Experiments with Semantic Similarity Measures based on LDA and LSA

NOBAL NIRAULA, RAJENDRA BANJADE, DAN STEFANESCU, VASILE RUS
DEPARTMENT OF COMPUTER SCIENCE
THE UNIVERSITY OF MEMPHIS
{rbanjade,nbnraula,dstfnscu,vrus}@memphis.edu

Abstract. We present in this paper experiments with several semantic similarity measures based on the unsupervised method Latent Dirichlet Allocation. For comparison purposes, we also report experimental results using an algebraic method, Latent Semantic Analysis. The proposed semantic similarity methods were evaluated using one dataset that includes student answers from conversational intelligent tutoring systems and a standard paraphrase dataset, the Microsoft Research Paraphrase corpus. Results indicate that the method based on word representations as topic vectors outperforms methods based on distributions over topics and words. The proposed evaluation methods can also be regarded as an extrinsic method for evaluating topic coherence or selecting the number of topics in LDA models, i.e. a task-based evaluation of topic coherence and selection of number of topics in LDA.

Keywords: semantic similarity, statistical methods, Latent Dirichlet Allocation

1 Introduction

We address in this paper the important task of finding how semantically similar two texts are. We employ a novel set of semantic similarity methods that rely on the probabilistic method Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordnan, 2003).

Semantic similarity is a widely used approach to the core problem of language understanding. It is an useful alternative to the true understanding approach which is intractable as it requires world knowledge. For instance, in dialogue-based Intelligent Tutoring Systems (ITS) it is important to understand students' natural language responses. One frequently used approach to address this issue is to compute how similar student responses are to benchmark, expert-articulated responses (Graesser, Olney,

adfa, p. 1, 2011.

Haynes, Chipman, 2005; Rus & Graesser, 2006). That is, the student response assessment task is being modeled as a text-to-text similarity problem.

Below, we show an example of a real student response from an ITS and the corresponding benchmark answer authored by an expert.

Student Response*: An object that has a zero force acting on it will have zero acceleration.*

Expert Answer: *If an object moves with a constant velocity, the net force on the object is zero.*

The student response above is deemed correct as it is semantically similar to the expert answer. A student response is deemed incorrect if it is not similar to the expert response. More nuanced categorizations are possible, e.g. a student response can be partially correct.

In this paper, we model the problem of semantic similarity as a binary decision problem in which a student response is deemed either correct or incorrect. We limit ourselves to such binary judgments because the primary scope of this work is to assess the novel semantic similarity methods based on the unsupervised method Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordnan, 2003). We plan to address more nuanced judgments of semantic similarity in the future. Also, the datasets that we used to evaluate the proposed methods only provide binary judgments.

It should be noted that this type of binary modeling has been extensively used in previously proposed semantic similarity tasks such as the Recognizing Textual Entailment task (Dagan, Glickman, and Magnini, 2004), the paraphrase identification task (Dolan, Quirk, & Brockett, 2004), or the student input assessment task (Rus & Graesser, 2006; McCarthy & McNamara, 2008).

The task of semantic similarity can be formulated at different levels of granularity ranging from word-to-word similarity to sentence-to-sentence similarity to document-to-document similarity or a combination of these such as word-to-sentence or sentence-to-document similarity. We propose in this paper novel solutions to the task of semantic similarity both at word and sentence level with an emphasis on sentence-level similarity. In particular, we rely on one probabilistic method, LDA (Blei, Ng, & Jordan, 2003), that automatically discovers a set of underlying topics, represented as distributions over words, in texts. That is, texts are regarded as distribution over topics. Words can be represented as a vector of contributions to topics in an LDA model.

The semantic similarity measures of larger texts, e.g. sentences, can be defined based on either individual word representations, e.g. by extending word-to-word similarity measures to sentence-to-sentence similarity (as in Lintean et al., 2010), or based on the representations of texts as distributions over topics (topics are distributions over words in the vocabulary). We propose here solutions based on both of these approaches. The LDA-based word-to-word semantic similarity measures are used in conjunction with greedy and optimal matching methods in order to measure similarity between larger texts such as sentences. The solutions based on the second approach, called text-to-text measures, are used directly to compute the similarity of two sentences.

For comparison purposes, we also report experimental results using an algebraic method, Latent Semantic Analysis (LSA, Landauer et al., 2007), that automatically derives meaning representations in the form of latent concepts. Like LDA, LSA is fully automated. Words are represented as vectors in an LSA-derived semantic space. The dimensions of this space are latent concepts. Similarity of individual words and texts are computed based on vector algebra. LDA has one conceptual advantage over LSA: LDA represents multiple meanings of a word explicitly while LSA does not.

We have experimented with a dataset compiled from dialogue-based intelligent tutoring systems as well as with the Microsoft Research Paraphrase corpus (Dolan, Quirk, & Brockett, 2004).

The rest of the paper is organized as in the followings. The next section provides an overview of related work. Then, we describe LDA and the semantic similarity measures based on LDA. The Experiments and Results section describes our experimental setup and the results obtained. We conclude the paper with Discussion and Conclusions.

2 Previous Work

The task of semantic similarity between two short texts, namely two sentences, has been addressed using various solutions that range from simple word overlap to greedy methods that rely on word-to-word similarity measures (Fernando & Stevenson, 2008;) to algebraic methods (Lintean, Moldovan, Rus, & McNamara, 2010) to machine learning based solutions (Kozareva & Montoyo, 2006).

The most relevant work to ours is by Lintean et al. (2010) who looked at the role of LSA (Landauer et al., 2007) in solving the paraphrase identification task. As already mentioned, LSA is a vectorial representation in which a word is represented as a vector in a low dimensionality space (300-500 dimensions or latent concepts; we use 300 dimensions in our experiments reported here). Computing the similarity between two words is equivalent to computing the cosine, i.e. the normalized dot product, between the corresponding LSA vectors.

Lintean et al. (2010) used LSA as a way to compute semantic similarity in two different ways. First, they used LSA to compute a word-to-word similarity measure which they combined with a greedy-matching method to obtain a sentence level similarity score. For instance, each word in one sentence was greedily paired with one word in the other sentence. An average of these word-to-word similarities was then assigned as the semantic similarity score of the two sentences. Second, LSA was used to directly compute the similarity of two sentences by applying the cosine (normalized dot product) of the LSA vectors of the sentences. The LSA vector of a sentence was computed by adding all the individual word vectors. We present results with these methods and, additionally, with a method based on optimal matching that only uses word-to-word LSA similarity.

LDA itself was occasionally used for computing the semantic similarity of texts. The closest use of LDA for a semantic similarity task was by Celikyilmaz, Hakkani-Tur, & Tur (2010) for ranking candidate answers to a question in Question Answering (QA). Given a question, they ranked candidate answers based on how similar these answers were to the target question. That is, for each question-answer pair they generated an LDA model which then they used to compute a degree of similarity (DES) that consists of the product of two measures: sim1 and sim2. Sim1 captures the word-level similarities of the topics present in an answer and the question. Sim2 measures the similarities between the topic distributions in an answer and the question. The LDA model was generated based solely on each question and candidate answers. As opposed to our task, in which we compute the similarity between sentences, the candidate answers in Celikyilmaz, Hakkani-Tur, & Tur (2010) are longer, consisting of more than one sentence. This particular difference is important when it comes to computing the semantic similarity based on LDA as the shorter the texts the sparser the distributions, in particular the distribution over topics, based on which the similarity is computed.

Similar to Celikyilmaz, Hakkani-Tur, & Tur (2010), we define several semantic similarity measures based on the topic and word distributions in LDA. We do use Information Radius as Celikyilmaz, Hakkani-Tur, & Tur (2010) and, in addition, propose similarity measures based on Hellinger and Manhattan distances.

Another use of LDA for computing similarity between texts, namely blogs, relied on a very simple measure of computing the dot product of topic vectors as opposed to a similarity based on distributions (Chen et al., 2012). Because using such topic vectors for short texts leads to very sparse topic vectors, we did not experiment and do not report results with similarity methods based on just topic vectors.

The work presented here extends our previous work on LDA-based semantic similarity (Rus, Niraula, & Banjade, 2013). To the best of our knowledge, LDA has not been used so far for addressing the task of paraphrase identification in the context of student responses in dialogue-based ITSs, which is the focus of our work.

3 LDA-based Similarity Measures

Latent Dirichlet Allocation (LDA; Blei, Ng,& Jordan, 2003) belongs to the broader category of methods called topic models. Topic models are based on the assumption that a relatively small set of latent topics underly natural language texts. The topics are groups of semantically related words. A word can belong to multiple topics. If one interprets each topic as a concept then LDA directly models polysemy which LSA does not. In LSA, each word has a unique vector representation. That is, multiple senses of the same word are mapped to the same representation in the reduced LSA space. Some argue that the LSA vector for a given word represents an average of all the senses of the word, while others argue that it represents the dominant, most frequent sense. Given this theoretical advantage of LDA over LSA when it comes to modeling word meanings, one wonders which one is better at tasks in which word meanings play a role such as sentence-level text-to-text similarity. This paper is a step towards understanding the strengths of LDA versus LSA.

It is important to add that LDA has been proposed to address several limitations of the earlier Probabilistic Latent Semantic Indexing model (pLSI; Hoffman, 1999). For instance, the pLSI model cannot handle unseen documents. Also, the number of parameters to be estimated in the pLSI models increases linearly with the number of documents leading to overfitting.

3.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model for collections of discrete items, i.e. words in our case. The only observed things are the words (denoted by *w*) in documents. All else are latent variables. LDA derives the parameters of the latent variables using only the observed words in the corpus. Thus, LDA captures significant intra-document statistical structure via mixing distributions.

We will use the notation in Blei, Ng, and Jordan (2003) to explain the basic LDA model. A word, denoted *w*, is a discrete unit entry in a vocabulary V whose elements are indexed {1,…,V}. A document is a sequence of N words denoted $\mathbf{w} = <w_1, w_2, \ldots, w_N,>$, where w_i is the *i*-th word in the document. A corpus D is a collection of documents $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

Documents are regarded as random mixtures of topics and a topic is a distribution over words in the vocabulary. LDA follows the following generative process for a document **w**.

(a) Choose a topic distribution $\theta \sim \mathrm{Dir}\ (\alpha)$; the dimensionality *k* (number of topics) of the Dirichlet distribution is given;
(b) For each of the N words w_i in **w:**
  (i) Select a topic z_i based on θ;
  (ii) Choose a word w_i using $p(w_i| z_i,\beta)$

LDA has two Dirichlet priors: α for document-topic distributions and β for topic-word distributions. These two priors, α and β, are also known as hyper-parameters for the document-topic and topic-word Dirichlet distributions. Although they can be vector valued, many LDA implementations use α and β as scalars to simplify and get symmetric Dirichlet priors. Currently most LDA users choose symmetric Dirichlet priors using some heuristics. One such heuristics is mentioned by Steyvers and Griffiths (2006): although the values of these priors depend on vocabulary size and the number of topics, setting $\alpha = 50/k$ and $\beta = 0.01$ worked well for many different text collections. We followed this latter approach in our work presented here.

LDA estimation includes learning the various distributions, e.g., the set of topics, the word probabilities for each topic, the topic mixture proportion of each document, and the topic of each word in each document. Estimation of the LDA parameters directly and exactly maximizing the likelihood of the whole data collection is intractable. Approximate estimation methods are used to solve the problem. The three popular methods reported in the literature are: variational methods (Blei et al, 2003), expectation propagation (Griffiths and Steyvers, 2004), and Gibbs sampling (Griffiths and Steyvers, 2004). We used in our work an implementation based on Gibbs sampling (i.e., JGibbLDA).

3.2 Number of Topics

The standard LDA model requires the specification of the number of latent topics in advance. That is, the number of topics is set by the user. Choosing the right number of topics is important as they determine the quality of the LDA model. Many believe that choosing the right value for the number of topics is more art than science.

One solution is to try a range of values and choosing the best number of topics according to some intrinsic criterion, such as the coherence of the topics, or according to some extrinsic criterion such as accuracy on a task, .e.g. paraphrase identification. We use in this paper as a starting point the topic coherence for selecting the number of topics (see next subsection). Furthermore, our experiments with using LDA for the task of paraphrase identification can be viewed as an extrinsic, task-based selection or validation of the number of topics.

Other methods to select the number of topics exist. Some rely on heuristics for selecting the number of topics. Nonparametric Bayesian models such as Hierarchical Dirichlet process were also proposed to automatically estimate the number of topics (Teh et al., 2004). The nonparametric models are not computationally efficient (Wallach et al., 2009).

3.3 Assessment of Topics

As mentioned, we used topic coherence as an intrinsic criterion to select the number of topics upfront. Newman et al. (2010) have explored techniques for measuring topic coherence and presented a comparative study of topic coherence evaluation using Wikipedia, Google n-gram dataset, and WordNet. The pointwise mutual information (PMI) method was best when compared to human judgments of topic coherence. They counted the frequency of the co-occurring words in a window of 10-word in Wikipedia corpus and 5 in case of Google 5-grams.

Similarly, we used the average PMI of the top 10 and also top 20 words to assess the quality of topic coherence. That is, we formed all possible pairs with the top 10 or 20 words in each topic (words in each topic are decreasingly ordered based on their contribution to the topic) and computed the PMI for each pair based on word frequencies derived from a Wikipedia-based corpus.

The PMI was calculated using 4,134,837 English-language Wikipedia articles dumped on January 3, 2013. It contained 1,284,156,826 tokens and 5,693,208 word types (i.e. unique words) counted after removing digits and punctuation and changing to lower case. After removing the stop words, the number of tokens was 672,542,579. We found that a 100-topic LDA model leads to highest average topic coherence (we varied the number of topics from 10 to 300, the typical dimensionality used in LSA spaces). Experimental results on the paraphrase identification task, which can be viewed as an extrinsic, task-based evaluation of topic coherence, confirmed that using k=100 topics is best. Given that measuring topic coherence based on the average

PMI of top 10 words recommends the same best number of topics as our task-based evaluation further supports the use of top 10 words PMI for measuring topic coherence (as suggested by Newman et al., 2010). The best coherence when using top 20 words is for a 20-topic LDA model. However, the average PMI for the 20 topics model is not significantly different from the 100 topics model.

In a way, our extrinsic, task-based validation of the number of topics is stronger than the validation based on human judgments provided by Newman and colleagues (2010) as they asked human judges to assess only a subset of the topics. Furthermore, it is not clear whether they asked the human judges to consider only top 10 words from each topic or not. They used the top 10 words only when computing the PMI.

3.4 LDA-based Semantic Similarity Measures

As we already mentioned, LDA is a probabilistic generative model in which documents are viewed as distributions over a set of topics and each word in a document is generated based on a distribution over words that is specific to each topic.

A first semantic similarity measure among words would then be defined as a dot-product between the corresponding vectors representing the contributions of each word to a topic. It should be noted that the contributions of each word to the topics does not constitute a distribution, i.e. the sum of contributions does not add up to 1. Assuming the number of topics T, then a simple word-to-word measure is defined by the formula below where we denote by *φ* distributions over words for a topic *t*.

$$LDA - w2w(w,v) = \sum_{t=1}^{T} \varphi_t\,(w)\varphi_t\,(v)$$

More global text-to-text similarity measures could be defined in several ways. Because a document is a distribution over topics, the similarity of two texts needs to be computed in terms of similarity of distributions. The Kullback-Leibler (KL) divergence defines a distance, or how dissimilar, two distributions p and q are as in the formula below.

$$KL(p,q) = \sum_{i=1}^{T} p_i \log \frac{p_i}{q_i}$$

If we replace p with θ_d (text/document d’s distribution over topics) and q with θ_c (text/document c’s distribution over topics) we obtain the KL distance between two documents (documents d and c in our example).

Furthermore, KL can be used to compute the distance between two topics using their distributions over words (φ_{t1} and φ_{t2}). The KL distance has two major problems. In case q_i is zero KL is not defined. Furthermore, KL is not symmetric which does not fit well with semantic similarity measures which in general are symmetric. That is, if text A is a paraphrase of text B that text B is a paraphrase of text A. The Information Radius (IR) measure solves these problems by considering the average of pi and qi as below.

The IR can be transformed into a similarity measure as in the following (Dagan, Lee, & Pereira, 1997):

$$SIM(p,q) = 10^{-\delta IR(c,d)}$$

All our results reported here for LDA similarity measures between two documents c and d are computed by multiplying the similarities between the distribution over topics (θ_d and θ_c) and distribution over words (φ_{t1} and φ_{t2}). For space reasons, we do not provide all the details.

The Hellinger distance between two distributions is another option that allows avoiding the shortcomings of the KL distance.

$$HD(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{1}^{T}(\sqrt{p_i} - \sqrt{q_i})^2}$$

The Hellinger distance varies from 0 to 1 and is defined for all values of p_i and q_i. A value of 1 means the distance is maximum and thus the distributions are very different. A value of 0 means the distributions are very similar. We can transform the Hellinger distance into a similarity measure by subtracting it from 1 such that a zero distance means a large similarity score and vice versa.

Lastly, we used the Manhattan distance between distributions *p* and *q* as defined below.

$$MD(p,q) = 2 \times (1 - \sum_{1}^{T} \min(p_i, q_i))$$

MD is symmetric, defined for any values of p and q, and ranges between 0 and 2. We can divide MD by 2 and subtract from 1 to transform it into a similarity measure.

4 From Word Representations to Text-to-Text Similarity

As mentioned, we focus in this paper on two categories of methods: those that rely on word-to-word similarity measures and those that compute similarity globally, i.e. avoiding word-to-word similarities. In LSA, text-to-text similarity can be computed directly using the global vectors of each sentence which are obtained by summing up the individual word vectors. In LDA, global text-to-text similarity measures can be computed using the distributions over topics and over words without the need for word-to-word similarity measures.

Word-to-word similarity measures can be expanded to work at text-to-text level using greedy (see Lintean et al., 2010) or optimal matching algorithms (Rus & Lintean, 2012). We experimented with a method that guarantees optimal overall best match using the job assignment algorithm, a well-known combinatorial optimization problem. The assignment problem can be formulated as finding a permutation π for which

$S_{OPT} = \sum_{i=1}^{n} w(s_i, t_{\pi(i)})$ is maximum where w(si,tπ(i)) is the fitness of worker si to job ti. Such an assignment is called optimum assignment. An algorithm, the Kuhn-Munkres method (Kuhn, 1955; Munkres, 1957), has been proposed that can find a solution to it in polynomial time.

In our case, we optimally match words in text T1 to words in text T2 based on how well the words in T1 fit the words in T2. The fitness between the words is nothing else but their word-to-word similarity according to some metric of word similarity, in our case LDA or LSA-based word-to-word measures.

5 Experimental Setup and Results

We present results with the previously described methods on the User Language Paraphrase Corpus (ULPC; McCarthy and McNamara, 2008) and additionally on the Microsoft Research Paraphrase corpus (MSRP; Dolan, Quirk, & Brockett, 2004). The ULPC corpus contains pairs of target-sentence and student response texts. These pairs have been evaluated by expert human raters along 10 dimensions of paraphrase characteristics. We used the ”Semantic Completeness” dimension that measures the semantic equivalence between the target-sentence and the student response on a binary scale, similar to the scale used in MSRP corpus. From a total of 1,998 pairs, 1,436 (71%) were classified by experts as being paraphrases. The data set is divided into three subsets: training (1,012 instances, 708-304 split of TRUE-FALSE paraphrases), validation (649 instances, 454-195 split), and testing (337 instances, 228-109 split). The average number of words per sentence is 15.

The MSRP corpus consists of 5,801 sentence pairs collected from newswire articles, 3,900 of which were labeled as paraphrases by human annotators. The whole set is divided into a training subset (4,076 sentences of which 2,753, or 67.5%, are true paraphrases), and a test subset (1,725 pairs of which 1,147, or 66.5%, are true paraphrases). A simple baseline for the MSRP corpus, the majority baseline when all instances are classified as positive, gives an accuracy and precision of 66.5% and perfect recall. The average number of words per sentence is 17 in this corpus.

We followed a training-testing methodology according to which we first trained to learn some parameters of the proposed model after which we used the learned values for the parameters on testing data. In our case, we learned a threshold for the text-to-text similarity score above which a pair of sentences is deemed a paraphrase and any score below the threshold means the sentences are not paraphrases. We report performance of the various methods using accuracy (percentage of correct predictions), F-measure (harmonic mean of precision and recall), and kappa statistics (a measure of agreement between our method’s output and experts’ labels while accounting for chance agreement).

We experimented with both word-to-word similarity measures and text-to-text similarity measures. The word-to-word similarity measures were expanded to work at sentence level using optimal matching. For LDA, we used the word-to-word measure and text-to-text measures described earlier. For LSA, we use the cosine between two words’ LSA vectors as a measure of word-to-word similarity. For LSA-based text-to-text similarity we first add up the word vectors for all the words in a text thus obtaining two text vectors, one for each text, and then compute the cosine between these two text vectors.

An important step in the process of obtaining the LSA vectorial representation is the derivation of the semantic space, i.e. discovering the latent dimensions or concepts, from a large enough corpus. In our work, we experimented with an LSA space computed from the TASA corpus (compiled by Touchstone Applied Science Associates), a balanced collection of representative texts from various genres (science, language arts, health, economics, social studies, business, and others). The TASA corpus contains 10,937,986 words with a vocabulary size of 91,897 after removing stop words.

We varied the number of topics for the LDA model and observed changes in performance. Fewer topics usually means semantically less coherent topics as more words with different meaning will be grouped under the same topic. Our experiments revealed that using just top 10 or 20 words for measuring topic coherence indicates the opposite: the fewer the topics the higher their semantic coherence, e.g. topics sets of size 100, 40, or 20, all have higher average topic coherence scores compared to 200- or 300-topic models. We concluded that the 100, 40, and 20 models yield results similar to higher 200 and 300 topics models. That is, using 100 topics models could be a good choice that balances a sufficiently large number of topics and good topic coherence when addressing sentence-level semantic similarity tasks such as paraphrase identification.

We also present results obtained using 300 dimensions for the LSA space, a standard value, and a similar number of topics for LDA (see column T=300 in Table 1). This number of dimensions has been empirically established by LSA researchers to deliver best results. We also present results for 100 dimensions to compare with the best LDA model which corresponds to 100 topics.

The results in Table 1 indicate that the best LDA-based methods rival the LSA based method. A combination of greedy matching and LDA word-to-word similarity yields best accuracy and F-measure results on the ULPC corpus while text-to-text similarity based on LSA yields best accuracy. The 100-topic LDA model produces similar accuracy results on ULPC and a higher kappa (kappa=37.89 for 100-topic model and kappa=34.40 for the 300-topic model).

Similarly, for the MSRP corpus the LDA models produce results very close to LSA. The 100-topic LDA model has a slightly better kappa score compared to the 300-topic model. The 100-topic models yields very similar accuracy score to the 300-topic model, and an identical F-measure score.

All the distance-between-distributions based LDA measures (top 3 rows in Table 1) yield modest results. This is mainly due to the sparsity of topic distributions in short texts compared to the size of the model in terms of number of topics. If a 100-topic model is used and the sentence has on average 15 words, in the best case scenario in which each word in the sentence corresponds to a unique topic, 85 of the remaining topics in the 100-topic model would have a probability of zero. This leads to small distances/large similarities between the corresponding topic distributions.

Method	*Accuracy/ Kappa/F-measure (T=300)*	*Accuracy/Kappa/F-measure (T=100)*	*Accuracy/Kappa/ F-measure (T=300)*	*Accuracy/Kappa/F-measure (T=100)*
LDA-IR	71.17/16.17/81.94	68.24/3.09/80.92	67.47/4.52/79.87	67.01/3.15/79.98
LDA-Hellinger	71.32/18.85/81.75	68.24/2.46/80.99	67.36/4.39/79.73	67.18/3.50/80.04
LDA-Manhattan	71.07/10.10/82.50	71.21/23.41/81.16	66.78/3.56/79.91	67.18/4.04/80.04
LDA-Greedy	**77.32/34.40/85.75**	**76.85/37.89/84.94**	73.04/35.01/81.31	73.10/34.27/81.32
LDA-Optimal	76.97/36.96/85.06	75.96/36.75/84.14	**73.27/36.74/80.71**	**73.15/36.86/80.71**
LSA-Greedy	77.22/33.82/85.73	*Same*	72.86/33.89/81.11	*same*
LSA-Optimal	77.12/36.80/85.24	*Same*	73.04/35.95/80.80	*same*
LSA	**77.47/37.54/85.50**	*Same*	**73.56/34.61/81.83**	*same*

Table 1. Results on ULPC (column 2 and 3) and MSRP (column 3 and 4) test data with LDA-based methods for various number of topics (T=100 represents the most coherent set of topics).

6 Discussion and Conclusions

We presented in this paper our work on defining semantic similarity measures at word and sentence level based on LDA. A measure based on word representations as vectors of topic contributions yields competitive results with the unsupervised algebraic method of LSA. Furthermore, Table 1 indicates that semantic similarity measures based on distances among distributions over words and topics (see the rows for LDA-IR, LDA-Hellinger, LDA-Manhattan) are not useful for short texts due to topic sparseness in short texts. We plan to investigate this issue in future work.

Acknowledgments

This research was supported in part by Institute for Education Sciences under awards R305A100875. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors'.

References

1. Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent dirichlet allocation, The Journal of Machine Learning Research 3, 993-1022.
2. Celikyilmaz, A., Hakkani-Tür, D., & Tur, G. 2010. LDA Based Similarity Modeling for Question Answering, NAACL-HLT. Workshop on Semantic Search, Los Angeles, CA, June 2010.
3. Chen, X., Li, L., Xiao, H., Xu, G., Yang, Z., Kitsuregawa, M. (2012). Recommending Related Microblogs: A Comparison between Topic and WordNet based Approaches. Proceedings of the 26th International Conference on Artificial Intelligence.
4. Dagan, I.; Glickman, O.; and Magnini, B. 2004. Recognizing textual entailment. In http://www.pascalnetwork.org/Challenges/RTE.
5. Dagan, I., Lee, L., Pereira, F.C.N. 1997. Similarity Based Methods For Word Sense Disambiguation, ACL, 1997, p. 56-63.

6. Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. COLING 2004, Geneva, Switzerland.
7. Fernando, S. and Stevenson, M. 2008. A semantic similarity approach to paraphrase detection. In Proceedings of the Computational Linguistics UK (CLUK 2008).
8. Graesser, A. C.; Olney, A.; Haynes, B. C.; and Chipman, P. 2005. Autotutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In Cognitive Systems: Human Cognitive Models in Systems Design. Mahwah: Erlbaum.
9. Griffiths, T.L. and Steyvers, M. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235.
10. Hofmann, T. 1999. Probabilistic latent semantic indexing. In Proceedings of SIGIR'99, pages 50–57.
11. Kozareva, Z. and Montoyo, A. (2006). Paraphrase Identification on the basis of Supervised Machine Learning Techniques. Proceedings of the 5th International Conference on Natural Language Processing (Fin-TAL 2006), pages 524-233.
12. Kuhn, H.W. 1955. "The Hungarian Method for the assignment problem", Naval Research Logistics Quarterly, 2:83–97, 1955. Kuhn's original publication.
13. Landauer, T.; McNamara, D. S.; Dennis, S.; and Kintsch, W. (2007). Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.
14. Lintean, M., Moldovan, C., Rus, V., & McNamara D. (2010). The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis. Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference. Daytona Beach, FL.
15. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., and McCallum, A. 2011. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262– 272. ACL.
16. Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics, volume 5(1), pages 32-38. Society for Industrial and Applied Mathematics.
17. Newman, D., Lau, J.H., Grieser, K.., and Baldwin, T. 2010. Automatic evaluation of topic coherence. In HLT-NACL, pages 100–108. ACL.
18. McCarthy, P.M. and McNamara, D.S. 2008. User-Language Paraphrase Corpus Challenge, online, 2008.
19. Rus, V. & Graesser, A.C. (2006). Deeper natural language processing for evaluating student answers in intelligent tutoring systems, Paper presented at the Annual Meeting of the American Association of Artificial Intelligence (AAAI-06), July 16-20, 2006, Boston, MA.
20. Rus, V. & Lintean, M. (2012). A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics, Proceedings of the Seventh Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, NAACL-HLT 2012, Montreal, Canada, June 7-8, 2012.
21. Rus, V., Niraula, N.B., and Banjade, R (2013). *"Similarity measures based on latent dirichlet allocation."*, In Proceedings of CICLing, 2013.
22. Steyvers, M. and Griffiths, T. 2006. Probabilistic topic models. Handbook of latent semantic analysis, 427(7):424–440.
23. Wallach, H., Mimno, D., and McCallum, A. 2009. Rethinking LDA: Why priors matter? Advances in Neural Information Processing Systems, 22:1973–1981.
24. Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. 2006. Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.