SimDoc: A Topic Sequence Alignment based Document Semantic Similarity Measure

Akshay Ratan[1], Sourish Dasgupta[1], Kunal Jha[1], Kandarp Joshi[1], Hitarth Kanakia[1], Michael Roeder[2], Jens Lehmann[2]

[1] DA-IICT, India;
{akshay, sourish, kunal, kandarp, hitarth, sourish}.rygbee@gmail.com
[2] AKSW Research Group, University of Leipzig, Germany
{roeder, lehmann}@informatik.uni-leipzig.de

Abstract. In this paper, we propose a *semantic document similarity* measure, called *SimDoc*, for computing semantic similarity between textual documents. We model the documents as topic sequences (generated using *Latent Dirichlet Allocation* (*LDA*) technique) and then calculate the alignment score between the sequences based on *Smith-Waterman* algorithm. However, topic-sequence alignment score is not very accurate since it largely generalizes the original document. Hence, we generate a *representative word-sequence* from the topic-sequence (using *Gibbs Sampling*) and fine-tune every dis-alignment cost with a lexicon based semantic distance. We use WordNet based similarity/relatedness measures for this purpose. A global document-level similarity score is then computed on all topic sequence pair local similarity score using *Root Mean Square Deviation* (*RMSD*) from an upper-bound baseline. We evaluated *SimDoc* using standard benchmark datasets.

1 Introduction

Document semantic similarity measures quantify the similarity in meaning and the semantic relationship between textual documents. Similarity measure is modeled as a function that maps two documents, that are represented in some formal structure, to a real space by considering certain feature parameters either based on *hypothesis of distributional semantics* (primarily keyword based statistical approach) [4] or *hypothesis of compositional semantics* (primarily based on formal semantics [29], or both [11]. Semantic relatedness and equivalence measures help in understanding and solving important NLP problems including question-answering [33], paraphrase identification [30], textual entailment [35], query reformulation [21], and automatic plagiarism detection [24].

Modeling similarity measure is a non-trivial research problem. This is because natural language texts mostly do not follow any rigid grammar structure and formal semantic theory. Moreover, there are several linguistic nuances that pose further modeling challenges (such as polysemy, underspecification, figurative phrases, etc.). The majority of modeling approaches are based on the hypothesis of distributional semantics, which states that words having similar

context have same semantic structure. One such approach is the *Vector-Space Model* (*VSM*), where documents are represented as term vectors [5]. The advantage of vector-based representation is that they are generated from text without the computational overhead involved in underlying linguistic analysis. Also, they significantly correlate with human annotated systems [11]. However, due to VSM's insensitivity towards language characteristics and other linguistic nuances (such as voices, phrases, coreference and word-sense ambiguities, etc.), more advanced generative models (such as pLSA [17], LDA [8], and hLDA [7]) were proposed. These techniques are based on *distributional language modeling* where the core idea is to quantify the similarity between documents in terms of the degree in which they share the same topic distribution.

In this paper, a topic-modeling based semantic textual similarity measure, called *SimDoc*, is proposed for matching two documents. As a first step, we generate a representative topic-sequence (based on *Latent Dirichlet Allocation* (*LDA*)) for each document. Next, we compute the topic-sequence alignment scores between all sequence pairs. For this, we have proposed a novel model that combines the *Smith-Waterman* algorithm [32] with *WordNet* [28] based semantic distance measures. We then find the *Root Mean Squared Deviation* (*RMSD*) of the best sequence alignment scores against an upper bound. Based on this value, we get the final score of semantic similarity between the two documents. The contributions of this paper are as follows:

- Modeling semantic document similarity as a topic-sequence alignment problem. Most contemporary approaches, that are based on template-based probabilistic models, treat documents as *bag-of-topic*. However, we argue that sequence alignment modeling is a novel and powerful technique for capturing the similarity in *thematic flow*. It also helps to model the *thematic incongruity* and *abrupt thematic transitions* when two different documents are compared. More details can be found in section 4.3.
- Combining *Smith-Waterman* edit-distance algorithm (which has been successfully applied in alignment problems in Bioinformatics) with *WordNet* based semantic similarity measures. This gives a lexicon based interpretation to the sequence alignment score.
- Empirically analyzing the difference between *strong* and *weak semantic similarity*, and how they are distinct from *strong semantic relatedness*. This is clearly demonstrates the adaptivity, in terms of re-usuability, of our proposed model. We hypothesise and statistically prove a relation between the three proposed frameworks. Also, empirically evaluating *SimDoc* extensively using benchmark datasets (20Newsgroup 18828, Reuters 21578, WebKB). We observe the mean average precision value over all datasets to be 79.48%.

The remaining sections of the paper are organised as follows: 2) *related work* where we describe in detail research work done in computing the measure of textual similarity; 3) *problem statement* which introduces the research problem, it's significance and challenges along with motivating running example; 4) *approach* wherein the architecture pipeline has been described. This section also discusses

the *SimDoc* framework in details; 5) *evaluation* of the model accuracy based on extensive experimental testing and analysis with standard benchmark datasets.

2 Related Work

Most work towards modeling of text similarity measures can be grouped into five broader approaches: (i) *string-based models*, (ii) *term vector-based models*, (iii) *topic-based models*, (iv) *lexicon-based distance models*, and (v) *measure-ensemble based models*.

String based models: A string-based similarity model uses string edit-distance measures (such as *levenshtein distance* [25], *Jaro-Winkler* [12] etc.), for approximate comparison between documents. Some approaches in this direction can be found, e.g., in [18,20]. Though edit-based distance measures are useful in comparing short strings where a small number of difference between the strings is expected, it falls short in performance when used for long text matching. Further, these string-based measures do not account for the semantics of text.

Term-vector based models: A significant number of approaches are based on term-vector modeling. In this technique, every document is represented as a vector of terms (represented by their term and inverse-document frequency). In this space, *Euclidean distance* and *Manhattan distance* can be used to compute the distance between two vector end points to measure the similarity of two texts. The *cosine similarity* addresses the similarity between two documents by calculating the cosine between two vectors. Another common approach is known as *Jaccard similarity*. It is calculated as the number of shared terms over the number of all exclusive terms in both strings. mapped the data similarity task to a clustering task. Documents in the same cluster are assumed to be similar to each other. Another information-theoretic measure for document similarity adapts the query retrieval to rate the quality of document similarity measure [2]. Though the vector space model is easy to implement and useful in measuring multi-level and partial text matching, there are considerable disadvantages as the model assumes independent relationship among the terms and does not take into account the semantic sensitivity between the pair of documents.

Topic generation based models: Models in this direction determine similarity between words according to distributional information gained from large corpora. Closer neighbouring words are thought to reflect more of the focus word semantics and so are weighted higher. *Latent Semantic Analysis* (*LSA*) techniques [22] are based on the assumption that words carry a latent structure in its usage and tries to solve the problem by deriving conceptual indices of those words. A matrix, that counts the occurrences of each word per paragraph, is maintained. Each row is considered a vector and the cosine angle between them gives the similarity between words. Some words that have similar co-occurence patterns are projected onto the same dimension. Hence, similarity metrics will group words with similar meaning into one topic. *Explicit Semantic Analysis* (*ESA*) [14] represents the meaning of the texts in a high-dimensional space of concepts derived from corpus. It is used to compute semantic relatedness be-

tween two arbitrary texts by representing terms (or texts) as high dimensional TF-IDF weighted vectors of concepts. Cosine metric gives similarity between the texts.

Lexicon-based distance models: These are one of semantic similarity models that are based on identifying the degree of similarity between words using information derived from semantic networks. These similarity measures can be divided roughly into two groups: measures of semantic similarity and measures of semantic relatedness. The five semantic based similarity measures incorporate measures based on information content (Resnik [31], Lin [26] and Jiang and Conrath [19]) and measures based on path length (Leacock and Chodorow [23] and Wu and Palmer [34]). The standard packages that cover knowledge-based similarity measures are WordNet [28] and Natural Language Toolkit (NLTK) [6].

Measure-ensemble based models: Techniques in this category use multiple, predominantly corpus or knowledge based similarity measures. Semantic Text Similarity (STS) determines the similarity of two texts from an amalgamation of semantic and syntactic information. [15] combines the latent knowledge from a large corpus and the semantic knowledge of an ontology. The word similarity was based on the LSA-similarity and a WordNet based boosting factor. Though the system performs well in giving a score in similar sentences mapped as active-passive, there is a significant scope for improvement. The system operates on a bag-of-words level and ignores the word positions. It becomes difficult to capture the flow of one topic to another inside a text as position of words inside the sentences are not taken into account. UKP [3] uses a simple log linear regression model based on training data, to combine multiple text similarity measures including string similarity, semantic similarity, text expansion mechanisms and measures related to structure and style.

3 Problem Statement

3.1 Problem Outline

Document similarity can be computed based on three linguistic aspects—*syntactic*, *lexical*, and *semantic*. Syntactic similarity, for a given language, implies resemblance in the grammar and structural relations between linguistic elements (eg.: words, phrases, sentences, etc.) in the documents. In contrast, lexical approaches focus on the similarity between word groups as either lexemes or strings. Semantic similarity quantifies the degree of equivalence in the meaning of central subject matter of two documents. We argue that in order to design an effective model for semantic document similarity, we cannot ignore the nuances that are innate in all the above three linguistic aspects. Due to this the current paper proposes a hybrid measure, called *SimDoc*, that is significantly accurate.

As a motivating instance of the problem, let us consider the two documents:
Document 1: *"The quick brown fox jumped over a lazy dog. The boy was clever to have dodged the fox in the chase".*
Document 2: *"The boy was very clever. He came first in the class, and also*

played sports dodging the ball promptly. The dog lay in the side watching the game."

Given this pair of document examples, a proposed similarity measure must give a quantitative value (say, on a scale of 0 to 1) for the degree of semantic similarity between these documents.

3.2 Problem Definition

Document (and sentence) similarity measures are usually defined on the real space via a function σ defined as follows:

Definition 1 (**Document Similarity Measure** (σ)):
$\sigma : D \times D \mapsto [a, b]$; where,

- D is the input textual document
- $a \in \mathbb{R}$ is the lower-bound.
- $b \in \mathbb{R}$ is the upper-bound.

Usually, the bounds are normalized into a finite interval. Certain desirable properties of σ are:

- *Positiveness*: $\forall D_i, D_j, \sigma(D_i, D_j) \geq 0$
- *Reflexiveness*: $\forall D_i, \sigma(D_i, D_i) = b$
- *Symmetry*[3]: $\forall D_i, D_j, \sigma(D_i, D_j) = \sigma(D_j, D_i)$
- *Maximality*: $\forall D_i, D_j, D_k; \sigma(D_i, D_i) \geq \sigma(D_j, D_k)$
- *Equivalence Invariance*: $\forall D_i, D_j, D_k;$
  $(\sigma(D_i, D_j) = b) \implies \sigma(D_i, D_k) = \sigma(D_j, D_k)$

3.3 Challenges

Modeling σ has several linguistic challenges that originate from semantic ambiguities related to: (i) polysemy[4], (ii) underspecification[5], (iii) pragmatics[6], (iv) anaphora[7], and (v) referent ambiguity[8]. Such ambiguities are largely due to uncertainty of the context and are quite difficult to generalize based on existing methods in distributional semantics.

[3] This property is not necessarily true and hence, not uniformally adopted [9].

[4] *"Manuel died in exile in 1932 in **England**"* - here *England* is a geographical location vs. *"**England** was being kept busy with other concerns"* - here *England* is a political entity.

[5] *"John showed the painting to every one but Joe did not."* - here it is not clear whether Joe showed the painting to some people or did not show the painting at all.

[6] "*I want my money right now, right here, all right?*" - emphasis on "*right*" adds special semantics

[7] "*John gave Joe **his** book*" - here it is not clear whether book belongs to John or Joe.

[8] *"John showed **his painting** to his teachers"* - here *"his"* painting might refer to the self-potrait of John or a painting made by John.

Apart from barriers due to linguistic nuances, we would also like to emphasize on three important yet distinct notions of document similarity for any given measure σ: (i) *weak document similarity* (*W-Sim*), (ii) *strong document similarity* (*S-Sim*), and (iii) *strong document relatedness* (*S-Rel*) [10]. We define each of them as follows:

Definition 2 (**Strong Document Similarity** (*S-Sim*)): *S-Sim* is a similarity perspective that exclusively captures the degree of *semantic-equivalence* in the central subject matter of two documents.

If two documents have high *S-Sim* score then it implies that the documents' content have similar narratives or facts. An example of high S-Sim will be the two single sentence documents: "*An author pens a book.*" and "*Poets write poems.*".

Definition 3 (**Strong Document Relatedness** (*S-Rel*)): *S-Rel* is a similarity perspective that exclusively captures the degree of *semantic-complementarity* in the central subject matter of two documents.

If two documents have high *S-Rel* score then it implies that the primary subject of discussion of one document is used as one of the predicates of the other document. However, they may not have high *S-Sim* score, if the central subject matter of one document is different from the other document. An example of high S-Rel will be the two single sentence documents: "*An author writes a book.*" and "*Books are sources of knowledge.*".

Definition 4 (**Weak Document Similarity** (*W-Sim*)): *W-Sim* is a similarity perspective that captures the degree of *semantic-relatedness* (which might also mean *semantic-equivalence*) in the central subject-matter of two documents.

If two documents have high *W-Sim* score then it implies that the documents have either high *S-Sim* score or high *S-Rel* score or both. It is to be noted that the general notion of document similarity in the community is loosely equivalent to the notion of *W-Sim*.

Union Hypothesis: *A document retreival process focusing on W-Sim should include all documents having high S-Sim and high S-Rel scores.*

The underlying implication of the above hypothesis is that, σ should be flexible enough to be used exclusively for *S-Sim* or *S-Rel*.

4 Approach

4.1 Outline

In the proposed measure, a document is modeled as *topic sequence*. *Topic*, in this context, means a group of words that are related to each other thematically (in the distributional semantics sense). As an example, "*house*" and "*rent*" can be within the same topic. A sequence of topics, thus, represents the transition chain from one theme to the other (i.e. the thematic flow of the document content). As an example, in the sentence: "*John loves to make ice-cream for Mary.*", we can assign the topic sequence instance: $\langle person, love, make, dessert, person \rangle$. It is quite evident that a *bag-of-topic* model fails to capture the thematic flow in the same way. The importance of accounting for such flow can be understood by

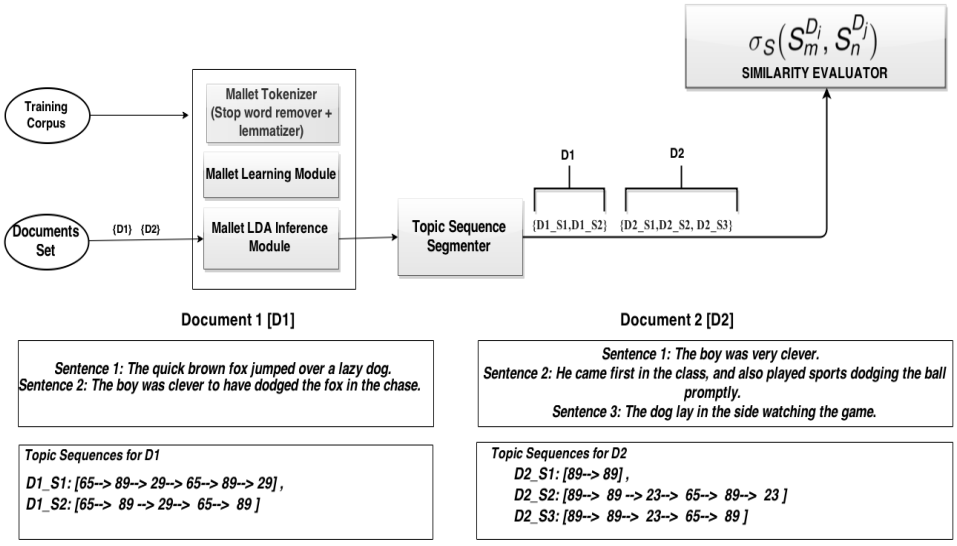

Fig. 1. *SimDoc* Architectural Pipeline

comparing the previous example sentence with the sentence: "*John loves the ice-cream made by Mary*" where the topic sequence is: $\langle person, love, dessert, make, person\rangle$. Given two topic sequences corresponding to two documents, we then try to compute the *semantic-similarity alignment* between them. This gives us a similarity score that measures *weak similarity* (*W-Sim*) between the documents. In a similar way, we compute *semantic-relatedness alignment* between topic sequences in order to find *strong relatedness*. However, in order to get a score for *strong similarity* (*S-Sim*), we compute the semantic-similarity alignment between original word sequences (rather than topic sequences). This is because a particular topic may include words that are both semantically related (eg: *author* and *book*) and semantically similar (eg: *author* and *poet*). In the subsequent sections we describe in details how topic-sequence is generated and how we compute *semantic alignment* both in the similarity as well as in the relatedness sense.

4.2 Topic Sequence Generation: Technique

In order to generate topics, we adopt a *Latent Dirichlet Allocation* (*LDA*) based topic modeling technique . LDA discovers probabilistically inter-related latent topics without any prior information about thematic content of the documents. In LDA, K topics (probabilistically distributed over a fixed vocabulary set of cardinality V) are assigned to a collection of words, and every document may (or may not) contain these topics in different proportions. We use *Mallet* [27] for implementing LDA. A training corpus is first fed into the *Mallet Tokenizer* (for

removal of stop-words and lemmatization) and then the LDA model is learned subsequently. After that, for a given document pair, each document is fed to the *Mallet LDA Inference Module* so as to get topic-sequence as output. The topic sequence is then passed through a *topic-sequence segmenter* module that splits the topic-sequence into a set of sub-sequences (or *segments*). We define segment boundary to be equivalent to a sentence boundary. Segmentation is important because of two reasons. First, to capture the locality of semantic similarity match, it is important to consider sentence-boundary based segments (rather than long topic-sequence). Identifying the locality helps us to contextualize the document similarity score. Also, it helps to measure document similarity in terms of top k sentence-level matches. In this way, we can also analyze whether two documents have similarity in their central thematic flow or otherwise (although, currently this is not within the scope of the paper). Secondly, if we take long topic-sequence of document without segmentation, then it may happen that dissimilarity in previous topic-sequence pairs can negatively bring down the score of subsequent similar topic-sequence pairs.

4.3 *SimDoc* Framework

Topic-Sequence Segment Alignment: In order to compute similarity between topic-sequence segments that are generated as described in previous section, we apply Smith-Waterman algorithm [32] to quantify the *topic alignment* between them. Smith-Waterman is a non-heuristic, dynamic programming algorithm that determines similarity measure between two sequences by computing minimum *local sequence alignment* score of character sequence. Here *local* implies alignment within a context window (or segment). For obtaining a best local alignment score between a pair of sequence segments, scoring starts at a local level. Mismatches on a global level are not accumulated by zero-ing out scores that are negative. In the context of *SimDoc*, character implies topics for measuring *weak similarity* (W-Sim) and *strong relatedness* (*S-Rel*). On the other hand, character implies original document words when we need to measure *strong similarity* (S-Sim). This is because when we look for strong similarity we need to find documents whose content are semantically equivalent (and not just thematically related).

Model Formulation: In this section we are going to give the formulation of *SimDoc*. We first describe some preliminary concepts as follows:

- $T^D_{p_S}$: p-th position topic of a topic-sequence segment S for a given document D.
- $\sigma_S\big(S^{D_i}_m, S^{D_j}_n\big)$: Semantic similarity between topic-sequence segment S_m of document D_i and topic-sequence segment S_n of document D_j on a scale: $(0, 1]$; where 1 is the maximum possible similarity between two segments.
- $G_{DEL/INS/SUB}$: *Gap-penalty scheme*. It is defined as: $G = O + \{(l-1) \times e\}$; where:

- O: *Deciding Factor* of minimum gap-penalty. It varies based on the type of topic-editing (i.e. *topic-insertion* (*INS*), *topic-deletion* (*DEL*), *topic-substitution* (*SUB*)) required for a given sequence segment comparison. If the best possible action during an editing step is *insertion* then it adds semantic incongruity as well as extension to the thematic flow of a topic segment in the other document, and hence, should be heavily penalized. However, if the best possible action is *deletion* then it shrinks the thematic flow, thereby forcing a abrupt topic-jump without adding any incongruity and thematic extension. On the other hand, if the best possible action is *substitution* then it only "*might*" add thematic incongruity in the other document without affecting the thematic flow and causing abrupt topic-jump. Hence, in our context, $O_{INS} > O_{DEL} = O_{SUB}$.
- l: Number of same topic-editing observed for a given sequence segment comparison. This is used to further penalize the occurance of an already observed disalignment.
- e: *Incremental-Penalty* for each l (explained above).

– M: *Match-Gain*. It is the reward assigned for every successful match between two topics (i.e. $T^{D_i}_{pS_m} = T^{D_j}_{qS_n}$).

We define the generic framework of $\sigma_S(S^{D_i}_m, S^{D_j}_n)$ by the following Bellman equations [9] (where $M = 2$):

$$\sigma_s\left(T^{D_i}_{pS_m}, T^{D_j}_{qS_n}\right) = \begin{cases} \sigma_s\left(T^{D_i}_{(p-1)S_m}, T^{D_j}_{(q-1)S_n}\right) + M, & \text{iff } T^{D_i}_m = T^{D_j}_n \\ max \begin{cases} \sigma_s\left(T^{D_i}_{(p-1)S_m}, T^{D_j}_{(q)S_n}\right) + G_{DEL}, & \text{iff } \textit{topic-del.} \\ \sigma_s\left(T^{D_i}_{(p)S_m}, T^{D_j}_{(q-1)S_n}\right) + G_{INS}, & \text{iff } \textit{topic-ins.} \\ \sigma_s\left(T^{D_i}_{(p-1)S_m}, T^{D_j}_{(q-1)S_n}\right) + G_{SUB}, & \text{iff } \textit{topic-sub.} \end{cases} \end{cases}$$

WordNet based Adjustment of Alignment Score: It should be noted that, for all the above three (*W-Sim,S-Sim,S-Rel*), if we directly apply the Smith-Waterman algorithm on topic/word sequence segments we will not be able to compute the true semantic document similarity. This is because Smith-Waterman based sequence similarity can understand only the broader thematic similarity of documents and hence, is incapable of computing a finer level semantic similarity score. To illustrate this with an example, let us take the single sentence documents: "*The quick brown fox jumped over a lazy dog. The boy was clever to have dodged the fox in the chase.*" and "*The boy was very clever. He came first in the class, and also played sports dodging the ball promptly. The dog lay in the side watching the game.*". In this case, the two documents are not semantically similar both in the sense of *S-Sim* and *W-Sim* (and hence, *S-Rel*). However, if we apply pure topic or word sequence alignment method then we get higher similarity score. In order to solve this problem we propose *SimDoc* where

[9] In the case of *S-Sim*, we use W^D_{pS}; where W is the original word in p-th position of the segment sequence S in document D.

we redefine the *gap penalty* (or editing cost) in Smith-Waterman algorithm by factoring the original edit cost with a similarity score computed from WordNet [28] based semantic relatedness/similarity metrics[10]. We used the average of the scores computed individually by *Resnik*[11], *Wu and Palmer*[12], and *Jiang and Conrath*[13] for measuring *W-Sim* and *S-Sim* and used *Hirst & St. Onge*[14] [16] metric for measuring *S-Rel*. Since topics are mostly quite abstract, once topic sequence segments are generated (as described in previous section), we then probabilistically generate corresponding *representative words* out of every topic using *Gibbs sampling technique* [1]. While computing *W-Sim* and *S-Rel*, we use these representative words for WordNet based similarity score.

In order to formally define the three notions of similarity we first redefine the *gap-penalty* (denoted as G') as follows:

$$G' = G \times \frac{1}{\sigma_{WN}\left(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}}\right)} \quad \text{where:}$$

- $\widehat{W_{p_S}^{D}}$: Most dominant word (i.e. *representative word*) of topic $T_{p_S}^{D}$ generated via *Gibbs sampling*
- G is the original *gap-penalty* (as defined in previous section)
- $\sigma_{WN}(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}})$: *WordNet* based semantic similarity between $\widehat{W_{p_{S_m}}^{D_i}}$ and $\widehat{W_{q_{S_n}}^{D_j}}$ of two segments S_m and S_n belonging to D_i and D_j.

For each of the three notions of similarity, G' is fine-tuned as follows:[15]

$$\begin{aligned} G'_{W\text{-}Sim} =& 3 \times G \div \left(\sigma_{Wu\text{-}Palmer}\left(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}}\right) \right. \\ & \left. + \sigma_{Resnik}\left(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}}\right) + \sigma_{Jiang\text{-}Conrath}\left(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}}\right) \right) \\ G'_{S\text{-}Rel} =& \frac{G}{\sigma_{\text{Hirst \& St.Onge}}\left(\widehat{W_{p_{S_m}}^{D_i}}, \widehat{W_{q_{S_n}}^{D_j}}\right)} \end{aligned}$$

[10] For this we used the *Ws4j* - Java implementation of WordNet.

[11] *Resnik*: Similarity is defined on the basis of information content of the most specific common subsumer of the two synsets

[12] *Wu Palmer*: Similarity is defined considering the graph depth of WordNet taxonomies and most common subsumer of the lexicons in comparison.

[13] *Jiang and Conrath*: Similarity is defined as the difference in Information Content of the two lexicons in comparison and the Information Content of their most specific common subsumer lexical entry.

[14] *Hirst & St. Onge: Relatedness* is defined by introducing the concept of allowable paths that may or may not be hyponymic.

[15] In the case of $G'_{S\text{-}Sim}$, we use $W_{p_S}^{D}$ where $W_{p_S}^{D}$ is the original p-th word in the segment sequence S in document D.

We use G' as the new *gap-penalty* in the Bellman equation that has been defined in previous section. This gives a fine-tuned, improved version of $\sigma_S(S_m^{D_i}, S_n^{D_j})$.

Global Document Similarity (*SimDoc*): In order to measure the global similarity (i.e. σ_{SimDoc}), we first select a sequence segment (say, S_m) from one document (say D_i) and match it against all the (unselected) sequence segments of the other document (say, D_j). We assign the minimum sequence segment alignment score (i.e. $min_{n=[1...N_j]}\{\sigma_S(S_m^{D_i}, S_n^{D_j})\}$) to S_m of D_i, where N_j is total number of sequence segments in D_j. This process is continued turn-wise (with a new selection of segment made from D_j and compared with unselected segments of D_i in the same way) till all the segments in both documents are selected.

After getting the set of minimum alignment scores we then compute the Root Mean Square Deviation (*RMSD*)[16] against the "*ideal*" baseline score. We define the ideal case as the maximum possible score (denoted θ_{max}) of 1. We compute RMSD with respect to θ_{max} because θ_{max} provides a uniform upper-bound for each matching sequence segment pair. This helps to generalize $\sigma_S(S_m^{D_i}, S_n^{D_j})$ even if a gold-standard prior baseline for individual segment matches is not available. The final *SimDoc* computation is done by the *Similarity Evaluator* (see fig. 1) as per the following equation:

$$\forall y_k \in \Big\{ max_m\big\{\sigma_S\big(S_m^{D_i}, S_n^{D_j}\big)\big\} \bigcup max_n\big\{\sigma_S\big(S_m^{D_i}, S_n^{D_j}\big)\big\}\Big\}$$

$$\sigma_{SimDoc}(D_i, D_j) = 1 - \sqrt{\frac{\sum_{k=1}^{K}\Big(\theta_{max} - y_k\Big)^2}{K}}$$

where K is the number of best matched sequence segment pairs. K is bounded as: $K = [min\{N_i, N_j\},\ (2 \times min\{N_i, N_j\}) - 1]$

5 Evaluation

5.1 Evaluation Goals and Metrics

Goal I. *W-Sim* Accuracy: In order to measure the *W-Sim* accuracy of *SimDoc*, we use the standard measures of *precision*, *recall*, *F-score*, and *rejection*, with respect to gold-standard human evaluated relevance-judgement. The objective is to evaluate accuracy on large textual documents. However, since we did not get direct relevance-judgement of document pairs, we adopted an indirect method of analyzing the accuracy. We chose datasets (described in the

[16] RMSD measures the deviation of values from a reference set. If $RMSD = 0$, it implies a perfect overlap.

next section) in which similar documents have been clustered based on human-judgement (clusters denoted as DC). We noted down the cluster-size of each DC (denoted as N_{DC}). For each document in DC, we use $\sigma_{W\text{-}Sim}$ to select top $(N_{DC} - 1)$ most similar documents. We then compute the *average accuracy*, in terms of the four measures, for each cluster DC (as per [13]). Then we compute the *mean average accuracy*, in terms of each measure, over all the clusters for a given benchmark dataset.

Goal II. *Union Hypothesis* Testing: In order to show that *SimDoc* obeys the *union hypothesis* (postulated in section 3.3), we first formulate a hypothesis testing measure, called *Hypothesis Support* (*H-Sup*), as follows:
Definition 5 (Hypothesis Support) (*H-Sup*): *H-Sup* for a given document D_i is the ratio of number of similar documents retrieved by $\sigma_{S\text{-}Sim}$ (denoted as $N^{D_i}_{S\text{-}Sim}$) and by $\sigma_{S\text{-}Rel}$ (denoted as $N^{D_i}_{S\text{-}Rel}$) that are also present in the retrieval set of $\sigma_{W\text{-}Sim}$ (denoted as $N^{D_i}_{W\text{-}Sim}$), with respect to $N^{D_i}_{W\text{-}Sim}$. The corresponding formula is:

$$H\text{-}Sup(D_i) = \frac{(N^{D_i}_{S\text{-}Sim} \cup N^{D_i}_{S\text{-}Rel}) \cap N^{D_i}_{W\text{-}Sim}}{N_{W\text{-}Sim}}$$

Definition 6 (Hypothesis Confidence) (*H-Conf*): *H-Conf* for a given document D_i is the ratio of number of similar documents retrieved by $\sigma_{S\text{-}Sim}$ (denoted as $N^{D_i}_{S\text{-}Sim}$) and by $\sigma_{S\text{-}Rel}$ (denoted as $N^{D_i}_{S\text{-}Rel}$) that are also present in the retrieval set of $\sigma_{W\text{-}Sim}$ (denoted as $N^{D_i}_{W\text{-}Sim}$), with respect to $(N^{D_i}_{S\text{-}Sim} \cup N^{D_i}_{S\text{-}Rel})$. The corresonding formula is:

$$H\text{-}Conf(D_i) = \frac{(N^{D_i}_{S\text{-}Sim} \cup N^{D_i}_{S\text{-}Rel}) \cap N^{D_i}_{W\text{-}Sim}}{(N^{D_i}_{S\text{-}Sim} \cup N^{D_i}_{S\text{-}Rel})}$$

We now measure the F-1 score of the *H-Sup and H-Conf*. The corresponding formula is given below:

$$HF\text{-}Score(D_i) = 1 \div (\frac{1}{H\text{-}Sup(D_i)} + \frac{1}{H\text{-}Conf(D_i)})$$

In order to measure *HF-Score* of *SimDoc* for an entire benchmark dataset, we compute the *Average HF-Score* over all the documents in the dataset.

5.2 Experimental Setup & Datasets

We use Mallet to assign topic sequence to the document. We initialize the parameters of *gap penalty* (G) of the alignment algorithm as: $l = 0$, $O = -2$(for *topic-insertion*); -1(for *topic-deletion*); -1(for *topic-substitution*), and $e = -1$ [17]. We also experimented with other parameters in *Mallet* [27] (such as tweaking number of representative topics generated for the corpus, and increasing

[17] Refer to Section 4.3

the iterations of Gibb's sampling to generate representative words) to fine-tune the performance. However, due to limited space, details of these experiments in topic-modeling are outside the scope of this paper. We evaluate *SimDoc* on three different benchmark corpus dataset. We describe each of them below:
20-Newsgroup [18]: This dataset is derived from the CMU Text Learning Group Data Archive. It involves newsgroups with a collection of 20,000 messages, collected from 20 different internet-news groups. The dataset includes 1000 messages from each of the twenty newsgroup, chosen at random and were partitioned on the basis of group name. We run our test specifically on the groups involving religion, medicinal science, electronics and hardware. For a diversified evaluation, we took all the possible combination pairs of documents.
Reuters 21578 [19]: One of the most widely used test collection for text categorisation evaluations, the data was collected by Carneige Group, Inc. and Reuters, Ltd. The documents are Reuters newswire stories and there are five different categories based on content theme. The five category sets are *Exchanges*, *Orgs*, *People*, *Places* and *Topics*. The first four above-mentioned categories correspond to named entities of the specified types and *Topics* categories are economic subject categories.
WebKB [20]: This dataset is derived from the *World Wide Knowledge Base* project of CMU text learning group. These webpages were collected from various computer science universities and manually classified into seven different classes: *student*, *faculty*, *staff*, *department*, *course*, *project*, and *other*.

5.3 Results & Analysis

Result I. *W-Sim* Accuracy: As shown in the Fig. 2, we obtain high Mean Average Precision (MAP) (20NewsGroup: 75.3%; Reuters: 82.8%; WebKB: 80.34%) and high Mean Average Recall (MAR) (20NewsGroup: 78.5%; Reuters: 82.80%; WebKB: 83.14%) for *SimDoc* with respect to *W-Sim*. Also, we observed that *SimDoc* works very accurately in giving a low similarity score in case of dissimilar document pair across all datasets. This has been clearly shown by the Mean Average Rejection (MA-Rejection) (20NewsGroup: 85.2%; Reuters: 82.5%; WebKB: 58.8%). However, we found certain anomalies in the behavior of *SimDoc* as per *W-Sim*. As an example, we observed that in a corpus which contained one document pertaining to *contraceptives* and another to *environment*, *SimDoc* incorrectly showed high semantic similarity. This is due to the co-occurence of the terms within very similar contexts (in the above example, the observed context, *sex* and *population* respectively, are strongly mutually related). However, in such cases, *SimDoc*, as per *S-Sim*, correctly showed high value of dissimilarity, and as per *S-Rel*, correctly showed high relatedness between the two documents.

[18] http://qwone.com/ jason/20Newsgroups
[19] https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection
[20] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data

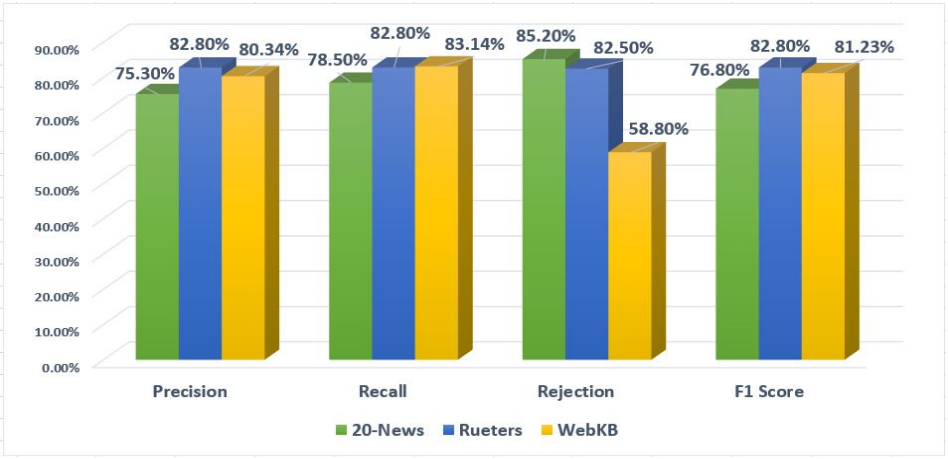

Fig. 2. Mean Average Accuracy of *SimDoc* with respect to *W-Sim*

Result II. *Union Hypothesis* Testing: The union set of correctly predicted results of *S-Sim* and that of *S-Rel* is compared to the correctly predicted results of *W-Sim*. We got a high *Mean Average HF-Score* (20NewsGroup: 66.7%; Reuters: 69.21%; WebKB: XXXX).

6 CONCLUSION

In this paper, we propose *SimDoc* - an adaptive topic-sequence alignment based document similarity measure. When run on standard benchmark datasets *SimDoc* shows high accuracy. We also empirically show that the accuracy of *SimDoc* significantly improves when it is boosted with WordNet-based similarity measures. We demonstrate through statistical support that *SimDoc* obeys the *union hypothesis* (given in section 3) about the relationship with *S-Sim*, *S-Rel*, *W-Sim*, and hence, that *SimDoc* is an adaptive measure. We hope to extend *SimDoc* with better topic-modeling techniques and improved upper ontology (such as DBpedia) based similarity measures.

References

1. C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
2. J. A. Aslam and M. Frost. An information-theoretic measure for document similarity. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 449–450. ACM, 2003.
3. D. Bär, C. Biemann, I. Gurevych, and T. Zesch. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1:*

Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 435–440. Association for Computational Linguistics, 2012.

4. M. Baroni and R. Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics, 2010.
5. J. Becker and D. Kuropka. Topic-based vector space model. In *Proceedings of the 6th International Conference on Business Information Systems*, pages 7–12, 2003.
6. S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.
7. D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.
8. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
9. D. G. Bridge. Defining and combining symmetric and asymmetric similarity measures. In *Advances in Case-Based Reasoning*, pages 52–63. Springer, 1998.
10. A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, 2001.
11. S. Clark, B. Coecke, and M. Sadrzadeh. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140, 2008.
12. W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003.
13. J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.
14. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.
15. L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese. Umbc ebiquity-core: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52, 2013.
16. G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332, 1998.
17. T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.
18. A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10, 2008.
19. J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.
20. S. Jimenez, C. Becerra, and A. Gelbukh. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the First Joint Conference*

on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 449–453. Association for Computational Linguistics, 2012.

21. R. Jones, D. Metzler, and F. Peng. Identifying and expanding implicitly temporally qualified queries, Apr. 10 2012. US Patent 8,156,111.
22. T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.
23. C. Leacock and M. Chodorow. Automated grammatical error detection. *Automated essay scoring: A cross-disciplinary perspective*, pages 195–207, 2003.
24. C.-H. Leung and Y.-Y. Chan. A natural language processing approach to automatic plagiarism detection. In *Proceedings of the 8th ACM SIGITE conference on Information technology education*, pages 213–218. ACM, 2007.
25. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
26. D. Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.
27. A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.
28. G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
29. R. Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.
30. C. Quirk, C. Brockett, and W. B. Dolan. Monolingual machine translation for paraphrase generation. In *EMNLP*, pages 142–149, 2004.
31. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.
32. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
33. M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers to non-factoid ques@incollectionlehmann2012framework, title=A Framework for Semantic-Based Similarity Measures for\ mathcal {ELH}-Concepts, author=Lehmann, Karsten and Turhan, Anni-Yasmin, booktitle=Logics in Artificial Intelligence, pages=307–319, year=2012, publisher=Springer tions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.
34. Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.
35. N. Zeichner, J. Berant, and I. Dagan. Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 156–160. Association for Computational Linguistics, 2012.