Natural Language Engineering **16** (1): 25–59. © Cambridge University Press 2009
doi:10.1017/S1351324909990167

Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words

TORSTEN ZESCH and IRYNA GUREVYCH

Ubiquitous Knowledge Processing Lab, Computer Science Department, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany
e-mail: {zesch,gurevych}@tk.informatik.tu-darmstadt.de

(*Received 21 December 2007; revised 28 January 2009; accepted 17 June 2009; first published online 9 September 2009*)

Abstract

In this article, we present a comprehensive study aimed at computing semantic relatedness of word pairs. We analyze the performance of a large number of semantic relatedness measures proposed in the literature with respect to different experimental conditions, such as (i) the datasets employed, (ii) the language (English or German), (iii) the underlying knowledge source, and (iv) the evaluation task (computing scores of semantic relatedness, ranking word pairs, solving word choice problems). To our knowledge, this study is the first to systematically analyze semantic relatedness on a large number of datasets with different properties, while emphasizing the role of the knowledge source compiled either by the 'wisdom of linguists' (i.e., classical wordnets) or by the 'wisdom of crowds' (i.e., collaboratively constructed knowledge sources like Wikipedia).

The article discusses benefits and drawbacks of different approaches to evaluating semantic relatedness. We show that results should be interpreted carefully to evaluate particular aspects of semantic relatedness. For the first time, we employ a vector based measure of semantic relatedness, relying on a concept space built from documents, to the first paragraph of Wikipedia articles, to English WordNet glosses, and to GermaNet based pseudo glosses. Contrary to previous research (Strube and Ponzetto 2006; Gabrilovich and Markovitch 2007; Zesch *et al*. 2007), we find that 'wisdom of crowds' based resources are not superior to 'wisdom of linguists' based resources. We also find that using the first paragraph of a Wikipedia article as opposed to the whole article leads to better precision, but decreases recall. Finally, we present two systems that were developed to aid the experiments presented herein and are freely available[1] for research purposes: (i) DEXTRACT, a software to semi-automatically construct corpus-driven semantic relatedness datasets, and (ii) JWPL, a Java-based high-performance Wikipedia Application Programming Interface (API) for building natural language processing (NLP) applications.

1 Introduction

A text's lexical cohesion is established by means of lexical–semantic relations between terms (Halliday and Hasan 1976; Morris and Hirst 1991). *Car* is related to *vehicle*, *prize* is related to *Nobel Prize*, and *tree* is related to *leaf*. In these examples, the

[1] http://www.ukp.tu-darmstadt.de/software

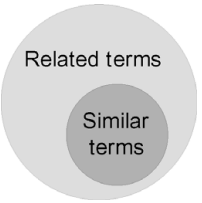

Fig. 1. Relationship between the set of similar terms and the set of related terms.

terms are connected by means of classical taxonomic relations like hyponymy (*car*, *vehicle*) or meronymy (*tree*, *leaf*). However, terms can also be connected through nonclassical relations (Morris and Hirst 2004). For example, *car* is related to *drive*, *Albert Einstein* is related to *Nobel Prize*, and *tree* is related to *green*. In the sentence 'Albert Einstein did not receive the Nobel Prize for his theory of relativity', the lexical cohesion of the sentence is almost fully established by nonclassical relations between the terms *Albert Einstein*, *receive*, *Nobel Prize*, and *theory of relativity*.[2]

Classical relations are encoded in linguistic knowledge sources like WordNet (Fellbaum 1998), and such knowledge sources can be used to determine the *similarity* between two terms. The algorithms used for determining the similarity are called semantic similarity (SemSim) measures (Budanitsky and Hirst 2006). However, even dissimilar terms can be semantically related. For example, *car* is not similar to *street*, while there is an obvious semantic relationship between the two terms. Another limitation is that similarity is only defined between terms of the same part-of-speech. Thus, SemSim measures cannot detect the relationship between, for example, *car* and *drive*.

Semantic relatedness (SemRel) measures overcome these limitations by using nonclassical relations or implicit connections between the descriptions of the terms in a knowledge source. Nonclassical relations are usually not encoded in knowledge sources, but recently there has been increased research in adding nonclassical and cross part-of-speech links to WordNet (Boyd-Graber *et al.* 2006). As SemRel measures make use of information beyond classical lexical semantic relations, they are able to determine whether two terms are in some way related, even if they are not similar or have different parts of speech. Thus, SemSim is a special case of the broader defined SemRel, i.e., two terms that are similar are also related, but the inverse is not true (see Figure 1).

Many natural language processing (NLP) applications, like word sense disambiguation (Patwardhan, Banerjee and Pedersen 2003), semantic information retrieval (Gurevych, Müller and Zesch 2007), information extraction (Stevenson and Greenwood 2005), finding real word spelling errors (Budanitsky and Hirst 2006), and computing lexical chains (Silber and McCoy 2002; Galley and McKeown 2003) rely

[2] Caveat lector – while humans also use multi-word expressions such as 'Nobel Prize' or 'theory of relativity' as building blocks of a document's lexical cohesion, we are currently limited to using single terms as the experimental unit in our study, due to limitations in existing algorithms and evaluation datasets.

on determining the cohesive structure of texts. Thus, improving the recognition of lexical cohesion within a document by detecting also related instead of only similar terms can have a major impact on these tasks.

Another important issue when determining the cohesive structure of a text is the knowledge source used. In recent work, there has been a shift from classical wordnets created by the 'wisdom of linguists' toward emerging knowledge sources created using the 'wisdom of crowds' like Wikipedia (Gabrilovich and Markovitch 2007; Zesch *et al.* 2007b) or Wiktionary (Zesch, Müller and Gurevych 2008b). In this article, we describe a set of experiments aimed at: (i) comparing the 'wisdom of linguists' with the 'wisdom of crowds' as a knowledge source for computing SemRel, (ii) comparing a set of SemRel measure types using different kinds of information from the knowledge sources that are evaluated using two different tasks (ranking wordpairs and answering word choice questions), and (iii) computing SemRel in two different languages (English and German).[3]

In this article, we focus only on knowledge based methods for computing semantic relatedness, as opposed to distributional methods (Weeds 2003). However, distributional methods showed competitive performance on some evaluation datasets (the interested reader may refer to Mohammad *et al.* 2007).

The article is structured as follows: In Section 2, we describe the state of the art in computing semantic relatedness as well as the knowledge sources used. Section 3 introduces the evaluation framework that was used for our experiments, whose results are presented in Section 4. Section 5 briefly explains the software tools that were used to conduct the experiments. The article concludes with a summary and future research directions in Section 6.

2 State of the art

In Section 2.1, we describe the state of the art in computing semantic relatedness, and categorize existing algorithms into four distinct types of measures, where each type exhibits unique properties. We then compare 'wisdom of linguists' based knowledge sources, and introduce 'wisdom of crowds' based knowledge sources such as Wikipedia, in Section 2.2. Section 2.3 describes the process of adapting semantic relatedness measures to the new knowledge source Wikipedia.

2.1 Semantic relatedness measures

A multitude of semantic relatedness measures working on structured knowledge sources have been proposed. Figure 2 gives an overview of the development of SemRel measures on a historic time scale. Measures can be categorized into *path based* (Rada *et al.* 1989; Wu and Palmer 1994; Hirst and St-Onge 1998; Leacock and Chodorow 1998; McHale 1998; Jarmasz and Szpakowicz 2003), *information content*

[3] Previously, research has focused much on English. However, using languages other than English for conducting semantic relatedness experiments can yield valuable insights about semantic relatedness itself as well as the adaptability of algorithms for computing semantic relatedness to other languages.

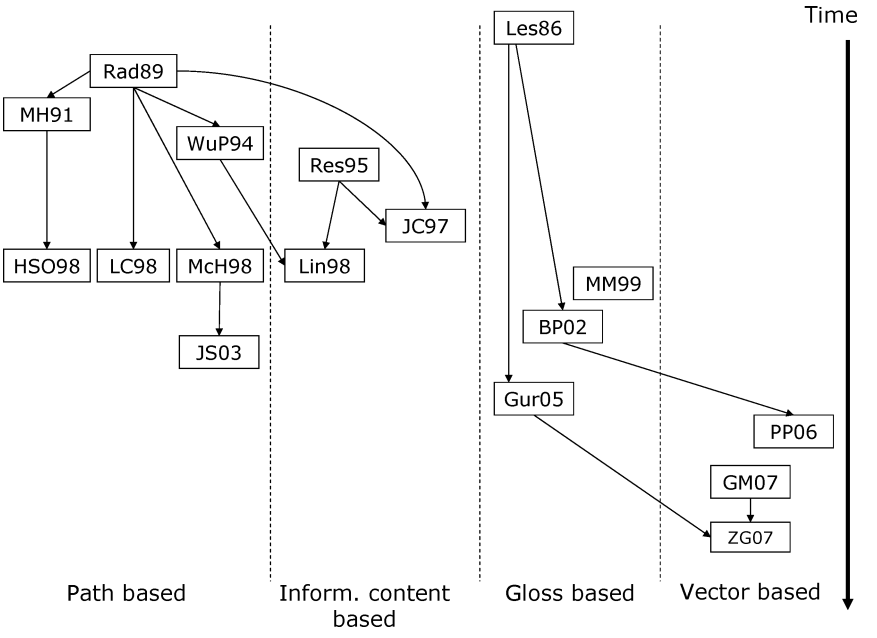

Fig. 2. Development of semantic relatedness measures on a historic time scale.

based (Resnik 1995; Jiang and Conrath 1997; Lin 1998), *gloss based* (Lesk 1986; Mihalcea and Moldovan 1999; Banerjee and Pedersen 2002; Gurevych 2005), and *vector based* (Patwardhan and Pedersen 2006; Gabrilovich and Markovitch 2007) approaches. The figure shows the recent shift from path based and information content based measures (using only explicitly modeled knowledge) to gloss based and vector based measures that are able to use information drawn from the description of a concept that is not explicitly modeled as a relation in a semantic graph structure.

2.1.1 Path based measures

Path based measures determine the length of the path between nodes representing concepts in a semantic graph structure (e.g., a wordnet, a thesaurus, or the Wikipedia category graph). The shorter the path, the higher the relatedness between concepts.

Rada *et al.* (1989) use the path length l between two nodes to compute semantic relatedness. This measure (Rad89) can either be a SemSim or SemRel measure depending on the type of edges that are allowed in a path. For example, if only edges corresponding to classical lexical semantic relations are allowed, Rad89 is a SemSim measure. However, if also edges corresponding to nonclassical relations are allowed, it is a SemRel measure. Rad89 can be computed as follows:

$$dist_{\text{Rad89}}(c_1, c_2) = l(c_1, c_2)$$

where *dist* means that the measure returns a distance value instead of a relatedness value, and $l(c_1, c_2)$ returns the number of edges on the path from c_1 to c_2. The

distance value can be easily transformed into a relatedness value by subtracting it from the maximum path length of the graph, $rel_{\text{Rad89}}(c_1, c_2) = l_{max} - l(c_1, c_2)$.

Jarmasz and Szpakowicz (2003) (JS03) adapt the Rad89 measure to Roget's thesaurus as a knowledge source based on the work of McHale (1998) (McH98). JS03 is also a relatedness measure as the relations in Roget's thesaurus are not restricted to classical lexical semantic relations.

As polysemous words may have more than one corresponding concept in a lexical semantic resource, the resulting semantic relatedness score between two words w_1 and w_2 can be calculated as

$$rel = \begin{cases} \min\limits_{c_1 \in C(w_1), c_2 \in C(w_2)} dist(c_1, c_2) \\ \max\limits_{c_1 \in C(w_1), c_2 \in C(w_2)} rel(c_1, c_2) \end{cases}$$

where $C(w_i)$ is the set of concepts that represent senses of word w_i. That means, the relatedness of two words is equal to the score of the least distant/most related pair of concept nodes, depending on whether the measure returns a relatedness $rel(c_1, c_2)$ or a distance $dist(c_1, c_2)$ value.

Leacock and Chodorow (1998) (LC98) normalize the path length with the depth of the graph,

$$sim_{\text{LC98}}(c_1, c_2) = -\log \frac{l(c_1, c_2)}{2 \cdot depth}$$

where *depth* is the length of the longest path from the root node of the taxonomy to a leaf node. The prefix *sim* means that the measure is a similarity measure in its original definition, as it was defined on WordNet using only taxonomic links. The scaling factor $2 \cdot depth$ assumes a tree-like structure, where the longest possible path runs from a leaf to the root node and back to another leaf node. The length of the path $l(c_1, c_2)$ is measured in nodes.

These simple path length methods do not take into account that concepts higher in the taxonomy are more abstract, i.e., that a path with a length of 1 between abstract concepts near the top of the taxonomy should yield a lower similarity value than a path of the same length between specific concepts on the leaf level of the taxonomy. Many measures have been proposed to overcome this limitation. For example, Wu and Palmer (1994) introduce a measure (WuP94) that uses the notion of a lowest common subsumer of two concepts $lcs(c_1, c_2)$. A *lcs* is the first shared concept on the paths from the concepts to the root concept of the hierarchy.

$$sim_{\text{WuP94}} = \frac{2 \cdot depth(lcs)}{l(c_1, lcs) + l(c_2, lcs) + 2 \cdot depth(lcs)}$$

WuP94 is a similarity measure, as a *lcs* is only defined in a taxonomy.

Hirst and St-Onge (1998) adapt a measure (HSO98) originally described by Morris and Hirst (1991) (MH91) to work with WordNet instead of Roget's Thesaurus (Roget 1962). Using the HSO98 measure, two words have the highest relatedness score, if (i) they are in the same synset, (ii) they are antonyms, or (iii) one of the words is part of the other (e.g., *car* and *car park*). In all other cases, relatedness depends on

the path between the concepts, where long paths and direction changes (upward, downward, horizontally) are penalized. The resulting formula is

$$rel_{\text{HSO98}}(c_1, c_2) = C - len(c_1, c_2) - k \cdot turns(c_1, c_2)$$

where C and k are constants, *len* is the length of the path and *turns* counts the number of direction changes in the path. The HSO98 measure is a relatedness measure as paths are not restricted to taxonomic links.

2.1.2 Information content based measures

Information content (IC) approaches are based on the assumption that the similarity of two concepts can be measured by the amount of information they share. In a taxonomy, the shared properties of two concepts are expressed by their lowest common subsumer *lcs*. Consequently, Resnik (1995) defines semantic similarity (Res95) between two nodes as the information content value of their *lcs*:

$$sim_{\text{Res95}}(c_1, c_2) = IC_{Res95}(lcs(c_1, c_2))$$

The information content of a concept can be computed as

$$IC_{Res95}(c) = -\log p(c)$$

where $p(c)$ is the probability of encountering an instance of c in a corpus. The probability $p(c)$ can be estimated from the relative corpus frequency of c and the probabilities of all concepts that c subsumes. This definition of IC is bound to the availability of a large corpus, and the obtained IC values are relative to that corpus. Hence, Seco *et al.* (2004) introduce the *intrinsic information content* which is computed only from structural information of the taxonomy and yields better results on some English datasets. It is defined as:

$$IC_{Sec04}(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(|C|)}$$

where $hypo(c)$ is the number of all hyponyms of a concept c and $|C|$ is the total number of concepts in the taxonomy.[4]

Jiang and Conrath (1997) define a measure (JC97) derived from Res95 by additionally using the information content of the concepts. The original formula returns a distance value, but can be easily transformed to return a relatedness value instead.

$$dist_{\text{JC97}}(c_1, c_2) = IC_{Res95}(c_1) + IC_{Res95}(c_2) - 2 \cdot IC_{Res95}(lcs)$$
$$sim_{\text{JC97}}(c_1, c_2) = 2 - (IC_{Res95}(c_1) + IC_{Res95}(c_2) - 2 \cdot IC_{Res95}(lcs))$$

[4] Intrinsic IC is equivalent to Resnik's definition of IC if we set the corpus frequency of each word to 1, and a word's frequency count is not divided between its concepts. Both definitions of IC yield similar results, indicating that ignoring the frequency information does not result in a performance loss. The depth scaling effect used by both definitions of IC seems to be more important than the frequency scaling.

Lin (1998) defines a universal measure (Lin98) derived from information theory.

$$sim_{\text{Lin98}}(c_1, c_2) = 2 \cdot \frac{IC_{Res95}(lcs)}{IC_{Res95}(c_1) + IC_{Res95}(c_2)}$$

The WuP94 measure is a special case of this formulation.

2.1.3 Gloss based measures

Dictionaries or wordnets usually contain short glosses for each concept that are used by gloss based measures to determine the relatedness of concepts.

Lesk (1986) introduces a measure (Les86) based on the amount of word overlap in the glosses of two concepts, where higher overlap means that two terms are more related.

$$rel_{\text{Les86}}(c_1, c_2) = |gloss(c_1) \cap gloss(c_2)|$$

where $gloss(c_i)$ returns the multiset of words in a concept's gloss.

Banerjee and Pedersen (2002) propose a more sophisticated text overlap measure (BP02) that additionally takes into account the glosses of related concepts forming an extended gloss *extGloss*. This overcomes the problem that glosses in WordNet are very short. The measure is defined as:

$$rel_{\text{BP02}}(c_1, c_2) = |extGloss(c_1) \cap extGloss(c_2)|$$

where $extGloss(c_i)$ returns the multiset of content words in the extended gloss.

Mihalcea and Moldovan (1999) take a similar approach in their word sense disambiguation system. They construct a linguistic context for each noun or verb sense c_i by concatenating the nouns found in the glosses of all WordNet synsets in the subhierarchy of c_i. The relatedness between two senses is then computed as the number of overlapping nouns in the corresponding contexts.

Gloss based measures cannot be used directly with semantic graph structures without textual definitions of concepts such as the German wordnet GermaNet (Kunze 2004). Therefore, Gurevych (2005) constructed pseudo glosses (Gur05) by concatenating concepts that are in close relation (synonymy, hypernymy, meronymy, etc.) to the original concept. This is based on the observation that most content words in glosses are in close relation to the described concept. For example, the pseudo gloss for the concept *tree (plant)* would be *'woody plant, ligneous plant, stump, tree stump, crown, treetop, limb, tree branch, trunk, tree trunk, bole, burl, ...'* showing a high overlap with its WordNet gloss *'a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown'*. The measure can be formalized as follows:

$$rel_{\text{Gur05}}(c_1, c_2) = |pseudoGloss(c_1) \cap pseudoGloss(c_2)|$$

where $pseudoGloss(c_i)$ returns the multiset of content words in the pseudo gloss.

2.1.4 Vector based measures

In this section, we focus only on SemRel measures where concept vectors are derived from a knowledge source, rather than on distributional vectors derived from

cooccurrence counts. Thus, we use the term *vector based measure* interchangeably with *concept vector based measure*.

Patwardhan and Pedersen (2006) represent a concept by a second-order gloss vector using WordNet glosses (PP06). They start with first-order context vectors, i.e., a vector of cooccurrence counts for each content word in a corpus. In this case, the corpus is the set of WordNet glosses. A second-order gloss vector $glossVector(c_i)$ is then constructed from the gloss of the target concept c_i by combining the first-order gloss vectors of words that appear in that gloss. For example, from the gloss of *tree (plant) 'a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown'*, the algorithm would construct a second-order gloss vector from the first-order gloss vectors of *plant*, *trunk*, *branches*, *crown*, etc. The relatedness of two concepts is then computed as the cosine of the second-order gloss vectors.[5]

$$rel_{\text{PP06}}(c_1, c_2) = \frac{glossVector(c_1) \cdot glossVector(c_2)}{|glossVector(c_1)|\,|glossVector(c_2)|}$$

Patwardhan and Pedersen (2006) also introduce a variant of this algorithm to compensate for short glosses, where a gloss is augmented with glosses of concepts that are in close relation to the original concept. This is conceptually close to the extended BP02 measure described in the previous section.

Gabrilovich and Markovitch (2007) introduce another vector based measure (GM07), where the meaning of a word w is represented as a high-dimensional concept vector $\vec{d}(w) = (d_1, \ldots, d_N)$. Each element d_i represents a document, and the value of d_i depends on the occurrence of the word w in this document. This is very similar to the approach by Qiu and Frei (1993) for constructing a similarity thesaurus used in information retrieval.

Gabrilovich and Markovitch (2007) derive the concept vector from Wikipedia articles $a_1, \ldots, a_N$, as each article focuses on a certain topic, and can thus be viewed as expressing a concept. The dimension of the concept vector is the number of Wikipedia articles N. Each element of the concept vector $\vec{d}$ is associated with a certain Wikipedia article a_i. If the word w can be found in this article, the word's tf.idf score (Salton and McGill 1983) in the article a_i is assigned to the concept vector element d_i. Otherwise, 0 is assigned.

$$d_i = \begin{cases} tf.idf(w), & w \in a_i \\ 0, & otherwise \end{cases}$$

As a result, the vector $\vec{d}(w)$ represents the word w in concept space. Semantic relatedness of two words can then be computed as the cosine of their corresponding concept vectors:

$$rel_{\text{GM07}}(w_1, w_2) = \frac{\vec{d}(w_1) \cdot \vec{d}(w_2)}{\left|\vec{d}(w_1)\right| \left|\vec{d}(w_2)\right|}$$

[5] This measure displays properties of of both gloss based and vector based approaches. It is categorized as a vector based measure, because the final relatedness computation relies on a vector representation that is only derived from glosses.

Generalizing the GM07 measure The GM07 measure relies on Wikipedia articles. However, the measure can be generalized to each knowledge source containing textual descriptions of concepts. WordNet contains glosses that can be seen as very short 'articles' describing the concepts expressed by WordNet synsets. For example, the term *car* is contained in more than 250 glosses of WordNet concepts including nouns (e.g., *polishing*, 'every Sunday he gave his car a good polishing'), verbs (e.g., *damage*, 'She damaged the car when she hit the tree'), and adjectives (e.g., *unfastened*, 'the car door was unfastened'). Each of these concepts leads to a nonzero entry in the resulting concept vector for *car*. When computing semantic relatedness, the whole vector is taken into account, and so are all the implicit relations to *car*-related concepts encoded in the glosses. In the case of GermaNet, that contains only few glosses, we need to construct pseudo glosses (as described in the context of the Gur05 measure) as a proxy for textual descriptions of a concept.

In general, we can define a concept vector based measure (ZG07) (Zesch and Gurevych 2007) that can also be applied to WordNet (by using glosses), GermaNet (by using pseudo glosses), or to any other knowledge source where we can retrieve or construct a textual description for each concept. In very recent work, the ZG07 measure was also adapted to entries in the wiki dictionary Wiktionary (Zesch *et al.* 2008). Wiktionary combines a collaborative construction approach with explicitly encoded lexical semantic relations, glosses, translations, etc. Thus, it provides a rich set of focused additional knowledge associated with each concept.

2.2 Knowledge sources

Computing semantic relatedness requires some knowledge source, e.g., dictionaries, thesauri, or wordnets. Dictionaries, like the Longman Dictionary of Contemporary English (Procter 1978), and thesauri, like Roget's Thesaurus (Roget 1962) or the Macquarie Thesaurus (Bernard 1986), have been employed for computing semantic relatedness (Morris and Hirst 1991; Kozima and Furugori 1993; Jarmasz and Szpakowicz 2003; Mohammad *et al.* 2007). However, most relatedness measures use a semantic wordnet. The English WordNet (Fellbaum 1998) is a well known representative, but also wordnets in other languages are available, e.g., GermaNet (Kunze 2004) for German. However, non-English wordnets are usually less developed containing fewer concepts and semantic relations. Recently, Wikipedia was discovered as a new knowledge source (Gabrilovich and Markovitch 2007; Zesch *et al.* 2007). A detailed analysis of Wikipedia's properties as a lexical semantic resource is given by (Zesch, Gurevych and Mühlhäuser 2007a). In the remainder of this section, we analyze the knowledge sources with respect to their applicability to each type of SemRel measures described in Section 2.1.

Gloss based measures Dictionaries, some wordnets, and Wikipedia usually contain definitions for each concept. Gloss based measures can then be directly applied. Thesauri or wordnets lacking glosses can use pseudo glosses (Gurevych 2005) as a substitute (see the description of the Gur05 measure in the previous section).

Vector based measures Concept vectors are derived from textual representations of concepts. Thus, they can be applied to each knowledge source that offers such a textual representation of a concept, e.g., glosses, pseudo glosses, or Wikipedia articles.

Path and IC based measures These two types of SemRel measures rely on a graph of concepts, where nodes represent concepts, and edges represent relations between these concepts. For example, in a dictionary, a term's definition contains other terms that can be found in the dictionary. This can be used to form a relationship graph between dictionary entries. Thesauri consist of a hierarchy of categories, where related terms are grouped together on the leaf level. Semantic wordnets group synonyms together (*synsets*) and link them by means of semantic relations between these *synsets* or lexical relations between single lexemes. The result is a graph with a backbone consisting of classical taxonomic relations.

Wikipedia contains a hierarchical category graph (Zesch and Gurevych 2007), having a single root node. This graph is known to resemble a thesaurus, where relations between nodes are not as strictly defined as in linguistically motivated wordnets, but their semantics is more of the kind 'broader term' or 'narrower term' (Voss 2006). Therefore, adapting path based and information content based SemRel measures to Wikipedia requires some modifications to the original definitions that are described in the following section.

2.3 Adapting path based and IC based measures to Wikipedia

In this section, we describe how path based and information content based SemRel measures originally defined on WordNet can be adapted to compute semantic relatedness based on a combination of the Wikipedia category graph (WCG) and a list of Wikipedia articles and redirects.

We start from the observation that Wikipedia articles are not organized in a hierarchical structure as required by path based and IC based measures. This role is filled by the WCG. Its nodes representing categories are connected by 'broader term' and 'narrower term' relations. Zesch and Gurevych (2007) performed a graph-theoretic analysis of the WCG and showed that it is a scale-free, small-world graph. The graph properties of WCG and WordNet are very similar, e.g., the depth of the WordNet hierarchy is 16, while it is 14 for the WCG.

However, the WCG alone cannot be used to compute the SemRel between terms, as its nodes usually represent generalized concepts or categories instead of simple terms. In ontological terms, it does not contain any instances, but only classes (e.g., the English WCG does not contain a node for 'seatbelt' but a category 'Automotive safety technologies'). Therefore, the WCG alone would not provide sufficient coverage for our experiments. We use the mutual links between Wikipedia articles and categories (see Figure 3) to connect articles to categories, and then utilize the hierarchy of categories to determine semantic relatedness.[6] While the number of

[6] We do not make any use of the links between articles.

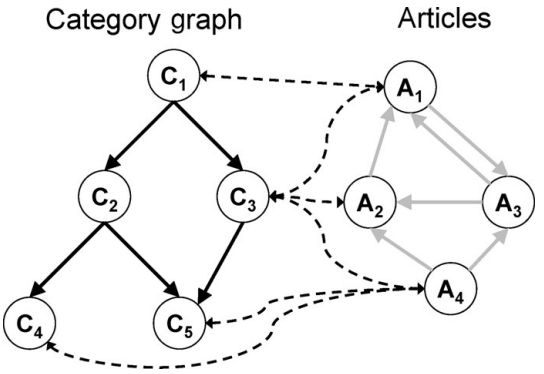

Fig. 3. Relations between Wikipedia articles and categories in the category graph.

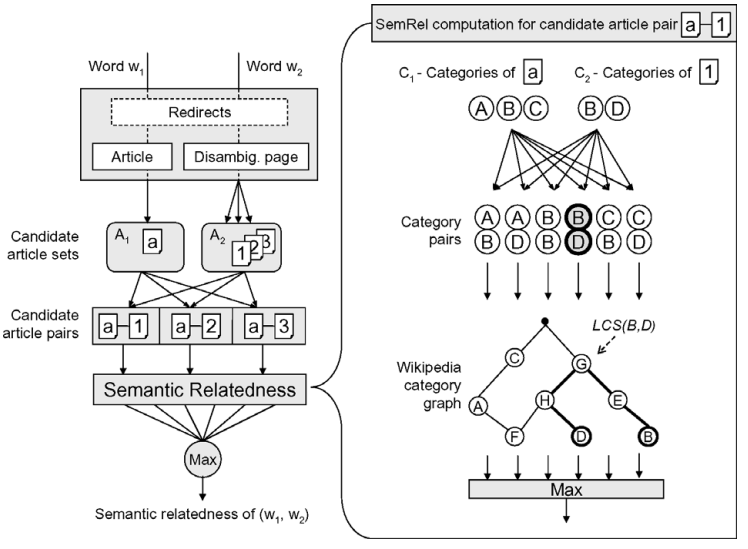

Fig. 4. Information flow for SemRel measures adapted to work on Wikipedia.

categories in Wikipedia (about 210,000 for the version from February 6th, 2007 used in this article) is comparable to the number of synsets in WordNet (about 120,000 in WordNet 3.0), the number of articles in the English Wikipedia is an order of magnitude higher providing much more coverage. In the remainder of this section, we give a formal description of the adaptation process that was first described in (Zesch *et al.* 2007).

To compute semantic relatedness of two words w_1 and w_2 using Wikipedia, we first retrieve the articles or disambiguation pages with titles that equal w_1 and w_2 (see Figure 4). If we hit a redirect page, we retrieve the corresponding article or disambiguation page instead. In case of an article, we insert it into the candidate article set (A_1 for $w1$, A_2 for w_2). In case of a disambiguation page, the page contains links to all encoded word senses, but it may also contain other links. Therefore, we only consider links conforming to the pattern ⟨Title (DisambiguationText)⟩ where '(DisambiguationText)' is optional – (e.g., 'Bank' or 'Bank (sea floor)') or to a

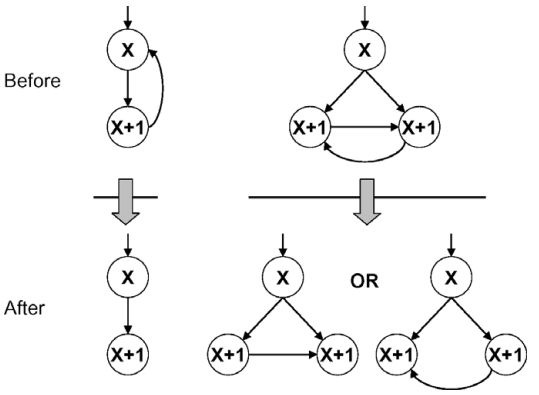

Fig. 5. Breaking cycles in the WCG.

person's name pattern. Following all such links gives the candidate article set. If no disambiguation links conforming to the pattern are found, we take the first link on the page, as most important links tend to come first, and add the corresponding article to the candidate set. We form pairs from each candidate article $a_i \in A_1$ and each article $a_j \in A_2$. We then compute $rel(a_i, a_j)$ for each pair. The output of the algorithm is the maximum SemRel value $\max_{a_i \in A_1, a_j \in A_2}(rel(a_i, a_j))$.

For computing the SemRel value $rel(a_i, a_j)$, we define C_1 and C_2 as the set of categories assigned to article a_i and a_j, respectively. We then determine the SemRel value for each category pair (c_k, c_l) with $c_k \in C_1$ and $c_l \in C_2$. We choose the maximum SemRel value among all pairs (c_k, c_l).

For the information content based measures, we need to compute the information content of a concept. We use intrinsic information content (IC_{Sec04}), relying on hyponym counts in the original definition. As links in the WCG are untyped, we define all 'narrower term' links (i.e., links to concepts deeper in the hierarchy) as pseudo hyponyms. Efficiently counting the hyponyms of a node requires to break cycles that may occur in the WCG. We perform a colored depth-first-search to detect cycles, and break them as visualized in Figure 5. A link pointing back to a node closer to the root node is deleted, as it violates the rule that downward links in the WCG typically express 'narrower term' relations. If the cycle occurs between nodes on the same level, we cannot decide based on that rule and randomly delete one of the links running on the same level. This strategy never disconnects any nodes from the graph.

Walkthrough example If we want to determine the semantic relatedness between the terms 'Zuse' (the last name of a famous computer pioneer) and 'Dijkstra' (the last name of another famous computer scientist), we first get the articles corresponding to these terms. For 'Zuse', we get redirected to the article 'Konrad Zuse'. For 'Dijkstra', we hit a disambiguation page, as there are a couple of famous persons called 'Dijkstra'. For the sake of the example, we consider only the first two persons mentioned on the disambiguation page 'Edsger W. Dijkstra' and 'Rineke Dijkstra'. We now form article pairs between the one article related

to 'Zuse' and each of the two persons called 'Dijkstra' yielding 'Zuse/Edsger W. Dijkstra' and 'Zuse/Rineke Dijkstra'. Then, we look at the categories assigned to each article. For the sake of the example, we only consider one category per article. The articles 'Zuse' and 'Edsger W. Dijkstra' both have the category *Computer pioneers*, while 'Rineke Dijkstra' has the category *Dutch photographers*. We now form category pairs yielding *Computer pioneers*/*Computer pioneers* and two times *Computer pioneers*/*Dutch photographers*. We then measure the semantic relatedness in terms of the path length between the categories in the hierarchy. Assume that for *Computer pioneers*/*Dutch photographers* it is 4, while for *Computer pioneers*/*Computer pioneers* it is clearly 0. If the maximum path length in the category graph is 8, then the semantic relatedness between *Computer pioneers*/*Dutch photographers* is 0.5, and between *Computer pioneers*/*Computer pioneers* it is 1. We now take the maximum of those values (i.e., 1) and select it as the value of semantic relatedness between 'Zuse' and 'Dijkstra', i.e., the terms are very highly related.[7]

Strube and Ponzetto (2006) take a similar approach to adapting some WordNet based measures to Wikipedia using the WCG. However, they use a different disambiguation heuristic. It relies on finding a common substring in links on disambiguation pages. As there is no Wikipedia editing principle that enforces a standardized vocabulary, this strategy sometimes fails even for closely related concepts (e.g., 'Bank (sea floor)' and 'Water' do not have any common substring). In our experiments (see Section 4), we found that the substring heuristic is used in less than 5% of cases. Thus, the disambiguation strategy almost fully relies on a fallback strategy that takes the first sense on a page that is considered to be the most common one. This strategy is not optimal for modeling the lexical cohesion of a text, as cohesion might be established by rare senses in certain domains. For example, in the sentence 'This tree spans the whole graph' the special sense of 'tree (graph theory)' contributes much to the lexical cohesion of the sentence, but cannot be determined using a heuristic that depends on the most common sense. However, better disambiguation strategies have to be developed, e.g., incorporating contextual information.

The adaptation of SemRel measures described in this section requires efficient access to different types of lexical semantic information available from Wikipedia. In Section 5.2, we present JWPL – a high performance Java-based Wikipedia API, that is especially suitable for turning Wikipedia into a knowledge source for NLP applications (Zesch *et al.* 2007a; Zesch, Müller and Gurevych 2008a).

3 Evaluation framework

The prevalent approaches for evaluating SemRel measures are (i) mathematical analysis (Lin 1998), (ii) application-specific evaluation (Gurevych and Strube 2004; Budanitsky and Hirst 2006), (iii) correlating semantic relatedness with human

[7] Note that we could not have used the category graph alone for computing semantic relatedness, as it contains neither a node 'Zuse' nor a node 'Dijkstra'.

judgments (Budanitsky and Hirst 2006), and (iv) solving word choice problems (Jarmasz and Szpakowicz 2003).

Mathematical analysis (Lin 1998) can assess a measure with respect to some formal properties, e.g., whether it is a metric,[8] but cannot tell us whether a measure closely resembles human judgments, or how it performs in a certain application.

The latter question is tackled by application-specific evaluation, where a measure is tested within the framework of a usually complex application. However, application specific evaluation entails influence of parameters besides the measure of semantic relatedness being tested. Gurevych and Strube (2004) evaluated a set of WordNet-based measures of semantic similarity for the tasks of dialog summarization, and did not find any significant differences in their performance. Rather, the performance of a specific measure is tightly correlated with the properties of the underlying knowledge source, as shown by Gurevych *et al.* (2007) when evaluating SemRel measures on an information retrieval task. Budanitsky and Hirst (2006) evaluated semantic relatedness measures on the task of real word error detection and found that the choice of a specific measure influences detection performance.

The remaining approaches, comparison with human judgments and solving word choice problems, are used in the present article to gain deeper insights into the nature of semantic relatedness, as well as the performance of measure types and their dependence on the knowledge source.

3.1 Correlation with human judgments

3.1.1 Evaluation measures

Semantic relatedness measures can be evaluated using two different correlation methods. The first method is to correlate the scores computed by a measure with the numeric judgments of semantic relatedness provided by humans. For example, on a 0–4 scale, where 4 is maximum relatedness, the pair 'car–automobile' might get a human judgment of 3.9 and a score of 3.7 from a measure. A second pair 'car–garden' only gets 1.1 (human) and 0.4 (machine). The Pearson product-moment correlation coefficient r can be employed as an evaluation measure. It indicates how well the results of a measure resemble human judgments, where a value of 0 means no correlation and a value of 1 means perfect correlation.

The second method is correlating word pair rankings. In a ranking task, a human and a measure would simply rank the pair 'car–automobile' higher than 'car–garden'. The ranking produced on the basis of the measure is compared to the one produced on the basis of human judgments. The quality of such a ranking is quantified by the Spearman rank order correlation coefficient ρ, where a value of 0 means no correlation and a value of 1 means perfect correlation.

Existing work on computing SemRel often employed Pearson correlation. However, this suffers from some limitations: First, Pearson correlation is very sensitive to outliers. Even a single outlier might yield fundamentally different results. This

[8] A metric fulfills the mathematical criteria: (i) $dist(c_1, c_2) \geq 0$; (ii) $dist(c_1, c_2) = 0 \Leftrightarrow c_1 = c_2$; (iii) $dist(c_1, c_2) = dist(c_2, c_1)$; and (iv) $dist(c_1, c_3) \leq dist(c_1, c_2) + dist(c_2, c_3)$.

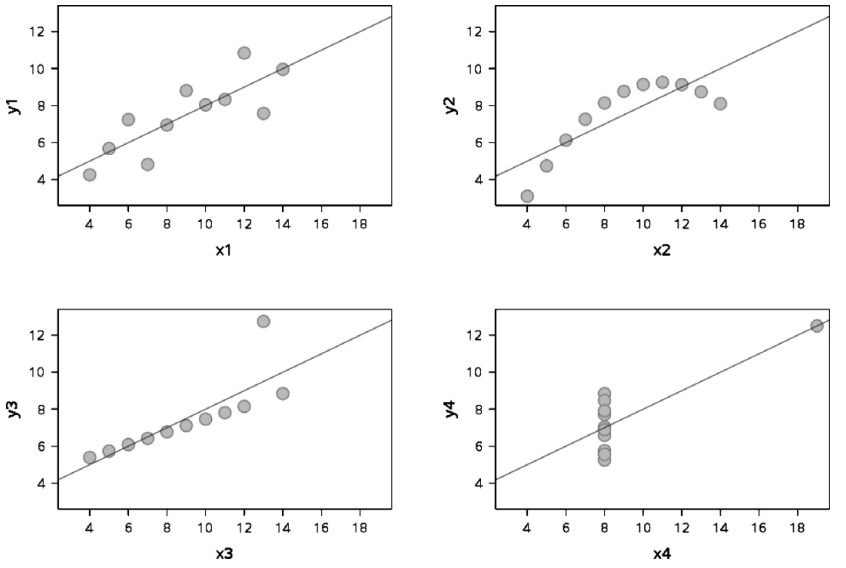

Fig. 6. 'Anscombe's quartet' showing relationships with exactly the same mean, standard deviation, regression line, and Pearson correlation of $r = 0.81$. (Adapted from http://en.wikipedia.org/wiki/Image:Anscombe.svg)

is visualized by Anscombe's quartet (Anscombe 1973), a set of four scatterplots showing relationships with exactly the same mean, standard deviation, regression line and Pearson correlation of $r = 0.81$ (see Figure 6). In the bottom figures, a single outlier is sufficient to disturb a perfect correlation (bottom left) or produce a high correlation in a fully nonlinear relationship (bottom right).

Second, Pearson correlation measures the strength of the linear relationship between human judgment and SemRel scores computed by a measure. If a relationship is not linear, results are flawed. For example, the upper right chart in Figure 6 shows a nonlinear relation that cannot be correctly measured using Pearson correlation.

Third, Pearson correlation requires the two random variables (the vectors) to be normally distributed and measured on interval scales. In Figure 6 only the variables in the upper left plot fulfill the prerequisite of being normally distributed. However, the real distribution of relatedness values is largely unknown. The values of most samples (small subsets of word pairs judged by humans) are not normally distributed. Recent findings (Budanitsky and Hirst 2006; Zesch and Gurevych 2007) even indicate that the relatedness values as perceived by humans are not interval scaled.

In contrast to these limitations, Spearman's rank correlation coefficient is robust against outliers, and can also measure the strength of nonlinear relationships. It does not pose the assumption of interval scales, i.e., it can be used for variables measured on the ordinal level. Additionally, Spearman's ρ does not make any assumptions about the distribution of the vectors being compared. However, there are also some disadvantages of using Spearman rank correlation to evaluate the performance of

Table 1. *Datasets used for evaluating semantic relatedness*

Dataset	Year	Language	# Pairs	POS[a]	Scores	# Subjects
RG-65	1965	English	65	N	0–4	51
MC-30	1991	English	30	N	0–4	38
Res-30	1995	English	30	N	0–4	10
Fin-353	2002	English	353	N, V, A	0–10	13/16
Fin1-153			153			13
Fin2-200			200			16
YP-130	2006	English	130	V	{0,1,2,3,4}	6
Gur-30	2005	German	30	N	{0,1,2,3,4}	24
Gur-65	2005	German	65	N	{0,1,2,3,4}	24
Gur-350	2006	German	350	N, V, A	{0,1,2,3,4}	8
ZG-222	2006	German	222	N, V, A	{0,1,2,3,4}	21

[a] The parts-of-speech of words in the dataset. 'N' means noun, 'V' means verb, and 'A' means adjective or adverb.

SemRel measures. From the statistical literature, it is known that Spearman's ρ tends to give higher values than Pearson's r for datasets with many tied ranks. For some applications, that rely on thresholding SemRel values, a measure that yields a perfect ranking might be of little use, if the differences between SemRel scores are too small to be sensibly thresholded.

For comparison with previous results, we report both Pearson's r and Spearman's ρ. Still, we recommend using Spearman rank correlation in future experiments, as this evaluation is more objective, if the performance of SemRel measures has to be evaluated intrinsically. As Pearson correlation and Spearman correlation are not directly comparable and might yield very different results under certain conditions, special care must be taken when comparing and interpreting such results.

3.1.2 Datasets

Evaluation datasets for correlation analysis are created by asking human annotators to judge the relatedness of presented word pairs. The gold standard score assigned to a word pair is the average score over all human judges (see Section 5.1 for a more detailed discussion on the creation of such datasets). For evaluation, a gold standard dataset is then correlated with the SemRel values computed by a particular measure.

An upper bound for the performance of a measure on a dataset is the inter-annotator agreement (InterAA), i.e., the amount of mutual agreement between human judges. InterAA is computed as the average pairwise Pearson correlation between human judges. As the distribution of Pearson's r is left-skewed, we cannot simply average the correlations, but have to use a Fisher Z-value transformation. In contrast to InterAA, the *intra*-annotator agreement (IntraAA) measures the agreement of a judge with herself over time. It is computed analogously to the inter-annotator agreement. Unfortunately, only very few experiments with intra-annotator agreement have been performed so far. Table 1 gives an overview of the datasets

Table 2. *Inter- and intra-annotator agreement on evaluation datasets. Missing values are not available from the references*

Dataset	Language	Correlation *r*	
		InterAA	IntraAA
RG-65	English	(.80)	.85
MC-30	English	–	–
Res-30	English	.90	–
Fin-353	English	–	–
Fin1-153	English	.73	–
Fin2-200	English	.55	–
YP-130	English	.87	–
Gur-30	German	–	–
Gur-65	German	.81	–
Gur-350	German	.69	–
ZG-222	German	.49	.65

for evaluating SemRel described below. Table 2 reports the InterAA and IntraAA values for those datasets.

Rubenstein and Goodenough (1965) created a dataset with 65 English noun pairs (RG-65). No InterAA was reported for this dataset, but Pirro and Seco (2008) repeated the experiment. The InterAA for native speakers is $r = 0.80$. A subset of the RG65 dataset has been used for experiments by Miller and Charles (1991) (MC-30). Resnik (1995) repeated the experiment (Res-30) and reported an InterAA of $r = 0.90$.

As creating datasets of this kind is time consuming and costly, most work on evaluating SemRel measures focused on such small-scale experiments restricted to nouns (Li, Bandar and McLean 2003; Budanitsky and Hirst 2006; Patwardhan and Pedersen 2006). This leads to an effect of overfitting algorithms to these specific datasets and the employed knowledge source. Many algorithms yield near human performance on these particular datasets using WordNet as a knowledge source. This is due to the strongly related word pairs in these datasets being only related by classical lexical semantic relations that are well modeled in WordNet.

We argue that previous evaluations restricted to those datasets were limited with respect to (i) the number of word pairs involved, (ii) the parts of speech of word pairs, (iii) approaches to select word pairs (manual versus automatic, analytic versus corpus based), and (iv) the kinds of semantic relations that hold between word pairs. However, an evaluation involving the aspects described above is crucial to understand the properties of a specific measure and the results obtained under certain experimental conditions (e.g., the knowledge source used). Discovering the true nature of SemRel requires larger datasets with word pairs from different part of speech that are connected by classical as well as nonclassical relations.

Finkelstein *et al.* (2002) created a larger dataset for English containing 353 word pairs (Fin-353). This dataset has been criticized in the literature (cf. Jarmasz and Szpakowicz 2003) for being culturally biased. Another problem with this dataset

is that it consists of two subsets, which have been annotated by different human judges. The first subset also contains the 30 word pairs from MC-30. We performed further analysis of their dataset and found that the InterAA differs considerably for the two subsets ($r = 0.73$ versus $r = 0.55$). Therefore, we treat them as independent datasets Fin1-153 and Fin2-200 henceforth.

Yang and Powers (2006) created a dataset (YP-130) that contains 130 verb pairs. They report a high InterAA of $r = 0.87$. As this dataset contains only verbs, the evaluation will be particularly informative about the ability of a SemRel measure to estimate verb relatedness.

Several German datasets have also been created (see Table 1). Gurevych (2005) conducted experiments with a German translation of the English RG-65 dataset (Gur-65). The subset of the Gur-65 dataset with the translated word pairs corresponding to the MC-30 dataset is called Gur-30.

As the Gur-65 dataset is small and contains only noun pairs connected by either synonymy or hyponymy, she conducted a follow-up study, and collected a larger dataset containing 350 word pairs (Gur-350). It contains nouns, verbs and adjectives that are connected by classical and nonclassical relations. However, word pairs for this dataset are biased toward strong classical relations, as they were manually selected. Thus, Zesch and Gurevych (2006) propose an approach to create word pairs from domain specific corpora using a semi-automatic process (see Section 5.1 for a more detailed description). The resulting ZG-222 dataset contains 222 domain specific word pairs that are connected by different kinds of lexical semantic relations. As human judgments on domain specific word pairs depend on the domain knowledge of the judges, the InterAA is relatively low ($r = 0.49$).

In Section 4.2, we describe the results of evaluating the SemRel measures introduced in Section 2.1 on the English and German datasets presented in this section.

3.2 Solving word choice problems

3.2.1 Evaluation measures

A different approach to evaluate the performance of SemRel measures relies on word choice problems (Jarmasz and Szpakowicz 2003; Turney 2006). A word choice problem consists of a target word and four candidate words or phrases. The objective is to pick the one that is most closely related to the target. An example problem is given below. There is always only one correct candidate, 'a' in this case.

beret

(a) round cap	(b) cap with horizontal peak
(c) wedge cap	(d) helmet

The relatedness between the target and each of the candidates is computed by a SemRel measure, and the candidate with the maximum SemRel value is chosen. We lemmatize the target and all candidates. This is especially beneficial for German words that can be highly inflected.

If two or more candidates are equally related to the target, then the candidates are said to be tied. If one of the tied candidates is the correct answer, then the problem is counted as correctly solved, but the corresponding score is reduced. We assign a score s_i of $\frac{1}{\#\text{ of tied candidates}}$ (in effect approximating the score obtained by randomly guessing one of the tied candidates). Thus, a correctly solved problem without ties is assigned a score of 1.

If a phrase or a multiword expression is used as a candidate and cannot be found in the knowledge source, we remove stopwords (prepositions, articles, etc.) and split the candidate phrase into component words. For example, the target 'beret' in the above example has 'cap with horizontal peak' as one of its answer candidates. The candidate phrase is split into its component content words 'cap', 'horizontal', and 'peak'. We compute semantic relatedness between the target and each phrasal component and select the maximum value as the relatedness between the target and the candidate. If the target or all candidates cannot be found in the knowledge source, a SemRel measure does not attempt to solve the problem. The overall score S of a SemRel measure is the sum of the scores yielded on the single problems $S = \sum_{wp_i \in A} s(wp_i)$, where A is the set of word choice problems that were attempted by the measure, and wp_i is a certain word choice problem.

Jarmasz and Szpakowicz (2003) used the overall score S for evaluation. However, this evaluation approach is problematic, as a measure that attempts more problems may get a higher score just from random guessing. Mohammad *et al.* (2007) used precision, recall and f-measure for evaluation. For word choice problems, recall is defined as $R = \frac{S}{n}$, where n is the total number of word choice problems. Under that definition, if the score S is 0 then a SemRel measure has a recall of 0, regardless of how many word choice problems were attempted. This stands in contrast to the use of precision and recall in information retrieval, where just retrieving all documents will always give a recall of 1, regardless whether they are relevant or not. For word choice problems, just attempting all problems will only yield a recall of 1 if all attempted problems are correctly solved at the same time. Thus, for this task, recall is of very limited value for judging about the performance of a semantic relatedness measure. We therefore decided not to use precision and recall, but evaluate the word choice problems using accuracy and coverage instead. We define accuracy as $Acc = \frac{S}{|A|}$, where S is the overall score as defined above and A is the number of word choice problems that were attempted by the SemRel measure. Coverage is then defined as $Cov = \frac{|A|}{n}$, where n is the total number of word choice problems. Accuracy indicates how many of the attempted problems could be answered correctly, and coverage indicates how many problems were attempted.

3.2.2 Datasets

The English dataset contains 300 word choice problems collected by Jarmasz and Szpakowicz (2003). We collected a German dataset from the January 2001 to December 2005 issues of the German language edition of Reader's Digest (Wallace and Wallace, 2005). We discarded 44 problems that had more than one correct candidate, and 20 problems that used a phrase instead of a single term as the

target. The remaining 1008 problems form our German word choice dataset, which is significantly larger than any of the previous datasets employed in this type of evaluation. Additionally, it is not restricted to synonym problems, but also includes hypernymy/hyponymy, and some candidates that are related to the target by nonclassical relations.

We tested human performance on a subset of 200 manually selected German word choice problems using 41 native speakers of German. Human coverage on word choice problems is always perfect, as the experimental setting did not allow to skip word choice problems. We found that human accuracy on this task strongly depends on the level of language competence of the subjects. Average accuracy was $Acc = 0.71$ with $\sigma = 0.10$. We also observed a Pearson correlation $r = 0.69$ between a subject's accuracy and her age (statistically significant, two tailed t-test with $\alpha = 0.01$). The highest accuracy was 0.91 (by the oldest subject), the lowest 0.45 (by the youngest subject).

4 Results and analysis

In this section, we first describe the configuration of measures used in the experiments in this article. We then present the results on the two evaluation tasks described in Section 3 (correlation with human judgments and solving word choice problems), and finally discuss the results.

4.1 Configuration of measures

WordNet We use WordNet 3.0 and the measures as available in the Perl WordNet::Similarity package (Patwardhan *et al.* 2003). For constructing the concept vectors used by the ZG07 measure, we treat each WordNet synset as a concept, and its gloss as the concept's textual representation. We access WordNet using the JWNL[9] WordNet API.

GermaNet We have adapted most WordNet based measures to GermaNet using the GermaNet-API[10] applied to GermaNet 5.0. We construct pseudo glosses for the Gur05 measure by concatenating the lemmas of all synsets that are reachable within a radius of three from the original synset. We use the same pseudo glosses as textual representations of synsets for constructing the concept vectors used by the ZG07 measure operating on GermaNet.

Wikipedia We use the English and German Wikipedia dumps from February 6th, 2007 that are offered by the Wikimedia Foundation.[11] We access Wikipedia using the JWPL Wikipedia API that is described in Section 5.2.

[9] `http://jwordnet.sourceforge.net/`
[10] `http://projects.villa-bosch.de/nlpsoft/gn_api/index.html`
[11] `http://download.wikipedia.org/`

Path based and information content based SemRel measures that were originally defined on WordNet are adapted to Wikipedia as described in Section 2.3. We redefine the LC98 measure as follows:

$$rel_{\text{LC98}}(c_1, c_2) = -\log \frac{(l(c_1, c_2) + 1)}{2 \times depth}$$

where *depth* is the depth of the Wikipedia category graph, and $l(c_1, c_2)$ is measured in edges.[12] The increase of the path length by 1 is needed as by definition $l(c_i, c_i)$ returns 0, and $log(0)$ is not defined.

For gloss based and vector based measures, we differentiate between considering the full Wikipedia article as the textual representation, or just the first paragraph. The first paragraph usually contains a definition of the concept described in that article. As some words in the latter parts of an article are likely to describe less important or even contradicting topics, we expect this refined measures to yield a better performance by trading in some coverage. In the following, we flag the measures using only the first paragraph with the suffix 'first'.

For the GM07 measure, we prune the concept space due to performance reasons by only considering Wikipedia articles as concepts if they contain at least 100 words and have more than five ingoing links and five outgoing links. As the GM07first measure only uses the first paragraph of each article, we do not need any performance tuning and consider all articles as concepts. In the experiments with the English path based measures applied to Wikipedia, we limited the search for a shortest path to five edges due to performance reasons.

4.2 Correlation with human judgments

In this section, we present the experimental results obtained on the task of correlating word pair rankings. The interannotator agreement is only given as an approximate indicator of human performance, but as its value expresses Pearson correlation between annotators, it cannot be directly compared with the Spearman correlation values presented in the charts. Tables 3 and 4 give an overview of the results on the English and German datasets. To ensure a fair comparison of the measures' performance, we only use the subset of word pairs that is covered by all measures. We do not consider the WuP94 measure on GermaNet with the ZG-222 dataset, as the measure covers too few word pairs.

In Section 3, we recommended Spearman correlation as a less problematic evaluation measure compared with Pearson correlation. The Les86 measure applied on the MC-30 dataset using WordNet as a knowledge source gives a good example

[12] It is a controversial question whether to count edges or nodes in a path. Both approaches have been used in the past. We recommend counting edges based on the following argumentation: If two concepts are identical, they are mapped to the same node. Both methods, measuring distance in edges as well as in nodes, will assign a distance of 0 in this case. If two nodes are direct neighbors, they are one edge but zero nodes apart. As a result, when measuring in nodes, there is no way to differentiate the distance of identical or neighboring concept nodes. This clearly puts measuring in edges in favor. Thus, we have redefined the original definitions of measures slightly from measuring nodes to edges, wherever necessary.

Table 3. *Correlations with human judgments on English datasets. Best Spearman correlation values for each dataset are in bold. Nonsignificant correlations are in italics (two tailed t-test, $\alpha = .05$). 'WN' means WordNet. 'Wiki' means Wikipedia*

Dataset			MC-30		RG-65		Fin1-153		Fin2-200		YP-130	
	Wordpairs used		30		65		144		190		80	
		Type[a]	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
	InterAA		—	.90	—	.80	—	.73	—	.55	—	.87
WN	Rad89	PL	.75	.76	.79	.79	.33	.38	.24	.36	.64	.74
	LC98	PL	.75	.80	.79	.84	.33	.34	.24	.31	.64	.74
	WuP94	PL	.77	.78	.78	.80	.38	.28	.24	.24	.67	.76
	HSO98	PL	.76	.67	.79	.73	.33	.35	.32	.35	.61	.70
	Res95	IC	.72	.79	.74	.81	.35	.36	.26	.31	.61	.69
	JC97	IC	.68	.69	.58	.71	.28	.34	*.10*	.28	.68	.61
	Lin98	IC	.67	.75	.60	.72	.27	.31	.17	.27	.67	.76
	Les86	G	**.78**	*.36*	.72	.34	.47	.20	.33	.28	.64	.39
	PP06	V	.54	.52	.62	.58	*.10*	*.11*	*.10*	*.09*	.39	.38
	ZG07	V	.77	.44	**.82**	.49	**.60**	.33	**.51**	.48	**.71**	.63
Wiki	Rad89	PL	*.33*	.36	.41	.40	.40	.40	.32	.32	*.07*	*.06*
	LC98	PL	*.33*	*.34*	.41	.41	.40	.38	.32	.33	*.07*	*.05*
	WuP94	PL	*.29*	*.27*	.31	.32	.34	.32	.18	.20	*.10*	*.12*
	Res95	IC	.51	.47	.40	.36	.30	.18	.22	.19	*.16*	*.13*
	JC97	IC	*.14*	*.20*	*.20*	*.14*	.29	.27	*.12*	*.14*	*.16*	*.15*
	Lin98	IC	.52	.50	.41	.43	.31	.25	.22	.20	*.16*	*.13*
	Les86first	G	*.21*	*.27*	*.07*	*.19*	.22	.25	*.07*	*.01*	*.00*	*.02*
	Les86	G	*.36*	.42	.25	.35	.36	.35	.16	.17	*.15*	*.16*
	GM07first	V	**.67**	.46	**.75**	.49	**.69**	.32	**.51**	.39	**.29**	.29
	GM07	V	.64	.59	.71	.54	.61	.50	.28	.29	**.29**	.30

[a] PL = path length, IC = information content, G = gloss, V = vector

(see Table 3). The measure shows a remarkable difference of 0.42 between a nonsignificant Pearson correlation of $r = 0.36$ and a Spearman rank correlation of $\rho = 0.78$, which is the best value on this dataset. The low r value is caused by a single word pair (car–automobile), where both words are mapped to the same synset in WordNet. Thus, they have identical glosses resulting in an extra-ordinary high overlap value that represents an outlier in the dataset. If we smooth the distribution by using the natural logarithm of the relatedness values returned by Les86, the resulting Pearson correlation coefficient increases to $r = 0.75$.

Wisdom of crowds versus wisdom of linguists Figures 7 and 8 show the best performing SemRel measures on all datasets. Contrary to previous research (Strube and Ponzetto 2006; Gabrilovich and Markovitch 2007; Zesch *et al.* 2007), 'wisdom of crowds' based knowledge sources are not generally superior to the 'wisdom of linguists' based knowledge sources. Most differences between the best performing SemRel measures on different knowledge sources are not statistically significant. The

Table 4. *Correlations with human judgments on German datasets. Best Spearman correlation values for each dataset are in bold. Nonsignificant correlations are in italics (two tailed t-test, $\alpha = .05$). 'GN' means GermaNet. 'Wiki' means Wikipedia. We do not consider the WuP94 measure on GermaNet with the ZG-222 dataset, as the measure covers too few word pairs*

Dataset			Gur-30		Gur-65		Gur-350		ZG-222	
Word pairs used		Type[a]	22		52		131		70	
			ρ	r	ρ	r	ρ	r	ρ	r
	InterAA		—	—	—	.81	—	.69	—	.49
GN	Rad89	PL	.78	.79	.82	.78	.51	.51	**.34**	.32
	LC98	PL	.79	.83	**.83**	.85	.51	.53	.29	.33
	WuP94	PL	.75	.77	.77	.76	.52	.54	—	—
	Res95	IC	.78	.81	.81	.83	.35	.39	*.14*	*.16*
	JC97	IC	**.84**	.87	.78	.81	.50	.49	*.17*	*.08*
	Lin98	IC	.77	.85	.78	.85	.47	.50	*.15*	*.14*
	Gur05	G	.67	.53	.76	.56	**.56**	.45	*.22*	*.17*
	ZG07	V	.72	.70	.78	.75	.54	.49	*.20*	*.23*
Wiki	Rad89	PL	.65	.62	.50	.48	.37	.39	.39	.42
	LC98	PL	.65	.60	.50	.49	.37	.40	.37	.40
	WuP94	PL	.62	.62	.47	.50	.34	.36	.38	.38
	Res95	IC	.59	.56	.54	.48	.40	.37	.38	.34
	JC97	IC	.53	.57	.26	.37	.31	.31	**.40**	.39
	Lin98	IC	.60	.58	.54	.49	.39	.36	.38	.38
	Les86first	G	*.07*	*.06*	*.11*	*.11*	.24	.29	.28	.32
	Les86	G	*.08*	*.14*	*.23*	*.24*	.34	.35	*.23*	*.21*
	GM07first	V	.39	.46	.43	.40	**.61**	.27	.36	*.25*
	GM07	V	**.68**	.66	**.65**	.59	.59	.48	.36	.34

[a] PL = path length, IC = information content, G = gloss, V = vector

only statistically significant differences[13] are observed for the English YP130 dataset (containing solely verb pairs) as well as the German Gur-30 and Gur-65 datasets (containing solely noun pairs connected by classical relations). In both cases, 'wisdom of linguists' resources outperform Wikipedia, but this is to be expected as classical similarity between nouns and verb similarity is better modeled by the linguists creating wordnets than by the crowds creating Wikipedia. However, it is worthwhile to note that collaboratively created knowledge sources are strongly competitive to linguistic knowledge sources on the majority of datasets.

Measure types Figures 9–12 show the maximum semantic relatedness value that is achieved by a particular measure type (path based, IC based, gloss based, and vector based as described in Section 2.1) using one of the available knowledge sources

[13] Two-tailed t-test with $\alpha = .05$

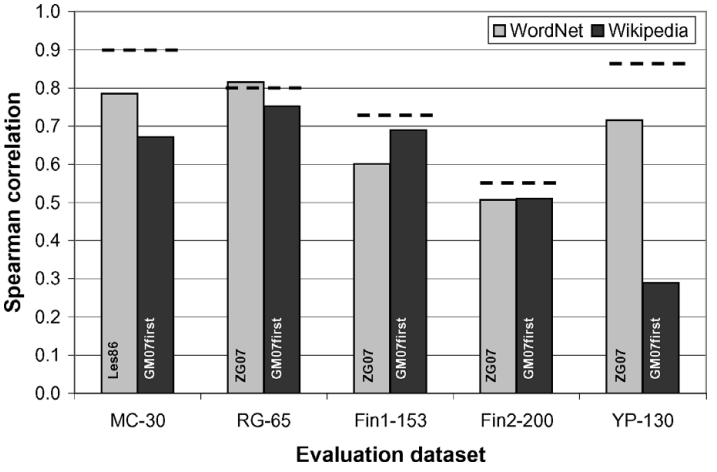

Fig. 7. 'Wisdom of linguists' versus 'wisdom of crowds' on English datasets. The dashed line indicates the approximate level of human performance (the InterAA) on the particular dataset. We show the measure with the highest Spearman correlation on a dataset.

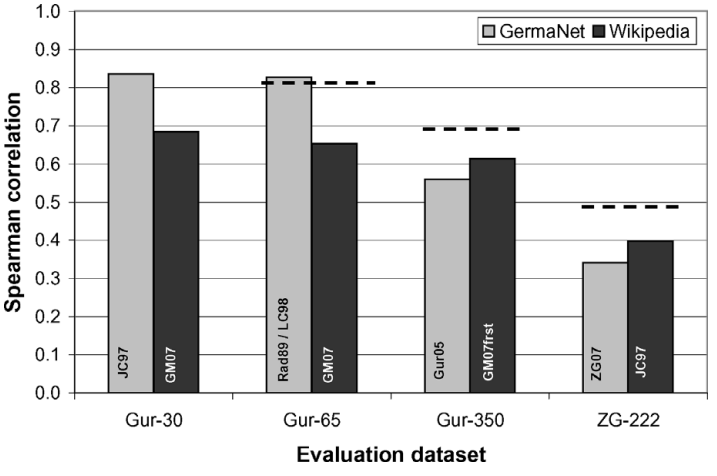

Fig. 8. 'Wisdom of linguists' versus 'wisdom of crowds' on German datasets. The dashed line indicates the approximate level of human performance (the InterAA) on the particular dataset. We show the measure with the highest Spearman correlation on a dataset.

(WordNet, English Wikipedia, GermaNet, and German Wikipedia).[14] The concept vector based measures GM07 and ZG07 consistently display superior performance compared to other measure types. Just in a few cases they show nonsignificant differences. Using Wikipedia as a knowledge source, the concept vector based measures outperform the other measure types by a wide margin on most datasets (Figures 10 and 12) Wikipedia as an encyclopedia encodes much implicit knowledge

[14] For this analysis, we aggregated all measures of a certain type, and only show the best results for each measure type.

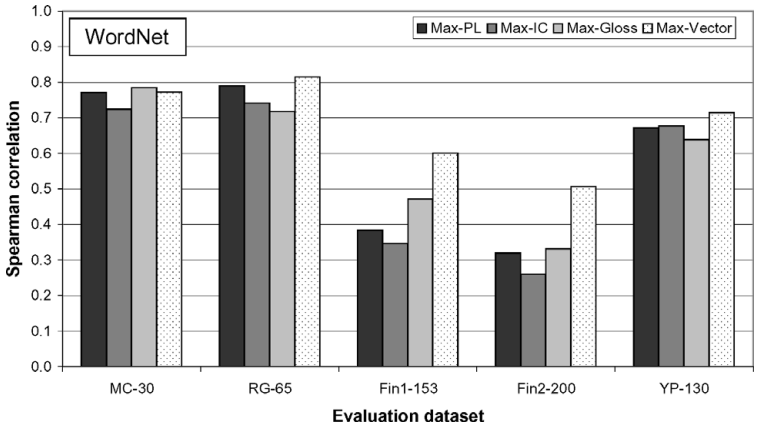

Fig. 9. Maximum SemRel value for each measure type using WordNet as knowledge source.

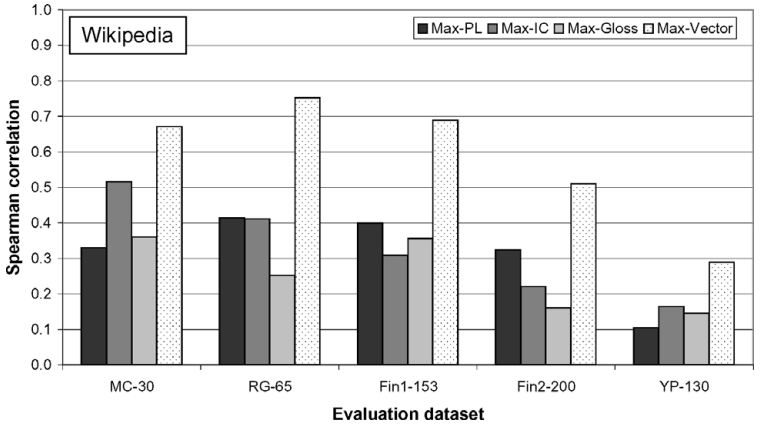

Fig. 10. Maximum SemRel value for each measure type using English Wikipedia as knowledge source.

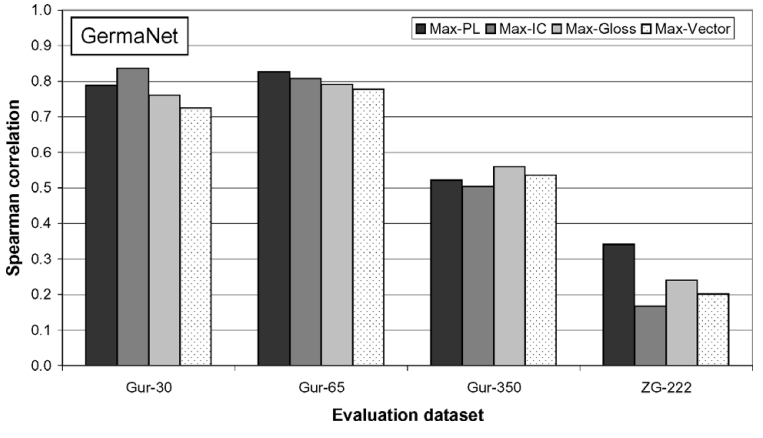

Fig. 11. Maximum SemRel value for each measure type using GermaNet as knowledge source.

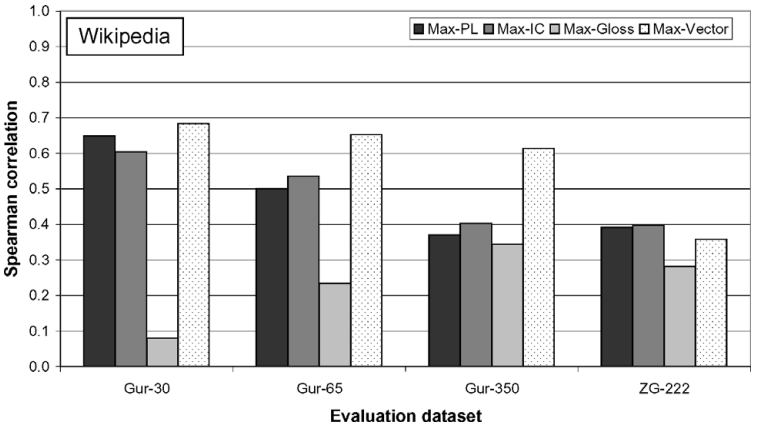

Fig. 12. Maximum SemRel value for each measure type using German Wikipedia as knowledge source.

in the article text. Thus, the concept vector based measures yield high performance improvements over the other measures types operating on Wikipedia.[15]

When using WordNet (Figure 9), the performance of vector based measures is only significantly superior to the other measure types for the Fin1–153 and Fin2–200 datasets. However, these datasets are particularly interesting as they contain also word pairs connected by nonclassical lexical semantic relations that are not fully covered by explicitly modeled WordNet relations. Here, the ZG07 measure using the implicitly encoded knowledge from the WordNet glosses is able to outperform the other measures.

In contrast to the findings on the English datasets, the performance of concept vector based measures using GermaNet (Figure 11) is only comparable to other measure types. From these results, we conclude that the performance gain, which can be obtained using concept vector based measures, depends on the amount of additional information attached to the concepts in the utilized knowledge sources. As GermaNet does not contain glosses, the concept vector based measures use pseudo glosses relying on GermaNet relations that are also used by other measures. Hence, we observe no performance gains for GermaNet.

Coverage of knowledge sources Figure 13 shows that the coverage of English word pair datasets is almost perfect for both types of resources: WordNet and Wikipedia. The coverage of the German datasets is generally lower. This is due to the fact that German resources are not as well developed as the English resources. Another reason is that German datasets contain more domain specific word pairs which are not always covered by a general purpose knowledge source. However, Wikipedia outperforms GermaNet by a wide margin on the German datasets with respect to

[15] Gloss based measures also take the full article text into account, but only those of the two articles being compared, whereas vector based measures draw knowledge from *all* articles texts in Wikipedia.

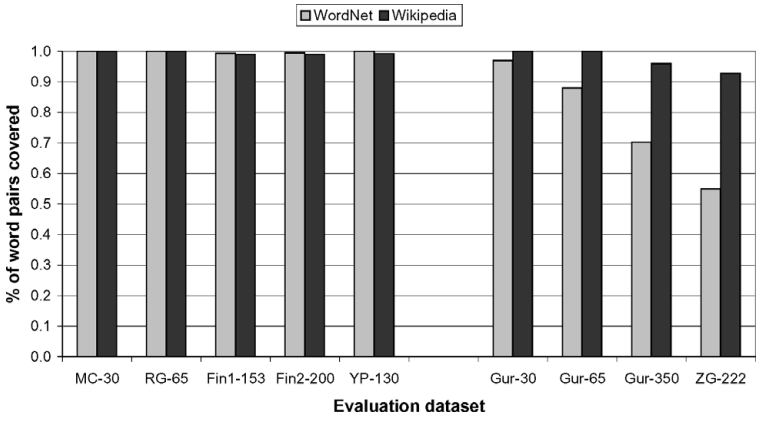

Fig. 13. Coverage of knowledge sources.

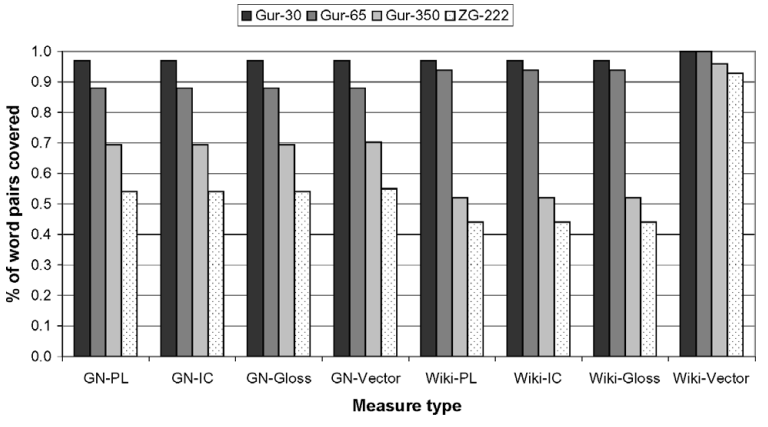

Fig. 14. Coverage of measure types on German datasets.

coverage. This demonstrates that the 'crowds' can outperform the linguists in terms of coverage due to the high number of Wikipedia contributors.

Figure 14 further analyzes the coverage of German datasets for the four types of SemRel measures on GermaNet and the German Wikipedia. When comparing the results of the vector based measures operating on both resources, we observe that the vector based measures using GermaNet (called 'GN-Vector' in Figure 14) do not have a coverage superior to other measures types. In contrast to that, the vector based measures using Wikipedia (called 'Wiki-Vector') have a higher coverage on all German datasets. The reason is that Wikipedia vector based measures make use of the article text, an information source that is not used by the other measure types. Vector based measures operating on GermaNet use pseudo glosses as the textual representation to create concept vectors. Thus, no additional information is used that other measures do not use as well.

Table 5. *Results on English word choice problems. Best values for accuracy and coverage for each knowledge source are in bold. 'WN' means WordNet. 'Wiki' means Wikipedia*

	Measure	Type[a]	Attempted	Score	# Ties	*Acc*	*Cov*
WN	Rad89	PL	196	121.9	25	.62	.65
	LC98	PL	196	123.1	25	.63	.65
	WuP94	PL	196	123.4	18	.63	.65
	HSO98	PL	205	152.1	21	.74	.68
	Res95	IC	184	104.6	42	.57	.61
	JC97	IC	132	80.5	1	.61	.44
	Lin98	IC	117	79.5	1	.68	.39
	Les86	G	279	184.2	15	.66	**.93**
	PP06	V	279	166.0	4	.59	**.93**
	ZG07	V	152	131.3	3	**.86**	.51
Wiki	Rad89	PL	226	88.33	96	.39	.75
	LC98	PL	226	88.33	96	.39	.75
	WuP94	PL	215	82.92	64	.39	.72
	Res95	IC	89	42.67	17	.48	.30
	JC97	IC	222	75.25	122	.34	.74
	Lin98	IC	89	41.83	19	.47	.30
	Les86first	G	172	52.50	23	.31	.57
	Les86	G	226	73.50	5	.33	.75
	GM07first	V	152	131.33	3	**.86**	.51
	GM07	V	288	165.83	2	.58	**.96**

[a] PL = path length, IC = information content, G = gloss, V = vector

4.3 Solving word choice problems

In this section, we present the experimental results obtained on the task of solving word choice problems. Tables 5 and 6 show the results on the English and the German dataset, respectively.

Wisdom of crowds versus wisdom of linguists Performance differs between the English and German dataset. On the English dataset, there is no difference between the best accuracy (0.86) obtained using Wikipedia or WordNet, and coverage is almost perfect for both knowledge sources (WordNet 0.93, English Wikipedia 0.96). On the German dataset, the best accuracy obtained using GermaNet (0.70) is slightly lower than that obtained using Wikipedia (0.86), but more interestingly GermaNet has only half the coverage (0.38) of the German Wikipedia (0.80). The reason for the different behavior seems to be that GermaNet is less developed than its English counterpart WordNet. However, we cannot compare the results directly as the English and German dataset might be of different difficulty for computational approaches introducing additional variability.

To analyze the dependency between human and computational performance, we extracted the 50 easiest and the 50 most difficult word choice problems from the subset of the German dataset that was used to obtain human performance scores (see

Table 6. *Results on German word choice problems. Best values for accuracy and coverage for each knowledge source are in bold. 'GN' means GermaNet. 'Wiki' means Wikipedia*

	Type[a]	Measure	Attempted	Score	# Ties	*Acc*	*Cov*
GN	PL	Rad89	386	214.4	35	.56	**.38**
	PL	LC98	386	212.9	40	.55	**.38**
	PL	WuP94	339	145.3	62	.43	.34
	IC	Res95	299	148.3	33	.50	.30
	IC	JC97	357	156.0	1	.44	.35
	IC	Lin98	298	152.5	1	.51	.30
	G	Gur05	255	177.7	13	**.70**	.25
	V	ZG07	304	193.3	3	.64	.30
Wiki	PL	Rad89	711	326.8	174	.46	.71
	PL	LC98	711	326.8	174	.46	.71
	PL	WuP94	486	268.0	60	.55	.48
	IC	Res95	424	258.0	58	.61	.42
	IC	JC97	711	322.0	45	.45	.71
	IC	Lin98	424	257.3	39	.61	.42
	G	Les86first	631	255.8	42	.41	.63
	G	Les86	711	269.3	12	.38	.71
	V	GM07first	268	230.0	0	**.86**	.27
	V	GM07	807	574.3	4	.71	**.80**

[a] PL = path length, IC = information content, G = gloss, V = vector

Section 3). Human performance on the easy problems is almost perfect ($Acc = 0.98$), while it drops to 0.33 on the difficult problems. The accuracy of computational methods is less affected as it drops only from 0.72 to 0.60 (obtained by the best measure – GM07). Similarly, the coverage drops from 0.86 for the easy problems to 0.70 for the hard problems. This is due to the fact that hard problems contain more domain specific vocabulary that is not covered by the knowledge sources.

We analyzed the results on the easy and hard word choice problems, and found that humans often fail if the candidate answers are similar with respect to spelling or pronunciation, or the candidate answers are strongly connotated with terms similar in spelling or pronunciation. The performance of computational measures is not easily influenced by such distractors. SemRel measures are more likely to fail, if the candidate answers are all semantically related to the target. Humans also fail because they do not know the meaning of rare words used as candidate answers well enough. Computational measures are subject to a similar effect, as the evidence provided by a knowledge source may not be sufficient for rare words.

Measure types Vector based measures display the best accuracy and coverage on the English as well as the German dataset. Only on the German dataset using GermaNet as a knowledge source, the ZG07 measure is slightly outperformed by other measures. This observation is very similar to our findings for the word pair

ranking task where we also did not observe performance gains using the ZG07 measure with GermaNet due to the missing glosses in GermaNet.

Overall, the GM07 measure yields the best coverage, while its modification GM07first that only uses the first paragraph of a Wikipedia article, yields the highest accuracy for English (0.86) and German (0.86), while the coverage is quite low in both cases (0.51 for English and 0.27 for German). This is consistent with our intuition described in Section 3.2.2. By utilizing this effect, it is possible to configure the GM07 measure according to whether accuracy or coverage is more important in a NLP application.

5 Tools

In this section, we present two software packages that have been developed to support the experiments presented in this article. The software is freely available for research purposes, and can be used to perform related research.

5.1 DEXTRACT

The ZG222 dataset described in Section 3 was created using the DEXTRACT tool, implementing a semi-automatic corpus-based approach for creating evaluation datasets for SemRel measures (Zesch and Gurevych 2006). DEXTRACT takes a corpus as input, and outputs an evaluation dataset containing a list of automatically generated word pairs. Because these word pairs are selected automatically, they cannot be biased toward strong classical relations beyond corpus evidence as it is the case with word pairs selected by humans. However, randomly generating word pairs from the corpus would result in too many unrelated pairs. Thus, words are assigned to word pairs according to their tf.idf weights. Then, a set of user defined filters is applied, normally including a stopword filter removing stopwords, and a part of speech based filter that forces the final evaluation dataset to contain a specified number of word pairs with certain part of speech combinations. A more detailed description of DEXTRACT is given by Zesch and Gurevych (2006). DEXTRACT is publicly available for research purposes from `http://www.ukp.tu-darmstadt.de/software/dextract`.

5.2 JWPL

The Wikipedia based SemRel measures used in this article are implemented based on JWPL, a high-performance Java-based Wikipedia API. JWPL operates on an optimized database that is created from the database dumps[16] available from the Wikimedia foundation. This allows for fast access to Wikipedia articles, categories, links, redirects, *etc*. Thus, the SemRel measures described in Section 2.1 can be efficiently implemented on top of JWPL. A more detailed description of JWPL can be found in (Zesch *et al.* 2008). JWPL is freely available for research purposes from

[16] `http://download.wikipedia.org/`

`http://www.ukp.tu-darmstadt.de/software/JWPL`. It has been successfully used for building various large scale NLP applications using Wikipedia as a knowledge source, e.g., in information retrieval (Gurevych *et al.* 2007).

6 Conclusions and future directions

In this article, we presented a comprehensive study aimed at computing semantic relatedness of word pairs. We categorized a large number of semantic relatedness measures into four distinct types: path based, information content based, gloss based, and vector based measures. We analyzed the performance of these measures on two evaluation tasks (correlation with human judgments, and solving Reader's Digest word choice problems) with respect to different experimental conditions such as: (i) the underlying knowledge source ('wisdom of linguists' versus 'wisdom of crowds'), (ii) the measure type, and (iii) whether results obtained for English can be confirmed for German.

Correlation with human judgments Contrary to previous research (Strube and Ponzetto 2006; Gabrilovich and Markovitch 2007; Zesch *et al.* 2007), we find that (i) 'wisdom of crowds' based resources are not generally superior to 'wisdom of linguists' based resources. We further find that (ii) concept vector based measures consistently display superior performance compared to other measure types, and (iii) that the results on German datasets confirm the results for English.

The restored competitiveness of 'wisdom of linguists' based resources is due to a generalized concept vector based measure (ZG07) introduced in this article. This measure is applicable to any knowledge source offering a textual representation of a concept. We showed how such textual representations can be inferred from semantic relations in wordnets without glosses. The performance gains that can be obtained with the generalized concept vector based measure strongly depend on the amount of additional information that the knowledge source offers in the textual representations.

Solving word choice problems As this task depends much on the coverage of a knowledge source, results are different for English and German. On the English dataset, we find (i) little differences between 'wisdom of linguists' or 'wisdom of crowds' knowledge sources. On the German dataset the 'crowds' outperform the 'linguists' by a wide margin due to the much higher coverage of the SemRel measures using the German Wikipedia. We find that (ii) concept vector based measures using Wikipedia as a knowledge source perform consistently well, and outperform all other measure types with respect to accuracy and coverage on the English as well as the German dataset. However, a more detailed analysis of the word choice datasets with respect to the expected difficulty for a SemRel measure is necessary before we can draw final conclusions.

Even if this present article could not confirm a superior performance of SemRel measures operating on knowledge sources created by the 'wisdom of crowds', it is worthwhile to note that the 'wisdom of crowds' is at least competitive to the 'wisdom

of linguists' on the majority of datasets. As collaboratively created knowledge sources such as Wikipedia are freely available for a wide range of languages, they can be used as a substitute for well-developed linguistic knowledge sources that are not available for many languages. In very recent research (Zesch *et al.* 2008), the wiki dictionary Wiktionary[17] was introduced as another promising knowledge source.

Finally, we presented two systems that were developed to aid the experiments presented herein and are freely available[18] for research purposes: (i) DEXTRACT, a software to semi-automatically construct corpus-based datasets, and (ii) JWPL, a Java-based high-performance Wikipedia API for building NLP applications.

Acknowledgements

This work was supported by the German Research Foundation under grant 'Semantic Information Retrieval from Texts in the Example Domain "Electronic Career Guidance" ', GU 798/1-2 and GU 798/1-3. Giuseppe Pirro and Nuno Seco kindly provided the data based on their repetition of the Rubenstein & Goodenough experiment. Dongqiang Yang kindly provided the human judgments for the verb dataset. Alistair Kennedy and Stan Szpakowicz kindly provided the English word choice problem dataset. We thank Christof Müller for his valuable contributions to this work and fruitful discussions. We are also grateful to the anonymous reviewers for their helpful and constructive comments.

References

Anscombe, F. J. 1973. Graphs in statistical analysis. *American Statistician* **27**: 17–21.

Banerjee, S., and Pedersen, T. 2002. An adapted lesk algorithm for word sense disambiguation using WordNet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 136–145, London: Springer Verlag.

Bernard, J. 1986. *The Macquarie Thesaurus*. Sidney, Australia: Macquarie Library.

Boyd-Graber, J., Fellbaum, C., Osherson, D., and Shapire, R. 2006. Adding dense, weighted, connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*, Jeju Island, Korea.

Budanitsky, A., and Hirst, G. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics* **32**(1): 13–47.

Fellbaum, C. 1998. *WordNet an Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., and Wolfman, G. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems* **20**(1): 116–31.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606–11, Hyderabad, India.

Galley, M., and McKeown, K. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pp. 1486–8, Acapulco, Mexico.

[17] `http://www.wiktionary.org`
[18] `http://www.ukp.tu-darmstadt.de/software`

Gurevych, I. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pp. 767–78. Jeju Island, Korea.

Gurevych, I., Müller, C., and Zesch, T. 2007. What to be? – electronic career guidance based on semantic relatedness. In *Proceedings of ACL*, pp. 1032–9, Prague, Czech Republic. Association for Computational Linguistics.

Gurevych, I., and Strube, M. 2004. Semantic similarity applied to spoken dialogue summarization In *The 22nd International Conference on Computational Linguistics (COLING)*, pp. 764–70, Geneva, Switzerland.

Halliday, M. A. K., and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hirst, G., and St-Onge, D. 1998. Lexical chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database and Some of Its Applications*, pp. 305–332. Cambridge, MA: The MIT Press.

Jarmasz, M., and Szpakowicz, S. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 111–20.

Jiang, J. J., and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, Taipei, Taiwan.

Kozima, H., and Furugori, T. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the sixth conference of the European chapter of the Association for Computational Linguistics*, pp. 232–9, Morristown, NJ.

Kunze, C. 2004. Computerlinguistik und Sprachtechnologie. In K. U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, and H. Langer (eds.), *Lexikalisch-semantische Wortnetze*, pp. 423–31. Berlin: Spektrum Akademischer Verlag.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*, pp. 265–83. Cambridge, MA: MIT Press.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pp. 24–6. Toronto, Canada.

Li, Y., Bandar, Z. A., and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* **15**: 871–82.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, pp. 296–304. Madison, WI.

McHale, M. 1998. A comparison of wordnet and roget's taxonomy for measuring semantic similarity. *CoRR*, cmp-lg/9809003.

Mihalcea, R., and Moldovan, D. I. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 152–8, Maryland, MD: Association for Computational Linguistics.

Miller, G. A., and Charles, W. G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* **6**(1): 1–28.

Mohammad, S., Gurevych, I., Hirst, G., and Zesch, T. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 571–80, Prague, Czech Republic: Association for Computational Linguistics.

Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* **17**(1): 21–48.

Morris, J., and Hirst, G. 2004. Non-classical lexical semantic relations. In *Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the ACL*, pp. 46–51. Boston, MA.

Patwardhan, S., Banerjee, S., and Pedersen, T. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 241–57, Mexico City, Mexico.

Patwardhan, S., and Pedersen, T. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pp. 1–8, Trento, Italy: Association for Computational Linguistics.

Pirro, G., and Seco, N. 2008). Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *OTM '08: Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE*, pp. 1271–88, Monterrey, Mexico.

Procter, P. 1978. *Longman Dictionary of Contemporary English.* Longman, London.

Qiu, Y., and Frei, H. P. 1993. Concept based query expansion. In *Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval*, ACM

Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* **19**(1): 17–30.

Resnik, P. 1995 Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–53, Montreal, Canada.

Roget, P. 1962. *Roget's International Thesaurus*, 3rd ed. L. V. Berrey, and G. Carruth (eds.), New York: Thomas Y. Crowell Co.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* **8**(10): 627–33.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Seco, N., and Hayes, T. V. J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence*, Valencia, Spain.

Silber, H. G., and McCoy, K. F. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Comput. Linguist.* **28**(4): 487–96.

Stevenson, M., and Greenwood, M. A. 2005. A semantic approach to ie pattern induction. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 379–86. Morristown, NJ: Association for Computational Linguistics.

Strube, M., and Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on ArtificialIntelligence (AAAI-06)*, pp. 1419–24, Boston, MA.

Turney, P. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 313–20, Sydney, Australia: Association for Computational Linguistics.

Voss, J. 2006. Collaborative thesaurus tagging the Wikipedia way. CoRR, abs/cs/0604036.

Wallace, D., and Wallace, L. A. 2001–2005. *Reader's Digest, das Beste für Deutschland*. January 2001–December 2005. Stuttgart: Verlag Das Beste.

Weeds, J. E. 2003. *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, East Sussex, UK: University of Sussex.

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the ACL*, pp. 133–8, Las Cruces, Mexico: Association for Computational Linguistics.

Yang, D., and Powers, D. M. W. 2006. Verb similarity on the taxonomy of WordNet. In *Proceedings of the Third International WordNet Conference (GWC-06)*, pp. 121–8, Jeju Island, Korea.

Zesch, T., and Gurevych, I. 2006. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the ACL-Workshop on Linguistic Distances*, pp. 16–24, Sydney, Australia: Association for Computational Linguistics.

Zesch, T., and Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pp. 1–8, Rochester, NY. Association for Computational Linguistics.

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007a. Analyzing and accessing Wikipedia as a lexical semantic resource. In G. Rehm, A. Witt, and L. Lemnitzer (eds.), *Data Structures for Linguistic Resources and Applications*, pp. 197–205. Tuebingen, Germany: Gunter Narr.

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007b. Comparing Wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 205–8. Rochester, NY: Association for Computational Linguistics.

Zesch, T., Müller, C., and Gurevych, I. 2008a. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Zesch, T., Müller, C., and Gurevych, I. 2008b Using Wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, pp. 861–7. Chicago, IL.