Evaluating the Results of Methods for Computing Semantic Relatedness

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. Due to the growing interest of researchers in areas such as Semantic Web, Information Retrieval and NLP, various approaches have been proposed in the literature 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 semantic relatedness methods. The limitations of the state of the art evaluation mechanisms prevent an effective evaluation and several works in the literature emphasize that the exploited approaches are rather inconsistent. In this paper we describe the limitations of the mechanisms used for evaluating the results of semantic relatedness methods. By taking into account these limitations, we propose a new methodology and new resources for comparing in an effective way different semantic relatedness approaches.

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 [18]. 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 antonymic[1] relation (such as the adjectives *bad* and *good*) are semantically

[1] Antonymy is the semantic relation which connects concepts with an opposite meanings.

A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816, pp. 447–458, 2013.

related and semantically dissimilar. According to [21], 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 can/are usually not explicitly stored in lexical ontologies. For example, from "Albert Einstein received the Nobel prize" we would be able to infer the existence of a relation between *Albert Einstein* and *Nobel prize*. This relation is not explicitly defined 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 [7]. The reasoning task where units of meaning are processed by the human mind in order to identify connections between concepts is referred in 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 that 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 or the TOEFL Synonym Question dataset, can be effectively used for evaluating the precision of SS methods: they connect terms by RT (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 this reason, the feedback of people 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 this feedback and evaluating the precision of SR methods is widely criticized by the same researchers who use it to analyze their results. In this paper we address these limitations and we propose an original approach for evaluating the precision of SR methods. More specifically, we focus on the idea of evaluating the results of SR methods which calculate the relatedness between the concepts included in Wikipedia. This choice is mainly due to the growing interest of the research community on the usage of Wikipedia as knowledge source for computing semantic

[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.

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 [20] 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 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 describe our proposal which is evaluated in Section 5; final considerations conclude the paper in Section 6.

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.

A possible approach, utilized for example in [11], evaluates SR methods according to a set of qualitative heuristics. The simplest heuristic takes into account if the evaluated measure is a metric, but in [8] the authors report a list of 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 quantify the accuracy of results. For this reason, even if this strategy is a useful tool for designing new SR methods, it cannot be used to have a significant comparison of state of the art mechanisms [3].

More concrete results can be obtained by integrating the SR methods in other systems such as metonymy resolution mechanisms [10], recommender systems [5], and approaches to text similarity [1]. In these cases, different SR methods are compared and evaluated according to the improvement produced by the integration of a specific SR method. However, it is quite clear that this strategy increases the difficulties in performing an extensive comparison of SR methods since: (i) different works face different tasks and use different datasets preventing, in this way, the repeatability of experimentations and (ii) the computed precision can be influenced by other components embedded in the adopted 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 [17]. 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: Miller and Charles used 30 pairs selected from the Rubenstein and Goodenough's deck of cards by using a larger set of judges [12]; Resnik used the feedback of 10 human evaluators for executing his experiments [16]. The idea of ordering and assigning a value to pairs of nouns was also replied for languages different from English. In particular, Gurevych replicated the experiment of Rubenstein and Goodenough by translating the 65 pairs into German [9]. The 65 pairs of nouns were a reference point for many studies and for this reason also Finkelstein et al. decided to start from these pairs for creating a larger dataset (referred in this paper as Related353 dataset) constituted by 353 word pairs [6]. In this case, the pairs were 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 researchers also worked on the task of generating larger sets of pairs of terms in a more automatic way. In [19], for instance, 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 coefficient 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 (i.e. the ranking where pairs are ordered according to the feedback of the humans and (ii) the ranking where pairs are ordered according to the results of a SR method) can be executed by the Spearman coefficient. These two coefficients are 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 report 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 described 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 [18]. For example, the term *love* can be associated to 6 synsets of Wordnet and, on the other hand, in Wikipedia the term *love* identifies 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 adopted 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 the meanings of the words when they annotated the pairs.

Uncovered Topics 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 degree of similarity. Following this idea, the authors used very general terms without taking into account the idea of choosing terms in different topics. Moreover there are not details about the semantic relations which involve the terms in the dataset. These limitations do not allow to generalize the computed results.

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 by following the idea that higher is the agreement more reliable is the collected feedback. According to the works in literature, the agreement of the available datasets is sufficient to evaluate the precision of SR methods. Actually, there is not a threshold for the required agreement between the judges and this is also true for domain-dependent datasets. However, also other features of the feedback collected from humans may greatly influence negatively the quality of the evaluation. We point out specifically the following points:

Pairs with Low Agreement. Different works use different strategies to manage the pairs 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 the 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 freely 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 motivating idea of 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 would require a huge load of work for ordering many pairs of terms. For this reason, the collection of the feedback for larger datasets like the Related353 dataset did not require the evaluators to order the pairs of terms. 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 a judge to be coherent with his previous votes. For this reason, when the judges only annotate pairs of terms with a number it is better to renounce to a very fine-grained scale in order to have more significant 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 author 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 community of people.

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 SR method. However, the correlation between the votes of humans and the results of a 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, as we said, does not directly compare the human votes with the results of a SR method, seems to be more robust. This shows again that by allowing the

evaluators to order concepts according to the degree of relatedness it is possible to acquire a more reliable feedback.

4 New Feedback for Evaluating Semantic Relatedness

In order to face the limitations described in the previous section we propose a new approach: collecting a different kind of feedback and, consequently, adopting a new way for evaluating the precision of SR methods. In our work we follow the idea that the source of the drawbacks of the state of the art datasets is primarily caused by the task assigned to the judges. In fact, as we said, other researchers showed that humans can judge the semantic relatedness by using a number only if the answer is quite obvious (for example, the terms are synonyms or the terms are completely unrelated). Our hypothesis is that humans can perceive the semantic relatedness, but they are not used to quantify it by using a number. Starting from this hypothesis we define a new approach for developing a new dataset (described in Section 4.1) which can be effectively used to evaluate SR methods (as described in Section 4.2).

4.1 Creating the Dataset

We decided to change the task used to collect the feedback by avoiding both expensive workload, such as the task of ordering a long sequence of pairs of terms, and tricky/noisy tasks, like the task of choosing a number to quantify the semantic relatedness among two terms. We ask to the judges to select the concept (from a set of proposed concepts) which is more related to a fixed/given concept, where each concept is associated to a specific Wikipedia page. By following this idea we defined the questions for the judges as triples of concepts. In particular, we defined a set of triples $T = (t_1, \ldots, t_m)$ where the triple $t_i = (target_i, c_{i1}, c_{i2})$ is constituted by a reference/fixed concept and two other concepts c_{i1}, c_{i2}. For each triple t_i, the judges have to decide which one among c_{i1} and c_{i2} is 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*. 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* is semantically more related to *Musician* then *Watch*. We also believe that there are some cases where humans cannot provide a response due to:

- Lack of knowledge. The judge may be not familiar with a concept or even a topic. In this case the judge may prefer to skip the question.
- Other possible ambiguities. In some cases two concepts may be (more or less) equally semantically related to the Target concept.

We decided to manage these two issues by allowing the judges to skip the evaluation of a triple, since we want to be able to identify the responses for which the judges are sufficiently confident.

By adopting this task we can better face some of the limitations presented in Section 3. First of all, by associating each term to a Wikipedia page we can overcome the limitation of having datasets composed by just terms. In our case the meaning of a term is specified by a specific page in Wikipedia. By associating the terms to Wikipedia pages we obtain two advantages: judges can take into account the real meaning of the concepts when they produce their responses and, moreover, also the evaluated SR method can exploit the Wikipedia page associated to the concept for computing the semantic relatedness. Another advantage of the proposed task is that it does not use 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 simplifying the task, we can (at least partially) face the shortage problem since we require lower efforts the evaluators.

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. This choice has an impact on the number of the possible covered relations since by selecting concepts in different topics it is more likely to pick concepts linked by many different semantic relations.

From a practical point of view, we used a Web application for collecting the feedback of 10 judges who had to provide their feedback for 420 triples in a month. By using our application the judges were allowed for each triple t_i to both take a look at the description of the concepts included in t_i (the gloss available in the corresponding Wikipedia page) and to select one among three possible responses: c_{i1}, c_{i2} and the *I DON'T KNOW* option (for skipping the response).

4.2 Evaluating the Precision of SR Methods

By following the idea that potential ambiguities can be discovered by taking into account the agreement among the judges, we defined a filter in order to throw out from our evaluation ambiguous triples. In order to reach this aim, for each triple t_i, we compute: an *agreement score* $A(t_i)$ and *indecision score* $I(t_i)$. In particular, $A(t_i)$ is equal to the maximum between (i) $A(t_i, c_{i1})$ which is the ratio between the number of judges who voted the concept c_{i1} and the total

number of judges and (ii) $A(t_i, c_{i2})$ which is ratio between the number of judges who voted the concept c_{i2} and the total number of judges. In the rest of the paper we will refer the concept which maximizes the agreement score $A(t_i)$ as c_{max_i}. On the other hand, $I(t_i)$ is computed as the ratio between the number of judges who skipped the triple (by choosing the *I DON'T KNOW* option) and the number of judges. We used these these two values in order to throw out from our analysis possible ambiguities. In particular, we use only the set of triples $FT = (t_1, .., t_n)$ characterized by: an agreement score higher or equal to 0.7 (i.e. we require that at least 7 of the 10 judges provided the same response); an indecision score lower or equal to 0.2 (i.e. we require that at maximum 2 judges skipped the question). In this way we throw out the triples where the responses of the judges are more or less equally divided between the two concepts as well as the triples for which the judges could not identify the most related concept. By following this strategy we removed only 27 triples and two examples of these triples are (*Mammal*, *Dolphin*, *Lion*) and (*Lifeguard*, *Holiday*, *Work*).

We use the triples in FT for evaluating the precision of the SR methods. Our metric, named *Order Count*, aims at checking if the evaluated SR methods order the concepts in the triples exactly as the judges did. In order to reach this aim, for each triple t_i, we use the evaluated SR method for computing the semantic relatedness between $target_i$ and c_{i1} (named $SR(target_i, c_{i1})$) and the semantic relatedness between $target_i$ and c_{i2} (named $SR(target_i, c_{i1})$)). In particular, our metric counts the number of times that the evaluated SR method computes a higher value for $SR(target_i, C_{max_i})$ of each triple t_i in FT.

5 Evaluation

In this section we focus on assessing if the proposed approach is useful to evaluate the precision of SR methods. In order to reach this aim we both implemented and modified some state of the art SR methods (as described in Section 5.1) in order to have a pool of SR methods. We compare our results to the state of the art approaches by estimating the significance of our results in Section 5.2.

5.1 The Evaluated SR Methods

Two methods, referred in this paper as *COUT* and *GDIN*, are proposed in [13]. The COUT metric describes each concept as a weighted vector of Wikipedia pages: given a concept of Wikipedia, the pages linked by the concept describe it and the weight of each page is equal to $log(|W| / |P|)$ where W is the set of pages in Wikipedia and P is the number of articles linked by the page. Given such representation of concepts, the COUT metric computes the semantic relatedness between two concepts as the cosine similarity between the two corresponding vectors. We modified the COUT approach by defining the *CIN* metric which represents a concept by means of the Wikipedia pages with a link to the concept. In this case, the weight of a page in the vector is equal to $log(|W| / |P|)$ where W is the set of pages in Wikipedia and P is the number of articles linked by the page and the cosine similarity is still used to compute the semantic relatedness.

The GDIN metric, on the other hand, adapts the Google Distance measure [4] in order to compute the semantic relatedness among the concepts a and b of Wikipedia as

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

where A is the set of pages with a link to concept a, B is the set of pages with a link to concept b, and W is the set of pages available in Wikipedia. We modified also the GDIN metric in order to produce the GDOUT method 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.

5.2 Results

The agreement among the judges is used in the literature in order to assess if datasets are reliable. We have specifically exploited the Fleiss' kappa for estimating the agreement among the judges. This analysis showed a very significant agreement among the evaluators (kappa=0.783) and that is why the pruning step (executed in order to identify ambiguities) removed only 27 triples from the initial set of 420 triples T.

We use also the Related353 dataset for computing the precision of the SR methods described in the Section 5.1. Since this dataset contains terms we cannot compute the semantic relatedness between fixed Wikipedia pages and to face this issue we adopt a strategy utilized in the literature. In particular, given a SR method and a pair of terms, we compute the semantic relatedness between all the Wikipedia pages associated to the two terms in the pair. The highest computed score is taken for the pair in order to compute both the Pearson and the Spearman coefficients. We use these two coefficients for ranking the SR methods and we show these rankings as well as the coefficients (reported in the parenthesis) in Table 1. In the same table we also report the results produced by using our approach/dataset.

Table 1. The results of the compared metrics

	Pearson	Spearman	Order Count
1	GDIN (0.555)	GOUT (0.5)	CIN (0.87)
2	GOUT (0.473)	COUT (0.497)	GDIN (0.85)
3	CIN (0.472)	GDIN (0.49)	COUT (0.83)
4	COUT (0.479)	CIN (0.48)	GDOUT (0.8)

It is interesting to observe that the Pearson and the Spearman correlations produce completely different results also if the coefficients are very near. However, the ranking produced by our approach is still different from the ones produced by the Pearson and the Spearman metrics.

In order to evaluate if our approach makes sense, we integrated the 4 SR methods in a Collaborative Filtering (CF) recommender system as described in [5]. We computed the *Mean Average Error* (*MAE*) of the predictions of the CF system by using the MovieLens dataset with a 5-cross fold validation evaluation. According to this analysis, the CF recommender system produces the most accurate results when the CIN method is integrated in the recommender (MAE=0.735). This confirms the result predicted by our approach by showing that our approach has sense. Moreover, it seems to show that the methods which represent a Wikipedia concept by using the incoming links are more precise than others.

6 Conclusion

In this paper we surveyed the main limitations of the state of the art mechanisms aimed at evaluating the precision of SR methods. Starting from this analysis we proposed a new approach and a new dataset for evaluating the accuracy of SR methods which compute the semantic relatedness between concepts of Wikipedia. Our results differ from the outcomes produced by the approaches described in the literature. However, we showed that our results makes sense and for this reason both the built dataset and the evaluation approach can be useful tools for evaluating SR methods. Future works will use crowdsourcing systems for collecting feedback from a larger set of judges in order to better evaluate the significance of the proposed approach. We are also interested in associating concepts defined in other knowledge sources (such as Wordnet) in order to extend our evaluation to mechanisms which use different knowledge sources.

References

1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Proceedings of the Sixth International Workshop on Semantic Evaluation, June 7-8, pp. 385–393. Association for Computational Linguistics, Montréal (2012)
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. 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)
8. Gracia, J.L., 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)
9. Gurevych, I.: Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 767–778. Springer, Heidelberg (2005)
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. 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)
12. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
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. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 1, pp. 448–453. Morgan Kaufmann Publishers Inc., San Francisco (1995)
17. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10) (October 1965)
18. 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)
19. 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)
20. Zesch, T., Gurevych, I.: The more the better? assessing the influence of wikipedia's growth on semantic relatedness measures. In: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation. European Language Resources Association, Valletta (2010)
21. 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)