table 3.1: Overview of measure characteristics.

Chapter 4

Label-relatedness Mechanism

The label-relatedness mechanism is responsible for determining the semantic relatedness between two labels, i.e. lexically expressed concepts. It is defined according to the results of an experiment described in this chapter. In this experiment a selection of state-of-the-art semantic relatedness measures is evaluated in the context of our Parliament thesaurus against human judgements of relatedness. The goal of this experiment is to identify the most promising semantic relatedness measure, i.e. the measure that correlates most closely to human judgements of relatedness. The most promising relatedness measure will be used for the low-level mechanism of the relatedness evaluation measure. The method and the results of this experiment are described and discussed in the subsequent sections of this chapter.

4.1 Method

This experiment follows a commonly accepted methodology [11, 15, 4, 21] which was first applied by Resnik [23]. Resnik assessed a relatedness measure by examining the correlation with human judgements of relatedness for a set of predefined term pairs. Taking human judgements of relatedness as the ground-truth is deemed to be correct by definition. The main drawback however lies in the difficulty of obtaining a large set of reliable, subject-independent judgements for comparison. It is therefore often only feasible to apply this methodology on a small scale, including just a few concept pairs.

In this experiment a selection of state-of-the-art semantic relatedness measures is evaluated against a set of human rated term pairs. The terms included in the set are terms from the Parliament thesaurus. Each term pair is assigned a relatedness score equal to the mean of nearly hundred human ratings. The human ratings were obtained in an electronic survey in which people were asked to rate each term pair on relatedness. A description of the thesaurus and its structure, the survey and its results and a motivation for the selection of relatedness measures together with some implementational details are given in the upcoming subsections.

4.1.1 Parliament thesaurus

A thesaurus can be defined as a type of controlled vocabulary in which relations between concepts are made explicit. The Parliament thesaurus contains nearly 8,000

terms oriented towards political themes such as welfare, healthcare, defence, culture, environmental aspects and so on.

The thesaurus groups all synonymous terms, i.e. terms characterizing one single concept, as a cluster. Such a cluster is built out of one preferred term and zero or more non-preferred terms. A term is designated as a preferred term when its lexical expression is preferred over the lexical expression of its synonymous terms. A preferred term is interconnected with its non-preferred terms through two different types of one-sided relations. A relation from a preferred to a non-preferred term is called a “use for” (UF) relation and its inverse relation is called a “use” (USE) relation. Such synonymous relations are illustrated in Figure 4.1. Around half of the terms of the thesaurus are preferred terms. An example of two terms that are considered synonymous is “aardolie” (petroleum) and “ruwe olie” (crude oil). The thesaurus denotes “aardolie” as the preferred term and “ruwe olie” as the non-preferred term. So the previously described cluster of terms is a structure that is actually quite similar to that of a synset of WordNet. The only difference is that a synset does not specify something like a preferred term. All terms are equal in a synset.

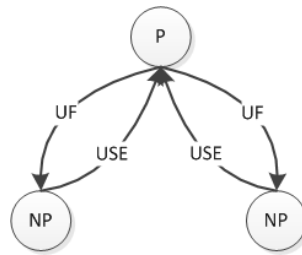


Figure 4.1: Synonymous relations.

All terms of the thesaurus are organised in a hierarchical structure. The hierarchical structure is defined by two types of one-sided relations occurring between preferred terms: the “broader term” (BT) relations, i.e. upward links, and the “narrower term” (NT) relations, i.e. downward links. The thesaurus does not have a single top term, but multiple, in total 32 top terms. These top terms represent the major topics of the thesaurus. For example, “welzijn” (welfare), “gezondheidszorg” (healthcare), “defensie” (defence), “cultuur” (culture) and “natuur” (nature) are one of the top terms. Every term of the thesaurus is organised under at least one of the top terms. It happens that a term has multiple “broader term” relations and that these relations occur between terms organised under different top terms. Besides these synonymous and hierarchical relations another type of relation exists between terms. These relations are called “related term” (RT) relations and occur between terms that are strongly related but that do not define a specific or generic relationship. The “related term” relations are two-sided and can, just like hierarchical relations, occur between terms organised under different top terms. The hierarchical and related relations are illustrated in Figure 4.2. Note that all relations can be represented as a single graph where relations are edges and terms are vertices. It should be noted that the hierarchical structure of the thesaurus is actually made of two types of relations, is-a and part-of relations, and that it does not make any distinction between the two.

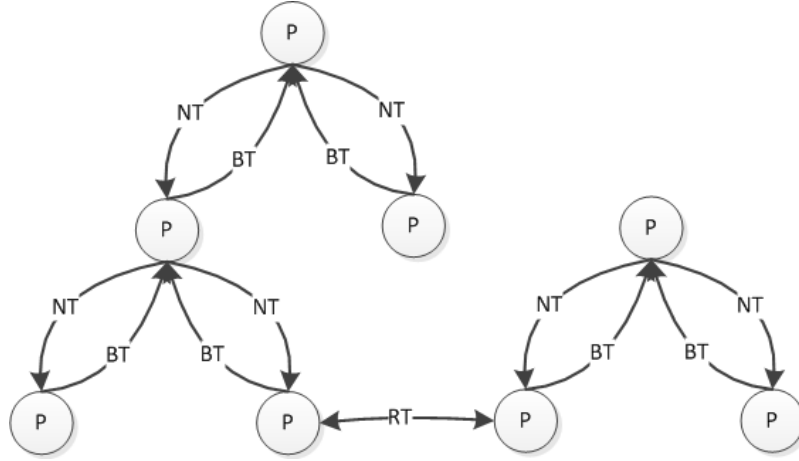


Figure 4.2: Hierarchical and related relations.

4.1.2 Survey

The conducted survey followed a methodology which was first introduced by Rubenstein and Goodenough [24]. In this survey Dutch speaking people were asked to rate 50 different term pairs on relatedness. The terms of the pairs originate from the Parliament thesaurus. Ratings were obtained in an electronic survey which was sent to relatives, acquaintances and colleagues. All persons taking part in the survey were given the same instructions. The survey participants were asked to rate each pair on a five-point scale. The higher the assigned rate, the stronger the perceived relatedness. The participants were explicitly asked not to rate a term pair when the meaning of one of the term was not known to them. This to prevent people from filling in rates that do not make any sense. A few examples were given to the survey participants to make sure that they understood the notion of relatedness. The pairs were picked with the goal to include as many different characteristics. This means varying in path lengths, types of relations, types of vertices and so on. Most of the terms were chosen manually. Only 20% of the pairs were assembled randomly.

The survey was filled in by 94 persons in total. A summary of the results is given in Appendix A in the form of a table. The statistics in this table are organised per term pair. For each term pair the mean and mean absolute deviation over all ratings are given as well as the number of times it was left unrated. The term pairs in the table are sorted by their mean in ascending order. In addition to this table an overview of the individual ratings is given in Figure 4.3. The individual ratings are plotted in a histogram. A bar in this histogram displays the number of times that each of the five ratings was assigned to a pair. The term pairs are ordered in the same way as the term pairs in the table of Appendix A.

Many interesting details can be observed in the overview of Appendix A and Figure 4.3. From the results we can see that the survey participants followed the instructions well since mostly term pairs containing complex terms were left unrated. The pair that was left unrated most of the time was “procesrecht” (procedural law) and “cassatie” (cassation). Both are quite abstract terms and are principally only used on a daily basis by people practising law. Another noteworthy observation is that 74% of

the term pairs received an average rating higher than 2.5 while one would expect this to be around 50%. This high percentage can be explained by the fact that the term pairs were not all assembled randomly. In totally 80% of the pairs were assembled manually. A pair was often formed by specifically paying attention to the relation of two terms as specified by the thesaurus. Another reasonable explanation for this high percentage is that the terms originate from one single domain specific thesaurus. It is also interesting to discuss the pairs that were assigned on average the highest and lowest relatedness scores. The pair “Vaticaanstad” (Vatican City) - “paus” (pope) received the highest relatedness score on average. This is most probably due to their strong relation, i.e. the Vatican City is a state that is ruled by the pope. The second highest score was assigned to “energiebedrijven” (power companies) - “elektriciteitsbedrijven” (electricity companies). A power company is generalisation of an electricity company. And thirdly highest placed is the pair “rijbewijzen” (driver licenses) - “rijbevoegdheid” (qualification to drive). These two terms are nearly synonymous. Now compare these pairs to the three lowest ranked term pairs: “makelaars” (broker) - “republiek” (republic), “roken” (smoke) - “dieren” (animals) and “boedelscheiding” (derision of property) - “gentechologie” (gene technology). The difference is that one can easily infer a relation between the terms of the highest placed terms, which is not the case for the lowest placed pairs.

Important is the degree of agreement among the raters. In this experiment human judgements are taken as the ground-truth, but if the humans mostly disagree, then the question is whether the means calculated over all ratings are representable degrees of relatedness. There are a number of statistics which can be used to determine the inter-rater agreement. Different statistics are appropriate for different types of measurement. Krippendorff’s alpha was used to determine the inter-rater agreement among the participants of the survey. This statistical measure was chosen among all others due to the following advantages:

- Applicable to any number of raters.
- Applicable to ordinal ratings.
- Can cope with missing data.

Krippendorff’s alpha returned an inter-rater agreement of 0.59. It is difficult to say something about this inter-rater agreement. Obviously there is some agreement. The question, however, is whether this degree of agreement is significant or not. According to a set of guidelines [12] this degree of agreement can be quantified as moderate. It should be noted that these guidelines are based on a personal opinion and that they are by no means universally accepted.

4.1.3 Selection

The relatedness measures examined in this experiment are basically selected on two criteria. First of all the measures should be the state-of-the-art. The selection only contains measures presented in Chapter 3. Secondly, the selected measures should conform to the available resources. Measures are only selected when they can be used to automatically determine a relatedness score between the terms of the Parliament

domain. The Parliament domain comes with a thesaurus where all terms are semantically organised into a graph structure. Unfortunately, the terms of the thesaurus are not provided with textual descriptions. This means that we can not make any use of relatedness measures relying on glosses (Tversky, Patwardhan and Banerjee & Pederson)¹. For this reason the selection only contains graph-based measures.

The following measures were selected: Rada, Leacock & Chodorow, Wu & Palmer, Resnik, Jiang & Conrath and Lin. The measures of Rada, Leacock & Chodorow and Wu & Palmer are all based on path lengths. The path length is calculated by summing up the weights of the edges in the path. The weights typically depend on the type of relation. The stronger the semantic relation, the lower the weight. The thesaurus specifies different semantic relations: hierarchical, synonymous and related term relations. It was chosen to implement two versions for both Rada's and Leacock & Chodorow's measure. One including only hierarchical and synonymous term relations, i.e. a similarity measure, and one including all types of term relations, i.e. a relatedness measure. The weights of the relations were chosen according to their semantic strength. A weight of 1 was assigned to both hierarchical and related term relations and a weight of 0 to synonymous term relations. It was decided to calculate the information content of the remaining three measures by using the alternative approach. This approach cuts the need for any corpus which will reduce some complexity in the results of this experiment. Finally, to guarantee that a path can always be found between any two terms one single top vertex was added on top of the 32 top terms of the thesaurus.

To make a proper comparison some of the selected measures were slightly adapted to ensure that all selected measures express their output as a relatedness score ranging from 0 to 1. Both Rada's and Jiang & Conrath's original formula were transformed from a distance into a similarity function. The transformed formula of Rada is defined as (4.1) where d equals the maximum depth. Jiang & Conrath's formula was changed according to the formulation of Seco et al. [27] and is defined as (4.2). The measure of Leacock & Chodorow was slightly modified to make sure that it only returns values between 0 and 1 (4.3).

$$rada(c_1, c_2) = 1 - \frac{\min_{p \in \mathcal{P}(c_1, c_2)} length(p)}{2d} \quad (4.1)$$

$$jiang-conrath(c_1, c_2) = 1 - \frac{2 * IC(c_{mscs}) - (IC(c_1) + IC(c_2))}{2} \quad (4.2)$$

$$leacock-chodorow(c_1, c_2) = \log \frac{\min_{p \in \mathcal{P}(c_1, c_2)} length(p) + 1}{2(d + 1)} \bigg/ \log \frac{1}{2(d + 1)} \quad (4.3)$$

Note that the formulas of Sussna and Hirst & St.Onge were not selected. Sussna's measure was not chosen because it depends on certain outdated particulars of WordNet not defined in the Parliament thesaurus². Hirst & St.Onge was not selected due to implementational reasons. The JUNG Framework³, an open-source library for graph

¹Using definitions of other resources, like a dictionary, will also not work in practice since one should then find a way to automatically map the definitions to the terms of the Parliament domain.

²The measure assumes that each type of relation has a predefined maximum and minimum weight.

³<http://jung.sourceforge.net/>

analysis, was used to implement the graph algorithms. This library contains several algorithms from graph theory. The problem however is that this library does not allow one to search for the shortest path while in the same time restricting the search space by excluding some paths corresponding to certain patterns.

Wu & Palmer’s measure and the information content measures all assume that each possible pair of vertices has exactly one most specific common subsumer. This is indeed true for graphs where each vertex is organised under exactly one parent vertex (excluding the root), but not for other graph structure like the Parliament thesaurus. Some vertices of the thesaurus have multiple parent vertices, which in theory means that it is possible for two vertices to have multiple most specific common subsumers. For the Parliament thesaurus it was therefore chosen to select the subsumer with the lowest number of hyponyms. This decision is based on the presumption that the more hyponyms a concept has the less information it expresses [27]. Note that in theory it is even possible to have more than one vertex with the lowest number of hyponyms. In this case the vertex is chosen with the term name that alphabetically precedes the term name of the other vertices. This ensures that for two given terms the measures always return the same result.

4.2 Results

The performance of each relatedness measure on the Parliament thesaurus is quantified by calculating the correlation between their outputs and the human judgements of relatedness. The correlations are shown in Table 4.1. In this table two types of correlation coefficients are given per measure: Pearson’s rho and Kendall’s tau. Both coefficients measure the association between two variables on a scale that varies from one to minus one. A positive monotonic association results in a correlation greater than zero and a negative smaller than zero. The closer the correlation is to zero, the weaker the monotonic association. The difference between the two coefficients is that Kendall’s tau is a rank correlation measure. It quantifies the degree of association by purely looking at the similarity of the orderings of the data when ranked by the variables. This is different to Pearson’s rho, which measures association by looking at the linear dependency between the variables, which goes beyond analysing the orderings.

From these results two aspects stand out, in particular: the relatively high correlation obtained for Rada’s and Leacock & Chodorow’s relatedness measure and the relatively poor performance of the remaining measures when compared to the results of earlier experiments conducted on the WordNet graph. The former observation is quite evident. Rada’s and Leacock & Chodorow’s relatedness measure are the only measures from our selection that also consider related term relations. Apparently a lot of relatedness information is contained in the related term relations, which is actually not very surprising due to its naming. This observation will be further discussed in the next section. The latter observation is less obvious. Seco et al. [27] evaluated all measures from our selection except Rada’s measure in a similar way on the WordNet graph against a test-bed of human judgements provided by Miller & Charles [16]. They reported significant higher correlations for the selected similarity measures. Their correlation results range from 0.74 (Wu & Palmer) to 0.84 (Jiang & Conrath). Their

	pearson	kendall
rada (similarity)	0.43	0.35
rada (relatedness)	0.73	0.55
leacock & chodorow (similarity)	0.49	0.36
leacock & chodorow (relatedness)	0.73	0.55
wu & palmer	0.39	0.33
resnik	0.45	0.37
jiang & conrath	0.48	0.41
lin	0.45	0.39

Table 4.1: Correlation of relatedness measures with human judgements of relatedness for the Parliament thesaurus.

results are in line with other similar relatedness studies on WordNet such as Budanitsky et al. [4]. The interesting question is whether we can attribute this difference in performance to one or more specific characteristics of the graph or test-bed. To gain insight into this question the selection of measures is examined under different circumstances. The selection of measures were tested on the category-graph of the Dutch Wikipedia for both the test-bed with terms from our Parliament domain and the test-bed of Miller & Charles. Each term occurring in one of the test-beds had to be mapped to the categories of the category-graph of the Dutch Wikipedia⁴. This mapping was done manually and induced an additional translation step for the terms of the test-bed of Miller & Charles. A term is mapped to one or more categories depending on whether it semantically matches to the title of a category or article: a category always has one parent category while an article can have more than one parent category. A semantically correct match could be found for almost all terms of the test-beds. Remaining terms were matched with Wikipedia articles or categories that most closely relates to the term in question. For instance, the Parliament term “onderverhuur” (sub-rent) was matched to the Wikipedia article with the title “huren” (renting). It should be noted that a similar study based on Wikipedia was conducted in the past by Strube et al. [29]. There are however some considerable differences. This study focusses on the the category-graph of the Dutch Wikipedia and not the English. Another difference is the way in which the terms of the test-beds are mapped. Their study mapped each term automatically. The drawback of this approach is that it may result in incorrect mappings of terms due to ambiguity. The measures were also tested on the WordNet graph for the Miller & Charles test-bed to confirm the results reported by Seco et al. [27]⁵. The results for all the different settings are shown in Table 4.2: setting 1 is the Parliament graph in combination with the Parliament test-bed, setting 2 and 3 is the Wikipedia graph in combination with respectively the Parliament and Miller & Charles test-bed and setting 4 the WordNet graph with the Miller & Charles test-bed. Note that Rada’s and Leacock & Chodorow’s relatedness measures are both missing from this table since they require a type of relation that is not explicitly specified in all

⁴This experiment uses the category-graph of the Dutch Wikipedia as of the 19th of June 2012.

⁵This experiment uses WordNet version 3.0.

three graphs.

	1		2		3		4	
	pearson	kendal	pearson	kendall	pearson	kendall	pearson	kendall
rada	0.43	0.35	0.59	0.42	0.54	0.50	0.63	0.55
leacock & chodorow	0.49	0.36	0.55	0.42	0.66	0.50	0.76	0.55
wu & palmer	0.39	0.33	0.27	0.15	0.30	0.19	0.72	0.51
resnik	0.45	0.37	0.49	0.39	0.68	0.50	0.77	0.52
jiang & conrath	0.48	0.41	0.57	0.38	0.57	0.48	0.81	0.60
lin	0.45	0.39	0.56	0.42	0.71	0.51	0.82	0.55

Table 4.2: Correlation of relatedness measures with human judgements of relatedness in all the different settings. Setting 1 is the Parliament graph in combination with the Parliament test-bed, setting 2 and 3 is the Wikipedia graph in combination with respectively the Parliament and Miller & Charles test-bed and setting 4 the WordNet graph with the Miller & Charles test-bed

Some principal numeric characteristics for all three graphs are reported in Table 4.3. In this table, the row referring to the total number of vertices is simply the total number of terms contained by the graph. The number of preferred vertices has to be perceived in slightly different ways according to the type of graph. For the parliament graph, this is simply the total number of preferred terms, whereas for the WordNet graph, this amount is equal to the total number of synsets. Note that the category-graph of Wikipedia contains no synonym structures at all and that the total number of vertices is therefore equal to the total number of preferred vertices. Most of the other characteristics are self-explanatory. The maximum depth is the maximum number of vertices that a vertex is away from the root vertex while the maximum width is the maximum number of child vertices over all vertices of the graph.

	parliament	wikipedia	wordnet
total vertices	7910	65994	179374
total preferred vertices	4063	65994	74374
maximum depth	8	15	19
maximum width	42	798	808
average depth	4.27	5.56	7.76
average width	2.19	4.15	4.08

Table 4.3: Numeric characteristics of the Parliament, Wikipedia and WordNet graph.

4.3 Discussion

The relatively good performance of Rada’s and Leacock & Chodorow’s relatedness measure was already pointed out in the previous section. The difference can be easily illustrated with scatter-plots. In Figure 4.4 and Figure 4.5 the ratings provided by

the human judgements are plotted against the ratings given by respectively Leacock & Chodorow's similarity and relatedness measure. In both plots the corresponding trend line is drawn in red. At this stage it was chosen to focus on Leacock & Chodorow's measure instead of Rada's since this measure is more recent and seems to provide better results on all the different settings. When observing the plots we can see that the data points in Figure 4.5 are generally closer to the trend line than the points in Figure 4.4. In Figure 4.4 it is shown that Leacock & Chodorow's similarity measure assigned rather low rates to some pairs which were obviously conceived as strongly related by the survey participants. The data points of these pairs have been encircled in light blue. Interestingly enough is that most of these pairs appear closer to the trend line when rated by Leacock & Chodorow's relatedness measure. This all has to do with the additional related term relations taken into consideration. Take for example the data point in the lower right corner of Figure 4.4 representing the pair "Vaticaanstad" (Vatican City) - "paus" (pope). The terms of this pair appear under different top terms. This consequently means that the shortest hierarchical path between the two terms is relatively long and therefore resulting in a low similarity rating. But when one also considers related term relations, like Leacock & Chodorow's relatedness measure, the terms appear to be strongly related. This is due to the fact that both terms are semantically connected with a related term relation in the thesaurus. Another interesting insight is that all the similarity measures fail to imitate human judgements of relatedness on basically the same pairs, i.e. the pairs represented by the data points encircled in light blue. This is actually not so strange since all other measures rely on the assumption that the degree of similarity between two terms is negatively correlated with the height of their most specific common subsumer. Take for example, again, the terms "Vaticaanstad" (Vatican City) and "paus" (pope). Since the shortest hierarchical path between the two terms is relatively long we can most probably assume that the most specific common subsumer lies high in the hierarchy (which actually appears to be true). The high placed most specific common subsumer therefore results in the assignment of a low similarity rating for the term "Vaticaanstad" (Vatican City) - "paus" (pope). With this in mind we can now easily deduce that besides Leacock & Chodorow's similarity also all other similarity measures from our selection will fail to determine a proper degree of relatedness for term pairs like "Vaticaanstad" (Vatican City) - "paus" (pope). It therefore seems crucial to also include these related term relations when determining the degree of relatedness. This all suggests that Leacock & Chodorow's relatedness measure is the most suitable for imitating human judgements of relatedness, at least for the Parliament domain.

In the previous section it was brought to the attention that our selection of measures performed relatively poor on the Parliament domain compared to other similar relatedness studies on WordNet. The question was raised whether this could be attributed to one or more specific characteristics of the graph or test-bed. To gain insight into this question the selection of measures was also examined under different circumstances. The results obtained are listed in Table 4.2. This overview clearly shows that the best results are obtained for setting 4, i.e. the WordNet graph in combination with the test-bed of Miller & Charles. The results returned for all the other settings are not as good as for setting 4 and are more or less similar. It seems however that this difference has nothing to do with some of the numeric characteristics of the graphs as listed in Table 4.3: there are no numbers for WordNet that considerably differ from the

ones reported for the other graph structures. It also seems unlikely that this difference is merely due to the different test-bed used: the results of setting 2 and 3, where the category-graph Wikipedia is tested on the two different test-beds, do not greatly differ. The most plausible explanation is that this difference is caused by the difference in semantics of the hierarchical structures. The hierarchical structure of the WordNet graph purely consists of is-a links whereas the hierarchical structures of the other graphs are not only built upon is-a links but also include part-of links. The Parliament thesaurus and the category-graph of Wikipedia do not make any distinction between the two types of links which can be considered as loosely in contrary to the strict hierarchical semantic structure of WordNet. This difference in semantics may have great influence on the performance of the measures since they were originally developed to work on hierarchical is-a graphs.

As mentioned earlier, the selection of similarity measures was also tested on the WordNet graph for the Miller & Charles test-bed in order to confirm the results reported by Seco et al. [27]. The results are summarised in Table 4.4. From the table we can see that the results returned by this study are nearly equal to the ones reported by Seco et al. The small discrepancies can be attributed to the different WordNet versions that were used ⁶.

pearson	this	seco
leacock & chodorow	0.76	0.82
wu & palmer	0.72	0.74
resnik	0.77	0.77
jiang & conrath	0.81	0.84
lin	0.82	0.81

Table 4.4: Correlations of similarity measures on WordNet for the Miller & Charles test-bed.

⁶Seco et al. used WordNet version 2.0 while this study uses version 3.0.

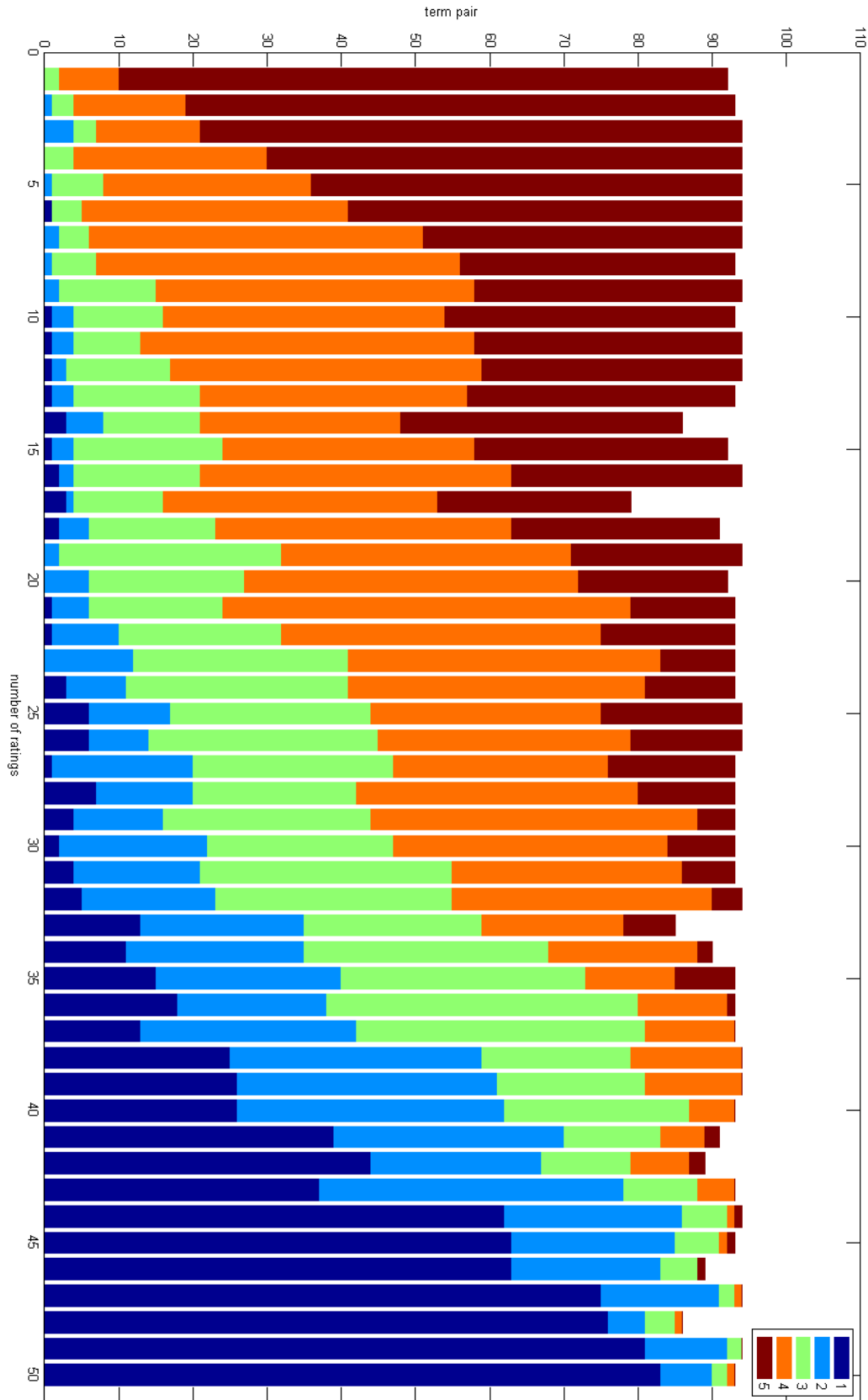


Figure 4.3: Overview of term pair ratings. It displays the number of times that each of the five ratings was assigned to a pair. The pairs are ordered in the same way as the pairs in the table of Appendix A.

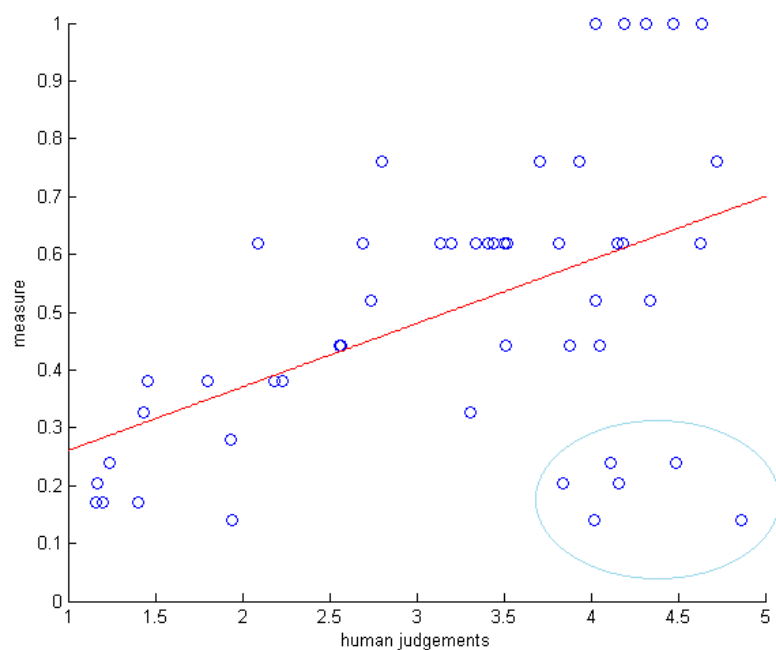


Figure 4.4: Human judgements versus Leacock & Chodorow's similarity measure.

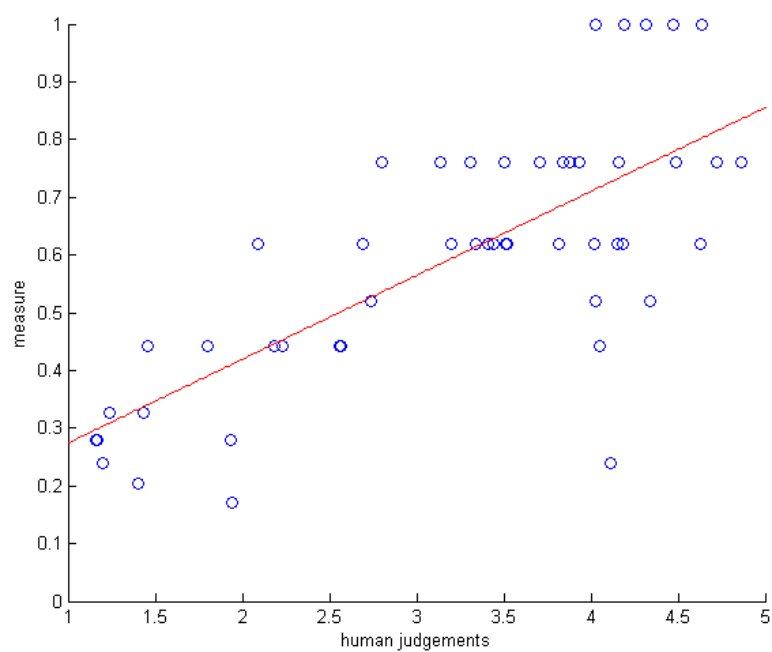


Figure 4.5: Human judgements versus Leacock & Chodorow's relatedness measure.

Chapter 5

Set-relatedness Mechanism

The set-relatedness mechanism operates on top of the label-relatedness mechanism and defines a procedure for comparing two sets of labels, i.e. the labels from the classifier and the labels from the expert judgements. The set comparison procedure quantifies the correctness of a classification based on measurements of relatedness provided by the low-level mechanism. It was chosen to use Leacock & Chodorow's relatedness measure for the low-level mechanism since it gave the best results as shown in the previous chapter. The set comparison procedure, as noted earlier, is based on a simplified version of Nowak's algorithm [17]. It is defined as (5.1). The algorithm calculates the degree of relatedness between the set C of classifier judgements and the set E of expert judgements in some kind of optimisation procedure. The procedure pairs every label of both sets with a label of the other set in a way that maximises relatedness: each label l_c from C is matched with a label l'_e from E and each label l_e from E with a label l'_c from C . The relatedness values of each of those pairs are summed up and divided by the number of labels occurring in both sets. This gives a value between 0 and 1. The higher the value, the more related the sets. Note that this definition does not really specify what to do in case of an empty set. This is due to the fact that in practice each document should be assigned at least one label. Documents are usually classified for information retrieval purposes. Leaving documents unclassified would not increase the ease of retrieving such documents. Moreover, one could argue that documents which cannot be classified with any relevant label simply do not belong to the population and therefore should not be included into the classification process. Nonetheless, a definition is given as follows. The relatedness evaluation measure assigns a score of 1 when both sets are empty and a score of 0 when only one of the sets is empty. The maximum score is assigned in the latter case because of two exactly matching sets.

$$\frac{\sum_{l_c \in C} \max_{l'_e \in E} rel(l_c, l'_e) + \sum_{l_e \in E} \max_{l'_c \in C} rel(l_e, l'_c)}{|C| + |E|} \quad (5.1)$$

Two examples illustrating the entire mechanism are given in Figure 5.1 and Figure 5.2. It illustrates the way in which the mechanism assigns a relatedness score to two different classifier judgements for the document shown in Figure 1.1. This document has been attributed the following expert judgements: “kernwapenopslag” (nuclear weapons storage) and “wapenbeheersing” (arms control). These expert judgements are compared to the judgements of two different classifiers. The classifier

of Figure 5.1 assigned the terms “kernwapens” (nuclear weapons) and “wapenbeheersing” (arms control) and the classifier of Figure 5.2 the terms “taakstraffen” (community services) and “wapenbeheersing” (arms control). The two figures show how the labels are paired and how a single relatedness score is determined by summing up the individual scores assigned to each of the pairs. From these illustration its easy to see that the classifier judgements of the second example are worse than the one of the first example due to the assignment of the obviously not so related term “taakstraffen” (community service).

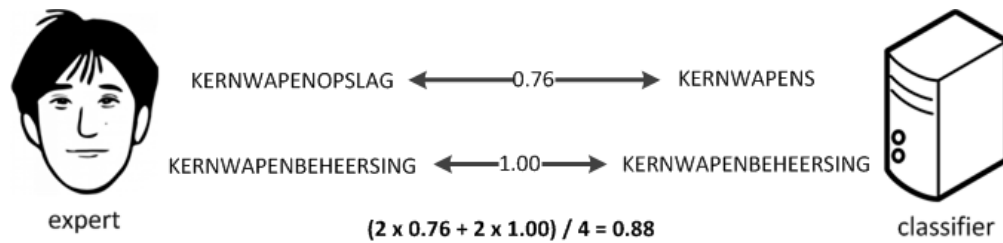


Figure 5.1: First example of mechanism.

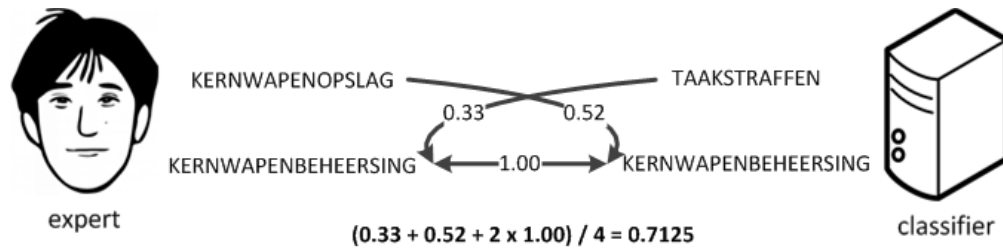


Figure 5.2: Second example of mechanism.

Now that the remaining mechanism of the relatedness evaluation measure is defined it is time to examine the measure in its entirety by conducting an experiment. The method and the results of this experiment are described and discussed in the subsequent sections of this chapter.

5.1 Method

This validation experiment evaluates our relatedness evaluation measure. Classification experts were asked in a survey to rate a few document classifications on correctness, i.e. the better the classification, the higher the rating. A description of the survey is given in the upcoming subsection. The expert evaluations are used as a ground-truth to compare the accuracy of the relatedness evaluation measure and the accuracy of a frequently used binary evaluation measure, the F-measure. This experiment should give an answer to the question whether or not our relatedness evaluation measure is more accurate than the well-known F-measure.

5.1.1 Survey

Three classification experts took part in this survey. These classification experts are all three employees involved in the development of classification systems at GridLine. The classification experts were asked to rate on an individual basis the quality of classification for 25 documents according to a five-point scale: very poor, poor, average, good and very good. They rated a document classification by comparing the classifier judgements to the content of the document and its expert judgements. Note the difference between the experts. The expert judgements come from domain experts employed by the Dutch government whereas the classification experts are employees of GridLine specialised in the development of classification systems. The means taken over all three individual expert ratings are considered as the ground-truth. An example of a document and its judgements is shown in Appendix B. This example document is also used as one of the 25 documents of the survey.

The results of the survey are listed in Appendix C in the form of a table summarising the results per document. For each document the ratings of all three experts, the mean and mean absolute deviation are given. The ratings have been translated from the nominal five-point scale to an ordinal five-point scale in which 1 equals very poor and 5 equals very good. The results show that the experts tend to agree. The experts assign quite similar rates often differing no more than one point on the scale, which means that they never assign totally opposite rates.

Krippendorff's alpha returns an inter-rater agreement of 0.68. Here again it is difficult to say something about this inter-rater agreement. According to a set of guidelines [12] this degree of agreement can be quantified as moderate. One should note however that there is a lot of subjectivity involved in the rating. The experts have to rate a classification (i.e. two sets possibly of unequal size) based on their own notions of relatedness. The obvious question to raise now is why the survey was not also presented to one or more domain experts? After all, it is the domain expert that has the best understanding of the thesaurus and its concepts. These people are extremely qualified in rating the thesaurus concepts on relatedness. The answer to all this is simple. The domain experts were simply not available. If such domain experts had been available I would have asked to participate not only in this survey but also in the term pair relatedness survey. Their ratings could have revealed interesting insights compared to those of non-experts.

5.2 Results

An in-depth overview of the results of the previously described experiment is shown in Table 5.1. The table gives the statistics per document. Three different types of statistics are indicated: the expert evaluation, the relatedness evaluation and the binary evaluation. The expert evaluations are the means that were reported in Appendix C. The relatedness and binary evaluations are the scores given by our relatedness evaluation measure and the F_1 -measure. The entries in the table are sorted by document id.

document id	expert	relatedness	binary
2554349	3.33	0.76	0
2591088	4	0.85	0
2614430	3.33	0.82	0.25
3833254	4.67	0.87	0.29
3872172	5	0.93	0.67
3877208	4.67	0.74	0
3880651	3.67	0.79	0
3924639	1	0.59	0
4183938	3.67	0.81	0.33
4197111	2.33	0.88	0
4285490	2	0.77	0.18
4318676	2.67	0.51	0
4320359	3.33	0.82	0
4339282	2.67	0.76	0
4354700	3	0.74	0
4369129	2	0.56	0
4386070	3.67	0.78	0
4414807	4	0.90	0.33
4434266	4	0.87	0
4463034	1.67	0.65	0
4509554	4	0.80	0
4594871	4.67	0.87	0.29
4595226	4.33	0.83	0.29
4604487	2	0.76	0
4616008	2.33	0.77	0.22

Table 5.1: Expert, relatedness and binary evaluations per document.

The results presented in Table 5.1 are used to form a general impression of the relatedness evaluation measure in comparison to the F-measure. The performance of both measures are quantified by calculating the correlation between their outputs and the expert evaluations considering all 25 document classifications. The correlations are shown in Table 5.2. In this table the same types of correlations coefficients are given as in the previous experiment: Pearson's rho and Kendall's tau.

	relatedness	binary
pearson	0.67	0.48
kendall	0.53	0.37

Table 5.2: Correlation of relatedness and binary evaluations with expert evaluations.

5.3 Discussion

The results from Table 5.1 clearly show that the experts do not only consider exact matches in evaluating a classification. The classifications of documents with identification number 2591088, 3877208 and 4434266 were rated as good while the classifier and expert judgements contain no labels in common. This can be interpreted from

the zero values which have been assigned by the F_1 -measure in the binary evaluations. From this we can deduce that the experts also consider other aspects when evaluating a classification. Two more aspects, both involving semantic relatedness, were identified. This results in the following set of aspects:

- Exact matches between the labels of the expert and classifier judgements.
- The semantic relatedness between the labels of the expert and classifier judgements.
- The semantic relatedness between the document content and the labels of the classifier judgements.

Our relatedness evaluation measure only take the former two aspects into account. This simply because it is not feasible to semantically analyse entire paragraphs of texts. Note, however, that this is still one aspect more than the F_1 -measure.

The individual document results of Table 5.1 are summarised in Table 5.2. The results in Table 5.2 indicate that for these 25 document classifications the relatedness evaluations correlate more with the expert evaluations than the binary evaluations of the F_1 -measure. The coefficients report an increase in correlation of at least 0.16 in favour of the relatedness evaluations. The difference in correlation can be easily illustrated with scatter-plots. In Figure 5.3 and Figure 5.4 the expert evaluations are plotted against respectively the relatedness and binary evaluations. In both plots the corresponding trend line is drawn in red. The less the data points deviate from the trend line, the stronger the correlation between the two quantities. With this in mind we can easily conclude that the data points of Figure 5.3 are closer to the trend line than the points of Figure 5.4. A factor that plays a significant role in this verdict are the zero entries of the F_1 -measure (encircled in light blue in Figure 5.4). These zero entries pull the trend line downwards, therefore enlarging the gap with the vertically higher placed data points. It is also noteworthy to point out that all of the 25 documents received a relatedness score higher than 0.51. Preferably one would like the trend line to go from the lower left corner to the upper right corner. Since these corners, point (0,1) and point (1,5), represent the lower and upper bounds of both a relatedness and expert evaluation.

From this we can conclude that the results of our relatedness evaluation measure look promising, but we should be aware of the fact that this experiment only covers 25 document classifications. This is a very small sample. Ideally this experiment should be carried out on a much larger scale. The problem however is that the manual evaluation of classifications is a time-consuming task. It is therefore unrealistic to ask classification experts to evaluate large samples of classifications. One should also pay attention to the edge cases of the relatedness evaluation measure. In theory the measure is also capable of handling empty sets. Empty sets could result in a score of zero. Such a score would significantly deviate from the the trend line formed by the vertically high placed data points.

The relatedness and binary evaluation measures both have their own pros and cons. Relatedness evaluations are capable of evaluating a classification at a much finer grain. The reliability of these evaluations however strongly depend on the external resource used to extract a relatedness score between two labels. Take for example our case

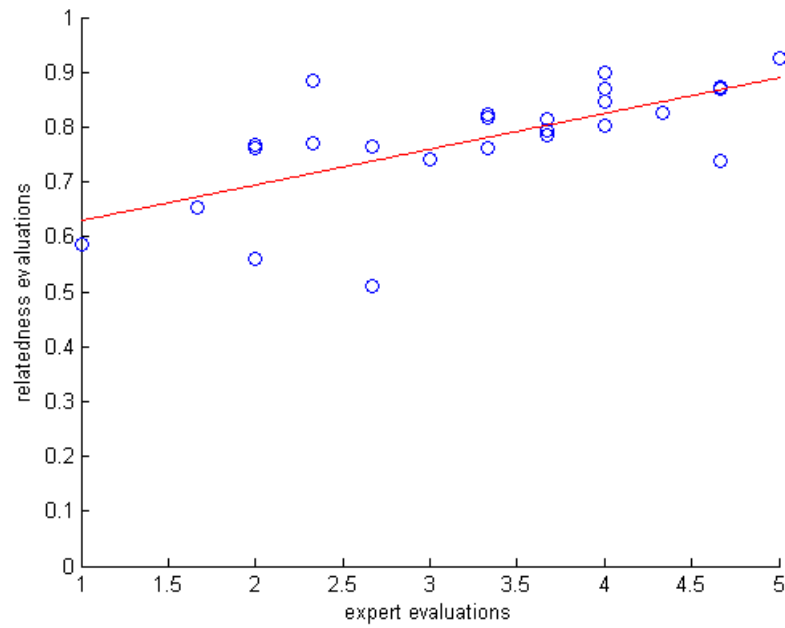


Figure 5.3: Expert versus relatedness evaluations.

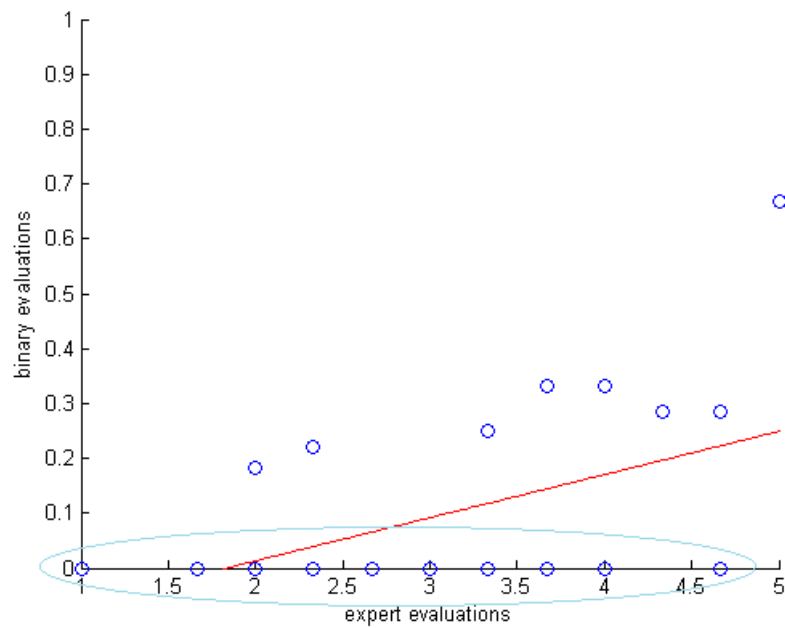


Figure 5.4: Expert versus binary evaluations.

where a relatedness score is calculated based on a semantic network. It is important that the concepts contained by the network are semantically well organised. The better the semantic structure of the network, the higher the reliability of the relatedness

evaluations. An advantage of a binary evaluation measure, like the F-measure, is its universal applicability. The measure is widely known and its exact definition can be easily looked up. Its meaning is therefore well understood. This cannot be said of the relatedness evaluation measure. It not only deals with a highly subjective matter, its output also heavily depends on the external resource used as pointed out earlier. One should thus be aware of the fact that a single relatedness evaluation on its own does not say much. Relatedness evaluations only make sense when comparing multiple classifications to each other. Assume that a certain number of documents were classified by two classifiers. The relatedness evaluation measure can then be used to calculate an average relatedness score for the classifications of each classifier. Since the scores of both classifiers are calculated under the same conditions we can safely compare these two quantities.

Chapter 6

Evaluation Framework

The evaluation framework is a workbench which has been developed for the purpose of automatic classifier assessment and was tailored to the needs of GridLine. It contains several techniques for measuring and analysing the performance of classifiers. In this chapter the features, competencies and components of the evaluation framework are described. The techniques implemented in the workbench are applicable on all common classification types as described in Chapter 2, i.e. binary, multi-class and multi-label classifiers.

6.1 Features

The workbench contains the necessary functionalities to manipulate and store the input and output of an evaluation. The features of the workbench can be controlled through a graphical user interface. This graphical interface is built in HTML and can therefore be displayed in every modern web browser. The different features of the workbench are organised into four pages, i.e. the library, study, dataset and evaluation page. This structure is basically deduced from the data model as shown in Figure 6.1. In this section an overview of the features is given per page.

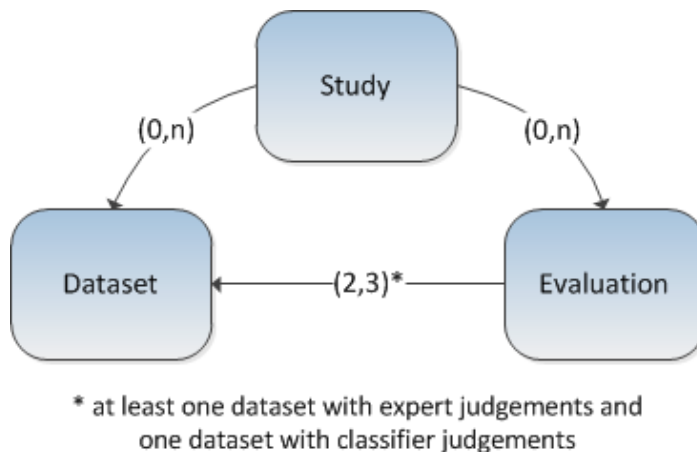


Figure 6.1: Data model.

6.1.1 Library

Each evaluation is organized in a so-called study. This study represents a classification problem. For example, our Parliament classification task can be seen as a separate classification study. All studies are shown in the main page, better known as the library page. This page, as shown in Figure 6.2, is used to manage and browse through the different studies. The functionalities of the this page can be summarised as follows:

- An overview of the different studies is given on this page. The overview of displayed studies can be manipulated by using the search bar or by sorting on a certain attribute. The number of studies displayed on one page is limited. The user can easily navigate through the different pages by using the “Previous” and “Next” buttons.
- A new study can be created by clicking on the “New” button in the right corner. The user is asked to specify several fields when creating a new study. This includes general attributes such as the name, date and description, but also at least one file specifying all classes of the classification task. This information is needed to define the number of true negatives in item-based evaluations.
- One can select a study by clicking on its identification number. A new page will open. This page is called the study page.

Quality Workbench				
Library				
Id or name...				New
Id ▼	Name	Type	Date	Description
1	study X	CLASSIFICATION	12/09/2012	a description
2	study Y	CLASSIFICATION	12/09/2012	a description
-- Previous Displaying 1 to 2 of 2 Next --				

Figure 6.2: Library page.

6.1.2 Study

A study is displayed in the study page. A snapshot of this page is shown in Figure 6.3. On this page one can manage all attributes related to a study. This not only includes general attributes such as the name, date and description, but also the datasets and evaluations associated to a study. One can load different datasets into a study. The datasets are either loaded as expert judgements or classifier judgements. An expert evaluation is flagged with the name “gold standard”. A dataset is loaded into the workbench by selecting a text file corresponding to be pre-specified format. The exact format can vary and depends on the selected reader. A user has the possibility to select between different readers. The workbench contains the flexibility to easily introduce new readers. By default the reader is selected that accepts text files where each line specifies an item and its associated classes. The format per line is defined as: *itemX;classX,classY,classZ;*. This line specifies that *itemX* is associated with *classX*, *classY* and *classZ*. The datasets are used as the inputs of an evaluation. An evaluation can be created after the user specified at least one dataset with expert judgements and one dataset with classifier judgements. The functionalities of the this page are given as follows:

- All attributes related to a study are given on this page. This includes general attributes, as well as its associated datasets and evaluations.
- A study can be removed by pressing the “Delete” button. It should be noted that removing a study also results in the deletion of all its associated datasets and evaluations.
- The general attributes of a study can be modified by clicking on the “Edit” button.
- An overview of the different datasets and evaluations associated with this study is given on this page. One can switch between the two overviews by clicking on one of the tabs. Both overviews can be manipulated in the same way as is in the library page.
- New datasets and evaluations can be created by clicking on the “New” button. The user will be asked to specify several fields.
- One can select a specific dataset or evaluation by clicking on its identification number. The user will be either forwarded to the dataset or evaluation page.
- The user can return to the library page by pressing the back button in the top right corner of the page.

Quality Workbench

Study

EditDelete

Name:study X

Type:CLASSIFICATION

Date:12/09/2012

Description:a description

DatasetsEvaluations

Id or name...Q

New

Id ▾	Name	Date	Description	Gold standard
1	expert judgement	13/09/2012	a description	true
2	classifier judgment X	13/09/2012	a description	false
3	classifier judgment Y	13/09/2012	a description	false

-- Previous

Displaying 1 to 3 of 3

Next --

Figure 6.3: Study page.

6.1.3 Dataset

The workbench loads and stores the judgements as a dataset of items. A user can inspect the items in the dataset page as shown in Figure 6.4. In our Parliament classification task, each item represents a document. Each document is associated with a set of labels, i.e. class names. These labels represent judgements. As mentioned earlier, a dataset can either contain expert or classifier judgements. An overview of the functionalities of this page is given as follows:

- All attributes related to a dataset are given on this page. This includes general attributes, as well as the content of the dataset.
- A dataset can be removed by pressing the “Delete” button. It should be noted that removing a dataset also results in the deletion of all its associated evaluations.
- The general attributes of a dataset can be modified by clicking on the “Edit” button.
- An overview of the entries in the dataset is given on this page. The overview can be manipulated in the same way as the other overviews.
- A dataset can be exported by using the “Export” button.
- The user can return to the study page by pressing the back button in the top right corner of the page.

Quality Workbench

Dataset

EditDelete

Name:expert judgement

Date:13/09/2012

Description:a description

☒ Gold standard

Name...Q

Export

Name ▼	Classes
001	[FRANCE, APPLE, GREEN, RIPE]
002	[GREEN, UNRIPE, PAPUA NEW GUINEA, BANANA]
003	[SPAIN, BROWN, ROTTEN, ORANGE]
004	[BROWN, KIWI, RIPE, NEW ZEALAND]
005	[GREEN, UNRIPE, PEAR, ENGLAND]
006	[GREEN, UNRIPE, STRAWBERRY, ITALY]
007	[THE NETHERLANDS, RASPBERRY, RIPE, RED]
008	[FRANCE, RIPE, PURPLE, GRAPE]
009	[APPLE, RIPE, RED, GERMANY]
010	[FRANCE, BROWN, ROTTEN, PEAR]
011	[APPLE, GREEN, RIPE, GERMANY]
012	[APPLE, GREEN, RIPE, GERMANY]

← Previous

Displaying 1 to 12 of 12

Next →

Figure 6.4: Dataset page.

6.1.4 Evaluation

The key page of this workbench is the evaluation page. This page is used to browser through all statistics regarding an evaluation. It is shown in Figure 6.6. A classification is evaluated by comparing it to the ground-truth, i.e. the expert judgements. The workbench allows the user to evaluate the classifier judgements of two classifiers at once. In this way a user can easily compare the performance of two classifiers. The user can choose between item-based or class-based evaluations by clicking on one of the tabs ¹. Different measures can be chosen for both types of evaluations including several traditional binary evaluation measures and a relatedness evaluation measure ². The relatedness evaluation measure is implemented according to the technique that has been described and validated in the previous chapters. This technique basically determines the quality of a classification by considering the distances between the concepts defined by a thesaurus. The user selects the desired measure from a selection box. When a measure is selected the user will be presented with a table of statistics. In item-based evaluations statistics are provided per item whereas in class-based evaluations statistics are provided per class. Aggregate statistics considering all items or classes are also given. For example, the F-measure can be calculated for each class and item individually as well as for all items and classes, as an aggregate, like the micro-average and macro-average. A list of all the functionalities in the evaluation page is given as follows:

- All attributes related to an evaluation are given on this page. This includes general attributes, as well as the statistics of the evaluation itself.
- An evaluation can be removed by pressing the “Delete” button.
- The general attributes of an evaluation can be modified by clicking on the “Edit” button. The user has also the possibility to change the parameter values of measures used in the evaluation. For example, the β in the F-measure. These parameter values can also be set when creating an evaluation.
- The statistics are organised in the usual overview and can be manipulated in the same way as the other overviews.
- A user can choose between item-based and class-based evaluations. In item-based evaluation statistics are provided per item whereas in class-based evaluations statistics are provided per class. Aggregate statistics considering all items and classes are also given.
- The workbench includes several measures for item-based and class-based evaluations. It includes both the traditional binary evaluation measures (accuracy, precision, recall and the F-measure) and a relatedness evaluation measure based on the technique described in the previous chapters. In order to calculate the relatedness measure one needs to provide the workbench with a file that specifies

¹The workbench constructs class-based evaluations from the item-based datasets.

²Note that the relatedness measure can be found under item-based evaluations.

the structure of a thesaurus. The user can provide a thesaurus when creating a new study ³.

- A short description including a mathematical definition is given for each measure.
- One can easily compare two classifier judgements in an evaluation.
- The aggregate and individual statistics can be easily exported by using the “Export” button. The user has the possibility to choose between different export formats. At this time the workbench contains a tabular and a graphical export format. The tabular format simply exports the statistics as a .csv file. This export format enables an extra degree of freedom. It allows the user to work on the data in a spreadsheet program. Therefore giving the user the ability to define custom data transformations and representations. The workbench currently only contains one single graphical output format. This graphical representation provides an overview of the individual statistics as an histogram. It plots the relative frequencies in ten intervals (bins) for the selected measure. An example of such a plot is given in Figure 6.5. Note that two types of bars are included in the histogram of this figure, i.e. one for each of the classifier judgements.
- The user can return to the study page by pressing the back button in the top right corner of the page.

³Note that this file replaces the file containing all classes since all classes can be extracted from the thesaurus, assuming that it contains all the classes.

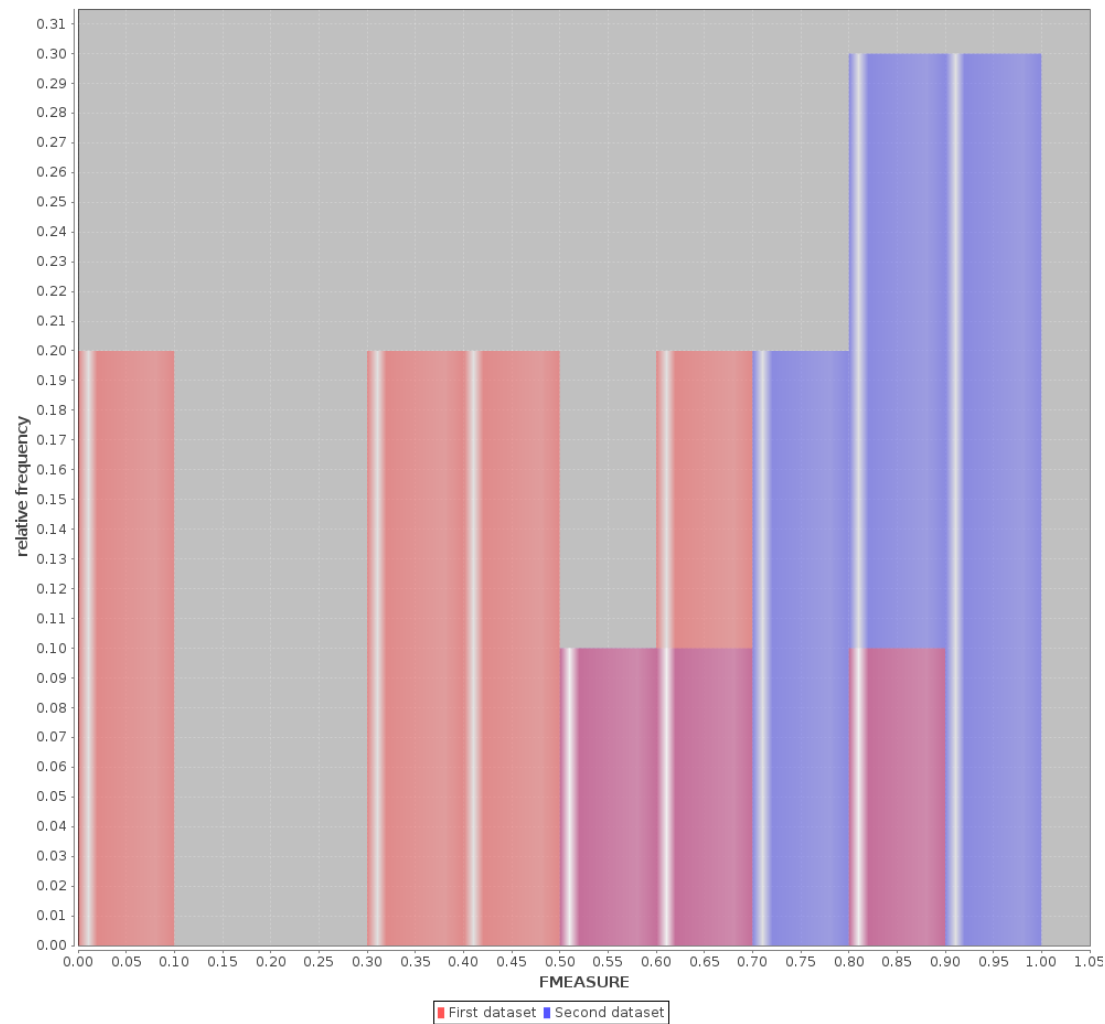


Figure 6.5: Graphical representation when exporting an evaluation as an histogram.

Quality Workbench

Evaluation

Name:

Date:

Description:

Gold standard:

First dataset:

Second dataset:

Items

Classes

Name...

Informations

Type	First dataset	Second dataset	Difference
micro-averaged f-measure	0.56	0.56	0.0
macro-averaged f-measure	0.52	0.52	0.0

Name ▼	First dataset	Second dataset	Difference
001	0.75	0.75	0.0
002	1.0	1.0	0.0
003	0.67	0.67	0.0
004	0.5	0.5	0.0
005	0.57	0.57	0.0
006	1.0	1.0	0.0
007	0.0	0.0	0.0
008	0.33	0.33	0.0
009	0.0	0.0	0.0
010	0.4	0.4	0.0

Figure 6.6: Evaluation page.

6.2 Competencies

In this section the core competencies of the evaluation workbench are outlined. The core competencies can be grouped into four classes, i.e. usability, extensibility, interoperability and robustness. These competencies are basically deduced from the most important requirements set by GridLine.

Usability

The workbench can be easily operated through a graphical user interface. It was designed with simplicity in mind. A minimalistic design is more inviting and easy to navigate compared to a heavy loaded design. The user interface is developed in a way that guides the user through a natural flow of operations. One first has to create a study and load in some datasets before starting an evaluation. The user is provided with feedback when necessary: input fields are validated on correct format, a confirmation window is shown for critical operations and the user is informed if an operation has succeeded or not. The different kinds of feedback mechanisms are illustrated in Figures 6.7, 6.8 and 6.9. Other user-friendly features include the possibility to manipulate the data overviews in all kinds of ways (by searching and sorting), the possibility to display the statistics in different formats (tabular or graphical) and the concise explanations given for each measure and their parameters.

The screenshot shows the 'New dataset' form in the 'Quality Workbench'. The form has several input fields with validation feedback:

- Name:** An empty text field with a red border and the message 'This field is required.' to its right.
- Date:** A text field containing 'XYZ' with a red border and the message 'Please enter a valid date (dd/mm/yyyy).' below it.
- Description:** A text field containing 'a description' with a green border.
- Dataset file:** A text field with a red border, a 'Browse...' button to its right, and the message 'This field is required.' to the right of the button.
- Dataset reader:** A dropdown menu showing 'STANDARD_CLASSIFICATION_READER' with a green border.
- Gold standard:** An unchecked checkbox labeled 'Gold standard'.

At the top right of the form are 'Cancel' and 'Save' buttons. The form is titled 'New dataset' and is part of the 'Quality Workbench' interface.

Figure 6.7: Field validation.

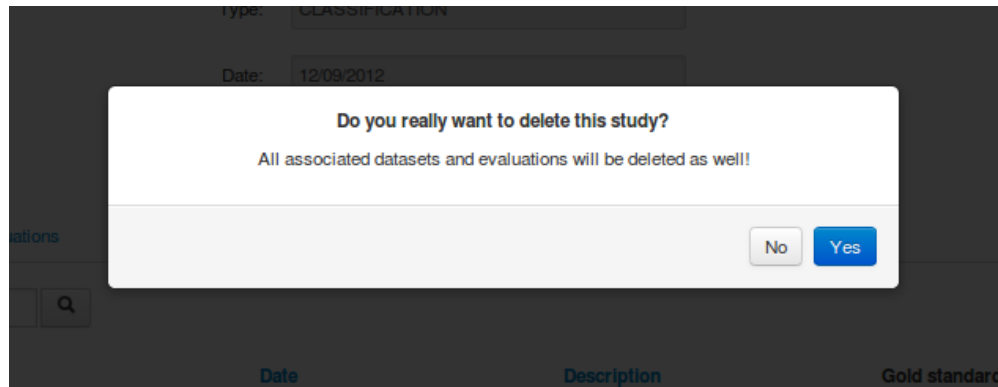


Figure 6.8: Confirmation window.

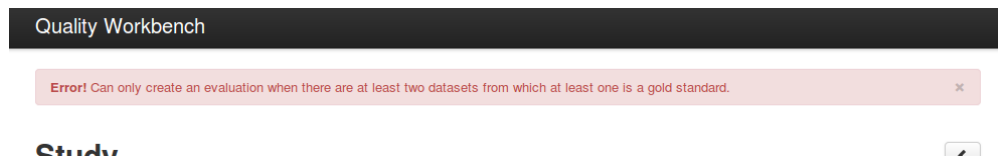


Figure 6.9: Operation feedback.

Extensibility

The workbench was designed to cope with future growth. Different aspects of the workbench can be extended with additional functionalities. These additional functionalities can be easily integrated into the workbench by following the specifications defined by interfaces. Almost all input and output functions adhere to interfaces. These interfaces makes it possible to extend the workbench with different types of input and output formats. A good example is the input specifying the thesaurus. Thesauri are often defined in different textual formats and therefore require different parsers. Another good example is the possibility to add new graphical output formats. A scatter-plot, for example, displays data in an other way than a histogram. These visualisations can result in different insightful observations. The workbench also allows to quite easily add new measures, especially binary evaluation measures. The reason for this is that the workbench only stores the fundamental variables of a binary evaluation, i.e. the number of true positives, false positives, true negatives and false negatives per item and class. This has two advantages. First of all, the workbench does not need to store a value for all the different binary measurements. Secondly, the user can easily change parameters such the β in the F-measure without the need to recalculate all true positives, false positives, true negatives and false negatives for each item and class. One can thus easily introduce new binary evaluation measures into the workbench without having to re-evaluate the data.

Interoperability

The workbench perfectly fits into the development process of GridLine as shown in Figure 6.10. It was chosen to implement the workbench as an independent piece of software capable of coping with the different classifiers of GridLine. This is supported by the flexible input functionalities of the workbench. The workbench, furthermore, allows to evaluate any type of classification problem, i.e. binary classifications, multi-class classifications and multi-label classifications. One should also note that the workbench is a web application. It therefore only needs to be installed on one single machine: the server. This allows users to approach the workbench through any machine connected with the server. The user also has the choice to operate the workbench with any modern web browser.

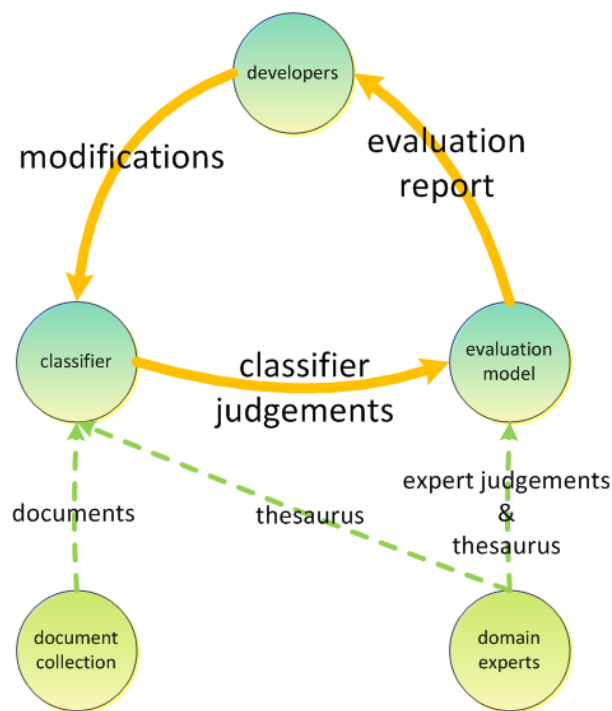


Figure 6.10: Development process.

Robustness

As mentioned earlier, all input provided to the workbench by the user is checked on correct format. Feedback is given on whether a certain input has been accepted or not. Since the input of a classification problem is provided in different files there might be some inconsistencies between the content described in the files. The workbench has been developed to cope with these inconsistencies. An overview of how the workbench handles these inconsistencies is given as follows:

- The workbench only evaluates the items that appear both in the expert and classifier judgements. This functionality allows the user to load all the expert judgements (gold standard) as one single dataset into the workbench. Note that when

comparing two classifier judgements at once only the items appearing in both judgements are taken into consideration.

- The workbench considers every class specified in one of the input files. This means that a class appearing in one of the judgements is always included, even if it does not appear in the file containing all the classes (or in the structure of the thesaurus). Note that, in the case of a thesaurus, our relatedness measure might try to calculate the semantic relatedness between classes that might not appear in the thesaurus. In this case the algorithm assigns a semantic relatedness value of zero.

6.3 Components

In this section some general implementation aspects are described. These aspects basically all relate to the architecture of the workbench.

6.3.1 Technologies

The workbench is built in Play ⁴. Play is a Java and Scala web application framework. It follows the Model-View-Controller pattern. This pattern splits the applications into three layers: a model, view and controller layer. All layers have different responsibilities. The model layer is responsible for representing the domain-specific information, the view for rendering the graphical user interface and the controller for responding to events triggered by users. Figure 6.11 shows the routing mechanisms between the layers of the workbench.

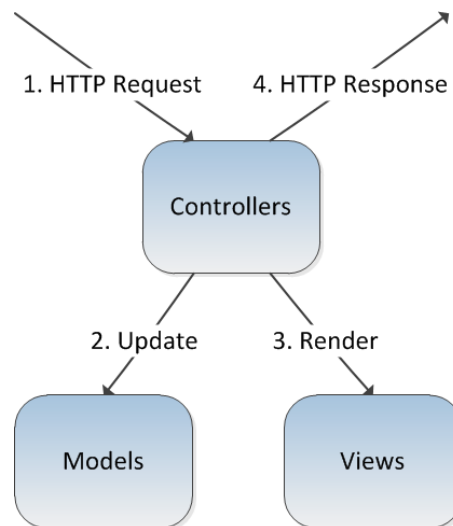


Figure 6.11: System routing.

The back-end of the workbench is entirely written in Java. The Hibernate library is used to map the Java object-oriented domain model to a traditional relational database.

⁴<http://www.playframework.org/>

All the data is stored in a MySQL database. The front-end is built upon several web-based technologies. It is styled with TwitterBootstrap ⁵, a tool-kit providing simple and flexible HTML, CSS and JavaScript for popular user interface components. Play itself makes use of an efficient templating system allowing to dynamically generate web pages. The template engine uses Scala as an expression language. A schematic view of all technologies used in the workbench is given in Figure 6.12.

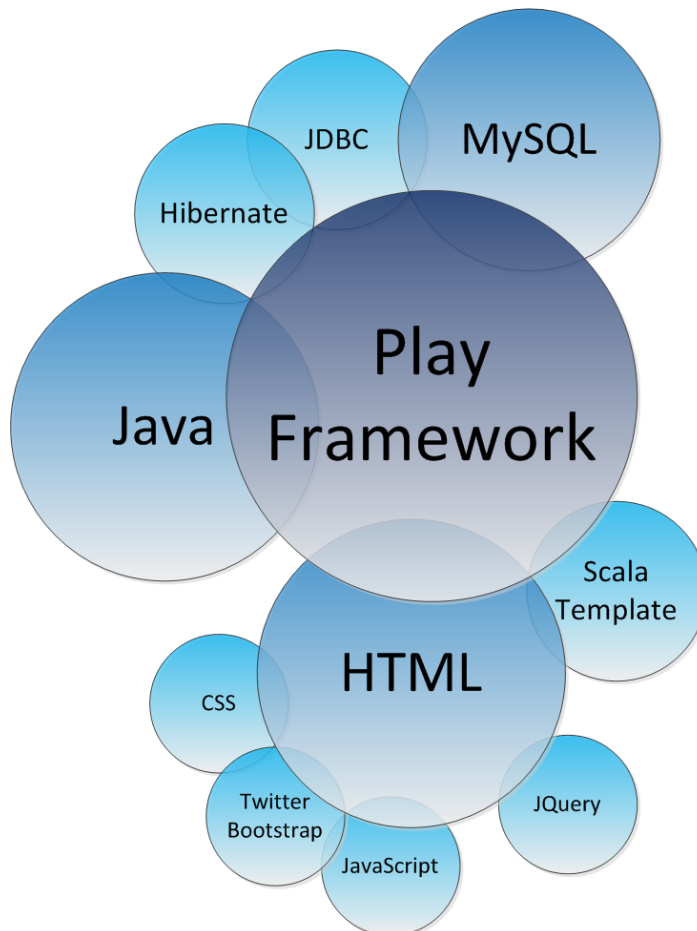


Figure 6.12: System technologies.

6.3.2 Modules

The system itself is divided into modules. This is shown in Figure 6.13. The modules Models, Views and Controllers are part of the three layer model of Play. The remaining modules are primarily components containing operational procedures. The code defining all input and output functionalities are all included into, respectively the Writers and Readers module. The algorithms used for evaluating a classification are located in the Evaluators module. These algorithms first construct item-based and class-based bi-partitions from the expert and classifier judgements after which the binary and re-

⁵<http://twitter.github.com/bootstrap/>

latedness evaluations are conducted. To reduce the computationally expensive tasks as much as possible it was chosen to only calculate intermediate statistics for the binary evaluations. This means that the workbench only stores the true positives, false positives, true negatives and false negatives for every item and class. The final result of a binary evaluation measure is then simply calculated on the fly as soon as a user asks to view it. As described earlier, this methodology has two advantages. First of all, the workbench does not need to store a value for all the different binary measurements. Secondly, the user can easily change parameters such the β in the F-measure without the need to recalculate all true positives, false positives, true negatives and false negatives for every item and class. For relatedness evaluations, however, no partial results are stored. The final result is computed and stored for each item at once. The relatedness evaluations rely on the Thesaurus module. This module contains all code necessary to calculate a relatedness value between two terms of a thesaurus. The Manager module is responsible for providing the statistics as shown to the user. This module delegates the calculation of item-based and class-based statistics to the Measures module. This operation is best illustrated with an example. Suppose that a user asks to view the item-based statistics of the F-measure. The Manager module will ask the F-measure class of the Measures module to calculate all item-based F-measure statistics. This means an F-measure for each item individually as well as the micro and macro aggregates including all items. These statistics are calculated from the (intermediate) results stored in the database.

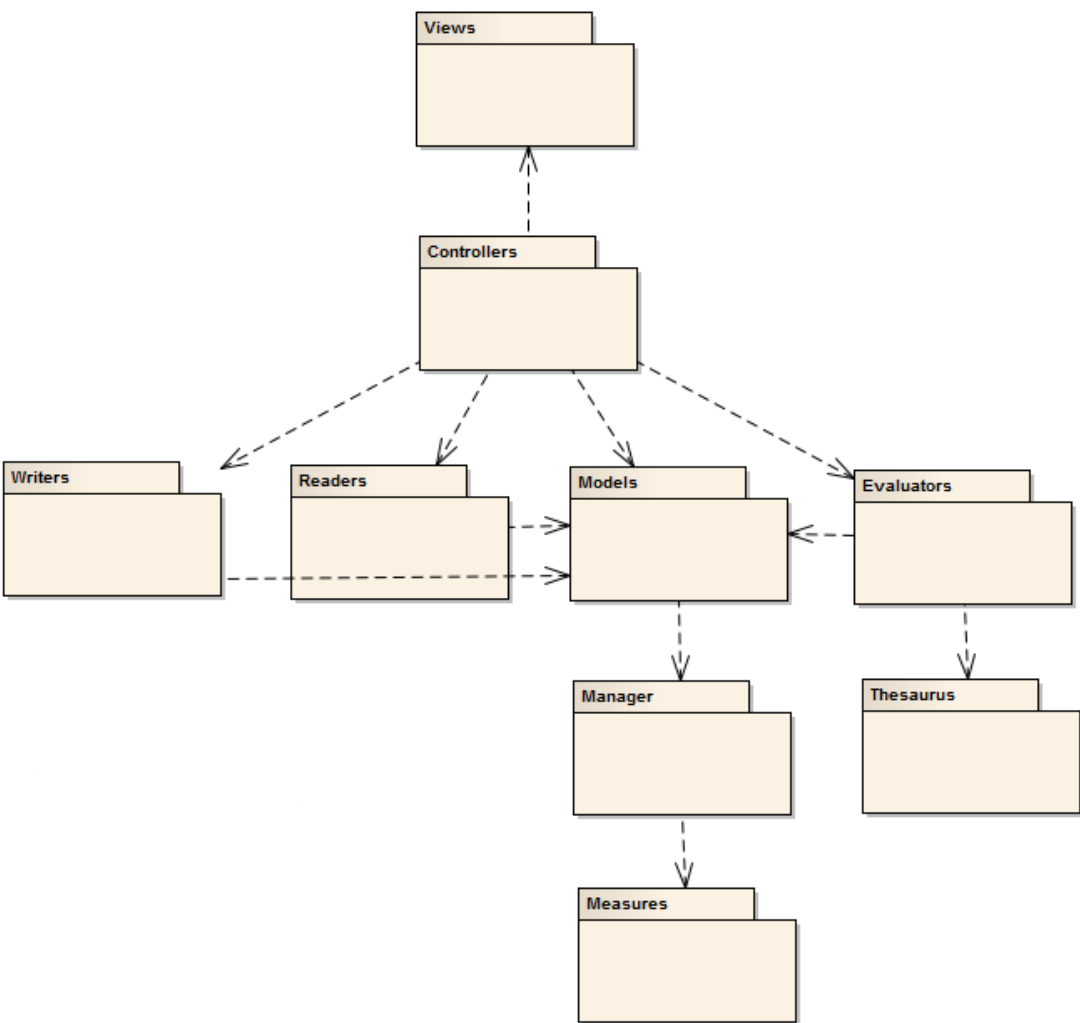


Figure 6.13: System modules.

Chapter 7

Conclusions

This chapter concludes this thesis. In Section 7.1 the research questions posed at the beginning of this work are recapitulated and in Section 7.2 the major contributions are summarised. Finally, some key directions for future research are given in Section 7.3.

7.1 Summary

The work in this thesis can be basically divided into two parts:

1. The research on evaluating classifications by means of semantic relatedness.
2. The specification and implementation of a framework for the evaluation of classifiers.

The research conducted in the first part takes place in the context of a real-world multi-label classification problem: annotating text documents originating from the Dutch Parliament with terms from a predefined thesaurus. The research consists of two stages. The first stage defines and validates the mechanism responsible for quantifying the relatedness between two labels whereas the second stage defines and validates the mechanism responsible for quantifying the relatedness between two sets of labels. The second mechanism works on top of the first mechanism. Each stage provides an answer to its corresponding research question posed in Chapter 1.

The first stage closely follows the research methodology of Resnik [23]. In this stage an accurate selection of state-of-the-art graph-based relatedness measures is evaluated against human judgements of relatedness. The most promising relatedness measure, i.e. the measure that correlates most closely with human judgements of relatedness is then used as the mechanism that is responsible for quantifying the relatedness between two labels. Given this situation, this thesis aims to answer the following research question:

How well do the different relatedness measures perform in accordance to human judgements of relatedness?

The research conducted in this stage is described in Chapter 4. A total of 8 different graph-based measures have been examined, including both similarity and relatedness

measures. The result of this research show that the relatedness measures, i.e. the measures including the term related links, outperform the similarity measures. Apparently a lot of relatedness information is contained in the related term relations, which is actually not very surprising due to its naming. One should therefore not limit oneself to the is-a structure of a semantic graph when quantifying relatedness between two terms. It is very well possible that strongly related terms appear far away in the is-a structure of the graph. The best results are therefore obtained by also including related term relations in determining the degree of relatedness. It was chosen to use Leacock & Chodorow's relatedness measure for the label-relatedness mechanism since it gave the best results.

In the second stage the set-relatedness mechanism is defined. This mechanism defines a procedure for comparing two sets of labels: the labels from the classifier judgements and the labels from the expert judgements. The set comparison procedure is based on a simplified version of Nowak's algorithm [17]. The set-relatedness mechanism that includes the label-relatedness mechanism and therefore also defines the relatedness evaluation measure is evaluated against the judgements of classification experts. Classification experts have been asked to rate a few document classifications on correctness, i.e. the better the classification, the higher the rating. The ratings that have been assigned are used to evaluate the output produced by the relatedness evaluation measure and also against one frequently used binary evaluation measure, the F-measure. This experiment gives an answer to our second research question:

How well does the relatedness evaluation measure perform according to the judgements of classification experts? Does it give more accurate evaluations than the commonly used F-measure?

The research leading to the answer of this question is described in Chapter 5. From the results reported in this chapter we can conclude that the results of our relatedness evaluation measure look promising. The scores attributed by our evaluation measure provide a more accurate evaluation than the scores of the F_1 -measure. It should be noted however that the experiment only covers 25 document classifications. This is a very small sample. Ideally this experiment should be carried out on a much larger scale. One should also be aware that the reliability of these evaluations strongly depends on the external resource used to extract a relatedness score between two labels. Take for example our case where a relatedness score is calculated based on a semantic network. It is important that the concepts contained by the network are semantically well organised. The better the semantic structure of the network, the higher the reliability of the relatedness evaluations. Another reason in favour of binary evaluation measures is their universal applicability. A binary measure like the F-measure is widely known and its exact definition can be easily looked up. Its meaning is therefore well understood. This all suggests that our relatedness evaluation measure should not be used instead of the traditional binary evaluation measures. It just provides a different point of view. The two types of measures are complementary and should be used side by side.

The knowledge on classifier assessment acquired in this study is applied in practice in the form of an evaluation framework. The evaluation framework is a workbench which has been developed for the purpose of automatic classifier assessment and was tailored to the needs of GridLine. It comprises several techniques for measuring and

analysing the performance of classifiers including both the common binary evaluation measures as well as our relatedness evaluation measure. The measures quantify the performance of a classifier by basically comparing the judgements produced by the classifier to the judgements of experts, i.e. the ground-truth. During development, special attention was paid to the usability, extensibility, interoperability and robustness of the workbench.

7.2 Contributions

The main contributions of this work include the research on an evaluation measure incorporating semantic relatedness and the specification and implementation of an evaluation framework. It can be summarised as follows:

1. The research on an evaluation measure incorporating semantic relatedness.
 - Detailed overview of state-of-the-art semantic relatedness measures.
 - Thorough examination of a relatedness evaluation measure. This examination takes place in the context of a real-world classification task.
2. The specification and implementation of an evaluation framework.
 - Detailed specification describing the evaluation framework (including its features, competencies and components).
 - Implementation of an evaluation framework for use in industry.

7.3 Future work

A few interesting areas of future work have been identified. The future work can be divided into work concerning the relatedness evaluation measure and work concerning the design and features of the evaluation framework.

7.3.1 Relatedness evaluation measure

Our relatedness evaluation measure computes degrees of relatedness based on a graph structure in which concepts are semantically organised. The measure therefore strongly relies on a semantic graph structure. The better the semantic structure, the higher the reliability of the relatedness evaluations. As apparent from the previous chapters, semantic networks can be defined in several ways. It is thus important to have an understanding of what kind of graph characteristics can positively or negatively influence the performance of our semantic relatedness measure. In Chapter 4 it was shown that the state-of-the-art relatedness measures perform significantly different on various graph structures and that this could be possibly due to the difference in strictness in the semantic meaning of the hierarchical graph relations. But are there any other factors that might positively or negatively influence the performance of graph-based semantic relatedness measures?

Another topic worth researching is the height of the weights assigned to the various types of relations in the semantic graph. In this work weights were assigned

intuitively. Synonymous term relations received a weight of zero and hierarchical and related term relations both a weight of one. The question however is whether we could determine more accurate weights based on sound scientific reasoning and explanation by performing user-studies for example.

7.3.2 Evaluation framework

The evaluation framework provides statistics in a tabular format. Statistics are either given per class or item or as an average over all classes or items. In this way a good global view is provided with the averages and in-depth details are provided by the individual class and item statistics. The problem is that the information provided by the individual statistics become less useful as the amount of classes or items increases. Take for example our Parliament classification task which includes nearly 8,000 terms. This amount is so large that it becomes somehow useless to look at individual cases. A possible solution to this problem is to present overviews of the individual statistics in visualisations. At this time the workbench only contains one such visualisation function as was presented in Chapter 6. The workbench however could be easily extended with additional visualisations. It would be interesting to explore if there are other, more effective ways of visualising the data. One could for instance consider using different types of charts or charts with interactive functionalities.

As mentioned before, the evaluations provided by the workbench do not only show aggregate statistics but also statistics per item and class. A user can easily get access to different types of measurements for each individual item and class. The missing feature however, is that a user is not able to directly inspect the classes associated with each of the items. For this the user needs to open the dataset page and browser through it. This requires the user to perform multiple operations. Preferably, one would like to provide this information to the user in one mouse click. For example, by showing a pop-up window when a user clicks on the identification label of an item. The same also more or less counts for a class. The difference however is that the workbench currently does not provide any overview which allows users to inspect the items associated with a class. This makes such a functionality even more valuable. The previously described feature is not very difficult to implement but will take up some time to implement since it requires changes in several front-end and back-end components of the workbench.

The workbench could also improve in user-friendliness. Although the workbench always informs the user if an action has succeeded or not, it not always give a proper description of the problem in case of failure. For example, when the insertion of a new dataset fails due to an improper format the following error is given: “Dataset has not been created”. This error is very general and does not give the reason of failure. Other improvements could be achieved by providing progress feedback for operations of long duration. Currently there is actually only one operations that can take up a significant amount of time: the creation of a new evaluation. Its duration depends on several factors, i.e. the number of items, the average amount of labels assigned to each of the items, the number of classifier judgements (one of two) and if a thesaurus is included or not. Improving and introducing some of the previously described features can drastically increase the user-friendliness of the workbench and are technically not difficult to implement. Investing in features like this should be given a high priority.

Bibliography

- [1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 805–810. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.
- [2] A.P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- [3] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, 2001.
- [4] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
- [5] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer Publishing Company, Incorporated, 2010.
- [6] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence*, pages 1606–1611, 2007.
- [7] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. *Advances in Knowledge Discovery and Data Mining*, pages 22–30, 2004.
- [8] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.
- [9] D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
- [10] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 13:305–332, 1998.
- [11] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Arxiv preprint cmp-lg/9709008*, 1997.

- [12] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.
- [13] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.
- [14] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.
- [15] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, volume 1, pages 296–304. San Francisco, 1998.
- [16] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.
- [17] S. Nowak, A. Llorente, E. Motta, and S. Rüger. The effect of semantic relatedness measures on multi-label classification evaluation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 303–310. ACM, 2010.
- [18] S. Nowak, H. Lukashevich, P. Dunker, and S. Rüger. Performance measures for multilabel evaluation: a case study in the area of image classification. In *Proceedings of the international conference on Multimedia information retrieval*, pages 35–44. ACM, 2010.
- [19] S. Patwardhan. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, 2003.
- [20] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. *Computational Linguistics and Intelligent Text Processing*, pages 241–257, 2003.
- [21] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, and C.G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.
- [22] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30, 1989.
- [23] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Arxiv preprint cmp-lg/9511007*, 1995.
- [24] H. Rubenstein and J.B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
- [25] H. Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123, 1998.

- [26] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
- [27] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, volume 16, page 1089. Citeseer, 2004.
- [28] X. Shen, M. Boutell, J. Luo, and C. Brown. Multi-label machine learning and its application to semantic scene classification. In *Proceedings of the 2004 International Symposium on Electronic Imaging (EI 2004)*, pages 18–22, 2004.
- [29] M. Strube and S.P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [30] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on Information and knowledge management*, pages 67–74. ACM, 1993.
- [31] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
- [32] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Machine Learning: ECML 2007*, pages 406–417, 2007.
- [33] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.
- [34] I.H. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, AAAI Press, Chicago, USA, pages 25–30, 2008.
- [35] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.
- [36] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 1999.

Appendix A

Survey Results: Term Pair Relatedness

In this appendix an overview of the results is given for the term pairs relatedness survey. This overview is shown in Table A.1. The table gives the statistics per term pair. Three different statistics are indicated: mean, deviation and unrated. The mean is the average over all ratings per term pair, the deviation is the mean absolute deviation over all ratings per term pair and the unrated value gives the number of times a term pair has not been rated. The term pairs in the table are sorted by their mean in ascending order.

first term	second term	mean	deviation	unrated
Vaticaanstad	paus	4.86	0.25	2
energiebedrijven	elektriciteitsbedrijven	4.72	0.43	1
rijbewijzen	rijbevoegdheid	4.64	0.55	0
kiezers	gemeenteraadsverkiezingen	4.63	0.5	0
aangepaste woningen	gehandicapten	4.49	0.6	0
voedselvoorziening	honger	4.47	0.58	0
natuurbehoud	oerwouden	4.34	0.57	0
verzorgingsstaat	welzijnsstaat	4.32	0.55	1
educatieve omroep	schoolradio	4.19	0.61	0
regerakkoord	portefeuilleverdeling	4.18	0.69	1
verpakkingen	productinformatie	4.16	0.59	0
bruggen	tunnels	4.15	0.64	0
bevelhebbers	bewind	4.11	0.69	1
windenergie	kerncentrales	4.05	0.61	0
procesrecht	cassatie	4.03	0.63	15
standbeelden	gedenktekens	4.03	0.7	2
Molukkers	Indonesië	4.02	0.82	8
filosofie	ethiek	3.93	0.69	3
defensiepersoneel	overheidspersoneel	3.88	0.66	0
belangenorganisaties	petities	3.84	0.65	2
landinrichting	streekplannen	3.82	0.57	1
waterwegen	meren	3.71	0.74	1
gezondheidszorg	maatschappij	3.52	0.74	1
automatisering	telecommunicatie	3.51	0.78	1
sportvisserij	watersport	3.5	0.94	0
groenten	vlees	3.44	0.87	0
beroepsbevolking	uitzendkrachten	3.41	0.91	1
Verkeer en Waterstaat	staat	3.41	0.91	1

inrichting voor jeugdigen	cellen	3.34	0.79	1
amateurkunst	muziekscholen	3.31	0.85	1
politiek	recht	3.2	0.81	1
onderwijs	welzijn	3.14	0.8	0
privaatrecht	faillissementen	2.8	0.96	9
nationaal inkomen	directe investeringen	2.74	0.84	4
welzijn	wereld	2.69	0.95	1
toerisme	consumptiegoederen	2.57	0.75	0
openbaar onderwijs	studentenverenigingen	2.56	0.82	1
e-mail	radio	2.23	0.84	0
onderverhuur	woonwagens	2.18	0.83	0
sport	overheid	2.09	0.72	1
Turken	Japan	1.94	0.79	2
Malta	Donau	1.93	0.87	5
wijkagenten	ambassadeurs	1.8	0.67	1
kernreactorongevallen	spoorwegpolitie	1.46	0.61	1
reclame	monumenten	1.43	0.57	0
inentingen	fouilleringen	1.4	0.56	5
onderwijsraad	zigeuners	1.24	0.39	0
boedelscheiding	gentehnologie	1.2	0.34	8
roken	dieren	1.17	0.29	0
makelaars	republiek	1.16	0.28	1

Table A.1: Overview of results from term pairs relatedness survey.

Appendix B

Example: Classified Document

In this appendix an example of a classified document is shown. Each document is identified with a unique number. A document is classified on its content. The content of a document typically contains a title and a body. The expert judgements are used to evaluate the classifier judgements. The classifiers judgements have been automatically assigned by a classifier of GridLine and the expert judgements have been manually assigned by domain experts employed by the Dutch government.

DOCUMENT ID: 4594871

CONTENT:

Nog geen beslissing nodig over plaatsing kernraketten.

Frinking analyseert 4 overwegingen van waaruit een Nederlands veiligheidsbeleid gevoerd zou kunnen worden. Zijn overwegingen betreffen: 1. Een volwaardig Nederlands NAVO-lidmaatschap. 2. Een goede uitvoering van de slotakte van Helsinki. 3. Het beschikken over nucleaire wapens en het uitvoeren van nucleaire gevechtstaken. 4. Het NAVO-dubbelbesluit over stationering van 572 kernraketten in Nederland. Dit laatste punt is volgens Frinking reeds een feit. De beslissing van Nederland zelf over de plaatsing hangt af van de voortgang van de onderhandelingen over wapenbeheersing in Genève. Nederland moet volgens Frinking een duidelijke tijdslimiet aangeven, waarbinnen zij de beslissing neemt. Bij een gunstige voortgang van de onderhandelingen in Genève is er geen aanleiding tot een voorlopige plaatsing. Bij voldoende succes in Genève bestaat de mogelijkheid dat plaatsing overbodig wordt. Alleen duidelijke onwil van de kant van de Sovjet-Unie kan aanleiding zijn een principe-besluit over plaatsing te nemen.

EXPERT JUDGEMENTS:

KERNWAPENOPSLAG
WAPENBEHEERSING

CLASSIFIER JUDGEMENTS:

MILITAIR MATERIEEL
WAPENS
INTERNATIONALE ORGANISATIES
WAPENBEHEERSING
KERNWAPENS

Appendix C

Survey Results: Document Classification Quality

In this appendix an overview of the results is given for the document classification quality survey. This overview is shown in Table C.1. The table gives the statistics per document. Three different types of statistics are indicated: the ratings by all three experts, the mean and the deviation. The mean is the average over all ratings per document and the deviation is the mean absolute deviation over all ratings per document. The mean and deviation are rounded to two decimals. The entries in the table are sorted by document id.

document id	expert 1	expert 2	expert 3	mean	deviation
2554349	4	3	3	3.33	0.44
2591088	4	4	4	4	0
2614430	4	3	3	3.33	0.44
3833254	5	4	5	4.67	0.44
3872172	5	5	5	5	0
3877208	5	4	5	4.67	0.44
3880651	4	3	4	3.67	0.44
3924639	1	1	1	1	0
4183938	4	3	4	3.67	0.44
4197111	3	1	3	2.33	0.88
4285490	2	2	2	2	0
4318676	2	3	3	2.67	0.44
4320359	3	4	3	3.33	0.44
4339282	3	3	2	2.67	0.44
4354700	3	3	3	3	0
4369129	2	1	3	2	0.67
4386070	3	4	4	3.67	0.44
4414807	5	4	3	4	0.67
4434266	4	3	5	4	0.67
4463034	1	1	3	1.67	0.88
4509554	3	4	5	4	0.67
4594871	5	4	5	4.67	0.44
4595226	5	4	4	4.33	0.44
4604487	3	2	1	2	0.67
4616008	2	2	3	2.33	0.44

Table C.1: Overview of results from document classification quality survey.

Appendix D

Glossary

In this appendix an overview of frequently used terms and abbreviations is given.

binary evaluation: a classification evaluation technique based on the variables of the confusion matrix, i.e. true positives, false positives, true negatives and false negatives

classifier: a classification system or algorithm

classifier judgements: classifications produced by a classifier

concept: an intuited object of thought

expert judgements: classifications produced by domain experts

multi-label classification: a variant of the classification problem where each item is assigned to one or more classes

relatedness evaluation: a classification evaluation technique incorporating the notion of semantic relatedness

relatedness measure: a measure that quantifies the degree to which a given pair of concepts is related

semantic relatedness: the degree to which a given pair of lexically expressed concepts is related according to human judgements

term: lexical expression of a concept

thesaurus: a controlled vocabulary in which relations between concepts are made explicit