Exploiting Semantic Relatedness Measures for Multi-label Classifier Evaluation

Christophe Deloo[*]
Delft University of Technology
Delft, The Netherlands
c.p.p.deloo@gmail.com

Claudia Hauff
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

ABSTRACT

In the multi-label classification setting, documents can be labelled with a number of concepts (instead of just one). Evaluating the performance of classifiers in this scenario is often as simple as measuring the percentage of correctly assigned concepts. Classifiers that do not retrieve a single concept existing in the ground truth annotation are all considered equally poor. However, some classifiers might perform better than others, in particular those, that assign concepts which are semantically similar to the ground truth annotation. Thus, exploiting the semantic relatedness between the classifier-assigned and the ground truth concepts leads to a more refined evaluation. A number of well-known algorithms compute the semantic relatedness between concepts with the aid of general-world knowledge bases such as WordNet[1]. When the concepts are domain specific, however, such approaches cannot be employed out-of-the-box. Here, we present a study, inspired by a real-world problem, where we first investigate the performance of well-known semantic relatedness measures on a domain-dependent thesaurus. We then employ the best performing measure to evaluate multi-label classifiers. We show that (i) measures which perform well on WordNet do not reach a comparable performance on our thesaurus and that (ii) an evaluation based on semantic relatedness yields results which are more in line with human ratings than the traditional F-measure.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
Keywords: semantic relatedness, classifier evaluation

1. INTRODUCTION

In this paper, we present a two-part study, that is inspired by the following real-world problem: Dutch Parliamentary papers[2] are to be annotated with concepts from an existing thesaurus[3] (the *Parliament thesaurus*). A multi-label classifier framework exists and each document can be automatically annotated with a number of concepts. Currently, the evaluation of the classifier is conducted as follows: the automatically produced annotations are compared to the ground-truth (i.e. the concepts assigned by domain experts) and the binary measures of precision and recall are computed. This means, that a document labelled with concepts which do not occur in the ground truth receives a precision/recall of zero, even though the assigned concepts may be semantically very similar to the ground truth concepts. As an example, consider Figure 1: the ground truth of the document consists of three concepts {*biofuel, environment, renewable energy*} and the classifier annotates the document with the concepts {*energy source, solar energy*}. Binary precision/recall measures evaluate the classifier's performance as zero, though it is evident, that the classifier does indeed capture the content of the document - at least partially.

Thus, we are faced with the following research question: *Can the evaluation of a multi-label classifier be improved when taking the semantic relatedness of concepts into account?*

To this end, we present two studies (Figure 1):

1. We investigate established semantic relatedness measures on the Parliament thesaurus. Are measures that perform well on WordNet or Wikipedia also suitable for this domain-specific thesaurus?
2. Given the best performing relatedness measure, we include the semantic relatedness in the evaluation of the multi-label classifier framework and investigate if such a semantically enhanced evaluation improves over the binary precision/recall based evaluation.

We find that the best performing measures on WordNet do not necessarily perform as well on a different thesaurus, and thus, they should be (re-)evaluated when a novel thesaurus is employed. Our user study also shows that a classifier evaluation, which takes the semantic relatedness of the ground truth and the classifier assigned concepts into account yields results which are closer to those of human experts than traditional binary evaluation measures.

[*] This research was performed while the author was an intern at GridLine.

[1] `http://wordnet.princeton.edu/`

[2] The documents come from the Dutch House of Representatives (de Tweede Kamer), which is the lower house of the bicameral parliament of the Netherlands.

[3] For more details see Section 3.

DIR 2013, April 26, 2013, Delft, The Netherlands.

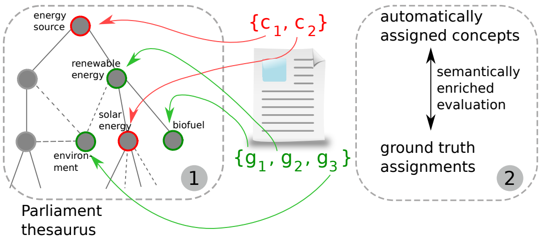

Figure 1: Overview of the two-step process: (1) we first investigate semantic relatedness measures on the Parliament thesaurus. Then, (2) given a document and its assigned ground truth concepts $\{g_1, g_2, g_3\}$ (by human annotators), we evaluate the quality of the classifier-assigned concepts $\{c_1, c_2\}$. The classifier evaluation takes the semantic relatedness between the concepts into account.

2. RELATED WORK

In this section, we first discuss semantic relatedness measures and then briefly describe previous work in multi-label classifier evaluation.

Several measures of semantic relatedness using a variety of lexical resources have been proposed in the literature. In most cases semantic relations between concepts are either inferred from large corpora of text or lexical structures such as taxonomies and thesauri. The state-of-the-art relatedness measures can be roughly organised into graph-based measures [11, 6, 19, 4, 16], corpus-based measures [17, 10] and hybrid measures [12, 5, 7, 1]. The latter combine information gathered from the corpus and the graph structure.

The majority of relatedness measures are graph-based and were originally developed for WordNet. WordNet is a large lexical database for the English language in which concepts (called synsets) are manually organised in a graph-like structure. While WordNet represents a well structured thesaurus, its coverage is limited. Thus, more recently, researchers have turned their attention to Wikipedia, a much larger knowledge base. Semantic relatedness measures originally developed for WordNet have been validated on Wikipedia. Approaches that exploit structural components that are specific to Wikipedia have been developed as well [14, 18, 3].

With respect to multi-label classifier evaluation, our work builds in particular on Nowak et al. [9]. The authors study the behavior of different semantic relatedness measures for the evaluation of an image annotation task and quantify the correctness of the classification by using a matching optimisation procedure that determines the lowest cost between the concept sets of the ground truth and of the classifier.

We note, that besides semantic relatedness measures one can also apply hierarchical evaluation measures to determine the performance of multi-label classifiers, as for instance proposed in [15]. We leave the comparison of these two different approaches for future work.

3. METHODOLOGY

Semantic Relatedness in the Parliament Thesaurus.

We first investigate the performance of known semantic relatedness measures on our domain-specific thesaurus (Figure 1 step (1)). 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. In order to evaluate the different measures, we employ an established methodology: we select a number of concept pairs from our thesaurus and ask human annotators to judge the relatedness of the concepts on a 5-point scale (where 1 means *unrelated* and 5 means *strongly related*). We consider these judgements as our ground truth and rank the concept pairs according to their semantic relatedness. Then, we also rank the concept pairs according to the scores they achieve by the different semantic relatedness measures. The agreement between the two rankings is evaluated with the rank correlation measure Kendall's Tau (τ) and the linear correlation coefficient (r).

The Parliament thesaurus contains nearly $8,000$ Dutch terms oriented towards political themes such as defense, welfare, healthcare, culture and environment. As is typical for a thesaurus, the concepts are hierarchically structured and the following three types of relations exist: hierarchical (narrower/broader), synonymy and relatedness. Fifty concept pairs were manually selected by the authors, with the goal to include as many different characteristics as possible, that is, concept pairs of varying path lengths, types of relations, etc. The human ratings were obtained in an electronic survey where Dutch speaking people were asked to rate the fifty concept pairs on their relatedness. As stated earlier, in the 5-point scale, the higher the assigned rating, the stronger the perceived relatedness.

The following relatedness measures were selected for our experiments: Rada [11], Leacock & Chodorow [6], Resnik [12], Wu & Palmer [19], Jiang & Conrath [5] and Lin [7]. The measures of Rada, Leacock & Chodorow and Wu & Palmer are all graph-based measures 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. Two versions of both Rada's and Leacock & Chodorow's approach were implemented: one including only hierarchical and synonymous relations, and one including all three types of thesaurus relations. The weights of the relations were chosen according to their semantic strength. A weight of 1 was assigned to both hierarchical and related concept relations and a weight of 0 to synonymous concept relations. The remaining three approaches, which are based on the concept of information content, were implemented using the approach of Seco et al. [13].

Multi-label Classifier Evaluation.

Having identified the best performing measure of semantic relatedness on the Parliament thesaurus, we then turn to the evaluation of the existing multi-label classifier framework (Figure 1 step (2)). Matching the concepts from the classifier with the ground truth concepts is performed according to a simplified version (which excludes the ontology and annotator agreement) of the procedure presented in [9]. Nowak et al. define a classification evaluation measure that incorporates the notion of semantic relatedness. The algorithm calculates the degree of relatedness between the set C of classifier concepts and the set E of ground truth concepts with an optimisation procedure. This procedure pairs every label of both sets with a label of the other set in a way that maximises relatedness: each label $l_c \in C$ is matched with a label $l'_e \in E$ and each label $l_e \in E$ is matched with a label $l'_c \in C$. The relatedness values of each of those pairs are summed up and divided by the number of labels occurring

Concept pairs		Av. rating	Std. Dev.
Vaticaanstad *Vatican City*	paus *pope*	4.86	0.25
energiebedrijven *power companies*	elektriciteitsbedrijven *electricity companies*	4.72	0.43
rijbewijzen *driver licenses*	rijbevoegdheid *qualification to drive*	4.64	0.55
	...		
boedelscheiding *derision of property*	gentechnologie *gene technology*	1.2	0.34
roken *smoke*	dieren *animals*	1.17	0.29
makelaars *broker*	republiek *republic*	1.16	0.28

Table 1: Shown are the three concept pairs from our annotation study achieving the highest and the lowest average rating respectively (in Dutch and English).

in both sets. This yields a value in the interval $[0, 1]$. The higher the value, the more related the sets. Formally:

$$\frac{\sum\limits_{l_c \in C} \max\limits_{l'_e \in E} rel(l_c, l'_e) + \sum\limits_{l_e \in E} \max\limits_{l'_c \in C} rel(l_e, l'_c)}{|C| + |E|} \tag{1}$$

To validate this measure we conduct a study with human experts: three expert users, who are familiar with the thesaurus and the documents, were asked to judge for twenty-five documents the relatedness between the ground truth concepts and the classifier assigned concepts (taking the content of the document into account) on a 5-point scale: *very poor*, *poor*, *average*, *good* and *very good.* It should be emphasised, that our expert users have not created the ground truth concepts (those were created by library experts employed by the Dutch government). The average rating taken over all three individual expert ratings are considered as the ground-truth. The expert evaluations are used to compare the performance of the relatedness evaluation measure and the performance of a frequently used binary evaluation measure (F-measure). We hypothesise, that the classifier evaluation, which takes the semantic relatedness of the concepts into account will correlate to a larger degree with the expert judgements than the traditional binary evaluation measure.

4. EXPERIMENTS & RESULTS

Semantic Relatedness in the Parliament Thesaurus.

Examples of concept pairs that were selected for the annotation study are shown in Table 1; in particular the three concept pairs yielding the highest human annotator relatedness scores and the lowest scores respectively are listed.

The performance of the relatedness measures on the Parliament thesaurus are listed in Table 2. From these results two aspects stand out: (i) the relatively high correlation obtained for Rada's and Leacock & Chodorow's relatedness measure, and, (ii) the relatively poor performance of the remaining measures.

Traditionally, semantic relatedness measures have been evaluated on WordNet, the most well-known manually created lexical database. Seco et al. [13] evaluated all measures from our selection (except Rada) in a similar way on the WordNet graph against a test-bed of human judgements provided by Miller & Charles [8]. They reported significant higher correlations for the selected relatedness measures. Their correlation results range from 0.74 (Wu & Palmer) to 0.84 (Jiang & Conrath) and are in line with similar studies on WordNet such as Budanitsky et al. [2]. We conclude that measures which perform best on WordNet are not performing as well on our domain-dependent Parliament thesaurus.

Measures	r	τ
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 2: Overview of the correlations of relatedness measures with human judgements of relatedness.

Classifier	Ground truth	Av. rating
toelating vreemdelingen	vreemdelingenrecht vreemdelingen procedures werknemers vluchtelingen	4.67
kinderbescherming kindermishandeling	jeugdigen gezondheidszorg	3.67

Table 3: Two examples of assigned classifier concepts vs. ground truth concepts and the average of the ratings obtained from the three experts users.

Multi-label Classifier Evaluation.

In Table 3 two examples of assigned classifier concepts vs. ground truth concepts are shown. Reported are also the average ratings obtained from the three expert users. Across all 25 evaluated documents, the mean rating was 3.28, indicating that the classifier framework performs reasonably well at assigning concepts related to the ground truth concepts.

Correlation	Semantically-enhanced	$\mathbf{F_1}$
r	0.67	0.48
τ	0.53	0.37

Table 4: Correlations between the expert ratings and the semantically-enhanced and the binary (F_1) classifier evaluation respectively.

The results of the second experiment are summarised in Table 4. Here, we employed Leacock & Chodorow's relatedness as it was our best performing approach (Table 2). The results indicate that for the annotated set of twenty-five documents, the relatedness evaluations correlate more with the expert evaluations than the evaluation based on F_1. The coefficients report an increase in correlation of at least 0.16 in favour of the relatedness evaluations. To emphasise the difference, we also present the scatter plots of the semantically-enhanced (Figure 2) and the binary, F_1 based, evaluation (Figure 3). In both plots, the corresponding trend line is drawn in red. It is evident, that in the binary case, the number of $F_1 = 0$ entries has a significant

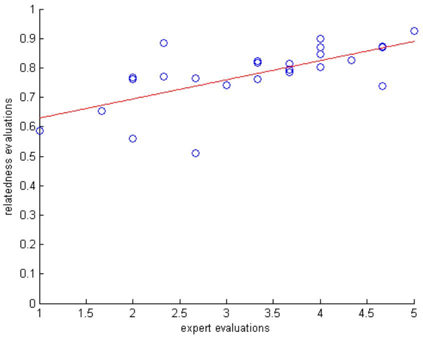

Figure 2: Expert versus relatedness evaluations.

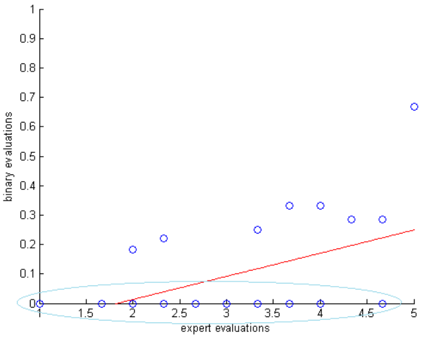

Figure 3: Expert versus binary evaluations.

impact on the obtained correlation. Note that the dispersion of relatedness evaluations in Figure 2 is higher at lower expert evaluations compared to higher expert evaluations. Whether this observation is to be attributed to noise is impossible to say due to the small size of the evaluation. We will investigate this issue further in future work.

5. CONCLUSIONS

In this paper, we have presented a two-step procedure to tackle a real-world problem: namely, the semantically-enhanced evaluation of multi-label classifiers that assign concepts to documents. We first investigated to what extent semantic relatedness measures that perform well on the most commonly used lexical database (WordNet) also perform well on another thesaurus (our domain-specific Parliament thesaurus). To this end, we conducted a user study where we let approximately 100 users annotate fifty concept pairs drawn from our thesaurus. We found that the results achieved on WordNet need to be considered with care, and it is indeed necessary to re-evaluate them when using a different source.

In a second step, we then exploited the semantic relatedness measure we found to perform best in the multi-label classifier evaluation. Again, we investigated the ability of such an evaluation measure to outperform a standard binary measure (F_1) by asking expert users to rate for a small set of documents the quality of the classifier concepts when compared to the ground truth concepts. Our results showed that an evaluation which includes the semantic relatedness of concepts yields results which are more in line with human raters than an evaluation based on binary decision.

Besides the issues already raised, in future work we plan to investigate in which graph/content characteristics WordNet differs from our thesaurus and to what extent these different characteristics can be employed to explain the difference in performance of the various semantic relatedness measures.

6. REFERENCES

[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, 2003.

[2] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[3] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007.

[4] 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.

[5] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. 1997.

[6] 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.

[7] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, volume 1, pages 296–304, 1998.

[8] G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

[9] S. Nowak, A. Llorente, E. Motta, and S. Rüger. The effect of semantic relatedness measures on multi-label classification evaluation. In *CIVR '10*, pages 303–310, 2010.

[10] S. Patwardhan. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, 2003.

[11] 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.

[12] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. pages 448–453, 1995.

[13] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, volume 16, page 1089, 2004.

[14] M. Strube and S. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419, 2006.

[15] A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 521–528. IEEE, 2001.

[16] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM '93*, pages 67–74. ACM, 1993.

[17] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[18] I. 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*, pages 25–30, 2008.

[19] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL '94*, pages 133–138, 1994.