Automatic and human scoring of word definition responses

Kevyn Collins-Thompson and Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, U.S.A. 15213-8213
{kct|callan}@cs.cmu.edu

Abstract

Assessing learning progress is a critical step in language learning applications and experiments. In word learning, for example, one important type of assessment is a definition production test, in which subjects are asked to produce a short definition of the word being learned. In current practice, each free response is manually scored according to how well its meaning matches the target definition. Manual scoring is not only time-consuming, but also limited in its flexibility and ability to detect partial learning effects.

This study describes an effective automatic method for scoring free responses to definition production tests. The algorithm compares the text of the free response to the text of a reference definition using a statistical model of text semantic similarity that uses Markov chains on a graph of individual word relations. The model can take advantage of both corpus- and knowledge-based resources. Evaluated on a new corpus of human-judged free responses, our method achieved significant improvements over random and cosine baselines in both rank correlation and label error.

1 Introduction

Human language technologies are playing an increasingly important role in the science and practice of language learning. For example, intelligent Computer Assisted Language Learning (CALL) systems are being developed that can automatically tailor lessons and questions to the needs of individual students (Heilman et al., 2006). One critical task that language tutors, word learning experiments, and related applications have in common is assessing the learning progress of the student or experiment subject during the course of the session.

When the task is learning new vocabulary, a variety of tests have been developed to measure word learning progress. Some tests, such as multiple-choice selection of a correct synonym or cloze completion, are relatively passive. In production tests, on the other hand, students are asked to write or say a short phrase or sentence that uses the word being learned, called the *target word*, in a specified way.

In one important type of production test, called a *definition production* test, the subject is asked to describe the meaning of the target word, as they understand it at that point in the session. The use of such tests has typically required a teacher or researcher to manually score each response by judging its similarity in meaning to the reference definition of the target word. The resulting scores can then be used to analyze how a person's learning of the word responded to different stimuli, such as seeing the word used in context. A sample target word and its reference definition, along with examples of human-judged responses, are given in Sections 3.3 and 4.1.

However, manual scoring of the definition responses has several drawbacks. First, it is time-consuming and must be done by trained experts. Moreover, if the researcher wanted to test a new hy-

Proceedings of NAACL HLT 2007, pages 476–483,
Rochester, NY, April 2007.

pothesis by examining the responses with respect to a different but related definition, the entire set of responses would have to be manually re-scored against the new target. Second, manual scoring can often be limited in its ability to detect when partial learning has taken place. This is due to the basic tradeoff between the sophistication of the graded scoring scale, and the ease and consistency with which human judges can use the scale. For example, it may be that the subject did not learn the complete meaning of a particular target word, but *did* learn that this target word had negative connotations. The usual binary or ternary score would provide no or little indication of such effects. Finally, because manual scoring almost always must be done off-line after the end of the session, it presents an obstacle to our goal of creating learning systems that can adapt quickly, within a single learning session.

This study describes an effective automated method for assessing word learning by scoring free responses to definition production tests. The method is flexible: it can be used to analyze a response with respect to whatever reference target(s) the teacher or researcher chooses. Such a test represents a powerful new tool for language learning research. It is also a compelling application of human language technologies research on semantic similarity, and we review related work for that area in Section 2. Our probabilistic model for computing text semantic similarity, described in Section 3, can use both corpus-based and knowledge-based resources. In Section 4 we describe a new dataset of human definition judgments and use it to measure the effectiveness of the model against other measures of text similarity. Finally, in Section 5 we discuss further directions and applications of our work.

2 Related Work

The problem of judging a subject response against a target definition is a type of text similarity problem. Moreover, it is a text *semantic* similarity task, since we require more than measuring direct word overlap between the two text fragments. For example, if the definition of the target word *ameliorate* is *to improve something* and the subject response is *make it better*, the response clearly indicates that the subject knows the meaning of the word, and thus should receive a high score, even though the response and the target definition have no words in common.

Because most responses are short (1 – 10 words) our task falls somewhere between word-word similarity and passage similarity. There is a broad field of existing work in estimating the semantic similarity of individual words. This field may be roughly divided into two groups. First, there are corpus-based measures, which use statistics or models derived from a large training collection. These require little or no human effort to construct, but are limited in the richness of the features they can reliably represent. Second, there are knowledge-based measures, which rely on specialized resources such as dictionaries, thesauri, experimental data, WordNet, and so on. Knowledge-based measures tend to be complementary to a corpus-based approach and emphasize precision in favor of recall. This is discussed further, along with a good general summary of text semantic similarity work, by (Mihalcea et al., 2006).

Because of the fundamental nature of the semantic similarity problem, there are close connections with other areas of human language technologies such as information retrieval (Salton and Lesk, 1971), text alignment in machine translation (Jayaraman and Lavie, 2005), text summarization (Mani and Maybury, 1999), and textual coherence (Foltz et al., 1998). Educational applications include automated scoring of essays, surveyed in (Valenti et al., 2003), and assessment of short-answer free-response items (Burstein et al., 1999).

As we describe in Section 3, we use a graph to model relations between words to perform a kind of *semantic smoothing* on the language models of the subject response and target definition before comparing them. Several types of relation, such as synonymy and co-occurrence, may be combined to model the interactions between terms. (Cao et al., 2005) also formulated a term dependency model combining multiple term relations in a language modeling framework, applied to information retrieval. Our graph-based approach may be viewed as a probabilistic variation on the *spreading activation* concept, originally proposed for word-word semantic similarity by (Quillian, 1967).

Finally, (Mihalcea et al., 2006) describe a text semantic similarity measure that combines word-word similarities between the passages being compared.

Due to limitations in the knowledge-based similarity measures used, semantic similarity is only estimated between words with the same part-of-speech. Our graph-based approach can relate words of different types and does not have this limitation. (Mihalcea et al., 2006) also evaluate their method in terms of paraphrase recognition using binary judgments. We view our task as somewhat different than paraphrase recognition. First, our task is not symmetric: we do not expect the target definition to be a paraphrase of the subject's free response. Second, because we seek sensitive measures of learning, we want to distinguish a range of semantic differences beyond a binary yes/no decision.

3 Statistical Text Similarity Model

We start by describing relations between pairs of terms using a general probability distribution. These pairs can then combine into a graph, which we can apply to define a semantic distance between terms.

3.1 Relations between individual words

One way to model word-to-word relationships is using a mixture of links, where each link defines a particular type of relationship. In a graph, this may be represented by a pair of nodes being joined by multiple weighted edges, with each edge corresponding to a different link type. Our link-based model is partially based on one defined by (Toutanova et al., 2004) for prepositional attachment. We allow directed edges because some relationships such as hypernyms may be asymmetric. The following are examples of different types of links.

1. **Stemming**: Two words are based on common morphology. Example: *stem* and *stemming*. We used Porter stemming (Porter, 1980).

2. **Synonyms and near-synonyms**: Two words share practically all aspects of meaning.
   Example: *quaff* and *drink*. Our synonyms came from WordNet (Miller, 1995).

3. **Co-occurrence**. Both words tend to appear together in the same contexts.
   Example: *politics* and *election*.

4. **Hyper- and hyponyms**: Relations such as "X is a kind of Y", as obtained from Wordnet or other thesaurus-like resources.
   Example: *airplane* and *transportation*.

5. **Free association**: A relation defined by the fact that a person is likely to give one word as a free-association response to the other.
   Example: *disaster* and *fear*. Our data was obtained from the Univ. of South Florida association database (Nelson et al., 1998).

We denote link functions using $\lambda_1, \ldots, \lambda_m$ to summarize different types of interactions between words. Each $\lambda_m(w_i, w_j)$ represents a specific type of lexical or semantic relation or constraint between w_i and w_j. For each link λ_m, we also define a weight γ_m that gives the strength of the relationship between w_i and w_j for that link.

Our goal is to predict the likelihood of a target definition $\mathcal{D}$ given a test response $\mathcal{R}$ consisting of terms $\{w_0 \ldots w_k\}$ drawn from a common vocabulary $\mathcal{V}$. We are thus interested in the conditional distribution $p(\mathcal{D} \mid \mathcal{R})$. We start by defining a simple model that can combine the link functions in a general purpose way to produce the conditional distribution $p(w_i|w_j)$ given arbitrary terms w_i and w_j. We use a log-linear model of the general form

$$p(w_i|w_j) = \frac{1}{Z} \exp \sum_{m=0}^{L} \gamma_m(i) \lambda_m(w_i, w_j) \quad (1)$$

In the next sections we show how to combine the estimate of individual pairs $p(w_i|w_j)$ into a larger graph of term relations, which will enable us to calculate the desired $p(\mathcal{D} \mid \mathcal{R})$.

3.2 Combining term relations using graphs

Graphs provide one rich model for representing multiple word relationships. They can be directed or undirected, and typically use nodes of words, with word labels at the vertices, and edges denoting word relationships. In this model, the dependency between two words represents a single inference step in which the label of the destination word is inferred from the source word. Multiple inference steps may then be chained together to perform longer-range inference about word relations. In this way, we can infer the similarity of two terms without requiring direct evidence for the relations between that specific pair. Using the link functions defined in Section 3.1,

we imagine a generative process where an author A creates a short text of N words as follows.

Step 0: Choose an initial word w_0 with probability $P(w_0|A)$. (If we have already generated N words, stop.)

Step i: Given we have chosen w_{i-1}, then with probability $1-\alpha$ output the word w_{i-1} and reset the process to step 0. Otherwise, with probability α, sample a new word w_i according to the distribution:

$$P(w_i|w_{i-1}) = \frac{1}{Z}\exp\sum_{m=0}^{L}\gamma_m(i)\lambda_m(w_i, w_{i-1}) \quad (2)$$

where Z is the normalization quantity.

This conditional probability may be interpreted as a mixture model in which a particular link type $\lambda_m(.)$ is chosen with probability $\gamma_m(i)$ at timestep i. Note that the mixture is allowed to change at each timestep. For simplicity, we limit the number of such changes by grouping the timesteps of the walk into three stages: early, middle and final. The function $\Gamma(i)$ defines how timestep i maps to stage s, where $s \in \{0, 1, 2\}$, and we now refer to $\gamma_m(s)$ instead of $\gamma_m(i)$.

Suppose we have a definition $\mathcal{D}$ consisting of terms $\{d_i\}$. For each link type $\lambda_m(.)$ we define a transition matrix $C(\mathcal{D}, m)$ based on the definition D. The reason $\mathcal{D}$ influences the transition matrix is that some link types, such as proximity and co-occurrence, are context-specific. Each stage s has an overall transition matrix $C(\mathcal{D}, s)$ as the mixture of the individual $C(\mathcal{D}, m)$, as follows.

$$C(\mathcal{D}, s) = \sum_{m=1}^{M}\gamma_m(s)C(\mathcal{D}, m) \quad (3)$$

Combining the stages over k steps into a single transition matrix, which we denote C^k, we have

$$C^k = \prod_{i=0}^{k} C(\mathcal{D}, \Gamma(i)) \quad (4)$$

We denote the (i, j) entry of a matrix A^k by $A^k_{i,j}$. Then for a particular term d_i, the probability that a chain reaches d_i after k steps, starting at word w is

$$P_k(d_i|w) = (1-\alpha)\alpha^k C^k_{w,d_i} \quad (5)$$

where we identify w and d_i with their corresponding indices into the vocabulary $\mathcal{V}$. The overall probability $p(d_i|w)$ of generating a definition term d_i given a word w is therefore

$$P(d_i|w) = \sum_{k=0}^{\infty} P_k(d_i|w) = (1-\alpha)(\sum_{k=0}^{\infty}\alpha^k C^k)_{w,d_i} \quad (6)$$

The walk continuation probability α can be viewed as a penalty for long chains of inference. In practice, to perform the random walk steps we replace the infinite sum of Eq. 6 with a small number of steps (up to 5) on a sparse representation of the adjacency graph. We obtained effective link weights $\gamma_m(i)$ empirically using held-out data. For simplicity we assume that the same α is used across all link types, but further improvement may be possible by extending the model to use link-specific decays α_m. Fine-tuning these parameter estimation methods is a subject of future work.

3.3 Using the model for definition scoring

In our study the reference definition for the target word consisted of the target word, a rare synonym, a more frequent synonym, and a short glossary-like definition phrase. For example, the reference definition for *abscond* was

abscond; absquatulate; escape; to leave quickly and secretly and hide oneself, often to avoid arrest or prosecution.

In general, we define the score of a response $\mathcal{R}$ with respect to a definition $\mathcal{D}$ as the probability that the definition is generated by the response, or $p(\mathcal{D}|\mathcal{R})$. Equivalently, we can score by $\log p(\mathcal{D}|\mathcal{R})$ since the $\log$ function is monotonic. So making the simplifying assumption that the terms $d_i \in D$ are exchangable (the bag-of-words assumption), and taking logarithms, we have:

$$\begin{aligned}\log p(\mathcal{D}|\mathcal{R}) &= \log\prod_{d_i\in\mathcal{D}} p(d_i|\mathcal{R}) \\ &= \sum_{d_i\in\mathcal{D}}\log[(1-\alpha)(\sum_{k=0}^{m}\alpha^k C^k)_{\mathcal{R},d_i}]\end{aligned} \quad (7)$$

Suppose that the response to be scored is *run from the cops*. In practical terms, Eq. 7 means that for our

example, we "light up" the nodes in the graph corresponding to *run, from, the* and *cops* by assigning some initial probability, and the graph is then "run" using the transition matrix C according to Eq. 7. In this study, the initial node probabilities are set to values proportional to the *idf* values of the corresponding term, so that $P(d_i) = \frac{idf(d_i)}{\sum idf(d_i)}$. After m steps, the probabilities at the nodes for each term in the reference definition $\mathcal{R}$ are read off, and their logarithms summed. Similar to an AND calculation, we calculate a product of sums over the graph, so that responses reflecting multiple aspects of the target definition are rewarded more highly than a very strong prediction for only a single definition term.

4 Evaluation

We first describe our corpus of gold standard human judgments. We then explain the different text similarity methods and baselines we computed on the corpus responses. Finally, we give an analysis and discussion of the results.

4.1 Corpus

We obtained a set of 734 responses to definition production tests from a word learning experiment at the University of Pittsburgh (Bolger et al., 2006). In total, 72 target words, selected by the same group, were used in the experiment. In this experiment, subjects were asked to learn the meaning of target words after seeing them used in a series of context sentences. We set aside 70 responses for training, leaving 664 responses in the final test dataset.

Each response instance was coded using the scale shown in Table 1, and a sample set of subject responses and scores is shown in Table 2. The target word was treated as having several key aspects of meaning. The coders were instructed to judge a response according to how well it covered the various aspects of the target definition. If the response covered all aspects of the target definition, but also included extra irrelevant information, this was treated as a partial match at the discretion of the coders.

We obtained three codings of the final dataset. The first two codings were obtained using an independent group, the QDAP Center at the University of Pittsburgh. Initially, five human coders, with varying degrees of general coding experience, were trained by the authors using one set of 10 example instances and two training sessions of 30 instances each. Between the two training sessions, one of the authors met with the coders to discuss the ratings and refine the rating guidelines. After training, the authors selected the two coders who had the best inter-coder agreement on the 60 training instances. These two coders then labeled the final test set of 664 instances. Our third coding was obtained from an initial coding created by an expert in the University of Pittsburgh Psychology department and then adjusted by one of the authors to resolve a small number of internal inconsistencies, such as when the same response to the same target had been given a different score.

Score	Meaning
0	Completely wrong
1	Some partial aspect is correct
2	One major aspect, or more than one minor aspect, is correct
3	Covers all aspects correctly

Table 1: Scale for human definition judgements.

Response	Human Score
depart secretly	3
quietly make away, escape	3
to flee, run away	2
flee	2
to get away with	1
to steal or take	0

Table 2: Examples of human scores of responses for the target word *abscond*.

Inter-coder agreement was measured using linear weighted kappa, a standard technique for ordinal scales. Weighted kappa scores for all three coder pairs are shown in Table 3. Overall, agreement ranged from moderate (0.64) to good (0.72).

4.2 Baseline Methods

We computed three baselines as reference points for lower and upper performance bounds.

Random. The response items were assigned labels randomly.

Coder pair	Weighted Kappa
1, 2	0.68
2, 3	0.64
1, 3	0.72

Table 3: Weighted kappa inter-rater reliability for three human coders on our definition response dataset (664 items).

Method	Spearman Rank Correlation
Random	0.3661
Cosine	0.4731
LSA	0.4868
Markov	0.6111
LSA + Markov	0.6365
Human	0.8744

Table 4: Ability of methods to match human ranking of responses, as measured by Spearman rank correlation (corrected for ties).

Human choice of label. We include a method that, given an item and a human label from one of the coders, simply returns a label of the same item from a different coder, with results repeated and averaged over all coders. This gives an indication of an upper bound based on human performance.

Cosine similarity using tf.idf weighting. Cosine similarity is a widely-used text similarity method for tasks where the passages being compared often have significant direct word overlap. We represent response items and reference definitions as vectors of terms using *tf.idf* weighting, a standard technique from information retrieval (Salton and Buckley, 1997) that combines term frequency (*tf*) with term specificity (*idf*). A good summary of arguments for using *idf* can be found in (Robertson, 2004). To compute *idf*, we used frequencies from a standard 100-million-word corpus of written and spoken English [1]. We included a minimal semantic similarity component by applying Porter stemming (Porter, 1980) on terms.

[1]The British National Corpus (Burnage and Dunlop, 1992), using American spelling conversion.

4.3 Methods

In addition to the baseline methods, we also ran the following three algorithms over the responses.

Markov chains ("Markov"). This is the method described in Section 3. A maximum of 5 random walk steps were used, with a walk continuation probability of 0.8. Each walk step used a mixture of synonym, stem, co-occurrence, and free-association links. The link weights were trained on a small set of held-out data.

Latent Semantic Analysis (LSA). LSA (Landauer et al., 1998) is a corpus-based unsupervised technique that uses dimensionality reduction to cluster terms according to multi-order co-occurrence relations. In these experiments, we obtained LSA-based similarity scores between responses and target definitions using the software running on the University of Colorado LSA Web site (LSA site, 2006). We used the pairwise text passage comparison facility, using the maximum 300 latent factors and a general English corpus (Grade 1 – first-year college).

Although LSA and the Markov chain approach are based on different principles, we chose to apply LSA to this new response-scoring task and corpus because LSA has been widely used as a text semantic similarity measure for other tasks and shown good performance (Foltz et al., 1998).

LSA+Markov. To test the effectiveness of combining two different – and possibly complementary – approaches to response scoring, we created a normalized, weighted linear combination of the LSA and Markov scores, with the model combination weight being derived from cross-validation on a held-out dataset.

4.4 Results

We measured the effectiveness of each scoring method from two perspectives: ranking quality, and label accuracy.

First, we measured how well each scoring method was able to rank response items by similarity to the target definition. To do this, we calculated the Spearman Rank Correlation (corrected for ties) between the ranking based on the scoring method and the ranking based on the human-assigned scores, averaged over all sets of target word responses.

Table 4 summarizes the ranking results. For

Method	Label error (RMS)	
	Top 1	Top 3
Random	1.4954	1.6643
Cosine	0.8194	1.0540
LSA	0.8009	0.9965
Markov	0.7222	0.7968
LSA + Markov	1.1111	1.0650
Human	0.1944	0.4167

Table 5: Root mean squared error (RMSE) of label(s) for top-ranked item, and top-three items for all 77 words in the dataset.

overall quality of ranking, the Markov method had significantly better performance than the other automated methods ($p < 2.38e^{-5}$). LSA gave a small, but not significant, improvement in overall rank quality over the cosine baseline. [2] The simple combination of LSA and Markov resulted in a slightly higher but statistically insignificant difference ($p < 0.253$).

Second, we examined the ability of each method to find the most accurate responses – that is, the responses with the highest human label on average – for a given target word. To do this, we calculated the Root Mean Squared Error (RMSE) of the label assigned to the top item, and the top three items. The results are shown in Table 5. For top-item detection, our Markov model had the lowest RMS error (0.7222) of the automated methods, but the differences from Cosine and LSA were not statistically significant, while differences for all three from Random and Human baselines were significant. For the top three items, the difference between Markov (0.7968) and LSA (0.9965) was significant at the $p < 0.03$ level.

Comparing the overall rank accuracy with top-item accuracy, the combined LSA + Markov method was significantly worse at finding the three best-quality responses (RMSE of 1.0650) than Markov (0.7968) or LSA (0.9965) alone. The reasons for this require further study.

[2] All statistical significance results reported here used the Wilcoxon Signed-Ranks test.

5 Discussion

Even though definition scoring may seem more straightforward than other automated learning assessment problems, human performance was still significantly above the best automated methods in our study, for both ranking and label accuracy. There are certain additions to our model which seem likely to result in further improvement.

One of the most important is the ability to identify phrases or colloquial expressions. Given the short length of a response, these seem critical to handle properly. For example, *to get away with something* is commonly understood to mean *secretly guilty*, not a physical action. Yet the near-identical phrase *to get away from something* means something very different when phrases and idioms are considered.

Despite the gap between human and automated performance, the current level of accuracy of the Markov chain approach has already led to some promising early results in word learning research. For example, in a separate study of incremental word learning (Frishkoff et al., 2006), we used our measure to track increments in word knowledge across multiple trials. Each trial consisted of a single passage that was either *supportive* – containing clues to the meaning of unfamiliar words – or not supportive. In this separate study, broad learning effects identified by our measure were consistent with effects found using manually-scored pre- and post-tests. Our automated method also revealed a previously unknown interaction between trial spacing, the proportion of supportive contexts per word, and reader skill.

In future applications, we envision using our automated measure to allow a form of feedback for intelligent language tutors, so that the system can automatically adapt its behavior based on the student's test responses. With some adjustments, the same scoring model described in this study may also be applied to the problem of finding supportive contexts for students.

6 Conclusions

We presented results for both automated and human performance of an important task for language learning applications: scoring definition responses. We described a probabilistic model of text seman-

tic similarity that uses Markov chains on a graph of term relations to perform a kind of semantic smoothing. This model incorporated both corpus-based and knowledge-based resources to compute text semantic similarity. We measured the effectiveness of both our method and LSA compared to cosine and random baselines, using a new corpus of human judgments on definition responses from a language learning experiment. Our method outperformed the *tf.idf* cosine similarity baseline in ranking quality and in ability to find high-scoring definitions. Because LSA and our Markov chain method are based on different approaches and resources, it is difficult to draw definitive conclusions about performance differences between the two methods.

Looking beyond definition scoring, we believe automated methods for assessing word learning have great potential as a new scientific tool for language learning researchers, and as a key component of intelligent tutoring systems that can adapt to students.

Acknowledgements

We thank D.J. Bolger and C. Perfetti for the use of their definition response data, Stuart Shulman for his guidance of the human coding effort, and the anonymous reviewers for their comments. This work supported by U.S. Dept. of Education grant R305G03123. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

D.J. Bolger, M. Balass, E. Landen and C.A. Perfetti. 2006. Contextual Variation and Definitions in Learning the Meanings of Words. (In press.)

G. Burnage and D. Dunlop. 1992. Encoding the British National Corpus. *English Language Corpora: Design, Analysis and Exploitation*. The 13th Intl. Conf. on Engl. Lang. Res. in Computerized Corpora. Nijmegen. J. Aarts, P. de Haan, N. Oostdijk, Eds.

J. Burstein, S. Wolff, and L. Chi. 1999. Using Lexical Semantic Techniques to Classify Free-Responses. *Breadth and Depth of Semantic Lexicons*. Kluwer Acad. Press, p. 1–18.

G. Cao, J-Y. Nie, and J. Bai. Integrating Word Relationships into Language Models. *SIGIR 2005*, 298–305.

P.W. Foltz, W. Kintsch, and T. Landauer. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2):285–307.

G. Frishkoff, K. Collins-Thompson, J. Callan, and C. Perfetti. 2007. The Nature of Incremental Word Learning: Context Quality, Spacing Effects, and Skill Differences in Meaning Acquisition Across Multiple Contexts. (In preparation.)

M. Heilman, K. Collins-Thompson, J. Callan and M. Eskanazi. 2006. Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension. *ICSLP 2006*.

S. Jayaraman and A. Lavie. Multi-Engine Machine Translation Guided by Explicit Word Matching. *EAMT 2005*.

T.K. Landauer, P.W. Foltz, and D. Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

LSA Web Site. http://lsa.colorado.edu

I. Mani and M.T. Maybury (Eds.) 1999. Advances in Automatic Text Summarization. MIT Press.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *AAAI 2006*

G. Miller. 1998. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11) 39–41.

D.L. Nelson, C.L. McEvoy, and T.A. Schreiber. 1998. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

M. Porter. 1980. An Algorithm for Suffix-stripping. *Program*, 14(3) 130–137. http://www.tartarus.org/martin/PorterStemmer

M. Quillian. 1967. Word Concepts: A Theory and Simulation of some Basic Semantic Capabilities. *Behav. Sci.*, 12: 410–430.

S. Robertson. 2004. Understanding Inverse Document Frequency: on Theoretical Arguments for IDF. *J. of Documentation*, 60:503–520.

G. Salton and C. Buckley. 1997. Term Weighting Approaches in Automatic Text Retrieval. *Reading in Information Retrieval*. Morgan Kaufmann.

G. Salton and M. Lesk. 1971. *Computer Evaluation of Indexing and Text Processing*. Prentice-Hall. 143 – 180.

K. Toutanova, C.D. Manning, and A.Y. Ng. 2004. Learning Random Walk Models for Inducing Word Dependency Distributions. *ICML 2004*.

S. Valenti, F. Neri, and A. Cucchiarelli. 2003. An Overview of Current Research on Automated Essay Grading. *J. of Info. Tech. Ed.*, Vol. 2.