# Exploiting Wikipedia for Evaluating Semantic Relatedness Mechanisms

Felice Ferrara and Carlo Tasso

Artificial Intelligence Lab,
Department of Mathematics and Computer Science,
University of Udine, Italy
{felice.ferrara,carlo.tasso}@uniud.it

**Abstract.** The semantic relatedness between two concepts is a measure that quantifies the extent to which two concepts are semantically related. In the area of digital libraries, several mechanisms based on semantic relatedness methods have been proposed. Visualization interfaces, information extraction mechanisms, and classification approaches are just some examples of mechanisms where semantic relatedness methods can play a significant role and were successfully integrated. Due to the growing interest of researchers in areas like Digital Libraries, Semantic Web, Information Retrieval, and NLP, various approaches have been proposed for automatically computing the semantic relatedness. However, despite the growing number of proposed approaches, there are still significant criticalities in evaluating the results returned by different methods. The limitations evaluation mechanisms prevent an effective evaluation and several works in the literature emphasize that the exploited approaches are rather inconsistent. In order to overcome this limitation, we propose a new evaluation methodology where people provide feedback about the semantic relatedness between concepts explicitly defined in digital encyclopedias. In this paper, we specifically exploit Wikipedia for generating a reliable dataset.

## 1 Introduction

The terms *semantic similarity* and *semantic relatedness* (on which we focus in this paper) have often been used as synonyms in the areas of Natural Language Processing, Information Retrieval and Semantic Web, but some researchers highlighted significant differences between these two concepts. The concept of semantic relatedness is defined in the literature as the extent to which two concepts are related by semantic relations [17]. On the other hand, a possible definition of semantic similarity describes it as the measure which quantifies the extent to which two concepts can be used in an interchangeable way. According to this definition two semantically similar entities are also semantically related, but two semantically related concepts may be semantically dissimilar [3]. For example, the concepts of *bank* and *trust-company* are semantically similar and their similarity implies that they are also semantically related, but two concepts related by an

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 105–117, 2014.

antonymic[1] relation (such as the adjectives *bad* and *good*) are semantically related and semantically dissimilar. According to [20], semantic similarity is a more strict relation since it takes into account a focused set of semantic relations which are often stored in lexical ontologies such as Wordnet. In Wordnet, for example, synonyms[2] are grouped in synsets and a hierarchical structure connects hyponyms and hypernyms[3]. On the other hand, the semantic relatedness between two concepts depends on all the possible relations involving them. For example, in order to compute the semantic relatedness between two Wordnet concepts, we should use all the available semantic connections by including, for example, meronomy[4] and antonymy. However, two concepts can be related by more complex semantic relations which are usually not explicitly stored in lexical ontologies. Think, for example, to the case of two concepts that are semantically related by means of a chain of more than one semantic relation, involving other 'intermediate' concepts. For example, the pair *pope* and *Italy* can be related through the chain *pope* $\rightarrow$ *Vatican City* $\rightarrow$ *Rome* $\rightarrow$ *Italy*. This kind of relations is not explicitly included in Wordnet as well as all the other possible relations which can be entailed between concepts which are not directly related by standard relations. Moreover, it has to be noticed that humans organize their knowledge according to complex schemas by connecting concepts according to their background knowledge and experience [8]. The reasoning task where units of meaning are processed by the human mind in order to identify connections between concepts is referred in the literature as *evocation* [2], which can be also defined as the degree to which a concept brings to mind another one. Evocation adds cross-part-of-speech links among nouns, verbs, and adjectives [14]. Since the human mind works under the influence of personal experience, the evocation process builds relations which may be not true in an absolute way (for instance the relations between emotions and objects/animals) and this is why these relations cannot be available in knowledge bases such as Wordnet.

Obviously, all these aspects must be considered when we have to plan the evaluation of methods aimed at automatically quantifying the semantic relatedness (*SR methods*) or the semantic similarity (*SS methods*). Thesaurus-like resources , such as the Roget dataset [1], can be effectively used for evaluating the precision of SS methods: they connect terms by TR (Related-Term) links and by UF (Used-For) links, however such links are just a few for each term, whereas many others could be entailed.

For the above reasons, the feedback provided by humans about the relatedness between pairs of terms is commonly used in order to evaluate the precision of SR methods. However, the methodology currently used for both collecting feedback and evaluating precision of SR methods is widely criticized, even by the same

[1] Antonymy is the semantic relation which connects concepts with an opposite meanings.
[2] Two terms are synonyms if they have the identical or very similar meaning.
[3] A hyponym shares a *type of* relationship with its hypernym.
[4] The meronomy denotes a *part of* relation.

researchers who use it to analyze their results. These limitations are addressed in this work and, more specifically, the paper has two goals:

- describing the limitations of the state of the art mechanisms. A survey of the limitations of the approaches utilized for evaluating the accuracy of SR methods is given.
- proposing a new evaluation approach. We propose a new procedure aimed at effectively evaluating the precision of SR methods which analyze the content of Wikipedia, one of the main examples of Digital Library 2.0.

The choice of focusing on this specific digital library is mainly due to the growing interest of the research community on the usage of Wikipedia as knowledge source for computing semantic relatedness. In fact, the large coverage of concepts and the support to multilinguism makes Wikipedia very attractive for developing SR methods. Moreover, other researches point out that the refinements of the Wikipedia articles do not significantly influences the results of SR methods [19] while new concepts can be easily introduced and connected to the existing ones.

The paper is organized as follows: in Section 2 we describe the state of the art, major drawbacks are illustrated in mechanisms used for evaluating the precision of SR methods while the drawbacks of these approaches are the object of Section 3; in Section 4 we propose a new approach for facing these limitations; final considerations conclude the paper in Section 5.

## 2 Evaluating SR Methods: State of the Art

As reported in [3], three main approaches have been proposed in the literature for evaluating the precision of SR methods.

The approach utilized in [12] evaluates SR methods according to a set of qualitative heuristics. The simplest heuristic takes into account if the evaluated measure is a metric; in [9] the authors report a list of other suitable features for SR methods such as domain independence, independence from specific languages, coverage of included words, and coverage of the meanings of each word. The heuristic-based strategy is the simplest one but it also does not provide very significant results since it cannot numerically quantify the accuracy of results. For this reason, even if this strategy is a useful tool for designing new SR methods, it is not an effective tool for comparison [3].

More concrete results can be obtained by embedding SR methods in other hosting systems such as text clustering systems [11], metonymy resolution mechanisms [10], and recommender systems [5]. In these cases, different SR methods are compared and evaluated according to the improvement produced by the integration of the specific SR method within a larger system. However, it is quite clear that this strategy increases the difficulty in performing an extensive comparison of SR methods since: (i) different works face different tasks and use different datasets so preventing the repeatability of experiments and (ii) the computed precision can be influenced by the other components in the embedding system.

In order to overcome these drawbacks, a more direct strategy can be implemented by comparing the feedback of a set of humans with the results produced by SR approaches. The feedback of volunteers has been collected in order to create datasets which have been used in the majority of the works where the precision of SR methods has been evaluated. The first experiments aimed at creating this kind of datasets was exploited by Rubenstein and Goodenough [16]. In their experiments they exploited a deck of 65 cards where on each card there was a pair of nouns written in English. The researchers asked to 51 judges both to order the 65 pairs of words (from the most related pair to the most unrelated one) and to assign a score in [0.0,4.0] for quantifying the relatedness of each pair of terms. This experiment was also replicated by other researchers in different settings. One of the most popular dataset is the Related353 dataset [6] which is constituted by 353 word pairs is annotated with an integer in $[0, 10]$ by two sets of evaluators (composed by 13 and 16 judges respectively). Other works focused on the task of defining similar datasets for specific domains: in the biomedical field, Pedersen et al. collected the feedback of medics and physicians in order to evaluate SR methods in that specific domain [15]. Other works focused on generating larger datasets in an automatic way: in [18], a corpus of document is analyzed in order to extract pairs of semantically related terms by following the idea that pairs of terms which appear frequently in the same document are probably semantically related.

The numeric scores acquired in these experiments have been extensively used for evaluating the precision of SR methods. In order to reach this aim the Pearson product-moment and the Spearman rank order correlation coefficients have been used. The Pearson product-moment is a statistical tool used to check if the results of a SR method resemble human judgments. On the other hand, the comparison of two rankings of the pairs (the ranking which order the pairs according to the feedback provided by humans and the ranking which order the pairs according to the result of a SR method) can be executed by the Spearman coefficient. Both these coefficients have a numerical value in $[-1, +1]$, where $-1$ corresponds to completely uncorrelated rankings (low precision) and, conversely, $+1$ corresponds to a perfect correlation (high precision).

## 3 Drawbacks of the State of the Art

The experiments proposed in the literature mainly use datasets constituted by pairs of terms annotated by a group of humans. However, this approach has many criticalities which are emphasized even by the same researchers who adopted it. In this section we illustrate these limitations by organizing the discussion in two parts: in Section 3.1, we focus on the characteristics of the collections of pairs of terms and, in Section 3.2, we describe the features of both the human feedback and the procedures exploited for computing the precision of SR methods.

### 3.1 Characteristics of the Pairs of Terms

The quality of the feedback collected in the experiments referred in Section 2 strongly depends on the task submitted to the volunteers. The following points summarize the main limitations:

- **Shortage**. The dataset proposed by Rubenstein and Goodenough is constituted by only 65 pairs of nouns which cannot be used to exploit an extensive analysis for generalizing the findings. This limitation is partially faced by the Related353 dataset which is constituted by 353 pairs.
- **Terms instead of concepts**. The datasets are build up by terms which do not identify concepts. On the other hand, SR methods compute the semantic relatedness among concepts such as the synsets of Wordnet or the pages of Wikipedia. The proliferation of senses in knowledge bases such as Wordnet and Wikipedia makes hard the task of manually associating a sense to each term included in a dataset [17]. Consider, for example, that the term *love* is associated to 6 synsets in Wordnet and, on the other hand, in Wikipedia the term *love* identifies several senses: an emotion as well as people, songs, fictional characters, and movies. For tackling this problem, it is possible to manually associate some of the terms of the considered dataset to the Wikipedia concept that, most probably, was considered by the evaluators. On the other hand, in order to avoid the need for manual disambiguation of terms, the semantic relatedness between all the possible senses of the two terms can be identified and fixed in the following way: the pair of senses with the highest semantic relatedness computed by the evaluated SR method is considered for assigning two specific senses to the two terms. Both these approaches are questionable since the judges were not conscious of all the various meanings of the words when they annotated the pairs.
- **Uncovered domains and semantic relations**. The datasets created by Rubenstein and Goodenough as well as the Related353 dataset were defined with the main goal of covering many possible degrees of similarity. Following this idea, the authors used very general terms without taking into account the idea of choosing terms in different domains. This is limitation which prevents the generalization of the results. In particular, we highlight that the information provided or extracted from a knowledge base may differ according to the given topic. We can imagine that in Wikipedia, for example, some topics are described better than others. It is also possible that different knowledge bases (such as Wikipedia, Wordnet or other ontologies) may provide better results in different domains. For this reason it would be interesting to have datasets where pairs of terms are associated to domains or at least to have datasets where several distinct domains are covered. Similarly, a more reliable approach should also take care of covering a sufficient set of semantic relations. In fact specific SR method could be adequate for catching a specific semantic relation but it could not work with other relations. This information is obviously missing also in datasets created in an automatic way.

### 3.2 Characteristics of the Feedback and Evaluation Procedure

The agreement among the evaluators is used in the literature for estimating the quality of the collected feedback: this follows the idea that higher is the agreement more reliable is the collected feedback. According to the literature, the level of agreement is sufficient to assess the precision of SR methods. However, there is not a threshold for the required agreement between the judges and this is also true for domain-dependent datasets. Moreover, also other features of the feedback collected from humans may greatly influence negatively the quality of the evaluation. More specifically, we identify the following points:

- **Pairs with low agreement.** Different works use different strategies to manage pairs ot terms with low agreement among judges. An example of these pairs is *(monk,oracle)* in the Related353 dataset which was annotated by 13 evaluators who returned the following votes $(7, 8, 3, 4, 4, 6, 5, 8, 6, 3, 4, 6, 1)$. In the majority of the works available in the literature these pairs are threaten exactly like the others, but in [15] the authors proposed to discard pairs with a very low agreement in order to have more significant results. Obviously, this idea can be applied only when the dataset is constituted by a large set of pairs. This is a very important issue since, as noticed in [3], the available datasets show a significant agreement only when the existence of the semantic relation is very clear (for instance the terms are synonyms or they are completely unrelated).
- **The choice of the scale.** The choice of the scale for collecting the feedback is a controversial point and has a strong impact on the agreement among the judges. By adopting a very fine-grained scale the judges have many possible choices and they can provide more accurate responses. This was the motivation for the approach proposed by Rubenstein and Goodenough who also asked people to order the pairs in order to have more coherent responses. In fact, by ordering the pairs each judge could assign a decreasing list of values to quantify the semantic relatedness. However, this mechanism does not scale up to a large set of pairs since it requires a huge workload for ordering many pairs of terms. For this reason, in the task for acquiring the feedback for larger datasets like the Related353 dataset it is not asked to the evaluators to order the pairs. In this case, the humans could not rely on the order imposed to the pairs for assigning a vote and, consequently, it was harder for them to be coherent with previously assigned votes. For this reason, when the judges only annotate pairs of terms with a number it is better to avoid very fine-grained scale in order to have more consistent responses.
- **Bias introduced by specific communities.** Different communities of evaluators may evaluate the semantic relatedness between two concepts according to different perspectives. This is clearly reported in [15] where the authors show that physicians and medics judged differently the semantic relatedness between terms in the field of biology. On the other hand, it makes sense to evaluate SR methods only on pairs where the feedback is not biased by the perspective of a specific community.

– **Metric robustness.** The Pearson coefficient is a statistical tool used to catch the strength of the linear correlation between the human judgments and the score computed by a specific SR method. However, the correlation between the votes provided by humans and the SR method can be nonlinear. Moreover, the Pearson correlation is based on the assumption that the two compared random variables are normally distributed, whereas the actual distribution of the relatedness values is at the moment unknown [3]. On the other hand, the Spearman coefficient, which does not directly compare human votes with the results of the SR method, seems to be more robust.

## 4 Toward a New Evaluation Strategy

In order to face the limitations described in the previous section we propose a new strategy for evaluating SR methods. In this section we describe our ongoing work (Section 4.1) as well as our future steps (Section 4.2).

### 4.1 New Resources and Procedures

As already mentioned, other researchers showed that humans can judge the semantic relatedness by using a numerical estimation only if the answer is quite obvious. In fact, the experiments described in Section 2 showed that the agreement among the judges was significantly hight only when the pairs were composed by two synonyms or by two completely unrelated terms. Our hypothesis is that humans can perceive the semantic relatedness, but they are not used to quantify it by using a number. The difficulty in acquiring reliable feedback from humans is mainly due to the problem of having datasets constituted by terms which may be polysemic, i.e. having multiple senses. Starting from this assumption, here we propose a new procedure for collecting more significant responses from the judges, by avoiding both expensive workload, such as ordering a long sequence of pairs of terms, and tricky/noisy tasks, such as selecting a numeric level to quantify the semantic relatedness among two terms.

Our proposal is to ask judges to select the concept (from a set of proposed concepts) which is most related to a given concept, where each concept is associated to a specific knowledge base (in our current work concepts are identified by Wikipedia pages). By associating each term to a concept of a knowledge base we can overcome the limitation of having datasets constituted by only terms. This approach allows to obtain two advantages: (i) the judges can take into account a unique specific meaning of the concepts when they produce their responses and (ii) the evaluated SR method can exploit the Wikipedia page associated to the concept for computing the semantic relatedness.

More technically, we defined the questions for the judges as triples $T = (t_1, \ldots, t_m)$, where the triple $t_i = (target_i, c_{i1}, c_{i2})$ is constituted by a target concept and two other concepts $c_{i1}, c_{i2}$. For each triple $t_i$, the judges have to identify which one among $c_{i1}$ and $c_{i2}$ is (in their views) more related to $target_i$. For example, given the triple $t$=(*Musician*, *Watch*, *Trumpet*), the evaluator can

select *Watch* or *Trumpet* as more related to *Musician*. The reader can notice that the proposed procedure does not depend on a specific scale for collecting the feedback and this also simplifies the work of the judges who have to select only the most related concept. By selecting the concept semantically more related to the target concept the judge orders the three concepts according to the relatedness to the target. By following the previous example, if a judge chooses *Trumpet* then he implicitly defines the ordered list of concepts (*Musician*, *Trumpet*, *Watch*) since *Trumpet* has been considered as more related to *Musician* than *Watch*. We then take into account the way the judges order the concepts in each triple for evaluating the precision of a SR method. In particular, the SR method can order the concepts in the triple $t_i = (target_i, c_{i1}, c_{i2})$ by computing the semantic relatedness between the target concept and the two concepts $c_{i1}$ and $c_{i2}$ and, consequently, it can produce a rank. In this way we compute the precision of the SR method according to the percentage of cases in which the SR method orders the concepts of the triples as the humans did.

However, we also believe that there are various cases where humans cannot provide a response: the judge may be not familiar with a concept or even a topic or two concepts may be (more or less) equally semantically related to the target concept. In order to manage these situations, the judges are allowed to skip the evaluation of a triple, since we are keen to identify the responses for which the judges are sufficiently confident. By taking into account the number of judges who skipped a triple, we can measure a degree of trustworthiness of the overall feedback acquired for a specific single triple. More specifically, for each triple we computed an *Indecision Score* as the ratio between the number of judges who skipped the triple and the total number of judges. By taking into account a maximum threshold on the Indecision Score, we then remove the triples for which there is a certain percentage of judges who did not provide a response. We also filter out the triples with a low agreement among judges, by following the idea that a low agreement can be the result of different evaluation perspectives. To this aim we computed, for each triple $t_i$, an *Agreement Score* as the maximum between (i) the ratio between the number of judges who selected the concept $c_{i1}$ and the total number of judges and (ii) the ratio between the number of judges who selected the concept $c_{i2}$ and the total number of judges. By requiring an Agreement Score higher than a certain threshold, we can remove ambiguities which may be introduced by different communities with different perspectives.

We identify for each triple a 'correct' order of the concepts by taking into account the order defined by the majority of the judges. For example, supposing that the concept *Trumpet*, in the triple $t$=(*Musician*, *Watch*, *Trumpet*), is more frequently selected than the concept *Watch*, then the order (*Musician*, *Trumpet*, *Watch*) is taken as the correct ranking. This 'correct ranking' is compared to the order computed by the evaluated SR method. In this way we define the *precision* of the evaluated SR method as the percentage of the correctly ordered triples by the SR method.

Obviously, the approach used to build the triples has a significant impact on the results. As we said in Section 3, one of the main drawbacks of the datasets

described in the literature depends on the number of domains and of different semantic relations included in the dataset. In order to face this issue, we have defined a specific set of templates for the triples, such as ($\langle TARGET\rangle$, $\langle Emotion_1\rangle$, $\langle Emotion_2\rangle$) and ($\langle TARGET\rangle$, $\langle Work_1\rangle$, $\langle Work_2\rangle$). Then, we create some triples by creating instantiating each template. For example, from the template ($\langle TARGET\rangle$, $\langle Emotion_1\rangle$, $\langle Emotion_2\rangle$), we can build the triple (*Love*, *Graditude*, *Jelausy*), the triple (*Clown*, *Humor*, *Fear*) and so on. We also include other triples by picking concepts from systems such as Delicious and Open Directory. In particular, tags, categories, and other terms are extracted from these systems in order to create new triples. By using stacks of Delicious and categories in Open Directory we also select concepts (that must be concepts of Wikipedia) belonging to different domains. In this way we face (at least partially) the problem of covering semantic relations in different domains. In our first experiments we collected the feedback of 10 judges and each of them evaluated 420 triples in a month. Since the Agreement Score measures the agreement among the judges only on a single triple, we evaluated the overall agreement among the judges by by means of the Fleiss'kappa [7]. The Fleiss'kappa allows us to measure the agreement of the judges over the entire set of triples and, according to this analysis, we have a significant agreement also over the entire set of triples (kappa=0.783). Then we filtered the triples by throwing out the triples with an Agreement Score lower then 0.7 (i.e. we require that at least 7 of the 10 judges provided the same response) and with an Indecision Score higher then 0.2 (i.e. we require that at maximum 2 judges skipped the question). As expected, after this filtering step, we have that the agreement among the judges increases (kappa=0.849). However, it is interesting to observe that our filtering interventions removed only 27 triples from the initial set of 420 triples. Two examples of these triples are (*Mammal*, *Dolphin*, *Lion*) and (*Lifeguard*, *Holiday*, *Work*). These two triples show the usefulness of the filtering step since we observed that many judges skipped the triple (*Mammal*, *Dolphin*, *Lion*) since *Dolphin* and *Lion* are both *Mammals*. The Indecision Score is used to discard this triple because, in this case, people could not find semantic relations for identifying which one of the two concepts is more related to the target concept. Similarly, if two concepts are completely unrelated to a given target concept, then judges cannot find semantic relations for answering. The Indecision Score allows us to remove these triples avoiding in this way potential ambiguities. In the case of the triple (*Mammal*, *Dolphin*, *Lion*), a part of the judges considered *Lifeguard* as someone you can meet during *Holiday* whereas another part of the judges considered *Lifeguard* as a *Work*. In this case, there is a low agreement due to the subjective way of perceiving the semantic relatedness. The Agreement Score allows us to remove such triples, enhancing in this way the significance of the dataset.

### 4.2 Ongoing Evaluation and Future Steps

At the moment we are comparing new SR methods (that we specifically designed in order to compute the semantic relatedness between the Wikipedia concepts) with other state of the art mechanisms.

In particular we defined new SR methods by extending some approaches proposed [13] where: (i) each Wikipedia concept is represented by its incoming pages (i.e. the pages with a link to the concept) and outcoming pages (i.e. the pages linked by the concept) and (ii) the semantic relatedness among two pages is computed by comparing their corresponding representations (larger is the number of shared incoming/outcoming links, higher is the similarity among the concepts). In particular, in this work we propose two metrics which will be referred as CIN and GDOUT in the rest of the paper.

The CIN metric describes each concept as a weighted vector of Wikipedia pages. In particular given a concept of Wikipedia, the pages which have a link to the concept describe it and the weight of each page (i.e. each component of the vector) is equal to $log(\frac{|W|}{|T|})$ where $W$ is the set of pages in Wikipedia and $T$ is the number of articles linked by the specific page (i.e. the specific component of the vector). Given such representation of concepts, the CIN metric computes the semantic relatedness between two concepts as the cosine similarity between the two corresponding vectors. In this way, the metric computes the semantic relatedness among two concepts according to the shared incoming pages. The metric assumes that the concepts having many outcoming pages are less specific and, for this reason, the semantic relatedness among two concepts is higher when the corresponding Wikipedia pages share many incoming pages with few outcoming pages.

On the other hand, the GDOUT metric is based on a different assumption. In fact, in this case, the semantic relatedness between two concepts is estimated by taking into account the number of outcoming pages shared between the corresponding Wikipedia pages. More technically, the GDOUT metric uses a variation the Normalized Google Distance [4] for computing the semantic relatedness between the concepts $a$ and $b$ as

$$GDOUT = 1 - \frac{log\left(max\left(|A|, |B|\right)\right) - log\left(|A \cap B|\right)}{log\left(|W|\right) - log\left(min\left(|A|, |B|\right)\right)}$$

where A is the set of pages linked by the concept a, B is the set of pages linked by the concept b, and W is still the set of pages available in Wikipedia.

We utilized our approach in order to evaluate the results produced by these two SR methods and we obtained that the CIN approach has a higher precision (the precision is equal to 0.87) than the GDOUT method (the precision is equal to 0.80 in this case).

At the moment we are working on utilizing the datasets constituted by pairs of terms and on embedding the SR methods in different systems in order to compare our evaluation approach with other evaluation approaches. In particular, we are interested in embedding the SR methods also in other different systems in order to verify if the results changes according to the system where the SR methods are integrated.

In order to promote a more exhaustive evaluation campaign of the SR methods proposed in the literature we are also working on other two possible extensions of our proposal.

First, we are interested in collecting a larger set of responses by utilizing crowdsourcing systems such as Amazon Mechanical Turk. In particular, we are interested in evaluating if different levels of agreement among the judges can be found by utilizing a new, different, and larger set of judges. These future experiments will allow us to better evaluate also the statistical relevance of our current results.

Second, we recognize that Wikipedia is not the only possible knowledge source which can be used for computing the semantic relatedness. For example, other works in the literature compute the semantic relatedness among the synsets of Wordnet. In order to have an exhaustive evaluation campaign we need to have triples constituted by the concepts defined in other knowledges sources such as Wordnet. We are evaluating two possible strategies. The first one is to repeat our work for constructing a new dataset which cover concepts of Wordnet. On the other hand, we have designed and developed an intelligent framework which can support the alignment of the concepts of Wikipedia to the synsets in Wordnet. By exploiting this tool we aim at associating the concepts in our dataset to Wordnet synsets.

## 5 Conclusion

Many tools and approaches which integrate the computation of SR among concepts have been proposed in the literature in order to improve the access to digital libraries [21]. On the other hand, in this paper, we (i) analyzed the limitations of the approaches traditionally utilized to evaluate the precision of SR methods and (ii) proposed a new approach for producing more reliable datasets and evaluations. Our first results about the agreement among the judges and the pruning of ambiguous triple seem promising.

Future works will investigate the usage of crowdsourcing systems for collecting larger set of responses from a larger set of judges. We will also study the problem of using concepts available in knowledge sources different from Wikipedia by associating concepts of Wikipedia to concepts of Wordnet.

## References

1. Roget's 21st century thesaurus, 3rd edn. (October 2012), `http://thesaurus.com/browse/dataset`
2. Boyd-graber, J., Fellbaum, C., Osherson, D., Schapire, R.: Adding dense, weighted connections to wordnet. In: Proceedings of the Third International WordNet Conference (2006)
3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. 32(1), 13–47 (2006)
4. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Trans. on Knowl. and Data Eng. 19(3), 370–383 (2007)

5. Ferrara, F., Tasso, C.: Integrating semantic relatedness in a collaborative filtering system. In: Proceedings of the 19th Int. Workshop on Personalization and Recommendation on the Web and Beyond, pp. 75–82 (2012)
6. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Trans. Inf. Syst. 20(1), 116–131 (2002)
7. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 1606–1611. Morgan Kaufmann Publishers Inc., San Francisco (2007)
9. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., Wang, X.S. (eds.) WISE 2008. LNCS, vol. 5175, pp. 136–150. Springer, Heidelberg (2008)
10. Hayes, J., Veale, T., Seco, N.: Enriching wordnet via generative metonymy and creative polysemy. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, pp. 149–152. European Language Resources Association (2004)
11. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 389–396. ACM, New York (2009)
12. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL 1998, vol. 2, pp. 768–774. Association for Computational Linguistics, Stroudsburg (1998)
13. Milne, D., Witten, I.H.: 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, pp. 25–30. AAAI Press (2008)
14. Nikolova, S., Boyd-Graber, J., Fellbaum, C.: Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools. In: Mehler, A., Kühnberger, K.-U., Lobin, H., Lüngen, H., Storrer, A., Witt, A. (eds.) Modeling, Learning, and Proc. of Text-Tech. Data Struct. SCI, vol. 370, pp. 81–93. Springer, Heidelberg (2011)
15. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40(3), 288–299 (2007)
16. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10) (October 1965)
17. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 2, pp. 1419–1424. AAAI Press (2006)
18. Zesch, T., Gurevych, I.: Automatically creating datasets for measures of semantic relatedness. In: Proceedings of the Workshop on Linguistic Distances, LD 2006, pp. 16–24. Association for Computational Linguistics, Stroudsburg (2006)

19. Zesch, T., Gurevych, I.: The more the better? assessing the influence of wikipedia's growth on semantic relatedness measures. In: Calzolari, N. (ed.) Proceedings of the Seventh International Conference on Language Resources and Evaluation. European Language Resources Association, Valletta (May 2010)
20. Zesch, T., Gurevych, I.: Wisdom of crowds versus wisdom of linguists; measuring the semantic relatedness of words. Nat. Lang. Eng. 16(1), 25–59 (2010)
21. Zhang, W., Feng, W., Wang, J.: Integrating semantic relatedness and words' intrinsic features for keyword extraction. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2225–2231. AAAI Press (2013)