The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis

Mihai Lintean and **Cristian Moldovan** and **Vasile Rus**
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152, USA
mclinten|cmldovan|vrus@memphis.edu

Danielle McNamara
Department of Psychology
Institute for Intelligent Systems
Institute for Intelligent Systems
The University of Memphis
dsmcnamara@memphis.edu

Abstract

In this paper, we investigate the impact of several local and global weighting schemes on Latent Semantic Analysis' (LSA) ability to capture semantic similarity between two texts. We worked with texts varying in size from sentences to paragraphs. We present a comparison of 3 local and 3 global weighting schemes across 3 different standardized data sets related to semantic similarity tasks. For local weighting, we used binary weighting, term-frequency, and log-type. For global weighting, we relied on binary, inverted document frequencies (IDF) collected from the English Wikipedia, and entropy, which is the standard weighting scheme used by most LSA-based applications. We studied all possible combinations of these weighting schemes on the following three tasks and corresponding data sets: paraphrase identification at sentence level using the Microsoft Research Paraphrase Corpus, paraphrase identification at sentence level using data from the intelligent tutoring system iSTART, and mental model detection based on student-articulated paragraphs in MetaTutor, another intelligent tutoring system. Our experiments revealed that for sentence-level texts a combination of type frequency local weighting in combination with either IDF or binary global weighting works best. For paragraph-level texts, a log-type local weighting in combination with binary global weighting works best. We also found that global weights have a greater impact for sententence-level similarity as the local weight is undermined by the small size of such texts.

Introduction

Assessing the semantic similarity of texts (words, sentences, paragraphs, documents) is an important step in many real-world applications ranging from summarization (Ibrahim, Katz, and Lin 2003) to educational systems (Graesser et al. 2007; McNamara et al. 2007) to automatic detection of duplicate bug reports in software testing (Rus et al. 2009).

For instance, in the intelligent tutoring system iSTART student-articulated sentences must be compared with benchmark sentences, i.e. sentences from a textbook. The closer in meaning the student-sentence and the benchmark the better as students are supposed to express in their own words (paraphrase) the benchmark sentence. This particular task is about finding the semantic similarity of two sentences. An example of a textbook sentence (T) and student paraphrase (SP; reproduced as typed by the student) in iSTART is provided below (from the User Language Paraphrase Challenge (McCarthy and McNamara 2008)):

T: *A glacier's own weight plays a critical role in the movement of the glacier.*

SP: *The weight of the glacier detemines how it will move.*

In a related task, automatic detection of student mental models in MetaTutor (Azevedo et al. 2008), an intelligent tutoring system that teaches students self-regulatory skills, a challenging task is deciding how similar a student-generated paragraph is to an ideal, expert-generated paragraph. The student-generated paragraphs are obtained from the prior knowledge activation (PKA) meta-cognitive activity in MetaTutor when students are prompted to write a paragraph outlining everything they know about a given learning goal, e.g. learn about the human circulatory system. In this case, the task is to assess how semantically similar two given paragraphs are.

One method to compute similarity between texts, such as the ones shown above, is to use Latent Semantic Analysis (LSA; (Landauer et al. 2007)). LSA represents the meaning of individual words using a vector-based representation. The similarity of two words can be computed as the normalized dot-product between corresponding vectors. Extending LSA to assess similarity of texts beyond word-level involves the use of local weighting, which quantifies the importance of words within the texts, and global weighting, which quantifies the importance of words in a large corpus, i.e. across many texts.

The choice of local and global weighting can have a significant impact on the overall performance of LSA-based semantic similarity methods (Dumais 1991; Landauer et al. 2007). Comparison among several weighting schemes for deriving the LSA-based representations of individual words have been done before (Dumais 1991). However, to the best of our knowledge the role of various weighting schemes for LSA-based similarity between texts the size of a sentence or more has not been investigated before. Furthermore, no study has been as extensive and conducted across several tasks as the one presented in this paper. It is important to assess the role of weighting schemes across texts of various sizes as some weights may behave differently depending on the size of the text. For instance, the local weight of raw fre-

Table 1: Example of ideal and student-generated paragraphs in MetaTutor.

Type	Paragraph
Ideal	The heart is a muscular organ that is responsible for pumping blood throughout the body through the blood vessels. The heart, blood, and blood vessels work together to carry oxygen and nutrients to organs and muscles throughout the body and carry away waste products. The circulatory system works with the system that makes hormones (the endocrine system), which controls our heart rate and other body functions. Blood carries oxygen from the lungs to all the other organs in the body. Metabolism occurs when blood delivers nutrients, such as proteins, fats, and carbohydrates, to our body.
Student	The circulatory system is composed of blood, arteries, veins, capillaries, and the heart. It's purpose is to supply blood flow and oxygen to the body and to pick up waste (carbon dioxide). Blood is either oxygen rich or poor. Oxygen poor blood needs to return to the lungs from the heart to get more oxygen. Once blood has generated through the body it's oxygen is depleted. Needs to get back to the heart so it can get back to lungs.

quency which counts the number of occurrences of a word in the text will be dominated by the global weight in texts the size of a sentence because in such texts raw frequency is most of the time 1. That is explained by the fact that words are not reused in a sentence while in a paragraph they are, e.g. for cohesion purposes.

In this paper, we present a comparison of 3 local and 3 global weighting schemes across 3 different standardized data sets related to semantic similarity tasks. For local weighting, we compare binary weighting, term-frequency, and log-type. For global weighting, we compare binary weight, inverted document frequencies (IDF) collected from the English Wikipedia, and entropy, which is the standard weighting scheme used by most LSA-based applications. We studied all possible combinations of these weighting schemes on four different tasks and corresponding data sets: paraphrase identification at sentence level using the Microsoft Research Paraphrase Corpus, paraphrase identification at sentence level using data from the intelligent tutoring system iSTART, and mental model detection based on student-articulated paragraphs in MetaTutor, another intelligent tutoring system. Accuracy, which is the percentage of a method's prediction that match the expected predictions suggested by experts, is used to compare the weighting schemes. Our experiments revealed that an IDF global weight usually helps more than using entropy weighting and that global weighting has a larger impact than local weighting for sentence-level texts.

The rest of the paper is organized as follows. The next section presents *Related Work* on weighting schemes for LSA and on extensions of word-to-word similarity measures to sentence and paragraph level. The *Latent Semantic Analyis* section describes in detail the LSA framework and type of weighting. Next, we present our experiments and results. The *Conclusions* sections summarizes the major findings and outlines plans for the future.

Related Work

There are two main lines of research that are directly related to our work. First, there is research regarding various weighting schemes used with LSA. Second, there is work on extending word-to-word similarity measures to larger text sizes such as sentences or paragraphs.

The basic LSA framework is used to represent the meaning of individual words. As it relies on word co-occurrences on large collections of documents, various techniques have been used to reduce the role of highly-frequent words, e.g. *the*, which do not carry much meaning (Dumais 1991; Berry, Dumais, and O'Brien 1995; Nakov, Popova, and Mateev 2001). These techniques combine local and global weights. However, this type of weighting is used to derive the LSA-based representation of individual words, which is different from our focus which is on the role of weighting on LSA-based representations of texts the size of a sentence or beyond. For instance, Dumais (Dumais 1991) has experimented with six weighting schemes that were derived from combining 3 local and 4 global schemes (not all combinations were explored). The most successful combination was based on the log of the local (within a document) term frequency and the inverse of the entropy of the term in the corpus. Nakov and colleagues (Nakov, Popova, and Mateev 2001) experimented with 12 combinations of more or less the same set of local and global weights and found similar results, i.e. a combination of log and entropy is best. It is important to note that Dumais and Nakov and colleagues focused on different tasks: information retrieval and text classification, respectively. Our work differs from theirs in two important aspects. First, we focus on text-to-text similarity tasks at sentence and paragraph level. Second, we experiment with weighting schemes to extend the LSA representation to sentences and paragraphs, after the derivation of the LSA representation of individual words. Previous research on local and global weighting schemes for LSA has focused on weighting before the derivation of the LSA representation, i.e. during the creation of the term-by-document frequency matrix which is the input to the LSA procedure to derive the LSA representation for words. The term-by-document matrix contains information about which word occurred in which document in a large collection of documents.

There has been an increasing interest recently to extend word-to-word similarity measures to sentence level and beyond. The recent developments were driven primarily by the creation of standardized data sets for the major text-to-text relations of entailment (RTE; Recognizing Textual Entailment corpus, (Dagan, Glickman, & Magnini 2004 2005)),

paraphrase (MSR; Microsoft Research Paraphrase corpus, (Dolan, Quirk, and Brockett 2004)), and more recently for elaboration (ULPC, User Language Paraphrase Challenge, (McCarthy and McNamara 2008)). For instance, (Corley and Mihalcea 2005) proposed an algorithm that extends word-to-word similarity metrics to a text-to-text semantic similarity metric based on which they decide whether two sentences are paraphrases or not. To get the semantic similarity between words they used the WordNet similarity package (Patwardhan, Banerjee, and Pedersen 2003). Then, they combined word-to-word similarity metrics using a weighted sum where the weight of each word is the inverted document frequency of the word. Their method only considered content words to compute similarity between texts because the WordNet similarity package only handles content words (nouns, verbs, adjectives, adverbs), a feature inherited from WordNet (Miller 1995). (Rus, Lintean, Graesser, and McNamara 2009) computed a semantic concept overlap score of two sentences by greedily matching each concept in one sentence to the most related concept, according to a WordNet-based word-to-word relatedness measure, in the other sentence. In a second method, concepts in one sentence were weighted by their importance which was estimated using their specificity. Specificity was derived based on inverted document frequency. Their goal was to decide whether two sentences are semantically equivalent, i.e. paraphrases, or not.

In this paper, we extended LSA-based word-to-word similarity metrics to text-to-text similarity metrics using a combination of local and global weights. We compared the different combinations of local and global weights by observing the accuracy of these methods on three text-to-text similarity tasks. Such extensive comparison of weighting schemes for LSA-based text-to-text similarity methods has not been done before to the best of our knowledge.

Latent Semantic Analysis

LSA is a statistical technique for representing meaning of words that relies on word co-occurrences to derive a vectorial representation for each word. It is based on the principle that the meaning of a word is defined by the company it keeps. Two words have related meaning if they co-occur in the same contexts. The co-occurrence information is derived from large collections of text documents. In a first step, a term-by-document matrix X is created in which element (i, j) contains a binary value, 1 if word i occurs in document j and 0 otherwise. More sophisticated weights could be used that indicate the importance of word i for document j, for instance, the raw frequency of word i in document j. All our experiments used the most common local-global weighting scheme for this case, which is global entropy with local log-type frequency. After the term-by-document matrix is created, a mathematical procedure, called Singular Value Decomposition (SVD), is applied resulting in three new matrices: T and D, which are orthonormal, and S, which is a diagonal matrix, such that $X = TSD^t$. The dimensionality of these matrices is then reduced by retaining k rows and columns corresponding to the highest k values in S. A new matrix $X' = T'X'D'^t$ can now be computed that is an approximation of original term-by-document matrix X in a reduced spaced of k dimensions. Usually, k takes values between 300 and 500. Every word in the initial collection of documents is characterized by a row (or vector) in the reduced matrix X'. These vectors supposedly characterize the words using so-called latent concepts, one for each of the k dimensions of the reduced space.

LSA-based Similarity of Texts

To compute how similar two words based on LSA vector representations, the cosine between the vectors must be computed. The cosine is the normalized dot product between the vectors. If we denote $V(w)$ the LSA vector of a word w then the cosine is given by Equation 1.

$$LSA(w_1, w_2) = \frac{V(w_1) * V(w_2)}{||V(w_1)|| * ||V(w_2)||} \quad (1)$$

A cosine value of 0 means there is no semantic similarity between words or paragraphs while 1 means they are semantically equivalent (synonyms).

The use of LSA to compute similarity of texts beyond word-level relies mainly on combining the vector representation of individual words. Specifically, the vector representation of a text containing two or more words is the weighted sum of the LSA vectors of the individual words. If we denote $weight_w$ the weight of a word as given by some scheme, local or global, then the vector of a text T (sentence or paragraph) is given by Equation 2. In Equation 2, w takes value from the set of unique words in text T, i.e. from the set of word types of T. If a word type occurs several times in a document that will be captured by the local weight ($loc - weight$). $Glob - weight$ in Equation 2 represents the global weight associated with type w, as derived from a large corpus of documents.

$$V(T) = \sum_{w \in T} loc - weight_w * glob - weight_w * V_w \quad (2)$$

To find out the LSA similarity score between two texts $T1$ and $T2$, i.e. $LSA(T1, T2)$, we first represent each sentence as vectors in the LSA space, $V(T1)$ and $V(T2)$, and then compute the cosine between two vectors as shown in Equation 1.

There are several ways to compute local and global weights. For local weighting, the most common schemes are: *binary*, *type frequency* and *log-type frequency*. *Binary* means 1 if the word type occurs at least once in the document and 0 if it does not occur at all. *Type frequency* weight is defined as the number of times a word type appears in a text, sentence or paragraph in our case. *Log-type frequency* weight is defined as log(1 + *type frequency*). It has been proposed by (Dumais 1991) based on the observation that type frequency gives too much weight/importance to very common, i.e. frequent, words. A frequent word such as *the* which does not carry much meaning will have a big impact although its entropy (described next) is low, which is counterintuitive. To diminish the frequency factor for such words, but not eliminate it entirely, the log-type weighting scheme was proposed.

As global weight, we started with a binary weight, similarly to local binary weight: 1 if the words exist in the text, 0 otherwise. The most commonly used global weight is entropy-based. It is defined as $1 + \sum_j \frac{p_{ij} log_2(p_{ij})}{log_2 n}$, where $p_{ij} = tf_{ij}/gf_i$, tf_{ij}= type of frequency of type i in document j, and gf_i= the total number of times that type i appears in the entire collection of n documents. We also used IDF, inverted document frequency, as a global weight.

Word Distribution Information from Wikipedia

Given the need for word distributional information in our global weighting schemes, i.e. gf_i and IDF, it is important to derive as accurate estimates of word statistics as possible. Accurate word statistics means being representative of overall word usage (by all people at all times). The accuracy of the estimates are largely influenced by the collection of texts where the statistics are derived from. Various collections were used so far to derive word statistics. For instance, (Corley and Mihalcea 2005) used the British National Corpus as a source for their IDF values. We chose Wikipedia instead because it encompasses texts related to both general knowledge and specialized domains and it is being edited by many individuals, thus capturing diversity of language expression across individuals. Furthermore, Wikipedia is one of the largest publicly available collections of English texts.

Extracting IDF values and word statistics from very large collections of text, such as Wikipedia, is non-trivial task. Due to space limitation we do not present the details of this step. We just mention that after considering only words that appear in at least two documents in the Wikipedia collection, then the number of distinct words is *2,118,550*. We have collected distributional information for this set of words and used it in our experiments.

Experiments and Results

We have explored all 9 possible combinations among local and global weighting schemes on three different datasets: Microsoft Paraphrase Corpus (MSR corpus), iSTART/ULPC, and PKA/MetaTutor. For each dataset, the task was to compute similarity between two texts and assess how well the LSA based predictions matched expert judgments. In the MSR and iSTART corpora, texts are the lengths of a sentence while the PKA data set contains texts the size of paragraphs. Details about each dataset will be provided in the next subsections.

We calculate LSA-based similarity between pairs of texts using all combinations of weighting schemes presented earlier and use logistic regression from WEKA machine learning toolkit (Witten and Frank 2005) to classify instances based on the LSA score. We report results using five performance metrics: accuracy, kappa measure, and weighted averages for precision, recall and f-measure. Accuracy is the percentage of correct predictions out of all predictions. Kappa coefficient measures the level of agreement between predicted categories and expert-assigned categories while also accounting for chance agreement. Precision is the percentage of correctly predicted instances out of all predictions. In case of multiple classes, as in the case of the MetaTutor dataset, precision is computed as average per class, same for recall. Recall is the percentage of correctly predicted instances having a certain class out of all actual instances that have that class. F-measure is calculated as the harmonic mean of precision and recall. Results were obtained using 10-fold cross-validation, except for MSR dataset which contains an explicit test subset, which we used.

We present three tables; each table coresponds to one dataset. Lines are specific to global weighting schemes, and on columns are listed each of the five evaluation measures grouped by local weighting schemes. We list all possible combinations of three global weighting schemes (binary weighting, entropy weighting, and idf weighting) with three local weighting schemes (binary weighting, type frequency weighting, and log-type frequency weighting).

Table 2 presents results on the MSR corpus, Table 3 reports results on the iSTART/ULPC corpus, while Table 4 shows results on the MetaTutor/PKA corpus.

Microsoft Research Paraphrase Corpus

First corpus we studied LSA on is the Microsoft Research Paraphrase Corpus (Dolan, Quirk, and Brockett 2004), which is a standard data set for evaluating approaches to paraphrase identification. Although it has its limitations (see (Zhang and Patrick 2005) and (Lintean inpress) for some discussions on it), the MSR Paraphrase Corpus has been so far the largest publicly available annotated paraphrase corpus and has been used in most of the recent studies that addressed the problem of paraphrase identification. The corpus consists of 5801 sentence pairs collected from newswire articles, 3900 of which were labeled as paraphrases by human annotators. The whole set is divided into a training subset (4076 sentences of which 2753, or 67%, are true paraphrases), and a test subset (1725 pairs of which 1147, or 66%, are true paraphrases). Average words per sentence number for this corpus is 17.

iSTART

Next corpus is related with MRS in the sense that it applies to the same problem of paraphrase identification. We experimented with the User Language Paraphrase Corpus (ULPC; (McCarthy and McNamara 2008)), which contains pairs of target-sentence/student response texts. These pairs have been evaluated by expert human raters along 10 dimensions of paraphrase characteristics. In current experiments we evaluate the LSA scoring system with the dimension called *"Paraphrase Quality bin"*. This dimension measures the paraphrase quality between the target-sentence and the student response on a binary scale, similar to the scale used in MSR. From a total of 1998 pairs, 1436 (71%) were classified by experts as being paraphrases. The average words per sentence number is 15.

Prior Knowledge Activation Paragraphs

Third corpus is a bit different than the previous two and was created to help evaluating methods that classify textual inputs given by students in an Inteligent Tutoring Environment. The corpus contains 309 paragraphs composed

Table 2: LSA results on the MSR dataset.

(global)	binary (local) Acc	Kap.	Prec	Rec.	F	type frequency (local) Acc	Kap.	Prec	Rec.	F	log-type frequency (local) Acc	Kap.	Prec	Rec.	F
binary	70.38	.247	.686	.704	.674	70.55	.244	.689	.706	.672	70.20	.237	.684	.702	.669
entropy	69.16	.202	.669	.692	.653	68.98	.199	.667	.690	.652	69.22	.205	.670	.692	.655
idf	69.85	.231	.679	.699	.667	69.74	.228	.667	.697	.665	69.85	.230	.679	.699	.666

Table 3: LSA results on the iSTART/ULPC dataset.

(global)	binary (local) Acc	Kap.	Prec	Rec.	F	type frequency (local) Acc	Kap.	Prec	Rec.	F	log-type frequency (local) Acc	Kap.	Prec	Rec.	F
binary	61.21	.196	.618	.612	.591	61.11	.194	.616	.611	.591	61.66	.205	.624	.617	.595
entropy	62.61	.228	.630	.630	.611	62.61	.227	.631	.626	.610	62.66	.228	.632	.627	.611
idf	63.21	.240	.638	.632	.616	63.21	.239	.639	.632	.616	63.16	.238	.638	.632	.615

by students when they are asked by to describe what they know about a complex science topic, in particular the circulatory system in biology. The tutoring system MetaTutor (Azevedo et al. 2008) tries to help learners activate a particular self-regulatory process called Prior Knowledge Activation (PKA), one of the self-regulatory processes that, if properly used, are believed to improve student's learning. The PKA paragraphs given by students are assumed to reflect student's knowledge about the current topic, in other words, the student's current mental model. A proper automatic evaluation of these paragraphs will help an interactive tutoring system to evaluate the student, measure its learning, and give feedback or act depending on student's current level of mental knowledge. The paragraphs in the corpus are labeled by human experts on three levels of mental models (MM): High (100 paragraphs), Medium (70 paragraphs) and Low (139 paragraphs). For this corpus we compare each student paragraph with one ideal paragraph which is considered as benchmark for a perfect mental model, representing the highest level of MM. This ideal paragraph was created by a human expert and contains summarized information about all important concepts encompased in the learning topic. The average number of words in a paragraph is 124.

Summary of Results

Table 5 presents the combination of local and global weighting schemes for which the best results in terms of accuracy were obtained for each dataset. For sentence-level texts a combination of type frequency local weighting in combination with either IDF or binary global weighting works best. For paragraph-level texts, a log-type local weighting in combination with binary global weighting works best. From the table, we can see that there is no clear winner of global and local weight combination across tasks and text sizes. That may be a result of different distribution of positive and negative examples in the three data sets and that on MSR we used a training-test form of evaluation while for the other we used 10-fold cross-validation.

Table 5: Weighting scheme combinations corresponding to best results for each dataset.

global×local	ULPC idf×type	MSR bin×type	PKA bin×log-type
accuracy	63.21	70.55	61.16
kappa	.239	.244	.354
precision	.639	.689	.473
recall	.632	.706	.612
f-measure	.616	.672	.534

We also conducted a repeated measures analysis of variance with the local weighting as a repeated measurement. The differences among the various local weightings were significantly different at $p<0.05$ with the exception of raw and binary local weighting for sentence-level texts. This could be explained by the fact that for such texts the raw frequency and the binary value for binary weighting coincides simply because words tend to occur only once in a sentence. That is, words are less likely to be re-used in a sentence as opposed to a paragraph.

Summary and Conclusions

The paper presented an investigation on the role of local and global weighting schemes on the accuracy of methods for computing similarity between texts the size of a sentence or paragraph. Our experiments revealed that for sentence-level texts a combination of binary local weighting in combination with either IDF or binary global weighting works best. For paragraph-level texts, a log-type local weighting in combination with binary global weighting

Table 4: LSA results on the MetaTutor/PKA dataset

	binary (local)					type frequency (local)					log-type frequency (local)				
(global)	Acc	Kap.	Prec	Rec.	F	Acc	Kap.	Prec	Rec.	F	Acc	Kap.	Prec	Rec.	F
binary	60.84	.347	.472	.608	.531	58.58	.318	.456	.586	.512	61.16	.354	.473	.612	.534
entropy	58.25	.308	.450	.583	.508	58.25	.310	.451	.583	.508	59.55	.334	.461	.595	.519
idf	60.19	.341	.465	.602	.525	59.87	.338	.463	.599	.522	59.55	.333	.461	.595	.519

works best. We also found that global weights have a greater impact for sententence-level similarity as the local weight is undermined by the small size of such texts. The experiments revealed that there is no clear winning combination of global and local weighting across tasks and text size, which is somehow different from earlier conclusions for different types of weigting in LSA, at word-level representations, that entropy and log-type is the best combination. We plan to further explore the role of local and global weighting in more controlled experiments in which we use the same distributions of positive and negative examples and same evaluation methodology, 10-fold cross-validation, across all data sets and text sizes.

References

Azevedo, R.; Witherspoon, A.M.; Graesser, A.C.; McNamara, D.S.; Rus, V.; Cai, Z.; Lintean, M.C. 2008. MetaTutor: An adaptive hypermedia system for training and fostering self-regulated learning about complex science topics. *Meeting of Society for Computers in Psychology*. Chicago, IL.

Berry, M.W.; Dumais, S.; and O'Brien, G. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37, 573-595.

Corley, C., and Mihalcea, R. 2005. Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.

Dagan, I.; Glickman, O.; and Magnini, B. 2004-2005. Recognizing textual entailment. In *http://www.pascal-network.org/Challenges/RTE*.

Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING*, Geneva, Switzerland.

Dumais, S. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23,229-236.

Graesser, A.C.; Penumatsa P.; Ventura M.; Cai, Z.; and Hu, X. 2007. *Handbook of Latent Semantic Analysis.* Erlbaum, Mahwah, NJ. chapter Using LSA in AutoTutor: Learning through Mixed-initiative Dialogue in Natural Language. (pp. 243-262)

Ibrahim, A.; Katz, B.; and Lin, J. 2003. Extracting structural paraphrases from aligned monolingual corpora *In Proceedings of the Second International Workshop on Paraphrasing, (ACL 2003).*

Landauer, T.; McNamara, D. S.; Dennis, S.; and Kintsch, W. 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Letsche, T.; and Berry, M.W. 1997. Large-scale information retrieval with latent semantic indexing. *Information Sciences*, 100, 105-137.

Lintean, M.; and Rus, V. (in press). Paraphrase Identification Using Weighted Dependencies and Word Semantics. to be published in *Informatica, An International Journal of Computing and Informatics.*

McCarthy, P.M.; and McNamara, D.S. 2008. *User-Language Paraphrase Corpus Challenge* https://umdrive.memphis.edu/pmmccrth/public/Paraphrase Corpus/Paraphrase_site.htm. Retrieved 2/20/2010 online, 2009.

McNamara, D.S.; Boonthum, C.; Levinstein, I. B.; and Millis, K. 2007. *Handbook of Latent Semantic Analysis.* Erlbaum, Mahwah, NJ. chapter Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms, 227-241.

Miller, G. 1995 WordNet: A Lexical Database of English. Communications of the ACM, v.38 n.11, p.39-41.

Nakov, P.; Popova, A.; Mateev, P. 2001. Weight functions impact on LSA performance. In *Proceedings of the Euro-Conference Recent Advances in Natural Language Processing*, 187-193.

Rus, V.; Lintean M.; Graesser, A.C.; and McNamara, D.S. 2009. Assessing Student Paraphrases Using Lexical Semantics and Word Weighting. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, Brighton, UK.

Rus, V.; Nan, X.; Shiva, S.; and Chen, Y. 2009. Clustering of Defect Reports Using Graph Partitioning Algorithms. In *Proceedings of the 20th International Conference on Software and Knowledge Engineering*, July 2-4, 2009, Boston, MA

Witten, I. H.; and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco.

Zhang, Y., and Patrick, J. 2005. Paraphrase Identification by Text Canonicalization In *Proceedings of the Australasian Language Technology Workshop 2005*, 160–166.