The More the Better? Assessing the Influence of Wikipedia's Growth on Semantic Relatedness Measures

Torsten Zesch and **Iryna Gurevych**

Ubiquitous Knowledge Processing Lab
Computer Science Department, Technische Universität Darmstadt
Hochschulstraße 10, D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

Wikipedia has been used as a knowledge source in many areas of natural language processing. As most studies only use a certain Wikipedia snapshot, the influence of Wikipedia's massive growth on the results is largely unknown. For the first time, we perform an in-depth analysis of this influence using semantic relatedness as an example application that tests a wide range of Wikipedia's properties. We find that the growth of Wikipedia has almost no effect on the correlation of semantic relatedness measures with human judgments, while the coverage steadily increases.

1. Introduction

Wikipedia is a multilingual, web-based, freely available encyclopedia, constructed in a collaborative effort of voluntary contributors. Wikipedia has arguably become the largest collection of freely available knowledge with currently approx. 9.25 million articles in more than 250 languages. This knowledge has been used in many areas of natural language processing (see (Medelyan et al., 2009) for an overview) to overcome the knowledge acquisition and coverage problems pertinent to conventional knowledge sources like WordNet (Fellbaum, 1998). However, most studies only performed their evaluation using a single Wikipedia snapshot available at the time of the experiments. Thus, it is largely unknown whether the results would have been different for other snapshots. The difference between snapshots might be significant, as all Wikipedia language editions grow very quickly. For example in 2008, the German Wikipedia has grown by over 150,000 articles (i.e. over 400 articles per day).[1] Figure 1 visualizes this development. In this paper, we investigate the influence of Wikipedia's growth on the performance of natural language processing tasks that use Wikipedia as a knowledge source.

In contrast to an unstructured corpus that just grows in size, Wikipedia is a structured resource that grows in different ways: (i) new articles, categories, or redirects are added, (ii) existing articles, categories, or redirects are corrected or augmented, or (iii) the links between articles or categories are changed. In our analysis, we need to ensure that all these aspects of Wikipedia's growth are tested. Thus, we selected the pervasive task of computing semantic relatedness for our studies, as the different approaches to computing semantic relatedness make use of different properties of Wikipedia like the article text, the article links, the category system, or the article titles and redirects. Additionally, computing semantic relatedness directly uses Wikipedia as a knowledge source, while more complex tasks would entail other influences. Furthermore, Wikipedia is increasingly used as a knowledge source for computing semantic relatedness (Gabrilovich and Markovitch, 2007; Nakayama et al., 2007; Ponzetto and Strube, 2007; Milne and Witten, 2008; Zesch et al., 2008b). Therefore, the impact of Wikipedia's growth on the performance of semantic relatedness is an important field of study on its own.

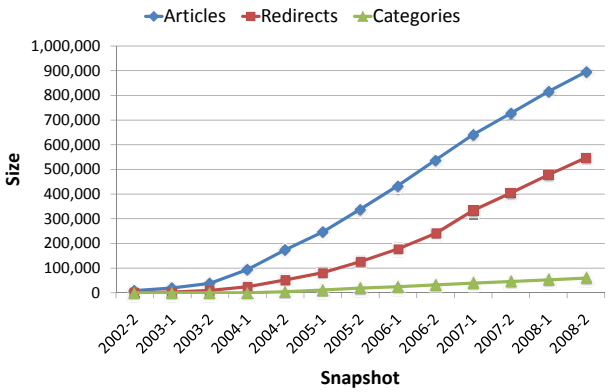


Figure 1: Growth of the German Wikipedia.

We use two approaches for the evaluation of semantic relatedness measures: (i) correlation with human judgments and (ii) solving word choice problems. The correlation with human judgments depends more directly on the performance of the semantic relatedness measures, while solving word choice problems is better suited to assess the coverage of a resource, as the available word choice datasets are significantly larger and contain more complex vocabulary than the word pair datasets used for measuring the correlation with human judgments.

The paper is structured as follows: We first describe related work in Section 2. We give a short overview of semantic relatedness measures in Section 3. We then explain our experimental setup in Section 4. We present the results in Section 5. Finally, we summarize our findings in Section 6.

2. Related Work

To our knowledge, there is no other study performing an in-depth analysis of the influence of Wikipedia's growth on the performance of tasks using it as a knowledge source.

[1] http://stats.wikimedia.org

Ponzetto and Strube (2007) replicated their semantic relatedness experiments, which were originally performed on an English snapshot from February 2006, on two more recent snapshots from September 2006 and May 2007. They found that the choice of the snapshot had no significant influence on the performance, but they did not report the influence on coverage. Furthermore, the number of snapshots used does not allow for general conclusions.

Buriol et al. (2006) create a graph representation of Wikipedia, and analyze the temporal evolution of the topological properties of this graph representation. They did not assess the consequences of the topological changes on NLP applications.

The growth of a resource is also an issue for corpus-based NLP approaches, where the size of the available corpus increases due to the growing Web and improved data processing capabilities. For corpus-based approaches, usually "more data are better data" (Church and Mercer, 1993), e.g. the quality of statistical machine translation continues to improve with increasing corpus size (Brants et al., 2007). However, these findings cannot be generalized to semantically structured corpora like Wikipedia.

3. Semantic Relatedness Measures

A multitude of semantic relatedness measures relying on structured knowledge sources have been proposed that can be categorized into: (i) path based, (ii) gloss based, (iii) concept vector based, and (iv) link vector based measures (Zesch and Gurevych, 2010).

Path based measures (Budanitsky and Hirst, 2006) rely on paths in a graph of concepts built from a knowledge source. Zesch et al. (2007) describe how state-of-the-art path based measures can be adapted to Wikipedia.

Gloss based measures (Lesk, 1986; Mihalcea and Moldovan, 1999; Banerjee and Pedersen, 2002) rely on word overlaps between definitions of concepts. Gloss based measures can be directly applied to Wikipedia, as each Wikipedia article represents a definition of a concept.

Concept Vector based measures (Patwardhan and Pedersen, 2006; Gabrilovich and Markovitch, 2007) use the textual description of a concept to construct a vector representation. The semantic relatedness between two words can then be computed as the cosine of their corresponding concept vectors.

Link Vector based measures (Nakayama et al., 2007; Milne and Witten, 2008) use the links between concepts to construct a vector representation. The semantic relatedness between two words can then be computed as the cosine of their corresponding link vectors.

Each measure type uses different properties of Wikipedia, and thus tests different kinds of growth in Wikipedia. For example, the path based measures are only affected by changes to the assignment of articles to categories and when new articles or categories are added. They are insensitive to changes made to the textual content, while these changes have a major influence on gloss based and the concept vector based measures. However, as the concept vector based measures draw knowledge from different articles in parallel, we expect them to be more robust to changes in Wikipedia than the gloss based measures. The link vector based measures are only influenced from changes to the article links or when new articles are added.

For our analysis, we select the most prototypical measure from each measure type. Note, that this is not necessarily the measure that yields the best results, as we are only interested in the changes of performance over time, not in the absolute scores.

From the path based measures, we select the simple path length measure by Rada et al. (1989). It uses the path length l between two nodes representing concepts in the knowledge source to compute semantic relatedness. We select this measure, as it is the most versatile path based measure imposing the least constraints on a resource. The measure (abbreviated as **Path**) is computed as follows:

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

where c_1 and c_2 are concepts, 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 l_{max} of the graph:

$$rel_{\text{Path}}(c_1, c_2) = l_{max} - l(c_1, c_2)$$

From the gloss based measures, we select the simple gloss overlap measure by Lesk (1986) abbreviated as **Gloss**. It is based on the amount of word overlap in the glosses of two concepts, where higher overlap means that two terms are more related.

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

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

From the concept vector based measures, we select the ESA measure (Gabrilovich and Markovitch, 2007) abbreviated as **ConceptVector**. The measure is based on concept vectors derived from Wikipedia articles $a_1, \ldots, a_N$, where N is the number of Wikipedia articles. 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 the Wikipedia concept space. Semantic relatedness of two words can then be computed as the cosine of their corresponding concept vectors:

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

Another vector based measure (abbreviated as **LinkVector**) was introduced by Milne (2007) and Nakayama et al. (2007). It makes use of the links between Wikipedia articles, but it does not measure path lengths like the path based measures. The *LinkVector* measure is based on the set of links that point to other articles (called 'targets'). The more targets two articles share, the higher their semantic relatedness. Links to targets are considered less significant if many other articles also link to the same target. For example, a link to a very common target like *automobile* is less important than a link to a more specific target like *Ethanol fuel*. Formally, the weight ω of a link is defined as:

$$\omega(s \rightarrow t) = \begin{cases} \log\left(\frac{N}{|T|}\right), s \in T \\ 0, \; otherwise \end{cases}$$

where T is the set all articles that link to t, and N is the number of Wikipedia articles. An article is then represented as a vector $\vec{l}$ of weighted outgoing links l. The semantic relatedness of two terms is computed as the cosine of the link weight vectors of the corresponding articles:

$$rel_{\text{LinkVector}}(a_1, a_2) = \frac{\vec{l}(a_1) \cdot \vec{l}(a_2)}{\left|\vec{l}(a_1)\right| \left|\vec{l}(a_2)\right|}$$

where a_i are Wikipedia articles corresponding to terms. An article corresponds to a term if its title or one of its redirects matches the term, or if the article is linked on a disambiguation page which title matches the term.

4. Experimental Setup

For our analysis, Wikipedia snapshots from different points of time are required. The Wikimedia Foundation only provides recent snapshots[2], and we are not aware of any other repository. However, the Wikimedia Foundation provides special snapshots that contain all revisions. From such snapshots, any past state of the Wikipedia can be reconstructed. For that purpose, we designed a data conversion tool that is able to create a snapshot of any Wikipedia language edition at any point in time since it was created. We access these snapshots using the JWPL Wikipedia API (Zesch et al., 2008a).

For a meaningful analysis, a large Wikipedia language version is necessary that provides sufficient coverage on the evaluation datasets. The English Wikipedia is the largest language edition, but unfortunately the recent snapshot containing all revisions is currently unavailable from the Wikimedia Foundation due to technical problems. The most recent available English snapshot containing all revisions was released in 2006, which would only provide a limited number of snapshots. Thus, we perform our experiments using the German Wikipedia which is the second largest edition that was created shortly after the English edition. Additionally, German evaluation datasets for semantic relatedness and word choice problems are available (Zesch et al., 2008b), which is not the case for most other languages.

[2] http://download.wikimedia.org/

		Number of		
Date	**Name**	**Articles**	**Redirects**	**Categories**
01.12.02	2002-2	8,596	658	0
01.06.03	2003-1	19,236	2,574	0
30.11.03	2003-2	37,999	9,397	0
30.05.04	2004-1	93,930	24,379	0
28.11.04	2004-2	173,837	51,765	4,180
29.05.05	2005-1	246,113	81,198	11,176
27.11.05	2005-2	338,887	126,050	19,114
28.05.06	2006-1	434,211	177,413	24,591
26.11.06	2006-2	537,868	240,271	31,936
27.05.07	2007-1	641,178	333,657	39,158
25.11.07	2007-2	727,186	404,431	45,889
25.05.08	2008-1	815,609	477,790	52,385
23.11.08	2008-2	895,136	547,244	59,453

Table 1: Growth of the German Wikipedia.

For our experiments, we created a snapshot of the German Wikipedia every 183 days (6 months) starting December 1st, 2002 (see Table 1). Each of the snapshots is used as a resource for computing semantic relatedness. We evaluate the performance of semantic relatedness measures using two evaluation approaches: (i) *correlation with human judgments* and (ii) *solving word choice problems*.

Correlation with Human Judgments Evaluation datasets for correlation with human judgments are created by asking human annotators to judge the semantic relatedness of presented word pairs. The gold standard score assigned to a word pair is the average score over all human judges. For evaluation, the gold standard dataset is then correlated with the scores computed by a particular semantic relatedness measure. We use the Spearman rank correlation coefficient ρ, where a value of 0 means no correlation and a value of 1 means perfect correlation.

We use two publicly available German datasets.[3] The Gur–65 dataset contains 65 word pairs from the English study by Rubenstein and Goodenough (1965) translated to German. This dataset only contains nouns, and human judgments rated the *similarity* between the words. The Gur–350 dataset contains 350 word pairs collected in a study by Gurevych (2005). This dataset contains nouns, verbs and adjectives connected by classical and non-classical relations (Morris and Hirst, 2004) that were rated by humans according to the *relatedness* between the words. It also contains a lot of domain-specific word pairs. Thus, this dataset will be more informative with respect to the coverage provided by a certain Wikipedia snapshot. We define **coverage** as the percentage of word pairs in the evaluation dataset for which a semantic relatedness measure using a certain Wikipedia snapshot is able to compute a score, i.e. both words could be found in Wikipedia (either as an article title or mentioned in the article text, depending on the kind of information used by a specific semantic relatedness measure).

Solving word choice problems A word choice problem (Jarmasz and Szpakowicz, 2003; Turney, 2006) consists of a target word and four candidate words or phrases. The

[3] http://www.ukp.tu-darmstadt.de/data/semantic-relatedness/

objective is to pick the one that is most closely related to the target.

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

There is always only one correct candidate, 'a' in this case. The semantic relatedness between the target 'beret' and each of the candidates is computed by a semantic relatedness measure, and the candidate with the maximum relatedness value is chosen. For preprocessing and handling of multiword expressions, we follow the approach outlined in (Zesch et al., 2008b).

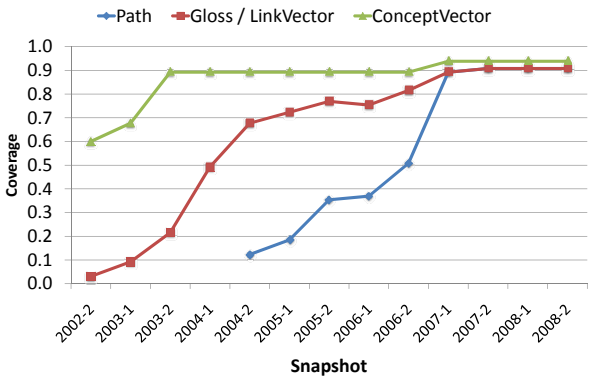
If two or more candidates are equally related to the target, the candidates are said to be tied. If one of the tied candidates is the correct answer, the problem is counted as correctly solved, but the corresponding score is reduced. We assign a score s_i of $\frac{1}{\text{\# of tied candidates}}$ (in 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.

We evaluate the word choice problems using accuracy and coverage. We define accuracy as $Acc = \frac{S}{|A|}$, where S is the sum of all scores s_i, and A is the number of word choice problems that were attempted by the semantic relatedness 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. The overall performance of a measure needs to take accuracy *and* coverage into account, as a measure might get a better coverage by sacrificing accuracy and vice versa. Thus, we define the combined evaluation metric $H = \frac{2 \cdot Acc \cdot Cov}{Acc + Cov}$, i.e. the harmonic mean of accuracy and coverage.

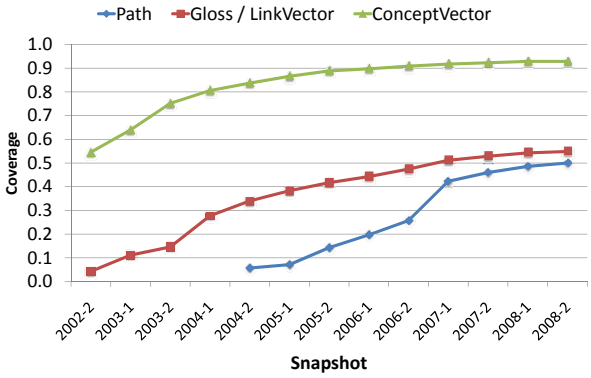
We use a dataset collected by Zesch et al. (2008b). It contains 1008 word choice problems from the January 2001 to December 2005 issues of the German-language edition of Reader's Digest (Wallace and Wallace, 2001-2005). As this dataset contains more complex vocabulary and is significantly larger than the available word pair datasets for correlation with human judgment, it is better suited to assess the coverage of a resource.

5. Results and Discussion

Correlation with Human Judgments Figure 2 shows the obtained coverage on the two datasets using the Wikipedia snapshots. As the *Gloss* and the *LinkVector* measure display equal coverage, we combined them in this figure. We find that the *ConceptVector* measure generally covers more word pairs than the other measures. The *ConceptVector* measure also displays high initial coverage even when using the quite small first snapshot from 2002. This is due to the special property of the *ConceptVector* measure that a word is covered, if it is contained in a Wikipedia article. This is in contrast to the other measure types, where the word has to appear as an article title or redirect. As Wikipedia article titles are mainly nouns or noun phrases, the coverage of verbs and adjectives contained in the Gur–350 dataset is limited for the *Path*, the *Gloss*, and the *LinkVector* measure.



(a) Gur–65

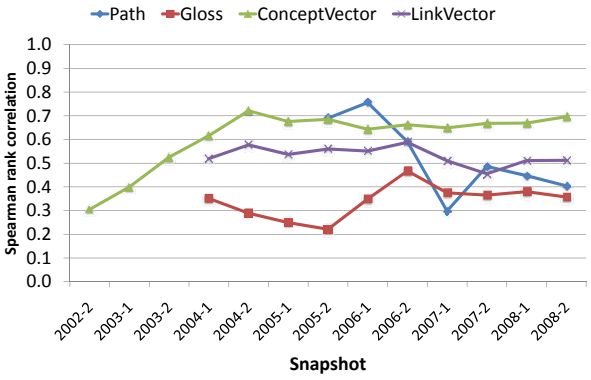


(b) Gur–350

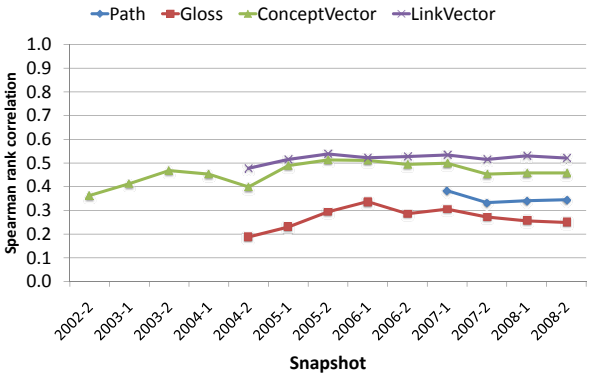
Figure 2: Coverage of measure types on the word relatedness datasets.

The *Path* measure does not cover any word pairs before the snapshot 2004-2, as it relies on the category system that was not added to Wikipedia until 2004. On the small Gur–65 dataset, the *Path* and the *Gloss* measure reach the same coverage as the *ConceptVector* measure when looking at the more recent snapshots. However, on the larger Gur–350 dataset that contains more domain-specific vocabulary, the *ConceptVector* measure still has a much higher coverage than the other measures; see Figure 2 (b). For the early snapshots, coverage rises steeply for all measures, while for the recent snapshots only small increases in coverage can be observed.

Figure 3 shows the obtained Spearman correlations between the human judgments in the gold standard dataset and the scores computed by the semantic relatedness measures. As correlation values, which are based on a small number of word pairs, are not reliable, we only present them if the coverage reaches at least 30% of the full dataset and at least 20 word pairs are covered. Thus, the lines in the chart which correspond to measure types with a low coverage do not extend over all the snapshots. Note that we generally cannot compare the correlation scores between single measures, as they were obtained on different subsets of the evaluation dataset due to the different coverage of the measures. Thus, the important information in this chart is the behavior of a single measure over time. On the



(a) Gur–65



(b) Gur–350

Figure 3: Correlation of measure types with human judgments.

Gur–65 dataset, as shown in Figure 3 (a), the correlation scores obtained by the *Path* measure differ much between the snapshots. However, these correlation scores are based on a very small number of word pairs, as shown in Figure 2 (a), and are quite unreliable. If we only look at the last four snapshots, where all measure types yield the same high coverage, results get more stable.

For the larger Gur–350 dataset, as shown in Figure 3 (b), all measures display almost stable correlation scores without statistically significant differences for the more recent snapshots. This means that Wikipedia's growth does not have significant effects on the task performance.

In the analysis presented above, the Spearman correlation scores are computed using as many word pairs as are covered by a certain snapshot. Thus, the analysis cannot tell us whether the growth of Wikipedia has an influence on the core set of word pairs covered by all snapshots. Thus, we perform an additional analysis where we only use a fixed number of word pairs covered by all snapshots. As we need a sufficient number of initially covered word pairs, we limit our analysis to the Gur–350 dataset and the *ConceptVector* measure. With this setting, even the initial snapshot from 2002 already covers over 50% of all word pairs in the Gur–350 dataset – cf. Figure 2 (b). Figure 4 visualizes the obtained results: In the beginning, the performance rises from snapshot to snapshot and then stays almost stable without statistically significant differences. This means that even extensive changes like re-structuring, extending and adding

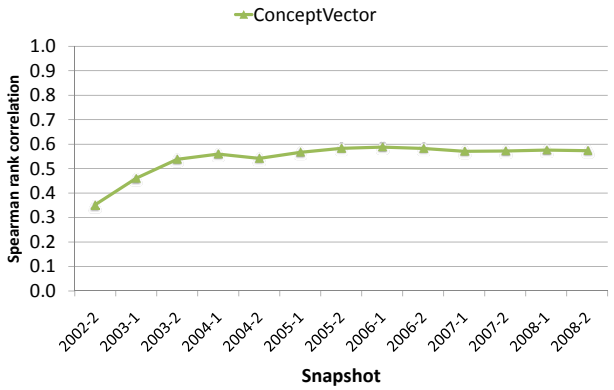


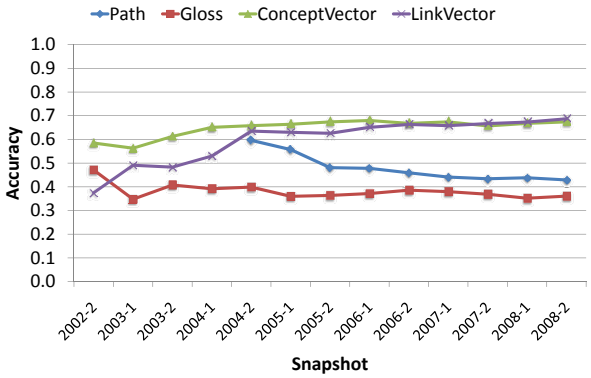
Figure 4: Performance of the *ConceptVector* measure using a fixed set of word pairs from the Gur-350 dataset.

articles do not have a significant influence on the performance of the *ConceptVector* measure on the initially covered word pairs. The *ConceptVector* measure is remarkably stable, as it draws knowledge from different articles in parallel and is thus not easily influenced by changes restricted to a subset of articles.

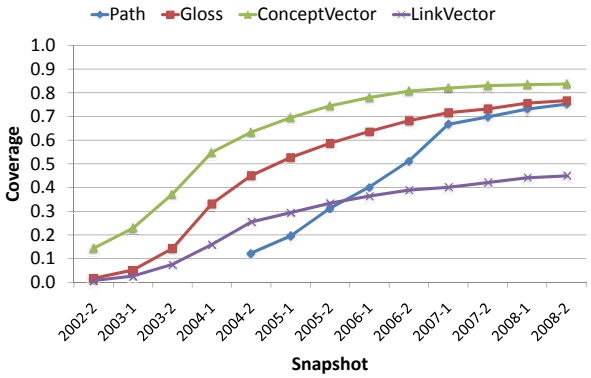
Solving Word Choice Problems Figures 5 (a), (b), and (c) compare the four measure types according to accuracy, coverage, and harmonic mean of accuracy and coverage. We find that the accuracy values of the *ConceptVector*, *LinkVector*, and *Gloss* measures are almost stable for later snapshots, while the *Path* measure shows a falling trend. However, the higher values for the *Path* measure are unreliable, as they are obtained on snapshots with a very low coverage. Thus, we can conclude that the growth of Wikipedia has almost no effect on its suitability for solving word choice problems. However, it has a positive effect on coverage as shown in Figure 5 (b), but coverage increases between the more recent snapshots are small showing a log-like trend. Note, that in this task the coverage of the *Gloss* measure is not equal to the coverage of the *LinkVector* measure (as it was the case for correlation with human judgments). The difference is due to the *LinkVector* measure, which returns 0 if two articles have no links in common. Zero scores often cause a measure not to attempt to solve a word choice problem, as this provides insufficient information for giving an answer. The *Gloss* measure only returns 0, if two articles do not share a single word, which happens less often.

The *Path* measure relying on the Wikipedia category graph only yields coverage comparable to the *Gloss* or *ConceptVector* measure when using very recent snapshots. The *LinkVector* measure generally shows a quite low coverage. As accuracy is almost constant and coverage rises, the combined performance values (H) in Figure 5 (c) are bound to coverage. The results on this task are consistent with the results obtained on the other evaluation task *correlation with human judgments*: Wikipedia's growth increases the coverage, while the accuracy is stable.

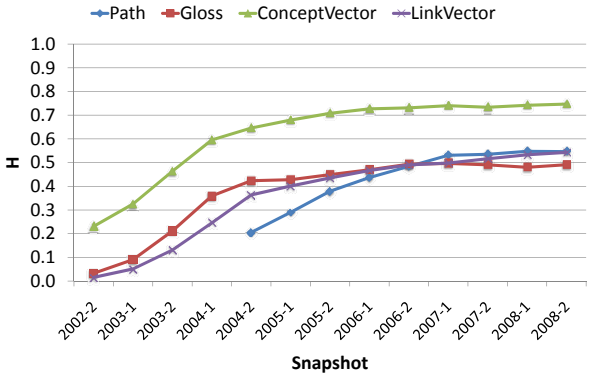
Overall, we can conclude that – as expected – the growth of Wikipedia has a positive effect on coverage.



(a) Accuracy



(b) Coverage



(c) Harmonic Mean – H

Figure 5: Performance of measure types when solving word choice problems.

Surprisingly, it has almost no effect on the suitability of Wikipedia as a resource for computing semantic relatedness. Especially for the *ConceptVector* measure, correlation values and coverage are quite high even for smaller snapshots. Thus, even small language-specific versions of Wikipedia can be used for computing semantic relatedness if there are no developed classical resources for a certain language. If the coverage provided by an older snapshot is already sufficient for a certain task, smaller (and thus computationally less demanding) Wikipedia snapshots can be used without negative effects on the task performance.

6. Summary

We analyzed the influence of the Wikipedia's growth on the performance of NLP applications using Wikipedia as a knowledge source. As Wikipedia is a structured resource that grows in different ways, we selected the task of computing semantic relatedness for evaluation. The different types of semantic relatedness measures (path based, gloss based, concept vector based, and link vector based) test a wide range of Wikipedia's properties.

We evaluated the performance of semantic relatedness using two tasks: correlation with human judgments and solving word choice problems. We created 6-monthly snapshots of the German Wikipedia that are used as knowledge sources for the relatedness measures. Our analysis performed on the German Wikipedia shows that the growth has little effect on the performance of semantic relatedness measures. It rises for the early snapshots providing very low coverage, and then stays stable even for the quite large more recent snapshots. This property, together with the increasing coverage, makes Wikipedia a valuable resource in the context of large-scale NLP applications, where coverage is one of the major criteria for overall performance.

Even if we selected semantic relatedness for evaluation, which directly assesses a wide range of different Wikipedia properties, other natural language processing applications might still display different behavior. Thus, we make the Wikipedia snapshots used in this study available upon request. We hope that this will foster research on the influence of Wikipedia's growth on other NLP tasks. Additionally, we will make the *TimeMachine* for creating the Wikipedia snapshots publicly available as part of the JWPL tool.[4] In future work, we plan to verify our results using the English Wikipedia (as soon as the required data gets available) and other NLP tasks.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Anouar Haha and Ivan Galkin for implementing the data conversion tool used for creating the Wikipedia snapshots.

7. References

Satanjeev Banerjee and Ted Pedersen. 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*, pages 136–145, London, UK. Springer-Verlag.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic.

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

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, 32(1):13–47.

Luciana Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. 2006. Temporal Analysis of the Wikigraph. In *Proceedings of Web Intelligence*, pages 45–51, Hong Kong.

Kenneth W. Church and Robert L. Mercer. 1993. Introduction to the special issue on Computational Linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611, Hyderabad, India, January.

Iryna Gurevych. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 767–778, Jeju Island, Republic of Korea.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's Thesaurus and Semantic Similarity. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 111–120, Borovets, Bulgaria.

Michael Lesk. 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*, pages 24–26, Toronto, Ontario, Canada.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Rada Mihalcea and Dan I. Moldovan. 1999. A Method for Word Sense Disambiguation of Unrestricted Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, College Park, Maryland, USA, June.

David Milne and Ian H. Witten. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*, pages 25–30, Chicago, USA.

David Milne. 2007. Computing Semantic Relatedness using Wikipedia Link Structure. In *Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC 2007)*, Hamilton, New Zealand.

Jane Morris and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In *Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the ACL*, pages 46–51, Boston.

Kotaro Nakayama, Takahiro Hara, and Shohiro Nishio. 2007. Wikipedia Mining for an Association Web Thesaurus Construction. In *Proceedings of International Conference on Web Information Systems Engineering (WISE)*, pages 322–334, Nancy, France, December.

Siddharth Patwardhan and Ted Pedersen. 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*, pages 1–8, Trento, Italy.

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge Derived from Wikipedia for Computing Semantic Relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(1):17–30.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Peter Turney. 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*, pages 313–320, Sydney, Australia.

DeWitt Wallace and Lila Acheson Wallace. 2001–2005. *Reader's Digest, das Beste für Deutschland*. Jan 2001–Dec 2005. Verlag Das Beste, Stuttgart.

Torsten Zesch and Iryna Gurevych. 2010. Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering*, 16(1):25–59, January.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. 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 2007)*, pages 205–208, Rochester, NY, USA.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008a. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. electronic proceedings.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008b. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, IL, USA.