© 2017 IJSRSET | Volume 3 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Similarity Measures to identify Text Similarity

Manpreet Singh Lehal
Department of Computer Science and IT, Lyallpur Khalsa College, Jalandhar, Punjab, India

ABSTRACT

Similarity and distance measures compute the similarity between words, sentences and documents into numeric value similarity scores and bring out the degree of parallelism or distance from one another. A number of similarity measures have been used by the researchers but their effectiveness differs from one language pair to another and also on the basis of quality of the corpus. Selection of right similarity measure is crucial to the performance of translation tasks and extraction of parallel data

Keywords: Parallel Data, EBMT, NGD

I. INTRODUCTION

Identifying similarity in the text is the first step towards different tasks of information retrieval and machine translation. In fact, the concept of translation is based on similarity and finding equivalent words. Similarity measures are the various functions which help to compute the degree of similarity between texts. The text can be in the form of two documents in the same language or in different language or it can be a set of queries and documents. A number of measures have been used and proposed by researchers and no single measure suits the nature/requirement of content. Research has shown that the combination of different similarity measures gives efficient outputs. The similarity measures are computed in the values of (0, 1).

II. LITERATURE REVIEW

Researchers used semantic as well as syntactic methods and tested different combinations of similarities on the basis of vectors, word order, parts of speech, questions, edit distance, knowledge based, corpus based. (Alexander Strehl, Joydeep Ghosh, 2000) evaluated four popular similarity measures (Euclidean, Cosine, Pearson correlation and extended Jaccard) in conjunction with several clustering techniques (random, self-organizing feature map, hyper-graph partitioning, generalized kmeans, weighted graph partitioning), on high dimensional sparse data of news and business web documents. (Mandreoli et al., 2002) presented a syntactic approach for searching identical sentences and phrases in accordance with EBMT system. It used the proposition that the sentences are similar when they retain similar kind of form and content. (N. Liu et al., 2004) explored the problem of non-orthogonal space in finding similarities. (F. Chen et al., 2004) developed Story Link Detection methods that easily determined whether two stories were about the same relations which generally depended on the cosine similarity measure between these two stories. (Bani-Ahmad et al., 2005) used the publication similarity measures. The publication similarity measures are broadly divided into text based and citation-based measures. This approach evaluated the publication measures for accuracy, separability, and independence.

(Achananuparp et al., 2008) addressed the challenges of variability of natural language expression and similarity of sentences at Semantic level. They investigated the performance of fourteen similarity measures based upon word overlap, TF-IDF and linguistic levels using six evaluation metrics. (Aliguliyev, 2009) worked upon summarization techniques and demonstrated that the summarization depended on the similarity measure and NGD measures performed better than Euclidean measure. (Bandyopadhyay & Mallick, 2013) prepared a novel shortest path-based hybrid measure by combining information content with gene ontology graph Gene Ontology is an acyclic representation of semantic connections between terms. (Y. Jiang et al., 2015) proposed a new similarity computation approach which used a feature-based technique to assess the semantic similarity using Wikipedia. Working on the formal representation of concepts in the Wikipedia, they designed a framework to find out the similarity. (Xia et al., 2015) proposed a method to learn similarities that are generally called as cosine similarity ensemble. This paper proposed a cosine similarity ensemble (CSE) method for learning similarity. The CSE method is not limited to measuring similarity using only pattern vectors that start at the origin. In addition, the thresholds of these separate cosine similarity learners are adaptively determined.

III. TYPES OF MEASURES

Similarity measures identify similarity on the basis of lexical information or semantic information and they have been divided into three major categories: string-based, corpus-based and knowledge-based. (Mihalcea et al., 2006) (Gomaa et al., 2013). The string-based group uses lexical information and is further divided into character-based and term-based. Knowledge-based and corpus-based similarity measures use semantic information such as Latent semantic analysis (LSA) (Deerwester et al., 1990), pointwise mutual information (Turney, 2001) or lexical databases such as WordNet . In this study, the main focus is on term-based similarity measures because they are relatively more efficient on high dimensional data such as documents, and for the most part, they are used as standard approaches in IR for addressing many document similarity measurement problems. Term based similarity measures use statistics derived from texts to compute their similarity. Such statistics include Term frequency, inverse document frequency, document length etc.

3.1 Corpus-Based similarity

It uses semantic information and identify similarity between words on the basis of information attained from a large corpus. A Corpus is a large collection of written or spoken texts that is used for language research as shown in **Figure 1**.

Hyperspace Analogue to Language (HAL) (Lund, 1995) (Lund & Burgess, 1996) utilized information based on word co-occurrences and formed a word-by-word matrix. As the text is analyzed, a focus word is placed at the beginning of a ten-word window that records which neighboring words are counted as co-occurring. Matrix values are accumulated by weighting the co-occurrence inversely proportional to the distance from the focus word; closer neighboring words are thought to reflect more of the focus word's semantics and so are weighted higher. HAL also records word-ordering information by treating the co-occurrence differently based on whether the neighboring word appeared before or after the focus word.

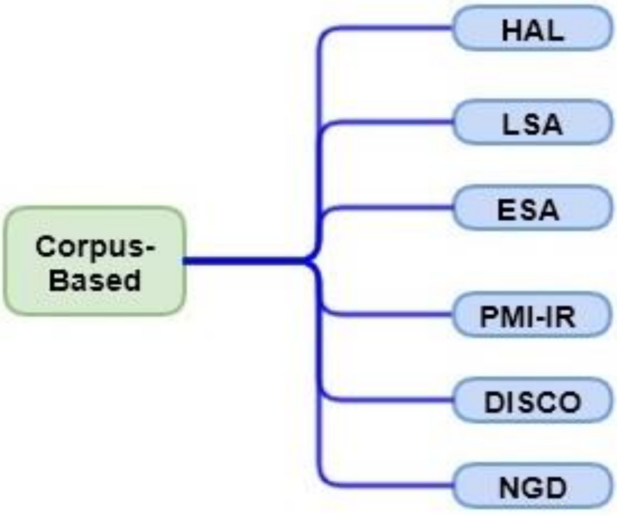

Figure 1 Corpus-Based Similarity Measures

Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) is another measure which forms a matrix and is based on the assumption that words having similar meanings occur in similar pieces of text. Vectors in the rows are compared using the cosine of the angle.

Generalized Latent Semantic Analysis (GLSA) (Matveeva et al., 2005) is an extension of the LSA approach and it uses term vectors in place of dual document terms. GLSA requires a similarity measure and a method of dimensionality reduction.

Explicit Semantic Analysis (ESA) (Gabrilovich et al., 2007) finds out the similarity between two random texts. In the Wikipedia-Based technique terms are converted into vectors and similarity is measured using cosine of the angle. A generalization of this CL-ESA represents documents as language independent vectors.

Pointwise Mutual Information - Information Retrieval (PMI-IR) (Turney, 2001) makes use of AltaVista's Advanced Search to calculate probabilities and works upon the frequency of co-occurrences of words.

Second-order co-occurrence pointwise mutual information **(SCO-PMI)** (Islam & Inkpen, 2008) takes into consideration the neighboring words in the target language and thus finds similarity for words that do not come together frequently.

Normalized Google Distance (NGD) (Cilibrasi & Vitanyi, 2007) finds out similarity using Google search engine and works on the idea that if two words are similar, they will occur together on many web pages.

Extracting **DIStributionally similar words using COoccurrences (DISCO)** (Kolb, 2009) calculates co-occurrences using Lin measure with a window of size ±3 words from large collection of texts. It works upon the notion that similar words occur in similar contexts.

3.2 Knowledge-Based Similarity

It is one of the semantic similarity measures that is based on identifying the degree of similarity between words using information derived from semantic networks like Wordnet (Miller et al., n.d.). WordNet is a large lexical database of Nouns, verbs, adjectives and adverbs grouped into synsets. Knowledge-based similarity measures are further divided into two groups as depicted in **Figure 2**

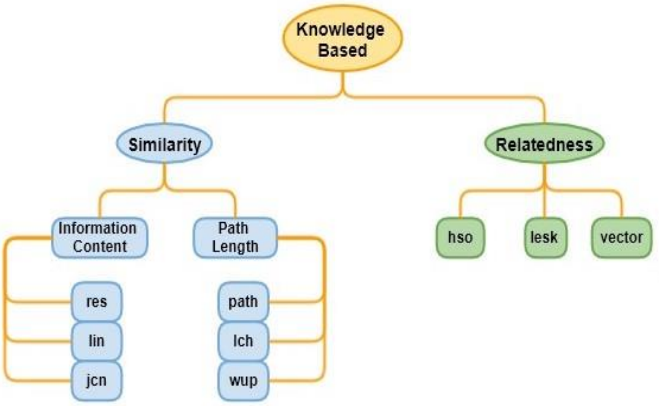

Figure 2 Knowledge-Based Similarity Measures

The concepts can be alike or similar to one another and they can be related to one another but not exactly similar. There are six measures of semantic similarity; three of them are based on information content: Resnik (res) (Resnik, 1995), Lin (lin) (Lin, 1998) and Jiang & Conrath (jcn) (Jiang & Conrath, 1997). The other three measures are based on path length: Leacock & Chodorow (lch) (Leacock & Chodorow, 1998), Wu & Palmer (wup) (Wu & Palmer, 1994) and Path Length (path).

3.2.1 BASED ON SIMILARITY

Path Length counts the edges between two words in the shortest path. The two words will be considered similar if the path between them is shorter as depicted on a thesaurus hierarchy graph. A thesaurus hierarchy graph is a tree drawn from a broad category of words to narrow category of words.

Leacock Chodorow (LCH) measures the negative log of the shortest path between two words divided by twice the total depth of the taxonomy and is an extended form of Path length similarity.

Wu and Palmer score denotes the similarity of two concepts based on their position, path length and the Information content of the Least Common Subsumer. The similarity is two times the depth of the two concepts' LCS divided by the product of the depths of the individual concepts.

Resnik came up with the concept of Information content which is the frequency count of concepts as found in a corpus of text. The similarity is based on the extent of common information. The more common is the information, the more similar are the words. Information Content is calculated for nouns and verbs, where the concepts are grouped in hierarchies.

Lin score combines the two-information content by taking the quotient between twice the IC of the concepts' LCS and the sum of the IC of the two concepts. It also utilizes thesaurus hierarchy depths like the Path Length Similarity. The results are dependent on the corpus which is used for information generation.

JCN, like Lin Similarity, uses the amount of information needed to state the commonality between the two concepts and the information needed to fully describe the terms. The similarity is calculated by taking the sum of the IC of the two concepts minus twice the IC of the concepts' LCS.

3.2.2 BASED ON RELATEDNESS

(Lesk, 1986) introduced gloss overlaps to perform word sense disambiguation based on the assumption that if the glosses of the concepts/words overlap, the concepts will be similar to one another. The **Lesk Algorithm** compares the glosses of the different senses of the word with the glosses of the neighboring words and selects the sense which has maximum overlaps. The limitation of the algorithm is that the dictionary glosses are rather short for developing comparisons. Absence of a single word will make a big difference in relatedness.

The HSO measure is path based, and establishes the relatedness between two concepts by trying to find a path between them that is neither too long nor that changes direction too often.

The vector measure creates a co–occurrence matrix of the glosses taken from a corpus and measures similarity using cosine. All the context vectors of the words in the gloss are averaged to obtain gloss vector.

3.3 String based similarity

String-based metrics consider the sentence as a sequence of characters and use string sequences to measure distance between two text strings for approximate string matching. It takes into account the intensity of the similarity between two strings and identifies the similar and dissimilar parts of the strings and factor them to generate the similarity. The metric that is used for measuring the distance between the text strings is called String metric and used for string matching and comparison. String-based Similarity is broadly classified into Character-based Similarity Measures and Term-based Similarity Measures as shown in **Figure 3**. The Character-based measures work upon on the contiguous chain of characters length which are present in both strings. Term- based measures depend on the term weights and frequencies.

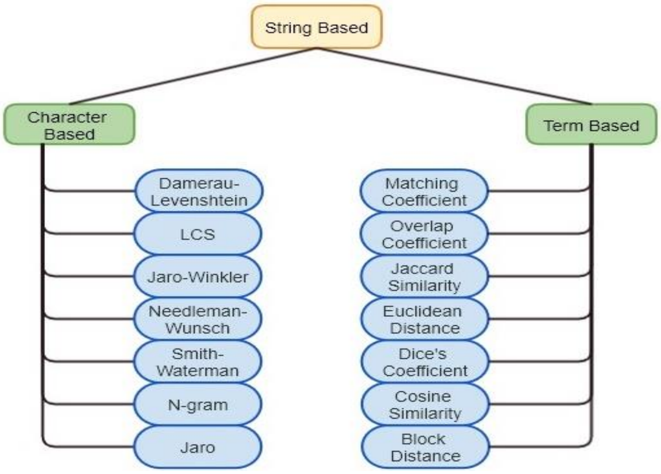

Figure 3 String based similarity Measures

3.3.1 CHARACTER BASED

Longest Common Sub String (LCS) algorithm considers that the similarity between two strings is based on the length of contiguous chain of characters that exist in both strings.

Damerau-Levenshtein defines distance between two strings by counting the minimum number of operations needed to transform one string into the other. The operations include insertion, deletion, substitution of a single character, or a transposition of two adjacent characters (Hall & Dowling, 1980) (Peterson, 1980).

Jaro is based on the number and order of the common characters between two strings; it utilizes the difference in spellings (Jaro, 1989) (Jaro, 1995). Jaro–Winkler is an extension of Jaro distance; it rates the strings using prefix length on a prefix scale (Winkler, 1990).

Needleman-Wunsch algorithm is an example of dynamic programming which finds the best alignment for the two sequences using global alignment. It will give best results if the two sequences are of similar length with a significant degree of similarity (Needleman & Wunsch, 1970).

Smith-Waterman is another example of dynamic programming which uses local alignment to find the best alignment over the conserved domain of two sequences. It is useful for dissimilar sequences which may have some similar motifs (Smith et al., 1981).

N-gram is a sub-sequence of n items from a given sequence of text. The algorithm compares the n-grams from each character or word in two strings and divides the number of similar n-grams by maximal number of n-grams to get the distance. (Barrón-Cedeno et al., 2010).

3.3.2 TERM BASED

The Dice similarity coefficient, simply Dice coefficient, is a statistical tool to measure similarity. It is the similarity between two sets of data. It is two times the number of common terms divided by the number of total terms as shown under:

The equation for this concept is:

$$2 \times \frac{|A| \cap |B|}{|A| + |B|}$$

where A and B are two sets and ∩ denotes the intersection of two sets, and is the number of common elements in the sets. (Yao et al., 2020)

Matching Coefficient is a statistical vector-based approach to measure the similarity and dissimilarity of elements. Given two strings with n binary attributes the coefficient is obtained by dividing number of matching attributes with the total number of attributes.(Heltshe, 1988)

Overlap coefficient measures the overlap between two finite sets. It is defined as the size of the intersection divided by the smaller of the size of the two sets. The two strings are similar if one is the subset of the other. (Vijaymeena & Kavitha, 2016)

The Block Distance or Manhattan Distance between two points (x_1, y_1) and (x_2, y_2) is represented as:

$$|x_1 - x_2| + |y_1 - y_2|$$

It measures the distance between two points in a grid like arrangement (Krause, 1986). It is measured as

$$\delta = \sum_{i=1}^{n} |x_i - y_i|$$

The Euclidean distance in either the plane or 3-dimensional space is simply the shortest distance between the two points. It is also called Pythagorean metric as it forms a right-angled triangle and is used to find the similarity between the two points. It helps to identify the sameness of vectors and hence find translation pairs in NLP. The higher the score, the less similar are the vectors.

Cosine similarity is a measure of the cosine of the angle between two non-zero vectors (arrays of the word count) projected in a multi-dimensional space, where both vectors are normalized to 1 and computes the similarity of documents independent of the size of the documents. The value of cosine of 0 degree is 1 and it is less than 1 for the angles between (0, pie) radians. The cosine similarity is used in positive space, where the output is clearly represented in binary forms of zeros and one.

Jaccard Similarity or Jaccard index is a measure to find similarity and difference of sample sets. Jaccard coefficient finds the similarity and is obtained by dividing the intersection by the union of the sets. Jaccard distance finds the dissimilarity and is obtained by subtracting the coefficient from 1. The value of dissimilarity will be 0.

IV. CONCLUSION

Each similarity measure has its own benefits and limitations. Hence researchers have experimented by using a combination of different measures to increase the efficiency of finding similarity between words/sentences/documents. Combining the scores of different similarity measures complement the features and give better results as this approach uses the best feature of each similarity measure.

V. REFERENCES

[1]. Achananuparp, P., Hu, X., & Shen, X. (2008). The evaluation of sentence similarity measures. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-540-85836-2_29

[2]. Alexander Strehl, Joydeep Ghosh, R. M. (2000). Impact of Similarity Measures on Web-page Clustering. Workshop of Arti Ial Intelligene for Web Searh, July 2000 by AAAI, 58--64.

[3]. Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2008.11.022

[4]. Bandyopadhyay, S., & Mallick, K. (2013). A new path based hybrid measure for gene ontology similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(1), 116–127.

[5]. Bani-Ahmad, S., Cakmak, A., Ozsoyoglu, G., & Hamdani, A. A. (2005). Evaluating Publication Similarity Measures. IEEE Data Engineering Bulletin.

[6]. Barrón-Cedeno, A., Rosso, P., Agirre, E., & Labaka, G. (2010). Plagiarism detection across distant language pairs. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 37–45.

[7]. Chen, F., Farahat, A., & Brants, T. (2004). Multiple similarity measures and source-pair information in story link detection. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 313–320.

[8]. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

[9]. Gabrilovich, E., Markovitch, S., & others. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. IJcAI, 7, 1606–1611.

[10]. Gomaa, W. H., Fahmy, A. A., & others. (2013). A survey of text similarity approaches. International Journal of Computer Applications, 68(13), 13–18.

[11]. Hall, P. A. V, & Dowling, G. R. (1980). Approximate string matching. ACM Computing Surveys (CSUR), 12(4), 381–402.

[12]. Heltshe, J. F. (1988). Jackknife Estimate of the Matching Coefficient of Similarity. Biometrics, 44(2), 447. https://doi.org/10.2307/2531858

[13]. Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 1–25.

[14]. Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414–420.

[15]. Jaro, M. A. (1995). Probabilistic linkage of large public health data files. Statistics in Medicine, 14(5–7), 491–498.

[16]. Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. ArXiv Preprint Cmp-Lg/9709008.

[17]. Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing & Management, 51(3), 215–234.

[18]. Kolb, P. (2009). Experiments on the difference between semantic similarity and relatedness. Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009), 81–88.

[19]. Krause, E. F. (1986). Taxicab geometry: An adventure in non-Euclidean geometry. Courier Corporation.

[20]. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic

analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.

[21]. Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database, 49(2), 265–283.

[22]. Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th Annual International Conference on Systems Documentation, 24–26.

[23]. Lin, D. (1998). Extracting collocations from text corpora. First Workshop on Computational Terminology, 57–63.

[24]. Liu, N., Zhang, B., Yan, J., Yang, Q., Yan, S., Chen, Z., Bai, F., & Ma, W. Y. (2004). Learning similarity measures in non-orthogonal space. International Conference on Information and Knowledge Management, Proceedings.

[25]. Lund, K. (1995). Semantic and associative priming in high-dimensional semantic space. Proc. of the 17th Annual Conferences of the Cognitive Science Society, 1995.

[26]. Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28(2), 203–208.

[27]. Mandreoli, F., Martoglia, R., & Tiberio, P. (2002). Searching Similar (Sub) Sentences for Example-Based Machine Translation. SEBD.

[28]. Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Generalized latent semantic analysis for term representation. Proc. of RANLP.

[29]. Mihalcea, R., Corley, C., Strapparava, C., Jiang, J. J., Conrath, D. W., Leacock, C., Chodorow, M., Wu, Z., Palmer, M., Hirst, G., St-Onge, D., others, Banerjee, S., Pedersen, T., Patwardhan, S., Hall, P. A. V, Dowling, G. R., Peterson, J. L., Jaro, M. A., … Inkpen, D. (2006). Approximate string matching. WordNet: An Electronic Lexical Database, 104(2), 1606–1611.

[30]. Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (n.d.). WordNet: An online lexical database. 1990. Int. J. Lexicograph, 3(4).

[31]. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453.

[32]. Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. Communications of the ACM, 23(12), 676–687.

[33]. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. ArXiv Preprint Cmp-Lg/9511007.

[34]. Smith, T. F., Waterman, M. S., & others. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, 147(1), 195–197.

[35]. Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. European Conference on Machine Learning, 491–502.

[36]. Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal, 3(2), 19–28.

[37]. Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

[38]. Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. ArXiv Preprint Cmp-Lg/9406033.

[39]. Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. Information Sciences, 307, 39–52.

[40]. Yao, A. D., Cheng, D. L., Pan, I., & Kitamura, F. (2020). Deep Learning in Neuroradiology: A Systematic Review of Current Algorithms and Approaches for the New Wave of Imaging Technology. Radiology: Artificial Intelligence, 2(2), e190026.