Evaluating Emergent Semantics in Folksonomies on Human Intuition

Thomas Niebler[1], Martin Becker[1], Daniel Zoller[1], Stephan Doerfel[2], and Andreas Hotho[1,3]

[1] Data Mining and Information Retrieval Group, University of Würzburg (Germany)
{niebler, becker, zoller, hotho}@informatik.uni-wuerzburg.de
[2] KDE Group, ITeG Research Center, University of Kassel (Germany)
doerfel@cs.uni-kassel.de
[3] L3S Research Center (Germany)

Abstract Semantic relations that closely resemble the human intuition of semantic relatedness, have been extracted automatically from sources, like text corpora, Wikipedia, or folksonomies, e.g., for constructing ontologies or for enhancing website navigation. Thereby, folksonomies are especially interesting since often, rich semantic structures emerge from the annotation of resources through users. In previous work however, proxies, like WordNet-based relatedness measures, have been used for evaluation, rather than relying directly on captured human intuition.
Here, we critically examine this form of evaluation and compare it to evaluations relying directly on human intuition. We find that WordNet-based measures hardly correlate with state-of-the-art human intuition datasets. Moreover, for meaningful results, the evaluation datasets must be domain-specific to provide a reasonably high overlap of words with the source for the extracted semantics. We demonstrate our results on two real world folksonomy datasets, using well-known evaluation datasets, as well as new, crowdsourced word similarity estimations. Overall, we argue that although directly evaluating semantic relatedness measures on human intuition may require collecting an adapted set of annotated samples, this form of benchmarking has clear advantages over the currently used WordNet-based measures, presenting a more realistic evaluation.

1 Introduction

The task of automatically extracting semantic relatedness between words or concepts that closely resembles actual human intuition (e.g., "ocean" is semantically more related to "water" than "tree" is to "car") has been tackled in a variety of studies, exploiting various data sources, such as tagging data [3,21], Wikipedia articles [27,12] and human navigational paths [26,20]. Semantic relations are helpful in the automatic construction of ontologies [1] or for enhancing website navigation [5]. For extracting such relatedness, folksonomies, or social tagging systems, are of special interest: [13] found that from the tagging process (users annotating resources with freely chosen keywords) a rich semantic structure emerges, from which semantic relatedness between tags can be captured. Consequently, folksonomies have been the focus of several studies [13,3,4,21,16].

Problem Setting. In previous work, methods for extracting semantic relatedness from folksonomies have mostly been tested using comparisons to measures computed on the WordNet taxonomy [8,18,3]: the information content measure by Resnik [23], the taxonomic shortest path distance on the WordNet graph, or the Jiang-Conrath measure [17], which is a combination of the aforementioned two approaches. Thereby, the Jiang-Conrath measure was found to correspond best with human intuition of semantic relatedness [7]. Using WordNet as a base has the advantage that a large number of words is covered, allowing the evaluation of semantics extracted from many datasets. In contrast, manually created human intuition datasets usually only cover a domain-specific and rather small set of word pairs (e.g., the WordSimilarity Test Collection [11] covers 353 pairs). However, the main goal of proposing a semantic relatedness measure is to model human intuition on semantic relatedness as closely as possible. Yet, WordNet only covers synonym relations and concept hierarchies, neglecting more subtle, yet intuitive relations such as between "pencil" and "paper" or "squirrel" and "tree", which are clearly not synonyms, but definitely semantically related. Thus, semantic relatedness measures based on WordNet can only cover certain aspects of semantic relatedness. Additionally, these measures and their validity as gold standards for extracted semantic relations have only been verified on two very small datasets by Rubenstein & Goodenough [24] and Miller & Charles [19]. Overall, these issues give reason to believe that despite their benefit regarding word coverage, WordNet-based measures may not always be the best and most realistic choice for evaluating methods for extracting semantic relatedness.

Approach. To investigate the suitability of WordNet-based evaluations in contrast to relying directly on captured human intuition, we evaluate existing methods for extracting semantic relatedness from folksonomies (cf., [8]) on two recent folksonomy datasets using both WordNet-based quality measures, as well as state-of-the-art human intuition datasets like WS-353 and a new dataset specifically collected for this work via crowdsourcing, and compare the results. Furthermore, we re-evaluate the WordNet-based measures on several state-of-the-art human intuition datasets. Finally, we analyze how the results derived from human intuition datasets depend on the covered vocabulary.

Findings and Contribution. We find that, while the ranking of the evaluated methods for extracting semantic relatedness from folksonomies stays roughly the same for both evaluation schemes, we observe subtle differences including some slightly differing ordering, as well as very small rank margins for WordNet-based evaluation. Additionally, we observe that the correlation between the WordNet-based semantic relatedness measures and human intuition datasets is very low, which indicates that they are not a useful model for human intuition. Our findings strongly favor evaluating semantic relatedness measures directly on hand-labeled human intuition datasets. Moreover, our experiments show that the results with human intuition based evaluation strongly depend on the coverage of the vocabulary in the dataset that semantic relatedness is extracted from. Nevertheless, we argue that using domain specific human intuition datasets for evaluating the quality of methods to extract semantic relatedness is preferable to the investigated WordNet-based measures. Thus, our contribution is threefold: 1) We evaluate

existing methods for extracting semantic relatedness on two recent folksonomy datasets, 2) we analyze and compare WordNet-based evaluation measures with evaluations relying directly on human intuition, and 3) we introduce a new human intuition dataset.

Structure. After covering related work in Section 2, we introduce basic definitions and methodology, including the different semantic relatedness measures and evaluation approaches, in Section 3. The datasets from which semantic relatedness is extracted and the human intuition datasets for evaluation are described in Section 4. In Section 5, we present our validation experiments and report their results, which we then thoroughly discuss together with their implications in Section 6. Section 7 concludes the paper.

2 Related Work

Our goal in this paper is to directly evaluate methods for extracting semantic relatedness from folksonomies based on human intuition rather than using WordNet based evaluation measures.

WordNet is a well-known lexical taxonomy of English words, verbs, adverbs and adjectives [10]. It is organized in so-called "synsets", which are sets of words which share the same meaning. These synsets thus represent semantic concepts. These synsets are themselves connected in a hierarchical tree structure, where the edges represent hypernymial, hyponomial, autonomial and synonymial relations between synsets. There has been much research about extracting semantic relatedness from this tree structure. Among others, the most well-known measures, which been evaluated in [7], are Hirst and St-Onge's [15] path-based measure, Resnik's information content based distance [23] and the Jiang-Conrath distance [17], which combines information content and path distance features. Evaluation of these measures was partially based on comparing the calculated similarity values with datasets of human intuition about word relatedness, namely RG65 [24] and MC30 [19]. WordNet based semantic relatedness measures are often used for evaluating other methods for extracting semantic relatedness.

In this paper, the evaluated methods for extracting semantic relatedness are based on vector-space based co-occurrences as introduced by [25]. In particular, they discern between first-order co-occurrence and second-order co-occurrence. Using first-order co-occurrence, two words are similar if they often appear together in a certain context, e.g., in a sentence or a tagging assignment. Second-order co-occurrence assumes that for words to be similar, the context in which they appear must be similar, i.e., if both words often occur with the same words in a specific context. The latter has been mentioned before in [24].

Regarding extracting semantic relatedness from *folksonomies*, Golder and Huberman [13] showed that folksonomies let stable semantic patterns emerge. This resulted in several works proposing methods for extracting semantic relatedness from folksonomies (e.g., [8], [18]). Cattuto et al. [8] applied the vector-space model from [25] on tagging data. They introduce several context-based relatedness measures for tags and evaluate them using the Jiang-Conrath and the taxonomic shortest-path distances on the WordNet taxonomy. Moreover, Markines et al.

evaluated similarity measures for social tagging systems, such as the cosine and the Jaccard measure [18]. They grounded their results on WordNet by applying Kendall's τ correlation coefficient between their similarity calculations and the corresponding similarities on WordNet using the Jiang-Conrath distance.

Regarding evaluation methods, both, [8] and [18], used WordNet and the Jiang-Conrath distance as a ground truth for semantic relatedness. However, for work on extracting semantic relatedness from other sources than folksonomies, it is more common to evaluate directly on human intuition of semantic relatedness by means of manually annotated datasets. Such a dataset is for example the WordSimilarity353 test collection (also called WS-353), which has been introduced in [11]. Work using such datasets for evaluation are for instance [27] and [11] who propose a measure consisting of a linear combination of context-based and WordNet-based semantic relatedness. Also, [12], [22], [14], [26] and [20] evaluate semantic relatedness on Wikipedia by using the WS-353 test collection. Note that all these works only cover Wikipedia as a source for semantic relatedness.

Up to this point, there has been no work which evaluates semantic similarity obtained from folksonomy data directly on human judgment. We will cover this gap in the remainder of this work.

3 Definitions and Methodology

In this section, we formally define folksonomies and review the similarity measures we use to extract semantic relatedness from folksonomy data. Lastly, we explain the two evaluation methodologies, which we compare in Section 5.

3.1 Folksonomy Definition

Folksonomies are the data structures underlying *social tagging systems*. In these systems, users collect resources and annotate them with freely chosen keywords, called tags. Examples are BibSonomy[4], for collecting web links and scholarly publications, Delicious[5] only for web links, FlickR[6] for images, and last.fm[7] for music. We recall the folksonomy definition from [8]:

Definition 1. *A folksonomy is a tuple* $\mathbb{F} := (U, T, R, Y)$*, where* U*,* T *and* R *are finite sets, whose elements are called* users*,* tags*, and* resources*, respectively.* Y *is a ternary relation between them, i.e.,* $Y \subseteq U \times T \times R$*. A* post *is a triple* (u, T_{ur}, r) *with* $u \in U$*,* $r \in R$*, and a non-empty set* $T_{ur} := \{t \in T | (u, t, r) \in Y\}$*.*

3.2 Measures for Semantic Similarity in Folksonomies

In the following, we will motivate and define four different semantic relatedness measures based on different tag context representations. All of these measures have been used in [8]. Because the authors found that FolkRank [16] yields similar results as direct co-occurrence, we do not include it in our study.

[4] `http://www.bibsonomy.org`
[5] `http://delicious.org`
[6] `http://www.flickr.com`
[7] `http://last.fm`

Context representations Given a folksonomy $\mathbb{F} = (U, T, R, Y)$, we first define three natural *first-order co-occurrence* context descriptions for tags, which can be used in the similarity measures we introduce afterwards. The context of a tag can be described by either its co-occurring tags in the same post, by the users who have used this tag, or by the resources that this tag has been assigned to. We employ a vector space model to represent words, i.e., tags, by their contexts, so our feature spaces are of the dimensions $|T|$, $|U|$, and $|R|$, respectively. Thus, each context description can also be represented as a matrix $C \in \mathbb{R}^{k \times |T|}$, where k denotes the dimension of the corresponding feature space. The column vectors c_j of these matrices represent the corresponding semantic contexts of each tag t_j, $j = 1, \ldots, |T|$.

Tag Context. The tag context matrix $C^{tag} \in \mathbb{R}^{|T| \times |T|}$ describes the context of a tag based on the postings in a folksonomy, i.e., the distinct annotations of resources by users.

$$C^{tag}_{ij} := |\{(u, r) \in U \times R \mid (u, t_i, r), (u, t_j, r) \in Y\}| \tag{1}$$

The entry in a matrix cell C^{tag}_{ij} in the tag context matrix is hence the number of posts, where a user assigned both tags t_i and t_j to some resource.

Resource Context. If we consider the resource context of tags, we obtain a matrix $C^{res} \in \mathbb{R}^{|R| \times |T|}$, which is defined as follows:

$$C^{res}_{ij} := |\{u \in U \mid (u, t_j, r_i) \in Y\}| \tag{2}$$

This measure counts how often a tag occurred with the same resource.

User Context. In the user context matrix $C^{user} \in \mathbb{R}^{|U| \times |T|}$, the entries describe how often a tag has been used by the same user.

$$C^{user}_{ij} := |\{r \in R \mid (u_i, t_j, r) \in Y\}| \tag{3}$$

Calculation of semantic similarity In order to calculate some kind of semantic similarity, we need a tag-to-tag relation. The easiest way to obtain such a relation is if we use the tag context matrix C^{tag}. This way, we interpret the co-occurrence count of a tag with another as the mutual similarity value (*co-occ*).

Another way to calculate similarity between two tags t_a and t_b is to compare the corresponding column vectors c_a and c_b of the context matrices defined above by applying the cosine similarity:

$$cossim(c_a, c_b) := \frac{\langle c_a, c_b \rangle}{\|c_a\| \cdot \|c_b\|} \in [0, 1] \tag{4}$$

Intuitively, the cosine similarity emphasizes the common context of tags. That is, if the value of $cossim(c_a, c_b)$ is close to zero, there is almost no common context of both tags and they are deemed very dissimilar. Whereas if this value is near 1, the tags frequently occur in the same context and can be interpreted as semantically very similar, if not synonymous.

3.3 Evaluation Methodology

This section describes the two evaluation methods we use in this paper. The first method was applied by [8] to evaluate semantic relatedness measures based on a comparison with WordNet, while the latter method was used in [7] for directly evaluating against human judgment.

Evaluation on WordNet. In [8], the authors evaluate the results of their similarity calculations on WordNet. For each word contained in the overlap T of WordNet and the top 10k tags from a Delicious dataset, they elect the most closely related tag according to their measures. For the resulting word-pairs, they calculate the corresponding taxonomic shortest path distance as well as the Jiang-Conrath distance on the WordNet graph. Then, for both distance measures, they calculate the average distance over all word-pairs. Formally this is:

$$eval(d_{ctxt}, d_{wn}) := \frac{1}{|T|} \sum_{t \in T} d_{wn} \left(t, \underset{t' \neq t}{\operatorname{argmax}}\ d_{ctxt}\left(t, t'\right) \right) \tag{5}$$

Here, d_{ctxt} is one of the context measures we defined in Section 3.2, while d_{wn} is one of the similarity measures defined on WordNet, which we use in this study: The *taxonomic shortest-path distance* looks for the shortest path in the WordNet taxonomy between two synsets. The *Jiang-Conrath distance* [17] (or short, JCN) is a combination of Resnik's information-theoretic measure [23] and the shortest-path distance.

Evaluation on Human Intuition of Similarity. A dataset containing human intuition on semantic relatedness contains word pairs with a manually assigned relatedness value. To gauge the quality of a new method for extracting semantic relatedness, we first calculate a semantic relatedness value for each of these word pairs using the new method. Now, we have two rankings: one assigned by humans and one assigned by our new method . For these two rankings, we calculate the correlation coefficient which represents the quality of the new method. Since an absolute similarity value is somewhat abstract and may be interpreted differently even by humans, we use the Spearman rank correlation coefficient ρ, because it only considers the relative placement of elements in a list instead of the actual values. This reflects the intuition given in [24], that similar words tend to have a more similar context and are thus placed higher in a similarity ranking. A high absolute correlation value near 1 means almost perfect correlation, i.e., that the extracted semantic fits well to human intuition whereas a correlation value near 0 means no correlation. If we cannot find a specific word from the evaluation dataset in our experiment datasets, e.g., because it has not been used, we leave out that word pair, since we cannot calculate a similarity value for it.

4 Datasets

In this section, we will describe the folksonomy datasets, which we will call the *experiment datasets*, as well as the datasets, which we performed our evaluation on. We will call the latter datasets the *evaluation datasets*.

Table 1. Sizes of the two folksonomy datasets before and after filtering.

Dataset	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$
BibSonomy	10,248	281,823	1,030,546	3,953,624
BibSonomy filtered	9,224	10,000	944,578	1,083,236
Delicious	1,951,207	14,782,752	118,520,382	1,026,152,357
Delicious filtered	1,884,280	10,000	92,715,855	797,796,374

Table 2. Base statistics for the human intuition similarity datasets.

Dataset	Pairs	Words	Score Range	RPP
RG65	65	48	0.0–4.0	15/36
MC30	30	39	0.0–4.0	38
WS-353	353	487	0–10	13-16
MTurk	287	499	1–5	23
MEN	3,000	751	0–50	50
Bib100	100	122	0–10	23

4.1 Folksonomy Datasets

In our work, we study two datasets of two public folksonomies. We use data from the social tagging system BibSonomy, which has a more academic and technical audience. The other dataset is a subset of the Delicious social tagging system, where the audience is focused on design and computers [28].

BibSonomy. The social tagging system BibSonomy [2] provides users with the possibility to collect bookmarks (links to websites) or references to scientific publications and annotate them with tags. We use a dump of BibSonomy from 2015.[8]

Delicious. Like BibSonomy, Delicious is a social tagging system, where users can share their bookmarks and annotate them with tags. We use a freely available dataset from 2011.[9]

Preprocessing. We filtered the tag assignments from both folksonomies. All tags not matching the regular expression `^\w+$`, i.e., all non-alphanumeric tags, were removed. Since their tags do not hold any meaning and have been added automatically, we also removed all tag assignments from the bot users `dblp`, `fbw_hannover`, `fbw` and `taggora` in BibSonomy. Also in BibSonomy, we removed all users which have been marked as spammers. Finally, we kept all tag assignments where the tag is contained in the top 10,000 occurring tags. The resulting and unfiltered dataset sizes are shown in Table 1.

4.2 Evaluation Datasets

Each evaluation dataset represents a collection of human collected scores about semantic relatedness. We will describe the datasets and its features in the following. Table 2 gives an overview over all evaluation datasets.

WordNet. WordNet is a large lexical taxonomy consisting of English nouns, verbs, adverbs and adjectives. The entities are organized in 117,659 synonym sets (synsets), which represent singular concepts. An important feature of WordNet is its hierarchical tree structure of synsets, where the edges represent hypernymial, hyponomial, autonomial and synonymial relations between synsets. While WordNet itself provides no relatedness information, there are many relatedness measures exploiting the tree structure of WordNet, such as the Jiang-Conrath distance [17] or the shortest-path distance (see [7]).

RG65. This dataset has been published for a very early judgment of human understanding of semantic similarity and context comparison [24]. The authors proposed a set of 65 word pairs generated from 48 nouns. Each pair has been given a rating between 0.0 and 4.0 by at least 15 raters.

MC30. Miller and Charles re-created a part of the RG65 dataset and selected a subset of 30 word pairs consisting of 39 words and collected ratings from 38 students on a 5-point scale from 0 to 4 [19].

WS-353. WordSimilarity-353[10] (WS-353) [11] consists of 353 pairs of English words and names. Each pair was assigned a relatedness value between 0.0 (no relation) and 10.0 (identical meaning) by 16 raters, denoting the assumed common sense semantic relatedness between two words. Finally, the total rating per pair was calculated as the mean of each of the 16 users' ratings. This way, WS-353 provides a valuable evaluation base for comparing our concept relatedness scores to an established human generated and validated collection of word pairs.

MTurk. The Mechanical Turk dataset is a collection of 287 word pairs and 499 words from New York Times articles [22]. The authors selected some pairs and asked workers from Amazon's Mechanical Turk to rate them on a scale of 1-5. Each word pair has been given 10 ratings.

MEN. The MEN Test Collection [6] contains 3,000 word pairs together with human-assigned similarity judgments, obtained by crowdsourcing using Amazon Mechanical Turk[11]. Contrary to WS-353, the similarity judgments are relative rather than absolute. Raters were given two pairs of words at a time. They were asked to choose the pair which pair of words was more similar. Each pair was rated 50 times, which leads to a score between 0 and 50 for each pair.

Bib100. Because the vocabulary of the mentioned evaluation datasets does not fit well to the vocabulary of our experiment datasets (see Section 5), we decided to create a new evaluation dataset[12] with a more fitting vocabulary to

[8] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/

[9] http://www.zubiaga.org/datasets/socialbm0311/

[10] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

[11] http://clic.cimec.unitn.it/~elia.bruni/MEN

[12] http://www.dmir.org/datasets/bib100

Table 3. Overlaps of the top tags from Delicious and BibSonomy with WordNet.

# top tags	100	500	1k	5k	10k
Cattuto et al. [8]	82%	80%	79%	69%	61%
Delicious	78%	75%	74%	61%	54%
BibSonomy	61%	59%	56%	45%	38%

Table 4. Results for the evaluation methods used in [8].

	Cattuto et al. [8]		Delicious		BibSonomy	
	JCN	Path	JCN	Path	JCN	Path
tag cosine	10.1	6.2	11.0	6.5	14.2	8.2
res cosine	9.1	6.2	10.8	6.4	14.0	8.1
user cosine	13.3	7.9	13.8	7.6	14.1	8.3
tag co-occ	12.3	7.4	13.1	7.6	13.7	8.1

the one used in BibSonomy and Delicious. We selected 122 words from the top 3,000 words of the BibSonomy dataset and combined them into 100 word pairs, which we subsequently each had judged 26 times for semantic relatedness using crowdsourced[13] scores between 0 (no similarity) and 10 (full similarity).

5 Experiments and Results

This section describes the experimental setup and states the results for each experiment. We first evaluate the folksonomy-based semantic relatedness measures on the BibSonomy and Delicious datasets. In the second experiment, we evaluate the Jiang-Conrath and taxonomic path distance measures on several human intuition datasets. The third experiment compares the results from the folksonomy-based semantic relatedness measures directly with the human judgment datasets. Finally, we assess the vocabulary overlap between the experiment and evaluation datasets.

5.1 Evaluation on WordNet

In our first experiment, we calculate the semantic relatedness based on the measures, which we defined in Section 3. We evaluate our results on WordNet analogously to [8] by comparing our results with the taxonomic path length and Jiang-Conrath distances as described in Section 3.3. Table 3 shows the overlap of several top-frequency tag subsets from both Delicious and BibSonomy with WordNet. As a trend, we can observe a higher overlap among the more frequent tags. Furthermore, the overlap of Delicious and WordNet is almost the same as the numbers which have been reported by [8], although we use a larger, more

[13] `http://www.dmir.org/semexp`

Table 5. Evaluation of the WordNet semantic similarity measures Jiang-Conrath and Path distance on different datasets with human intuition of similarity. When evaluated on more recent datasets, both measures yield poor correlation with human judgment.

	RG65	MC30	WS-353	MTurk	MEN	Bib100
JCN	0.776	0.826	0.292	0.364	0.366	0.389
Path	0.781	0.724	0.308	0.345	0.391	0.284
pairs	65	30	347	243	2,606	83

recent dataset. This tells us that over time, the semantic structure and frequency of words in the Delicious folksonomy have been very stable.

Table 4 shows the results of the WordNet evaluation setting. The Delicious results are very similar to the reported results in [8], i.e., tag cosine similarity based on resource and tag context finds closer nearest concepts according to Jiang-Conrath and path distance than cosine similarity based on user context or direct first-order co-occurrence. Similar experiments on BibSonomy have not been reported in [8]. In our investigation, however, we find that the BibSonomy results show little to no variation between the different context descriptions. We will see in the course of this work, that BibSonomy yields a similar ranking of the investigated similarity measures, but the evaluation setting has to be changed.

5.2 Evaluation of WordNet-based Measures on Human Intuition

Because the Jiang-Conrath and shortest path distances have only been evaluated on the RG65 and MC30 datasets, we wanted to extend this evaluation to the other evaluation datasets presented in Section 4 to assess the quality of the WordNet based measures as a representation of human intuition of semantic relatedness. Table 5 gives the results for the evaluation of the WordNet based measures on human intuition. It is obvious that, while the correlation with RG65 and MC30 is very strong, evaluation performance decreases greatly when using the other evaluation datasets, though the number of evaluable pairs is very high for all datasets, which validates our results.

5.3 Evaluation on Human Intuition of Similarity

In the following, we evaluate the folksonomy data using our context measures directly on human intuition of similarity, which is represented by the evaluation datasets from Section 4. For this, we calculate the Spearman rank correlation coefficient between the evaluation dataset and the results from the context measures, because the important dimension are not the absolute relatedness values themselves, but their order, i.e., the similarity strength of a pair.

Table 6 shows the results for evaluation on human intuition datasets for the relatedness measures defined in Section 5.1. We omitted the results for RG65 and MC30 because for both datasets, the number of evaluable pairs is very small. The performance order as reported in Table 4 of the results shown in Table 6

Table 6. Results for evaluation on the human similarity intuition datasets for the Delicious and BibSonomy folksonomy dataset. Results for RG65 and MC30 are omitted, since the amount of common pairs, which can be used for evaluation, is very small.

	Delicious				BibSonomy			
sim	WS-353	MTurk	MEN	Bib100	WS-353	MTurk	MEN	Bib100
tag cosine	0.454	0.504	0.581	0.640	0.395	0.596	0.436	0.621
res cosine	0.338	0.479	0.502	0.591	0.392	0.583	0.431	0.578
user cosine	0.227	0.452	0.303	0.193	0.208	0.437	0.226	0.307
tag co-occ	0.555	0.631	0.733	0.655	0.510	0.656	0.566	0.694
#pairs	194	105	1400	94	151	57	373	100

Table 7. Results for evaluation on WS-353 in other literature. Note that all of these works conduct their experiments on Wikipedia data, e.g., the link network.

paper	reported correlation
WikiRelate [27]	0.55
ESA [12]	0.75
WikiGame paths [26]	0.76
TSA [22]	0.8

Table 8. Vocabulary comparison between WS-353 and the two folksonomy datasets.

	Delicious					BibSonomy				
measure	100	500	1k	5k	10k	100	500	1k	5k	10k
word overlap	17	61	99	240	302	16	50	80	193	248
pair overlap	0	18	37	136	194	5	14	26	114	151

is partially even more clearly visible, e.g., tag cosine similarity shows a better correlation than resource cosine similarity. The co-occ measure is an exception and shows superior performance compared to all other measures. This might be due to vocabulary choices, because in BibSonomy, many closely related pairs also co-occur very frequently.

5.4 Vocabulary Comparisons

Because the results in Section 5.3 are mediocre compared to other papers, which achieve much higher correlation values on WS-353 (see Table 7), we investigated the reasons for why BibSonomy and Delicious do not yield more competitive results, although it has been shown that folksonomies exhibit strong regularities in tag usage, which should allow to extract semantic relatedness [13]. In Table 8, we show how many words and pairs from WS-353 are contained in different subsets of the most frequent tags of the experiment datasets. For both BibSonomy and Delicious, one must consider the top 5k tags to achieve a reasonable amount

of evaluable pairs. Also, another big part of the pairs is only matchable when the whole vocabulary is taken into account.

6 Discussion

We will now discuss the results from Section 5: We compare the soundness of the evaluation methods and discuss issues of evaluation on human intuition datasets.

6.1 Comparison of Evaluation Approaches

While WordNet makes it possible to evaluate hierarchical relations between concepts, it is not explicitly designed to calculate semantic relatedness. Nevertheless, similarity measures, such as the Jiang-Conrath or the path distance measure, exploit the WordNet graph. [7] even found relatively large correlation with human intuition on semantic relatedness by comparing against two relatively small datasets. However, we saw in Table 5 that correlations are much smaller when comparing with the larger, more recent datasets WS-353, MEN, MTurk and Bib100. Since the goal of extracting semantic relations is to find measures that are well aligned with human intuition, our results on WordNet-based metrics are strong evidence against their suitability for evaluating extracted semantics.

A possibility to still use the evaluation method depicted in Formula 5 would be to replace the WordNet-based values d_{wn} by humans judgments. However, the $argmax$ component yields different word pairs for each semantic relatedness measure we examine. Thus, it is necessary to collect new human judgements every time we evaluate a new measure. In addition, similarity values should be supported by a minimum number of human ratings. Now, the original approach calls for 10,000 words to average over. Assuming a minimum of 10 judgements per word pair, this would require an overall of 100,000 judgments.

Overall, since WordNet based measures are hardly correlated with human judgements and adopting the evaluation method by Cattuto et al. to use human judgements is expensive, we argue to evaluate directly on datasets containing human judgements on semantic relatedness using the Spearman correlation coefficient as introduced in Section 3. While corresponding datasets are smaller in size than the WordNet taxonomy and thus are not covering every possible relation between words, they provide a quicker and more realistic setting for evaluating semantic relatedness measures.

6.2 Size and Vocabulary of Human Intuition Datasets

After we have argued for the use of datasets covering the actual human intuition on semantic relatedness, we now discuss some issues, that must be considered when using that approach. First and foremost, the size of the evaluation datasets is important. Table 2 shows the sizes of the used evaluation datasets in this study. The MEN collection is by far the biggest dataset, consisting of 3,000 word pairs. The second largest dataset is WS-353, containing 353 pairs.[14] Compared to the

[14] actually, only 352, since `cash - money` is contained twice, once in reverse.

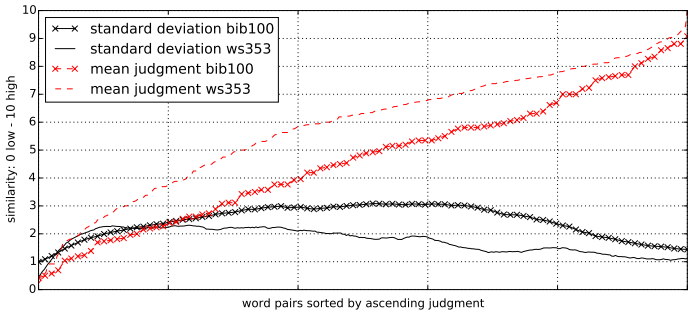

Figure 1. Mean rater scores (dashed) and smoothed standard deviations (continuous) in Bib100 (crosses) and WS-353 (no crosses), each sorted by ascending score.

sizes of our experiment datasets, and thus to the number of word pairs for which similarity judgments can be extracted, the evaluation datasets are rather small.

Secondly, human raters exhibit deviations and uncertainty in their judgments. Figure 6.2 shows the mean distribution of the word-pair similarities with smoothed standard deviations in Bib100 and WS-353. For word-pairs with a mean similarity between 2 and 7.5, the standard deviation of ratings is about 3 and lower otherwise. This could be interpreted as insecurity of raters about the extent of similarity of words, which are neither obviously related nor clearly unrelated.Interestingly enough, a very similar rating behavior can be observed for the WS-353 dataset.

Finally, our results show that not only size matters, but for evaluating semantic relatedness measures the evaluation datasets must also contain "the right" vocabulary – i.e., yield a high overlap with the experiment datasets. Table 8 shows this overlap between the WS-353 dataset and subsets of the top tags of the experiment datasets Delicious and BibSonomy.It takes the top 5k tags in each dataset to find a reasonably large coverage. Another big part of the vocabulary is contained among second half of the top 10k tags. With the creation of Bib100, we were able to show that it is possible to achieve more competitive results by creating a dataset with a more fitting vocabulary. The suitability of the Bib100 vocabulary for both folksonomies can especially be seen in the high percentage of found word pairs in Table 7.

Moreover, the competitive results of Bib100 are another argument in favor of using such human intuition datasets to evaluate semantic relatedness: It is very easy to construct a domain-specific dataset by picking a set of representative, frequently used words from the experiment dataset vocabulary, combine them into pairs, and to have them evaluated by humans, e.g., through crowdsourcing. They thus provide a cheap, easy, and fast option to judge semantic relatedness, while yielding a plausible evaluation scenario, as they rely directly on the explicitly expressed human intuition of semantic relatedness.

7 Conclusion

Since previous work has mainly evaluated methods for extracting semantic relatedness from folksonomies on WordNet based measures, we have extended this work by directly evaluating on human intuition datasets such as WS-353 or a new human intuition dataset collected specifically for this work using crowdsourcing. We compared both evaluation approaches and found that the semantic relatedness measures underlying the WordNet based evaluation hardly correlate with human intuition. Thus, we argue that although directly evaluating semantic relatedness measures on human intuition may require collecting an adapted set of annotated samples, this form of benchmarking has clear advantages over the currently used WordNet-based measures, presenting a more realistic evaluation.

Regarding future work, it is interesting to further investigate which part of the vocabulary of the dataset used for extracted semantic relatedness can be considered mature enough to be included in the evaluation dataset to be annotated by humans. Additionally, it is an open question of how to know when enough users have annotated a word-pair for the annotation to be judged accurate. It might even make sense to incorporate corresponding accuracy weightings into the evaluation procedure. Because Doerfel et al. showed in [9] that actual usage of a tagging system differs from posting behaviour, it would be interesting to exploit usage data to extract semantic relatedness.

References

1. Benz, D., Hotho, A.: Position paper: Ontology learning from folksonomies. In: Proc. LWA 2007. pp. 109–112 (2007)
2. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The social bookmark and publication management system BibSonomy. The VLDB Journal 19(6), 849–875 (Dec 2010)
3. Benz, D., Hotho, A., Stumme, G.: Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In: Proceedings of the 2nd Web Science Conference (WebSci10). Raleigh, NC, USA (2010)
4. Benz, D., Körner, C., Hotho, A., Stumme, G., Strohmaier, M.: One tag to bind them all: Measuring term abstractness in social metadata. In: Proc. ESWC (2011)
5. Berendt, B.: Using site semantics to analyze, visualize, and support navigation. Data Mining and Knowledge Discovery 6(1), 37–59 (2002)
6. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR) 49, 1–47 (2014)
7. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguists 32(1), 13–47 (2006)
8. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Proc. ISWC 2008. pp. 615–631 (2008)
9. Doerfel, S., Zoller, D., Singer, P., Niebler, T., Hotho, A., Strohmaier, M.: What users actually do in a social tagging system a study of user behavior in bibsonomy. Transactions on the Web (2016), accepted
10. Fellbaum, C.: WordNet. Blackwell Publishing Ltd (2012)
11. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proc. of the 10th international conference on World Wide Web. ACM (2001)

12. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. IJCAI. pp. 1606–1611 (2007)
13. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
14. Hassan, S., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: Proceedings of AAAI Conference on Artificial Intelligence (2011)
15. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database (1998)
16. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Proc. ESWC. pp. 411–426 (2006)
17. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy (1997)
18. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of the 18th international conference on World wide web. pp. 641–650. ACM (2009)
19. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language & Cognitive Processes 6(1), 1–28 (1991)
20. Niebler, T., Schlör, D., Becker, M., Hotho, A.: Extracting Semantics from Unconstrained Navigation on Wikipedia. KI - Künstliche Intelligenz pp. 1–6 (2015)
21. Niebler, T., Singer, P., Benz, D., Körner, C., Strohmaier, M., Hotho, A.: How tagging pragmatics influence tag sense discovery in social annotation systems. In: Advances in Information Retrieval, LNCS, vol. 7814, pp. 86–97 (2013)
22. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time: computing word relatedness using temporal semantic analysis. In: Proc. of the 20th WWW conf. pp. 337–346. ACM (2011)
23. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy (1995)
24. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10), 627–633 (1965)
25. Schütze, H., Pedersen, J.: A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing & Management pp. 307–318 (1997)
26. Singer, P., Niebler, T., Strohmaier, M., Hotho, A.: Computing semantic relatedness from human navigational paths: A case study on wikipedia. IJSWIS (2013)
27. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proc. AAAI (2006)
28. Zubiaga, A., Fresno, V., Martinez, R., Garcia-Plaza, A.P.: Harnessing folksonomies to produce a social classification of resources. IEEE Transactions on Knowledge and Data Engineering 25(8), 1801–1813 (2013)