A Comparison of Corpus-Based and Structural Methods on Approximation of Semantic Relatedness in Ontologies

Tuukka Ruotsalo, Aalto University, Finland

Eetu Mäkelä, Aalto University, Finland

ABSTRACT

In this paper, the authors compare the performance of corpus-based and structural approaches to determine semantic relatedness in ontologies. A large light-weight ontology and a news corpus are used as materials. The results show that structural measures proposed by Wu and Palmer, and Leacock and Chodorow have superior performance when cut-off values are used. The corpus-based method Latent Semantic Analysis is found more accurate on specific rank levels. In further investigation, the approximation of structural measures and Latent Semantic Analysis show a low level of overlap and the methods are found to approximate different types of relations. The results suggest that a combination of corpus-based methods and structural methods should be used and appropriate cut-off values should be selected according to the intended use case.

Keywords: *Latent Semantic Analysis, Ontologies, Semantic Relatedness, Semantic Web, Structural Measures*

INTRODUCTION

Ontologies are the backbone of Semantic Web information systems. They are designed to provide a shared understanding of a domain and support knowledge sharing and reuse (Fensel, 2004). Recently, attention has been devoted to using ontologies to improve the performance of information retrieval (Castells et al., 2007) and extraction systems (Ruotsalo et al., 2009), and to support tasks such as query expansion (Kekäläinen & Järvelin, 2000), knowledge-based recommendation (Ruotsalo & Hyvönen, 2007), word sense disambiguation (Ide & Véronis, 1998), and text summarization (Lin & Hovy, 2000).

The ontologies used by such systems are often light-weight general purpose concept ontologies that provide conceptualizations suitable to be used in many domains and applications, but without a manual effort they can not be expected to explicate all the relations required in specific sub-domains (Chandrasekaran et

DOI: 10.4018/jswis.2009100103

al., 1999). For example, a user searching for objects annotated with the concept *flu* on a health portal could be offered articles indexed with the concepts *respiratory infection* or *pneumonia*. On the other hand, the user could be interested in news related an ongoing flu epidemic with related content indexed with concepts such as *vaccinations*, *nutrition* or *medication*.

Avoiding manually tailoring the ontologies, but still enabling such functionalities can be enabled through augmenting relations by estimating relatedness of concepts. Estimates of semantic relatedness can be obtained by making use of structural measures that approximate the relatedness based on the structure of the ontology (Budanitsky & Hirst, 2006). On the other hand, the mentioned applications deal with unannotated corpora that can be used as a source for learning the relations (Landauer et al., 1998; Blei et al., 2003).

While good results have been obtained using both of the approaches (Landauer et al., 1998; Budanitsky & Hirst, 2006), a comprehensive empirical comparison of the approaches has not been reported. To address this, we compare the performance of a widely used corpus-based method, Latent Semantic Analysis (Landauer et al., 1998), and two well-known ontological structural measures, a conceptual measure proposed by Wu & Palmer (1994), and a path-length measure by Leacock & Chodorow (1998).

We report results of a large user study comparing these approaches in semantic relatedness approximation. The focus of the study is to (1) determine the accuracy of the methods, (2) determine the difference between corpus-based methods and structural measures, and (3) identify the strengths and weaknesses of the methods in potential application scenarios.

We show that good accuracy can be achieved using both types of methods, but the methods provide clearly distinct approximations. The results suggest that the approaches are complementary. Structural measures alone can be adequate in scenarios such as information extraction, where synonymy and hyponymy relations suffice (Califf & Mooney, 1999). The combination of methods could be beneficial in scenarios such as information retrieval or word sense disambiguation, where an extensive word context is found to be important (Kekäläinen & Järvelin, 2000; Sussna, 1993). In addition, the results suggest that the performance of the methods are dependent on the correct combination of the methods and assignment of appropriate cut-off values to ensure optimal performance.

The rest of this paper is structured as follows. The following section introduces the semantic relatedness approximation methods used. Section 3 describes the empirical study. The results of the study are presented in section 4. Finally, we conclude with a summary of results, a discussion of shortcomings, and suggestions for future work.

Semantic Relatedness Approximation

In essence, semantic relatedness answers the question: "How much does the meaning of a concept A have to do with the meaning of a concept B?". According to Budanitsky & Hirst (2006) semantic relatedness, or its inverse semantic distance, is a more general concept than similarity. For example, the concepts *bank* and *trust company* are similar, but dissimilar entities may also be semantically related by some other relationship such as associative (*student – school*) or meronymy (*car – engine*).

In this study, approximating semantic relatedness between concepts is defined as determining a relation $r(c, c', w)$ between two concepts c and c' in an ontology. Each relation has a rank $w \in [0,1]$, that indicates the semantic relatedness of the concepts. The rank having a value 1 indicates a strong semantic relatedness and the rank having a value 0 indicates no semantic relatedness. To approximate the rank of the concept pairs, we use two measures that are based on distances of concepts in subsumption hierarchies of lightweight ontologies (Leacock & Chodorow, 1998; Wu & Palmer, 1994). In addition, we use Latent Semantic Analysis (LSA) (Landauer et al., 1998) to approximate the relations based on a text corpus. The methods

are referred as $rel_{LC}(c,c')$, $rel_{WP}(c,c')$, and $rel_{LSA}(c,c')$ respectively.

Leacock-Chodorow Path-Length Measure

Subsumption hierarchies are the backbone of ontologies. For this reason, several measures that use this structure as a source for measuring semantic relatedness have been developed. A simple way to compute semantic relatedness in a subsumption hierarchy is to view it as a graph and identify relatedness with path length between the concepts.

The Leacock & Chodorow (1998) measure is a structural relatedness measure based on path lengths. It is a function of the length of the shortest path in a hierarchy. Formally, for concepts c and c' it is defined as

$$rel_{LC}(c,c') = -log\frac{l(c,c')}{2\times maxdepth(C)},$$

where $l(c,c')$ is a function that returns the smallest number of nodes in the path connecting c and c' (including c and c' themselves) and $maxdepth(C)$ is a function that returns the maximum depth in nodes of all the subsumption hierarchies in the ontology.

Wu-Palmer Conceptual Measure

Despite their simplicity, an acknowledged problem with the path-length measures is that they typically rely on the notion that links in the taxonomy or subsumption hierarchy represent uniform distances (Budanitsky & Hirst, 2006). However, some sub-taxonomies (e.g., biological categories) are often much denser than others, and therefore path length measures tend to give less accurate results.

The Wu & Palmer (1994) measure is a conceptual relatedness measure between a pair of concepts in a hierarchy. It takes into account the fact that two classes near the root of a hierarchy are close to each other in terms of path length but can be very different conceptually, while two classes deeper in the hierarchy can be separated by a larger number of nodes and can still be closer conceptually. Formally, Wu-Palmer measure for concepts c and c' is:

$$rel_{WP} = \frac{2\times l(lcs(c,c'),r)}{l(c,lcs(c,c'))+l(c',lcs(c,c'))+2\times l(lcs(c,c'),r)},$$

where $l(c,c')$ is a function that returns the smallest number of nodes on the path connecting c and c' (including c and c' themselves), $lcs(c,c')$ is a function that returns the lowest common superconcept of concepts c and c', and r is the root concept of the ontology.

A Running Example

To illustrate the function of the structural measures we use an example ontology depicted in Figure 1. The example ontology consists of an subsumption hierarchy of concepts for an air vehicle domain. Computing semantic relatedness between the concepts *seaplanes* and *sailplanes* in this ontology using the Leacock-Chodorow measure results in an equal relatedness rank as for the concepts *helicopters* and *aircraft*, because they are both sister concepts:

$$rel_{LC}(seaplanes,sailplanes) = -log\frac{3}{2\times 4}\approx 0.43$$

$$rel_{LC}(helicopters,aircraft) = -log\frac{3}{2\times 4}\approx 0.43$$

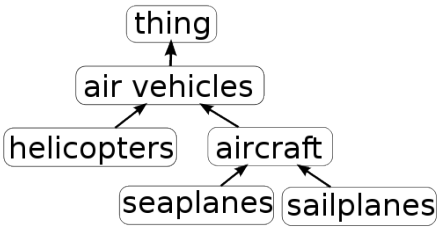
The semantic relatedness for the concepts *helicopters* and *seaplanes* using the Leacock-Chodorow measure is smaller because they are further from each other in the ontology:

$$rel_{LC}(helicopters,seaplanes) = -log\frac{4}{2\times 4}\approx 0.30$$

Comparing these results to those given by the Wu-Palmer measure shows how it considers the depth that the concepts are placed in the hierarchy.

The Wu-Palmer measure for the concepts *seaplanes* and *sailplanes* is

Figure 1. An example ontology



$$rel_{WP}(seaplanes, sailplanes) = \frac{2 \times 3}{2 + 2 + 2 \times 3} = 0.60,$$

while for the concepts *aircraft* and *helicopters* it returns

$$rel_{WP}(aircraft, helicopters) = \frac{2 \times 2}{2 + 2 + 2 \times 2} = 0.50$$

This is because the concepts aircraft and helicopters are closer to the root concept than the concepts seaplanes and sailplanes.

Latent Semantic Analysis

Statistical methods, like Latent Semantic Analysis (LSA) (Landauer et al., 1998), extract and represent the contextual meaning of terms by applying statistical computations to a large corpus of unstructured text.

LSA operates on a vector space model, where a term-document matrix describes the occurrences of terms in documents. The rows of this matrix correspond to terms and the columns correspond to documents. If a term occurs in the document, its term weight in the document vector is a non-zero value given by a weighting scheme. We use tf-idf (Salton & Buckley, 1988) weighting, where rare terms are up-weighted to reflect their relative importance.

Transposing the term-document matrix results in a matrix where each term is associated with a vector of documents containing the occurrence contexts for that term. These vectors provide co-occurrence information that can be used to determine the conceptual distance between terms and sets of terms.

The document vectors could be used directly to determine relatedness of the terms, for example, using a cosine distance (Manning et al., 2008). However, the problem with such direct estimation is that the matrix can be sparse and result in poor estimation (Manning et al., 2008). LSA is a way of reducing this problem by finding latent semantic relations between the terms. LSA reduces the term space by calculating a lower-rank approximation of the document-term matrix. This is done in a way that minimizes the squared error for each number of reduced dimensions.

Consider an example of two terms: *money* and *deposit*. If due to sparseness, they do not appear in the same documents, a direct distance measure does not find a relation between them. However, if the documents in which these terms do appear are similar with respect to other terms (e.g. *bank*, *loan*), the lower rank approximation may combine the documents in latent document space. Measuring relatedness in this latent document space now relates the original terms *bank* and *deposit*.

Formally, LSA can be defined as follows. Let X be a $d \times t$ document-term matrix that describes the documents and the occurrences of each term in these documents. The singular value decomposition (SVD) of X is defined as

$$X = U\Sigma V^T,$$

such that U and V^T are orthogonal matrices and Σ is a diagonal matrix.

SVD makes it possible to translate the matrices to a lower dimensional space with optimal fitting. First, a cut-off value k is set. Then, the first k largest values in Σ are kept and the rest set to zero. Now, composing the matrices by matrix product results in a matrix $\hat{X}$ which approximates the original document-term matrix X in a lower dimensional space (Landauer et al., 1998). For details of the computation, we encourage the readers to see examples by Manning et al. (2008) or Landauer et al. (1998).

The matrix $\hat{X}$ can then be used to calculate relatedness between two terms. We used the cosine measure, where the dot product between two vectors of $\hat{X}$ reflects the extent to which two terms have a similar occurrence pattern in the vector space. Formally,

$$rel_{LSA}(c, c') = \frac{\hat{t}_c \cdot \hat{t}_{c'}}{|\,\hat{t}_c\,||\,\hat{t}_{c'}\,|}$$

where c and c' are concepts and $\hat{t}_c$ and $\hat{t}_{c'}$ the corresponding latent concept space vectors. In our study, the terms appearing in the documents were not directly used in the original term-document matrix. Instead, we mapped each term to a concept in the ontology that contained an equivalent label. This concept-document matrix was then used in the computation.

Comparison of methods in Semantic Relatedness Approximation

Two proper approaches to evaluate methods that approximate semantic relatedness exist (Budanitsky & Hirst, 2006). In case the intended end-use application is known, an option is to evaluate the performance of the methods as a part of that particular application.

In this study however, we are interested of the performance of the methods in a general setting, where a particular end-use application is not known. In such a setting, comparing the ranks given by the methods to gold standard ranks assessed by humans acquired from a large user study is the best way to evaluate the performance of the methods.

Next we will discuss the data, data pre-processing methods, sampling and evaluation methods, and describe the experimental setup used in our user study.

Data

Two kinds of data was used for this study: a lightweight ontology and a text corpus.

The general Finnish ontology (YSO)[1] (Hyvönen et al., 2008) is a lightweight ontology based on the general Finnish keyword thesaurus YSA[2]. The transformation of YSA to YSO was done with care by hand with the following procedure to ensure the coherence of the subsumption hierarchies.

First, an upper ontology was created for the ontology. The upper ontology of the YSO is based on the DOLCE ontology, where enduring, perduring and abstract concepts are separated (Gangemi et al., 2002). Second, the ambiguity of broader-term relations was solved. The subsumption relations were specified based on the original broader term relations. For example, the concept *graduate schools* had a broader term *universities*. However, because graduate schools are not a kind of universities, the concept *graduate schools* was placed in its correct place in the hierarchy, under *educational institutions*. Third, the meaning of polysemous and homonymous concepts were specified and, if required, a new concept for a specific sense of a term was created. Finally, the concepts were organized as subsumption hierarchies under the upper ontology. In this process, more than 1000 concepts and 6000 relations were added into the ontology. After transformation, YSO contains some 26,000 concepts and some 24,000 subsumption relations. In this study we used the version of YSO published in 2007.

The corpus used by the LSA consists of 883 randomly selected articles from the Finn-

ish news paper ”Helsingin Sanomat” from the year 2007. The articles are all written in Finnish. For this study we used both the text in the headings and in the body of the articles. The corpus was selected because it contains news articles, reviews and columns and is therefore relatively domain-neutral or at least represents a general news domain.

Data Pre-Processing and Implementation

The corpus contained many other terms than the ones found in the ontology, such as names of individual persons. For this study, we only used terms for which a corresponding label could be found in the ontology.

A particular concept may have several labels. For example in the ontology used in this study, the concept *academic education* has a term space that contains multiple term correspondences for the concept, such as “academic education”, “higher education”, “higher level education” and “university education”. Therefore, occurrence of any of these terms was used to indicate an occurrence of the concept.

Finnish is a morphologically rich and complex language. Therefore, the terms in the corpus and in the term-space of the ontology were lemmatised with the Omorfi lemmatiser [3].

LSA is based on dimensionality reduction as explained in the previous section. The term-document matrix is decomposed using SVD and then reconstructed with a lower number of target dimensions. An initial test was performed to find the optimal number of target dimensions k for the matrix Σ. The LSA was run in a way that k was iterated from one to a natural cut-off point, the total number of documents in the document collection. The LSA showed an optimal performance with regards to the gold standard when the number of target dimensions was set between 120 and 150. For the actual study, LSA was then run with 150 target dimensions.

The ranks returned by the methods were normalized. LSA and Wu-Palmer measure return semantic relatedness as a real number between 0 and 1. The rank given by Leacock-Chodorow measure was normalized to have the same scale.

LSA was implemented using MTJ, a Java-based matrix calculation API[4], and the structural measures using the Java-based Semantic Web framework Jena [5].

Sampling

A gold standard requires a sample of concept pairs that enable non-biased performance evaluation for all compared methods. The sample should (1) treat all methods equally and (2) retain the distribution of concept pairs.

Two possible sources of bias were identified. First, an information source bias that is caused by the unavoidable fact that the structural and the corpus-based methods use different datasets. The terms can appear in the corpus, but not in the ontology and vice versa. Second, a sampling bias that is caused by the unequal proportion of the concept pairs between the compared methods.

We minimized the possibility of information source bias by restricting the sampling to the terms that were mentioned in the intersection of the terms in the corpus and the terms in the term space of the ontology. After this, the possible number of concept pairs was found to be 4477528. A random sample from such a population would lead to too large sample to be used in a user study. Therefore, stratified sampling was used.

Stratified sampling minimizes the possibility of sampling bias. It groups members of the population into subgroups and the actual sampling is performed from each subgroup. We grouped concept pairs based on two criteria: the method and the rank given by the method for a concept pair. The rank for concept pairs is given on an interval. Therefore we divided the interval into ten bins, the first with values between 0.0 and 0.1, the second with values between 0.1 and 0.2 and so on. This ensures that each method had a representative amount of concept pairs from each rank level.

The obvious choice to sample from these subgroups would be a proportionate stratification, where the sample size of each stratum is proportionate to the population size of the stratum. However, the number of concept pairs that are rated close to non-related is much larger than the number of concept pairs rated closely related. In addition, different methods give different ranks for individual concept pairs. Therefore, we decided to apply disproportionate stratification.

A fixed size sample of 200 random concepts was sampled from each bin. For example, in the case of LSA we first run the method and then sampled 200 concept pairs from the relatedness measure interval between 0.0 and 0.1 given by the method, 200 pairs from the relatedness measure interval between 0.1 and 0.2 and so on. This was repeated for each bin for each method.

It has been shown in previous studies of Rubenstein and Goodenough (1965) and Miller and Charles (1991) that subjects tend to use the dominant sense of the word when assigning relevance judgements for concept pairs. Therefore, we removed concept pairs that contained a polysemous concept to avoid the bias caused by humans using an inappropriate sense in the user study. This resulted in a sample of 3168 concept pairs that have a maximally equal representation for each method on each bin.

This full sample was used in the main study. Our first research goal was to investigate how results given by the three methods differ. Such differences can be obtained by comparing the concept pairs determined by different methods on different relatedness levels. We sampled subsets of the full sample by first ranking the concept pairs based on the rank assigned by each individual method. For each method we sampled 100, 200, 400, 800 and 1600 top-ranked concept pairs from the full sample of 3168 concept pairs. This sampling principle is known as sampling based on cutoff value (CV).

Measuring the performance of the methods based on the full sample directly would suffer from bias caused by stratified sampling because the full sample does not retrain original proportions of the stratums. For example, the equal sampling from stratums would cause underestimating the error where a method makes mistakes in lower relatedness levels and the proportion of such concept pairs would be dominant in a random sample. Therefore, we used post-stratification where scaling factors, based on the proportions of the stratums in the original data were assigned for each bin.

We also collected a smaller sample to measure inter-annotator agreement. This small sample was created by sampling 10 concepts from each stratum of 200 concepts. The samples are further refereed as full, full with cutoff, scaled full and small sample.

Experimental Setup

Human ranks of 15 participants were collected for all together 3168 concept pairs. The participants were students and faculty in the Department of Media Technology at the Helsinki University of Technology. The participants were explicitly asked to judge concept pairs as related in case of any relation and not only inclusive or subsumption relation. Each of the concept pairs was judged on a binary scale (related / non-related). We collected four individual opinions for each concept pair. The relatedness value for a concept pair was set as an average of these ranks (0, 0.25, 0.5, 0.75 or 1). This captured the fact that the relatedness of some of the concept pairs were more vague among the annotators.

Evaluation Methods

According to our research questions, we measured two things: (1) the accuracy of the methods in semantic relatedness approximation, and (2) the difference between corpus-based methods and structural measures.

Because the methods were measured against a gold standard collected from multiple annotators, we first ensured the concordance of the annotators. We used Cohen's Kappa (Cohen, 1960) as a measure for inter-annotator agreement. Kappa measures concordance between

classifiers or annotators using nominal data, varying between -1.0 and 1.0. Cohen's Kappa was run for the concept pairs in the small sample that were annotated by all of the participants. The Kappa measure showed a substantial agreement between the users (Kappa = 0.68).

The accuracy of methods in semantic relatedness approximation was measured using generalized precision and generalized recall originally proposed by Kekäläinen and Järvelin (2002). These measures take into account the fact that the distance between human rank and the rank given by the method are not on a binary scale, but are measured on an interval. Ehrig and Euzenat (2005) have defined the measure in the scope of ontology matching, where the generalized precision and recall are calculated based on an overlap function between a gold standard and the result given by a method. Generalized precision gP and generalized recall gR are defined as follows:

$$gP(A,G) = \frac{overlap(A,G)}{\mid A \mid},$$

$$gR(A,G) = \frac{overlap(A,G)}{\mid G \mid},$$

where G is the set of concept pairs in the gold standard and A is the set of concept pairs given by the method.

The overlap function returns the value 1 if the score in the gold standard and the score given by the method are the same (Ehrig & Euzenat, 2005). The overlap function can now be defined as the difference between the score given by the gold standard $G(c,c')$ and the score given by the method $A(c,c')$ for each concept pair: $1- \mid G(c,c') - A(c,c') \mid$. Intuitively, the generalized precision measures the proportion of error between the gold standard and the method with respect to the number of concept pairs retrieved, and the generalized recall measures the proportion of error between the gold standard and the method with respect to all concept pairs in the gold standard. If all and only all of the concept pairs are retrieved, the generalized precision and generalized recall becomes equal and can be called generalized accuracy gA. Generalized precision, recall and accuracy were determined on the scaled full sample.

It is also interesting if the methods not only perform differently in terms of accuracy, but actually approximate different kinds of relations. We solicited the difference in the kind of relations that the methods approximate by using Jaccard (1901) coefficient. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the concept sets returned by the methods:

$$J(C,C') = \frac{C \cap C'}{C \cup C'},$$

where C and C' are sets of concepts returned by the methods compared.

Real life use cases often aim at finding the best relations and using those in the application. The sets of these relations are based on cut-off values. Thus, Jaccard coefficient was determined on the full sample with cutoff. In addition, we sampled examples from the intersection of the compared sets. These examples were used in a qualitative comparison to find the kind of relations that the methods rank with a certain rank, but are only found by either one of the compared methods.

Because the samples are not normally distributed, which was also checked with the normality test of Shapiro and Wilk (1965), the statistical significance of the results were ensured by using the Friedman test (Conover, 1998; Hull, 1993). Friedman test is a non-parametric test based on ranks and is suitable for comparing more than two related samples. The statistical significance between method pairs was then ensured using a paired Wilcoxon Signed-Rank test (Wilcoxon, 1945) with Bonferonni correction as a post-hoc test. All of the results reported in the next section are statistically significant ($p<0.000001$).

Results

This section presents the results of the experiments. First, we discuss the performance of the methods in semantic relatedness approximation. Further, we compare the differences of the results given by the different methods.

Performance of the Methods

The generalized precision-recall curves of all three methods are shown in Figure 2. The Figure shows that when all concept pairs of the full sample are analyzed LSA performs best (Accuracy 0.84), Wu-Palmer second best (Accuracy 0.74) and Leacock-Chodorow third (Accuracy 0.53). It is notable that the generalized precision of the methods increases when the recall increases. This is due to the fact that we use generalized precision and generalized recall. The performance improvement indicates that the methods are fractionally better in approximating non-related concepts than approximating related concepts. In other words, the measures give better approximation for the concepts that have a human rank 0 in the gold standard, than for the concepts that have rank other than 0 in the gold standard.

Figure 4 shows generalized recall and Figure 3 generalized precision for each of the methods on different ranks given by the methods. These indicate in which rank levels the different methods perform accurately and on which levels they fail. Generalized precision and generalized recall are computed for each bin separately. The figures show that LSA has high performance on rank levels 0.0 to 0.1 on both generalized precision and generalized recall. However, the generalized recall rapidly decreases already on rank level 0.2. This means that LSA is very accurate in approximating non-related concept pairs and therefore shows good overall performance, as the gold standard judges most concept pairs in the original data as non-related.

Table 1 shows the distribution, generalized precision and generalized recall of the relations found by the methods on different rank levels. The structural measures are as accurate or more accurate than LSA on rank levels above 0.2, but their ability to filter out less related concept pairs is weaker. Wu-Palmer measure gives relatively high generalized precision on all rank levels, while Leacock-Chodorow seems to sacrifice precision for recall. LSA approximates most of the concepts to have a rank between 0.0 and 0.1, while for the other methods the distribution is more even. While the structural measures overall perform less accurate than LSA, they perform better in situations where only the relations ranked above a rank 0.2 by the methods are measured. Such cases are typical when cut-off values are used to filter only the top ranked relations to be used in real life applications. Therefore we also run the experiments with seven cut-off values from 0.3 to 0.9.

Table 2 shows the generalized precision and generalized recall of each of the methods on the sets based on the cut-off values. The performance of the Wu-Palmer measure is superior in terms of precision. However, the Leacock-Chodorow measure also achieves good overall performance because of a good recall also on lower (0.3-0.5) cut-off values. LSA performs moderately in terms of precision, but has low recall. This suggests that structural measures with appropriate cut-off values give best performance.

The performance measures that have been used so far still do not reveal a possible differences in the relations that the different methods are able to approximate. A rationale behind this phenomena was investigated by comparing the results of the Wu-Palmer, the structural method that achieved highest precision, and LSA on subsets of the top ranked relations. The relations returned by these methods were analyzed qualitatively and the overlap of the results of the methods was measured. The results of this comparison are discussed in the next section.

Differences in Performance

The difference between the LSA and the Wu-Palmer measure in performance on different CVs can be obtained from the Jaccard coef-

Figure 2. Generalized precision of the methods on 10 generalized recall levels

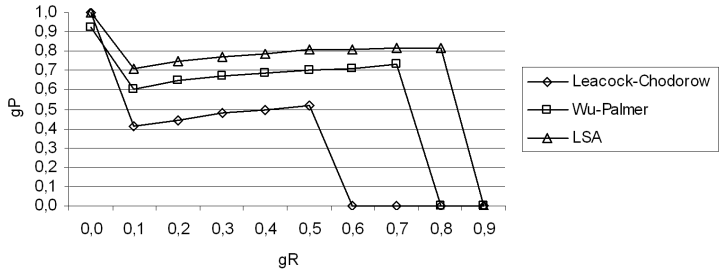


Figure 3. Generalized precision of the methods on different rank levels (not cumulative)

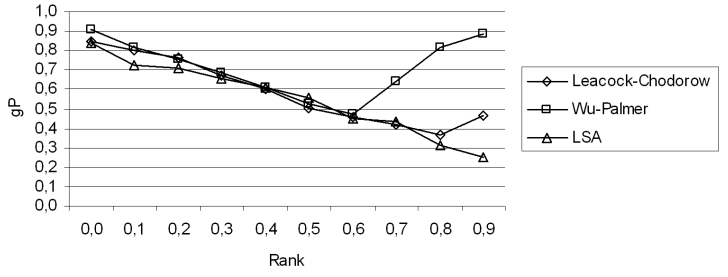
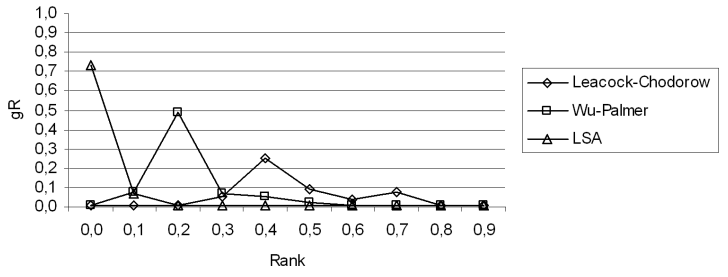


Figure 4. Generalized recall of the methods on different rank levels (not cumulative)



ficient values shown in Table 3. The Jaccard coefficient shows a moderate overlap between the structural measures and low overlap between the Latent Semantic Analysis and the structural measures. For example, on CV 400 the Jaccard coefficient for the Wu-Palmer and the Leacock-Chodorow is 0.47 and for the LSA and the structural measures 0.04 and 0.08 respectively. This indicates that the structural measures and the corpus-based methods are complementary. In possible end-use applications, both should be

Table 1. Number of relations found (thousands) N(K), generalized precision (gP) and generalized recall (gR) for the compared methods on different rank levels. The values within a bin are absolute (not cumulative) and values smaller than 0.01 are rounded to 0.01.

Method	LSA			WP			LC		
Rank	N(K)	gP	gR	N(K)	gP	gR	N(K)	gP	gR
0.9 - 1.0	4.3	0.25	0.01	0.03	0.89	0.01	124.9	0.46	0.04
0.8 - 0.9	1.5	0.31	0.01	1.1	0.82	0.01	138.6	0.37	0.04
0.7 - 0.8	2.2	0.44	0.01	5.6	0.64	0.01	834.8	0.42	0.28
0.6 - 0.7	3.9	0.45	0.01	49.3	0.47	0.01	423.4	0.46	0.15
0.5 - 0.6	6.2	0.56	0.01	194.7	0.53	0.02	768.1	0.51	0.19
0.4 - 0.5	12.7	0.61	0.01	383.1	0.61	0.05	1857.1	0.60	0.31
0.3 - 0.4	31.1	0.66	0.01	444.1	0.68	0.07	326.5	0.67	0.11
0.2 - 0.3	101.3	0.71	0.01	2941.7	0.76	0.49	4.0	0.76	0.01
0.1 - 0.2	436.2	0.72	0.01	450.8	0.81	0.08	0.01	0.78	0.01
0.0 - 0.1	3878.0	0.84	0.40	6.8	0.91	0.01	0.01	0.85	0.01
Total	4477.5	0.84	0.84	4477.5	0.74	0.74	4477.5	0.53	0.53

Table 2. Generalized precision (gP) and generalized recall (gR) for the compared methods on sets based on different cut-off levels

Method	LSA		WP		LC	
Cut-off	gP	gR	gP	gR	gP	gR
0.9	0.34	0.01	0.89	0.01	0.43	0.01
0.8	0.33	0.01	0.82	0.01	0.39	0.02
0.7	0.36	0.01	0.67	0.01	0.41	0.10
0.6	0.39	0.01	0.50	0.01	0.42	0.15
0.5	0.44	0.01	0.53	0.03	0.45	0.23
0.4	0.51	0.01	0.58	0.08	0.52	0.48
0.3	0.59	0.01	0.62	0.15	0.53	0.53

used to obtain good approximation of semantic relatedness.

The Wu-Palmer measure was better in terms of performance than the other structural measure Leacock-Chodorow. Therefore only the Wu-Palmer and the LSA were further compared by qualitatively analyzing the relations they approximate. We investigated sample CV 400 by looking at the concept pairs found by only one of the methods. In other words, either concept pairs that were found by the LSA and not the Wu-Palmer or the other way around. The sample CV 400 was chosen because it already has a relatively high (0.47) Jaccard coefficient for the Wu-Palmer and the Leacock-Chodorow, but a low (0.08 and 0.07) Jaccard coefficient for the LSA and the structural measures.

A systematic sample of concept pairs including human rank and having a rank above 0.6 by the LSA, but not by Wu-Palmer are

Table 3. Jaccard similarity coefficient for pairs of methods. Methods are compared pairwise on different CV points given by each method

CV method pair	100	200	400	800	1600
LC / LSA	0	0.02	0.04	0.07	0.43
WP / LSA	0.08	0.06	0.08	0.17	0.52
LC / WP	0.2	0.39	0.47	0.39	0.82

shown in Table 4. LSA determines relations that are non-hierarchical and can be far away from each other in terms of path length. Good examples shown in Table 4 are concept pairs such as *product / production costs*, *guerilla / child soldier*, *books / shelf* and *flu / nutrition*. An interesting notion is that based on the news corpus that we used, the LSA assigns high rank (0.97) for the concept pair *update / ASP*, where *ASP* stands for an abbreviation for a form of financing subvention of Finnish government for young people planning to buy their first apartment. Such an update was recently made. This is an example of a relation relevant for the domain at a specific time, but clearly one that should be approximated based on the corpus rather than included in the ontology. Similar concept pairs that are relevant for the domain based on the documents are *flu / nutrition* and *guerilla / child soldier*. These are concepts that are only valid in the case of a specific document collection or corpus.

Although low in number, LSA also finds relations that have a common super concept, but are not found by the Wu-Palmer measure. An example of such a relation is *wizard / giants* that have a common super concept *mythic creature*. A possible explanation why LSA performs better than the Wu-Palmer measure is that the hierarchy where these concepts are in is relatively flat and therefore the Wu-Palmer measure approximates a low value for the concept pair. LSA also finds relations that might be relevant for some specific documents, but are difficult to interpret. For example, relations such as *studios / gardener*, *church / plain* and *leather / persistence* may have relevance in terms of an individual news article, but would probably not be beneficial in many end-use applications. The Wu-Palmer measure determines relations that are close based on the subsumption hierarchy. It is notable that LSA only found very few of these relations (Jaccard 0.08 in CV 400 and 0.17 in CV 800).

A systematic sample of concept pairs including human rank and having a rank above 0.7 by the Wu-Palmer, but not the LSA, are shown in Table 5. The examples show that the Wu-Palmer relies on the subsumption hierarchy on a specific depth. All of the concept pairs shown in Table 5 are placed on depth greater than six in the subsumption hierarchy. Wu-Palmer also suffers of a relatively low recall on the rank levels above 0.3. A possible explanation is that the Wu-Palmer measure achieves a high precision by restricting the analysis to concepts relatively deep in the hierarchy. This also causes it to sacrifice recall for precision. On the other hand, the precision of the relations that Wu-Palmer determines is the highest among the compared methods. It also gives accurate approximation for relations, such as *change / boiling* that are difficult to interpret in the scope of possible end use applications, but are found related by the human annotators.

In summary, the Wu-Palmer measure seems remarkably reliable when it assigns a high rank for a concept pair. However, it fails to approximate almost all of the relevant concept pairs that LSA ranked high. LSA seems to be useful in finding relations between concepts that are related, but for which the relation is difficult to obtain using only the ontology graph. In addition, LSA approximates relations that are dependent on the domain and time, but useful in case of the particular document collection or

Table 4. A systematic sample of concept pairs ranked to CV of 400 by the Latent Semantic Analysis and not ranked to CV of 400 by the Wu-Palmer measure

Concept 1	Concept 2	Method	Human
product	production costs	1.0	0.5
Irish	immigrants	0.99	0.5
guerilla	child soldier	0.99	1.0
update	ASP	0.97	0.25
wizard	giants	0.96	0.75
flu	nutrition	0.93	1.0
drinks	chemicals	0.93	0.0
books	shelf	0.90	1.0
studios	gardener	0.82	0.0
foundations	organist	0.81	0.0
suicide attack	population group	0.79	0.5
church	plain	0.69	0.0
symbols	rose	0.68	0.0
leather	persistence	0.66	0.0
sick	disease	0.63	1.0

Table 5. A systematic sample of concept pairs ranked to CV of 400 by the Wu-Palmer measure and not ranked to CV of 400 by the Latent Semantic Analysis

Concept 1	Concept 2	Method	Human
parents	father	0.92	1
masonry	construction work	0.89	1
minorities	population group	0.88	1
document	application	0.84	1
expression	crying	0.84	0.5
measurement	weighing	0.84	1
sportsman	jockey	0.75	1
novel	story	0.82	1
trial	preliminary investigation	0.81	0.75
stone	marble	0.81	0.75
turkey	chicken	0.81	0.75
anecdote	fairy tale	0.77	0.5
windows	stairs	0.77	0.5
near relative	role	0.77	0
change	boiling	0.71	1

corpus. LSA also seems to find subsumption relations, when the subsumption hierarchy itself does not contain enough information to relate the concepts. On the other hand, LSA makes much more mistakes even when it assigns a rank over 0.9.

Conclusion and Discussion

Our goal in this paper was to measure and compare the performance of structural measures and corpus-based methods in approximating semantic relatedness in light-weight ontologies, and to identify the the strengths and weaknesses of the methods in possible application scenarios. Two structural measures by Wu and Palmer (1994) and Leacock and Chodorow (1998), and a corpus-based method Latent Semantic Analysis (Landauer et al., 1998), were compared.

The experimental results show that neither corpus-based method or structure-based measures alone dominate. LSA showed the best performance for the whole dataset. However, both of the structural measures had substantially better performance than LSA when cut-off values were used. Further analysis revealed that LSA and Wu-Palmer measure approximated very different kinds of relations. In addition, we found that the performance of the compared methods varies on different rank levels. LSA is superior in filtering out the non-relevant relations, and is able to find relations in which the structural measures fail.

A combination of LSA and structural measures can be useful in applications such as information retrieval or word sense disambiguation, where an extensive word context is important (Kekäläinen & Järvelin, 2000; Sussna, 1993). Structural measures alone may suffice in scenarios such as information extraction, where synonymy and hyponymy relations are found to be most useful (Califf & Mooney, 1999). In summary, depending on the intended use case, a combination of structural measures and corpus-based methods should be selected and appropriate cut-off values set.

With respect to the size of the empirical study this is, up to our knowledge, the most comprehensive study that evaluates semantic relatedness measures against human relevance assessments. Although Budanitsky and Hirst (2006) compared larger number of methods, they report that their results were obtained using an inadequate sample. In Budanitsky and Hirst (2006) the concept pairs were selected based on their distribution with respect to human ranks. Such a sample can be used as a study to measure human ranks. However, it can be biased when applied to measurement of computational methods that should generalize over an ontology or a corpus.

We used a light-weight ontology developed on a basis of a thesaurus that may have a different concept distribution and lexical coverage compared to other lexical databases, such as WordNet (Miller, 1995). On the other hand, the lightweight ontology used in this study contained more than 26,000 concepts and was ensured to have coherent subsumption hierarchies, which makes the study more fair for the structural measures. We compare the performance of the methods to human ranks acquired in a large user study. A limitation of our analysis is that the concept pairs were annotated by humans on a binary scale. However, we determined a very high value of Cohen's Kappa that showed substantial inter-annotator agreement. In addition, we used the averages of the binary votes of four annotators. Although the measurement accuracy may contain some bias because only five level judgements (values of 0, 0.25, 0.5, 0.75 or 1), the bias is the same for all of the methods, and all of the comparisons are statistically significant.

One of the strengths of the LSA method is that it is able to approximate semantic relatedness based on a corpus in an unsupervised manner. This makes it a good choice to supplement the often limited lexical coverage of an ontology. A limitation of our study is that we restricted the concept pairs to the intersection of concepts appearing in the corpus and in the ontology. On the other hand, the purpose of the study was to measure the accuracy and differences between the methods in the context of ontologies. Because YSO has more than

26,000 concepts, the lexical coverage should be acceptable.

The good performance achieved using hybrid methods proposed by Jiang & Conrath (1997), Resnik (1995), and D. Lin (1998) suggests that a hybrid approach that combine corpus statistics and knowledge-based measures into one method could be a promising direction. Such methods weight the ontology paths based on their mutual information content observed from the corpus and show improved accuracy compared to straightforward path-length measures. However, our results suggest that there is actually a low overlap between the relations found based on the corpus statistics and the relations found based on the ontology structure.

Budanitsky and Hirst (2006) have studied the performance of structural methods using human ranks in a similarity task. They found that Leacock-Chodorow measure along with the hybrid methods proposed by D. Lin (1998), and Jiang and Conrath (1997) performed most accurately. However, as they note, the data used in their study was collected for a similarity task and therefore does not indicate the performance of the methods in semantic relatedness approximation. In addition, they note that the data in their study was not necessarily representative. Therefore, it is difficult to compare their results to the ones obtained in our study.

Our study concentrated on any type of semantic relatedness. However, different kinds of semantic relatedness can be identified. Turney (2006) makes a difference between attributional and relational semantic similarity. For example, the concept pair *mason / stone* is relationally similar or analogous to the concept pair *carpenter / wood* as opposite to attributional similarity that refers to synonymy. Turney (2006) proposes Latent Relational Analysis to approximate relational similarity. This is an important future research direction especially on application areas, such as information retrieval and question answering, where analogous concept pairs could be used to increase accuracy of the retrieval methods (Nakov & Hearst, 2008).

Semantic relatedness approximation has also been included in natural language engineering tasks. LSA been used for hyponymy extraction (Cederberg & Widdows, 2003), topic structure extraction (Valle-Lisboa & Mizraji, 2007), and applied in the area of information retrieval (Deerwester et al., 1990). Coccaro & Jurafsky (1998) combined LSA with n-gram language model and showed improvement in speech recognition. Maguitman et al. (2005) present a graph-based similarity measure to detect similar web-pages. The performance of different methods and combinations of methods in such application cases would be a natural future research direction.

Recent research has proposed Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), as probabilistic variants of LSA. The performance of these methods could be better than LSA. LDA can be used especially for small corpora where the usage of prior information can increase the accuracy.

Future research could explore the findings of this paper in hybrid methods and apply the methods to real life application cases. It would be interesting to measure the performance of LSA without restricting it to the concepts found in ontologies. A hybrid approach where both the terms appearing in the term space of the corpus and the concepts appearing in the ontology could be used in the computation. In this way, LSA could benefit from the additional terms that are not available in ontologies. For example, concept correspondence for proper names can be limited in the ontologies, but could improve the accuracy of LSA. As discussed, LSA approximates also relations that have no meaningful interpretation in terms of the end-use applications. Therefore, LSA could be used to approximate particular types of relations by using the ontology structure as a background knowledge. For example, restricting the approximation of concept pairs to roles and named entities, could reveal useful relations. Another interesting research direction could be to incorporate reasoning in the LSA computation. Constructing the concept-document matrix using reasoning, where occurrences of concepts would imply the occurrences of other concepts

through subsumption reasoning, would enable knowledge-based LSA.

Another possible research direction is to develop methods that are able to both adjust the weights of the paths in the ontology graph based on corpus statistics, and use a meta-classifier that selects the most appropriate prediction method for each concept pair. Such methods would benefit from corpus statistics as an adjustment of the existing ontology graph, and would be able to approximate relations that may not be found using only structural measures.

ACKNOWLEDGMENT

This research is part of the Research and Development project SMARTMUSEUM (Cultural Heritage Knowledge Exchange Platform) sponsored under the Europeans Commission's 7th Framework (FP7-216923).

REFERENCES

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 2003. doi:10.1162/jmlr.2003.3.4-5.993

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47. doi:10.1162/coli.2006.32.1.13

Califf, M. E., & Mooney, R. J. (1999). Relational learning of pattern-match rules for information extraction. In *AAAI '99/IAAI '99: Proceedings of The Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference* (pp. 328-334). Menlo Park, CA: American Association for Artificial Intelligence.

Castells, P., Fernandez, M., & Vallet, D. (2007, February). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *19*(2), 261–272. doi:10.1109/TKDE.2007.22

Cederberg, S., & Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (pp. 111-118). Morristown, NJ: Association for Computational Linguistics.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, *14*(1), 20–26. doi:10.1109/5254.747902

Coccaro, N., & Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP-98)* (pp. 2403-2407). ASSTA. (Volume 6)

Cohen, J. (1960, April). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Conover, W. J. (1998). *Practical nonparametric statistics*. New York: John Wiley & Sons.

Deerwester, S., Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Ehrig, M., & Euzenat, J. (2005, October). Relaxed precision and recall for ontology matching. In Integrating ontologies '05, proceedings of the K-CAP 2005 workshop on integrating ontologies (Vol. 156). CEUR-WS.org.

Fensel, D. (2004). *Ontologies: A silver bullet for knowledge management and electronic commerce* (2nd ed.). Heidelberg, Germany: Springer.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE. In *EKAW '02: Proceedings of the 13th international conference on knowledge engineering and knowledge management ontologies and the semantic web* (pp. 166-181). London: Springer-Verlag.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1-2), 177–196. doi:10.1023/A:1007617005950

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 329-338). New York: ACM.

Hyvönen, E., Viljanen, K., Tuominen, J., & Seppälä, K. (2008, June 1-5). Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In 5th european semantic web conference (eswc 2008) (pp. 95–109). Springer.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, *24*(1), 2–40.

Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Jiang, J. J., & Conrath, D. W. (1997, September). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the international conference research on computational linguistics (ROCLING X)* (pp. 19-33).

Kekäläinen, J., & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval*, *1*(4), 329–344. doi:10.1023/A:1009983401464

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, *53*(13), 1120–1129. doi:10.1002/asi.10137

Landauer, T., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*(1), 259–284. doi:10.1080/01638539809545028

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification . In Fellbaum, C. (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge, MA: MIT Press.

Lin, C. Y., & Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on computational linguistics* (pp. 495-501). Morristown, NJ: Association for Computational Linguistics.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics* (pp. 768-774). Morristown, NJ: Association for Computational Linguistics.

Maguitman, A. G., Menczer, F., Roinestad, H., & Vespignani, A. (2005). Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web* (pp. 107-116). New York: ACM.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (1st ed.). Cambridge, UK: Cambridge University Press.

Miller, G. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38*(11). doi:10.1145/219717.219748

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28. doi:10.1080/01690969108406936

Nakov, P., & Hearst, M. A. (2008, June). Solving relational similarity problems using the Web as a corpus. In *Proceedings of ACL-08: HLT* (pp. 452-460). Columbus, OH: Association for Computational Linguistics.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on artificial intelligence* (pp. 448-453). San Francisco: Morgan Kaufmann Publishers Inc.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633. doi:10.1145/365628.365657

Ruotsalo, T., Aroyo, L., & Schreiber, G. (2009). Knowledge-based linguistic annotation of digital cultural heritage collections. *IEEE Intelligent Systems*, *24*(2), 64–75. doi:10.1109/MIS.2009.32

Ruotsalo, T., & Hyvönen, E. (2007, September). A method for determining ontology-based semantic relevance. In *Proceedings of the international conference on database and expert systems applications (DEXA 2007)*. Regensburg, Germany: Springer.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *3*(52).

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM '93: Proceedings of the second international conference on information and knowledge management* (pp. 6774). New York: ACM.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416. doi:10.1162/coli.2006.32.3.379

Valle-Lisboa, J. C., & Mizraji, E. (2007). The uncovering of hidden structures by latent semantic analysis. *Information Sciences*, *177*(19), 4122–4147. doi:10.1016/j.ins.2007.04.007

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. doi:10.2307/3001968

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics* (pp. 133-138). Morristown, NJ: Association for Computational Linguistics.

ENDNOTES

1 Latest version of the ontology is available in RDF(S) from: http://www.yso.fi
2 http://vesa.lib.helsinki.fi/ysa/
3 http://home.gna.org/omorfi/omorfi/
4 http://rs.cipr.uib.no/mtj/
5 http://jena.sourceforge.net/

Tuukka Ruotsalo is a PhD student at the Aato University's Department of Media Technology. His research interests include knowledge-based methods for information retrieval, recommendation, and annotation of media content. Ruotsalo has an MSc in information systems science from the University of Jyväskylä.

Eetu Mäkelä is a PhD student at the Aato University's Department of Media Technology. His research interests include knowledge-based methods for information access and user interface technology. Mäkelä has an MSc in computer science from the University of Helsinki.