Constructing a Linearly Combined Similarity Measure with High Accuracy for Assessing the Similarity between Linguistic Items

Xiaolan Cui, Shuqin Cai, Yuchu Qin

Abstract—**Selecting local similarity measures and weighting their contributions to construct a linearly combined similarity measure with high accuracy is a key problem in assessing the similarity between linguistic items. Focusing on this problem, a number of approaches have been presented during the past few decades. Each approach can construct a linearly combined measure with high accuracy in its specific case. However, constructing such a measure for arbitrary cases remains a challenge. In this paper, an approach for constructing different linearly combined measures with high accuracy in different cases is proposed. This approach uses the Pearson correlation coefficient between the computed and judged similarities to quantify the accuracy of a linearly combined measure. For different cases, different local measures are selected and different weights are assigned by maximizing this coefficient. Thus the approach can ensure high accuracy in arbitrary cases. The effectiveness of the approach is theoretically proved and a set of experiments are carried out to verify the result of this proof. The proof and experiment results show that the linearly combined measure constructed by the approach has high accuracy and the weight assignment and local measure selection ways are helpful to improve the accuracy of the linearly combined measure.**

***Index Terms*—Similarity Assessment; Linear Combination; Similarity Measure; Linguistic Item; Weight; Accuracy.**

I. Introduction

The similarity of a pair of linguistic items refers to the degree of the nearness or proximity of them. It indicates how near the two linguistic items are since they share some aspects of their features [1]. Similarity is useful for many applications dealing with textual data such as data integration [2, 3], information retrieval [4, 5], knowledge extraction [6, 7], and ontology alignment [8, 9]. The core issue of using similarity in these applications is how to quantify the similarity between linguistic items. Focusing on this issue, a number of similarity measures have been proposed during the past few decades. These measures can be divided into sense-level, word-level, and sentence-level measures based on the linguistic levels they can be used to [10].

Sense-level measures mostly work on lexical databases like WordNet [11] and BabelNet [12]. They often consider lexical databases as semantic networks and calculate similarities on the basis of the structural attributes like path length and depth of these semantic networks. A comprehensive review about WordNet-based measures was presented in [13]. Word-level measures have attracted the most attention and gained the most popularity over the past decade among the three levels' measures [10]. Various word-level measures have emerged and were comprehensively surveyed in [14]. Sentence-level measures can be grouped into string-based, knowledge-based, and corpus-based measures [15]. Among them, string-based measures assess the similarities based on string sequences and character compositions, and knowledge-based (corpus-based) measures firstly split sentence into words and then calculate the semantic similarities of word pairs according to specific knowledge bases (corpuses) and finally linearly combine the similarities of word pairs to obtain the similarity of sentences.

In some practical applications of similarity measures, one may encounter the following case: Obtain the overall similarity of a pair of linguistic items by selecting two or more similarity measures at identical or different linguistic levels to respectively assess the similarities of each pair of the features of the two linguistic items and computing the weighted sum of the assessed similarities. As one example, the practical application in [16] defines the overall similarity of two words as a weighted sum of the similarities of their synonym sets, of their distinguishing features, and of their semantic neighborhoods, which are all calculated by word-level measures. As another example, in the practical application in [17], the overall similarity of two sentences is defined as a weighted sum of the similarities of their word semantics and of their word orders, which are assessed using a sense-level measure and a sentence-level measure, respectively. As can be seen from the two examples, two key questions in the assessment of the similarity between two linguistic items are: (1) How to select two or more similarity measures to respectively assess the similarities of each pair of their features? (2) How to assign weights to the selected measures to obtain an overall similarity measure? Since each selected measure is usually called a local similarity measure and the overall measure is often named linearly combined similarity measure in similarity assessment, the two questions can be summarized as one question: How to select local similarity measures and assign weights to the selected measures to construct a linearly combined similarity measure?

For this question, a number of researchers have mentioned or attempted to solve it in their studies. For examples, Rodríguez and Egenhofer [16] listed the linearly combined similarity measures with three groups of fixed weights and chose the one with the highest accuracy as the final linearly combined measure. Their experimental result showed that the chose linearly combined measure can obtain high accuracy in

Xiaolan Cui, School of Management, Huazhong University of Science and Technology, Wuhan 430074, China

Shuqin Cai, School of Management, Huazhong University of Science and Technology, Wuhan 430074, China

Yuchu Qin, School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

some specific cases. In [17], Li et al. designed a local similarity measure for word senses and a local measure for word orders and used fixed weights to aggregate these local measures to obtain a linearly combined similarity measure. In [18], Li et al. presented ten linearly combined similarity measures with ten groups of fixed weights and selected the one obtaining the highest accuracy as the ultimate linearly combined measure. Li et al. provide more candidate groups of weights and linearly combined measures to find out a linearly combined measure that obtains the highest accuracy. The found out measure however can hardly obtain high accuracy in any situations. Different from the studies of Rodríguez and Egenhofer [16] and Li et al. [17, 18], the study of Islam and Inkpen [19] used equal weights to construct a linearly combined similarity measure for calculating the similarity of sentences, which achieved high accuracy in specific cases. From the listed examples, it can be seen that existing solutions to the question can ensure high accuracy in specific situations but not in arbitrary situations. This is because the local measures and weights in these solutions are not simultaneously adjustable to maximize the accuracy.

To address this limitation, this paper proposes an adjustable approach that is capable of constructing a linearly combined similarity measure with high accuracy for assessing the similarity between linguistic items. Like most of existing measure construction approaches (e.g. the approaches [16−19]), this approach also leverages the Pearson correlation coefficient between the similarities of a certain number of randomly selected samples computed by the linearly combined measure and the similarities of these samples judged by a certain number of domain experts (the similarity of each sample is the mean value of the similarities of this sample judged by a certain number of domain experts) to quantify the accuracy of the linearly combined measure. For arbitrary cases, the approach can select different local similarity measures and assign different weights to the selected local measures to maximize the Pearson correlation coefficient so that it is capable of constructing a linearly combined measure with high accuracy.

The rest of the paper is organized as follows. An overview of related work is provided in Section II. The details of the proposed approach are explained in Section III. Section IV evaluates the effectiveness of the approach via theoretical proof and experimental verification. Conclusions are drawn in Section V.

II. Related Work

Generally, a linearly combined similarity measure for two linguistic items is defined as a weighted sum of two or more local similarity measures for the features of these two linguistic items. So the construction of a linearly combined similarity measure mainly includes the design or selection of local similarity measures and the assignment of the weights of local similarity measures. During the past few decades, a number of approaches for constructing a linearly combined similarity measure have been presented. These approaches can be classified into four groups on the basis of the selection way of the local similarity measures and the assignment way of the weights.

The first group consists of the approaches using fixed local measures and weights to construct a linearly combined measure. Representative examples for such approaches are the approaches presented by Rodríguez and Egenhofer [16], Li et al. [17, 18], and Islam and Inkpen [19]. By aggregating the local measures for words' synonym sets, distinguishing features, and semantic neighborhoods, the approach of Rodríguez and Egenhofer [16] constructed the following linearly combined measure for two words:

$$Sim(W_1,W_2) = \alpha Sim_{\mathbf{S}}(W_1,W_2) + \beta Sim_{\mathbf{F}}(W_1,W_2) + \gamma Sim_{\mathbf{N}}(W_1,W_2) \quad (1)$$

where α, β, and γ ($0 \le \alpha, \beta, \gamma \le 1$ and $\alpha+\beta+\gamma = 1$) respectively weight the contributions of the local measures $Sim_{\mathbf{S}}(W_1, W_2)$, $Sim_{\mathbf{F}}(W_1, W_2)$, and $Sim_{\mathbf{N}}(W_1, W_2)$, which are respectively used to assess the similarities of the synonym sets, distinguishing features, and semantic neighborhoods of words and are all instantiated by Tversky's measure [20]. In the experiment, Rodríguez and Egenhofer evaluated the accuracies of the linearly combined measures with three groups of fixed weights (α=β=γ=0.333), (α=γ=0.500; β=0), and (α=γ=0; β=1) and selected the linearly combined measure with the highest accuracy for a specific case. The experimental result suggested that the selected linearly combined measure can obtain high accuracy in each specific case. But the local measures and weights in this approach are not adjustable to maximize the accuracy. Similar to Rodríguez and Egenhofer's approach [16], the approaches of Li et al. [17, 18] and Islam and Inkpen [19] evaluated the accuracies of the linearly combined measures with fixed local measures and weights and selected the linearly combined measure obtaining the highest accuracy to assess the similarity of two words or two sentences.

In the second group of approaches, fixed local similarity measures and adjustable weights are used to establish a linearly combined similarity measure. Representative approaches in this group are the approaches of Petrakis et al. [21], Furlan et al. [22], and Li et al. [23]. The approach of Petrakis et al. [21] used a maximum function, which can be seen as a weight assignment function, to construct a linearly combined measure between words:

$$Sim(W_1,W_2) = \max\{Sim_{\mathbf{N}}(W_1,W_2), Sim_{\mathbf{D}}(W_1,W_2)\} \quad (2)$$

where W_1 and W_2 are not synonyms, and $Sim_{\mathbf{N}}(W_1, W_2)$ and $Sim_{\mathbf{D}}(W_1, W_2)$ are two local measures used to compute the similarities of the semantic neighborhoods and descriptions of W_1 and W_2 and are all calculated using Maedche and Staab's measure [24]. As can be seen from Expression (2), Petrakis et al.'s approach is essentially a weight assignment approach, which can obtain high accuracy in some specific cases. As this expression does not aim to maximize the accuracy of the linearly combined measure, it cannot ensure high accuracy in arbitrary situations. In Furlan et al.'s approach [22], two fixed local similarity measures are used to compute the string and semantic similarities and two adjustable weights are assigned to the two local measures, respectively. Furlan et al. pointed out that the two weights can be determined by experiment, but the details of such experiment are not explained. Li et al.'s approach [23] presented some local measures and a weight assignment function to construct a linearly combined measure

for a pair of words. This function assigned either a weight 0 or 1 to each local similarity measure according to some specific conditions. Their experiment result showed high accuracy and efficiency of the constructed linearly combined measure. However, no evidence has shown that the accuracy of the measure can remain high in different applications.

The third group consists of the approaches using adjustable local measures and fixed weights to construct a linearly combined measure. A typical example of these approaches is proposed by Jiang et al. [25]. This approach combined weighting and maximum functions to construct a linearly combined measure of two words:

$$Sim(W_1,W_2) = f(Sim_{\mathbf{S}}(W_1,W_2), Sim_{\mathbf{G}}(W_1,W_2), Sim_{\mathbf{A}}(W_1,W_2), Sim_{\mathbf{C}}(W_1,W_2)) \quad (3)$$

where f is a weighting or maximum function and $Sim_{\mathbf{S}}(W_1, W_2)$, $Sim_{\mathbf{G}}(W_1, W_2)$, $Sim_{\mathbf{A}}(W_1, W_2)$, and $Sim_{\mathbf{C}}(W_1, W_2)$ are four local measures used to assess the similarities of the synonyms, glosses, anchors, and categories of the two words W_1 and W_2 and are all computed through making a selection from Rodríguez and Egenhofer's measure [16] and Petrakis et al.'s measure [21]. It can be seen from Expression (3) that this approach is essentially a local measure adjustment approach, which chooses the local measures on the basis of the accuracy of the constructed linearly combined measure. The approach is capable of obtaining high accuracy in some specific cases, but it also cannot ensure high accuracy in arbitrary cases since the weights of the local measures do not aim to maximize the accuracy.

In the fourth group of approaches, both local similarity measures and weights are adjustable when establishing a linearly combined similarity measure. A typical example is Akmal et al.'s approach [26]. In this approach, two local measures in a linearly combined measure for two words are selected from Wu and Palmer's measure [27], Lin's measure [28], Dice's coefficient measure [29], Jaccard's coefficient measure [30], confidence measure [30], overlap coefficient measure [30], van der Weken et al.'s measure [31], Cosine measure, and Tversky's measure [20] based on the accuracy (the Pearson correlation coefficient between the computed and judged similarities) of the linearly combined measure. The weights of these two local measures were assigned by minimizing the residual sum of squares between the similarities of a certain number of randomly selected samples which are assessed by the linearly combined measure and the similarities of these samples judged by a certain number of domain experts:

$$\mathrm{rss}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{L},\boldsymbol{J}) = \sum_{i=1}^{N}\left[\sum_{j=1}^{2} w_j Sim_j(W_{i,1},W_{i,2}) - J(W_{i,1},W_{i,2})\right]^2 \quad (4)$$

where $Sim(W_1, W_2) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{L}$ (where vector $\boldsymbol{w} = [w_1, w_2]^{\mathrm{T}}$ is the weight vector and vector $\boldsymbol{L} = [Sim_1(W_1,W_2), Sim_2(W_1,W_2)]^{\mathrm{T}}$ is the local similarity measure vector) is a linearly combined similarity measure for two words, $\boldsymbol{J}$ is a column vector whose elements are the judged similarities, $(W_{i,1}, W_{i,2})$ $(i = 1, 2,\ldots, N)$ are N samples, $Sim_j(W_{i,1}, W_{i,2})$ $(j = 1, 2)$ is the j-th local similarity of $(W_{i,1}, W_{i,2})$, and $J(W_{i,1}, W_{i,2})$ is the judged similarity of $(W_{i,1}, W_{i,2})$. The approach offers a feasible way to adjust the local measures and the weights of local measures on the basis of the accuracy of the linearly combined measure. However, the residual sum of squares between the computed and judged similarities is not appropriate for quantifying the accuracy of a similarity measure since it just provides a sum of the differences. One cannot know if each difference is uniformly spread through the samples (the accuracy is high in this case) or if it is concentrated in a subset of the samples (the accuracy is low in this case). In practical applications, the Pearson (or Spearman) correlation coefficient between the computed and judged similarities is more appropriate than the residual sum of squares for quantifying the accuracy of a similarity measure. This is because a highly accurate similarity measure does not mean that the similarities computed by this measure have very small errors with the similarities judged by domain experts, but means that the computed similarities highly correlate with the judged similarities (if the judged similarities increase, the computed similarities increase in the same magnitude).

This paper continues the line of research in the fourth group of approaches and proposes an approach for constructing a linearly combined similarity measure with high accuracy for assessing the similarity between linguistic items. Compared to the existing approaches, the most important characteristic of the proposed approach is that the linearly combined measure constructed by it can have high accuracy in arbitrary cases. This is because the approach can adjust the local measures and weights to maximize the Pearson correlation coefficient of the similarities assessed by the linearly combined measure and the similarities judged by domain experts, while there is yet no evidence that the local measures and their weights in the existing approaches can simultaneously be adjusted to maximize such correlation coefficient.

III. Measure Construction Approach

It is common practice for a linguistic item to be described by a limited number of features. With such description, a local similarity measure is designed for each feature and a linearly combined similarity measure is defined as a weighted sum of all local similarity measures. Formally, let I_1 and I_2 be two linguistic items, $f_i(I_1)$ and $f_i(I_2)$ be respectively the i-th features of I_1 and I_2, and $Sim(f_i(I_1), f_i(I_2))$ be a similarity measure for $f_i(I_1)$ and $f_i(I_2)$ (i.e. the i-th local similarity measure). Then an overall similarity measure (i.e. a linearly combined similarity measure) for I_1 and I_2 is defined as:

$$Sim(I_1,I_2) = \sum_{i=1}^{n} w_i Sim(f_i(I_1), f_i(I_2)) \quad (5)$$

where $w_1, w_2,\ldots, w_n$ are respectively the weights of the local measures $Sim(f_1(I_1), f_1(I_2))$, $Sim(f_2(I_1), f_2(I_2)),\ldots,$ $Sim(f_n(I_1), f_n(I_2))$ such that $0 \le w_1, w_2,\ldots, w_n \le 1$ and $w_1 + w_2 +\ldots+ w_n = 1$. As can be seen from Expression (5), the value of $Sim(I_1, I_2)$ is determined by the values of $w_1, w_2,\ldots, w_n$ and $Sim(f_1(I_1), f_1(I_2))$, $Sim(f_2(I_1), f_2(I_2)),\ldots,$ $Sim(f_n(I_1), f_n(I_2))$. That is, the assignment of the n weights $w_1, w_2,\ldots, w_n$ and the selection of n local similarity measures to respectively assess $Sim(f_1(I_1), f_1(I_2))$, $Sim(f_2(I_1), f_2(I_2)),\ldots,$ $Sim(f_n(I_1), f_n(I_2))$ directly affect the accuracy of the linearly combined similarity measure $Sim(I_1, I_2)$. Hence, How to assign the n weights and how to select the n local measures, where both the weights and

measures can ensure high accuracy of $Sim(I_1, I_2)$, are two key questions in the construction of $Sim(I_1, I_2)$. In this section, the details of the solutions to these two questions are firstly explained. Then an algorithm for constructing a linearly combined measure with high accuracy is designed based on the explanations.

A. Assignment of Weights

Generally, the accuracy of a measure is quantified by the Pearson correlation coefficient between the similarities of a certain number of randomly selected samples which are computed by this measure and the similarities of the selected samples which are judged by a certain number of domain experts (the similarity of each sample is the mean value of the similarities of this sample judged by a certain number of domain experts). The greater this Pearson correlation coefficient, the higher the accuracy of the measure is. As a result, the n weights w_1, w_2,…, w_n can be assigned by maximizing the Pearson correlation coefficient between the similarities of a certain number of randomly selected samples computed by $Sim(I_1, I_2)$ and the similarities of these samples judged by a certain number of domain experts.

Formally, let N be the number of the randomly selected samples, $J(I_{i,1}, I_{i,2})$ (i = 1, 2,…, N) be the judged similarity of the i-th sample $(I_{i,1}, I_{i,2})$, $\boldsymbol{X} = [Sim(f_1(I_{i,1}), f_1(I_{i,2})), Sim(f_2(I_{i,1}), f_2(I_{i,2})),\ldots, Sim(f_n(I_{i,1}), f_n(I_{i,2}))]^{\mathrm{T}}$ be a matrix consisting of $n\times N$ local similarities, $\boldsymbol{Y} = [J(I_{i,1}, I_{i,2})]^{\mathrm{T}}$ be a vector consisting of the N judged similarities, and $\boldsymbol{\omega} = [\omega_1, \omega_2,\ldots, \omega_n]^{\mathrm{T}}$ be a vector. The Pearson correlation coefficient between the computed similarities and the judged similarities is the Pearson correlation coefficient between $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\begin{aligned}\mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{Y}) &= \frac{\boldsymbol{\omega}^{\mathrm{T}}\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\boldsymbol{\omega}^{\mathrm{T}}\mathrm{cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{\omega}}\sqrt{\mathrm{cov}(\boldsymbol{Y},\boldsymbol{Y})}} \\ &= \frac{\boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\boldsymbol{XY}}}{\sqrt{\boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}\boldsymbol{\omega}}\sqrt{\Sigma_{\boldsymbol{YY}}}}\end{aligned} \tag{6}$$

where cov is a covariance function and $\Sigma_{\boldsymbol{XX}}$, $\Sigma_{\boldsymbol{XY}}$, and $\Sigma_{\boldsymbol{YY}}$ are respectively the following $n\times n$, $n\times 1$, and 1×1 matrices:

$$\Sigma_{\boldsymbol{XX}} = \begin{bmatrix} \mathrm{cov}(S_1,S_1) & \mathrm{cov}(S_1,S_2) & \cdots & \mathrm{cov}(S_1,S_n) \\ \mathrm{cov}(S_2,S_1) & \mathrm{cov}(S_2,S_2) & \cdots & \mathrm{cov}(S_2,S_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(S_n,S_1) & \mathrm{cov}(S_n,S_2) & \cdots & \mathrm{cov}(S_n,S_n) \end{bmatrix} \tag{7}$$

$$\Sigma_{\boldsymbol{XY}} = \begin{bmatrix} \mathrm{cov}(Sim(f_1(I_{i,1}), f_1(I_{i,2})), J(I_{i,1}, I_{i,2})) \\ \mathrm{cov}(Sim(f_2(I_{i,1}), f_2(I_{i,2})), J(I_{i,1}, I_{i,2})) \\ \vdots \\ \mathrm{cov}(Sim(f_n(I_{i,1}), f_n(I_{i,2})), J(I_{i,1}, I_{i,2})) \end{bmatrix} \tag{8}$$

$$\Sigma_{\boldsymbol{YY}} = \left[\mathrm{cov}(J(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2}))\right] \tag{9}$$

where S_j (j = 1, 2,…, n) are $Sim(f_j(I_{i,1}), f_j(I_{i,2}))$. To solve $\boldsymbol{\omega}$ that can maximize $\mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{Y})$, a canonical correlation analysis method [32] is used and the solving process is as follows.

Firstly, let $\boldsymbol{\alpha} = \sqrt{\Sigma_{\boldsymbol{XX}}}\boldsymbol{\omega}$ and $\boldsymbol{\beta} = \sqrt{\Sigma_{\boldsymbol{YY}}}$. Then Expression (6) can be converted to the following expression:

$$\mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{Y}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}^{-1/2}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1/2}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}}} = \frac{\boldsymbol{\gamma}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}}} \tag{10}$$

According to the Cauchy-Schwarz inequality, the following inequality is achieved:

$$\boldsymbol{\gamma} \le \sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}^{-1/2}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}\Sigma_{\boldsymbol{XX}}^{-1/2}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}} \tag{11}$$

where $\Sigma_{\boldsymbol{YX}}$ is the following $1\times n$ matrix:

$$\Sigma_{\boldsymbol{YX}} = \begin{bmatrix} \mathrm{cov}(J(I_{i,1}, I_{i,2}), Sim(f_1(I_{i,1}), f_1(I_{i,2}))) \\ \mathrm{cov}(J(I_{i,1}, I_{i,2}), Sim(f_2(I_{i,1}), f_2(I_{i,2}))) \\ \vdots \\ \mathrm{cov}(J(I_{i,1}, I_{i,2}), Sim(f_n(I_{i,1}), f_n(I_{i,2}))) \end{bmatrix}^{\mathrm{T}} \tag{12}$$

According to Expression (10) and Expression (11), the following inequality is obtained:

$$\mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{Y}) \le \frac{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}^{-1/2}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}\Sigma_{\boldsymbol{XX}}^{-1/2}\boldsymbol{\alpha}}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}}} \tag{13}$$

As can be seen from Expression (13), the maximum value of the correlation coefficient $\mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{Y})$ is attained if and only if $\boldsymbol{\alpha}$ is the eigenvector with the maximum eigenvalue for the matrix $\Sigma_{\boldsymbol{XX}}^{-1/2}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}\Sigma_{\boldsymbol{XX}}^{-1/2}$. Therefore, the solution is: $\boldsymbol{\omega}$ is an eigenvector with the maximum eigenvalue for the matrix $\Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}$.

Now although the solving process of the vector $\boldsymbol{\omega}$ ends, the elements in the solved $\boldsymbol{\omega}$ are not the real weights. This is because some of these elements may be smaller than 0 and the sum of the elements that are not smaller than 0 is often not equal to 1. To solve the real weights, $\boldsymbol{\omega}$ is normalized as follow: For all $\omega_j < 0$ (j = 1, 2,…, n), let $w_j = 0$ (this indicates that $Sim(f_j(I_1), f_j(I_2))$ make no contribution to $Sim(I_1, I_2)$ in this case, i.e. f_j is a non-distinguishing feature of I_1 and I_2 in this case) and so a new linearly combined measure is obtained. Now re-solve the new vector $\boldsymbol{\omega}$ until all $\omega_j \ge 0$. Finally, let $w_j = \omega_j/(\omega_1+\omega_2+\ldots+\omega_n)$. A vector $\boldsymbol{w} = [w_1, w_2,\ldots, w_n]^{\mathrm{T}}$ whose elements are real weights obtained by such normalization is a vector that can maximize $\mathrm{pcc}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{Y})$ because:

$$\begin{aligned}\mathrm{pcc}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{Y}) &= \left(\boldsymbol{w}^{\mathrm{T}}\Sigma_{\boldsymbol{XY}}\right)\Big/\left(\sqrt{\boldsymbol{w}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}\boldsymbol{w}}\sqrt{\Sigma_{\boldsymbol{YY}}}\right) \\ &= \frac{(\boldsymbol{\omega}\Big/\sum_{j=1}^{n}\omega_j)^{\mathrm{T}}\Sigma_{\boldsymbol{XY}}}{\sqrt{(\boldsymbol{\omega}\Big/\sum_{j=1}^{n}\omega_j)^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}(\boldsymbol{\omega}\Big/\sum_{j=1}^{n}\omega_j)}\sqrt{\Sigma_{\boldsymbol{YY}}}} \\ &= \frac{\boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\boldsymbol{XY}}}{\sqrt{\boldsymbol{\omega}^{\mathrm{T}}\Sigma_{\boldsymbol{XX}}\boldsymbol{\omega}}\sqrt{\Sigma_{\boldsymbol{YY}}}} = \mathrm{pcc}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{Y})\end{aligned} \tag{14}$$

B. Selection of Local Measures

Naturally, the n local measures for respectively assessing $Sim(f_1(I_1), f_1(I_2))$, $Sim(f_2(I_1), f_2(I_2))$,…, $Sim(f_n(I_1), f_n(I_2))$ can be selected also through maximizing the Pearson correlation coefficient between the similarities of a certain number of randomly selected samples computed by $Sim(I_1, I_2)$ and the similarities of these samples judged by a certain number of domain experts.

Assume $Sim_{1,1}$, $Sim_{1,2}$,…, Sim_{1,m_1} are m_1 measures which can be selected to assess $Sim(f_1(I_1), f_1(I_2))$, $Sim_{2,1}$, $Sim_{2,2}$,…, Sim_{2,m_2} are m_2 measures which can be selected to assess $Sim(f_2(I_1), f_2(I_2))$,…, $Sim_{n,1}$, $Sim_{n,2}$,…, Sim_{n,m_n} are m_n measures which can be selected to assess $Sim(f_n(I_1), f_n(I_2))$.

Then $m_1m_2\ldots m_n$ linearly combined measures for assessing the similarity of I_1 and I_2 are obtained:

$$\begin{cases} Sim_1(I_1,I_2) = w_{1,1}Sim_{1,1} + w_{1,2}Sim_{2,1} + \cdots + w_{1,n}Sim_{n,1} \\ Sim_2(I_1,I_2) = w_{2,1}Sim_{1,1} + w_{2,2}Sim_{2,1} + \cdots + w_{2,n}Sim_{n,2} \\ \vdots \\ Sim_{m_1m_2\cdots m_n}(I_1,I_2) = w_{m_1m_2\cdots m_n,1}Sim_{1,m_1} + w_{m_1m_2\cdots m_n,2}Sim_{2,m_2} \\ \qquad\qquad + \cdots + w_{m_1m_2\cdots m_n,n}Sim_{n,m_n} \end{cases} \tag{15}$$

If N randomly selected samples $(I_{i,1}, I_{i,2})$ $(i = 1, 2,\ldots, N)$ and their judged similarities $J(I_{i,1}, I_{i,2})$ are given, the values of the weights $w_{k,1}$, $w_{k,2},\ldots$, $w_{k,n}$ $(k = 1, 2,\ldots, m_1m_2\ldots m_n)$ and the linearly combined similarities $Sim_k(I_{i,1}, I_{i,2})$ will be computed successively. Let $\boldsymbol{U} = [Sim_k(I_{i,1}, I_{i,2})]^{\mathrm{T}}$ be a vector which consists of the N linearly combined similarities and $\boldsymbol{V} = [J(I_{i,1}, I_{i,2})]^{\mathrm{T}}$ be a vector which consists of the N judged similarities. The Pearson correlation coefficients between the computed similarities of the N samples $Sim_k(I_{i,1}, I_{i,2})$ and the judged similarities of the N samples $J(I_{i,1}, I_{i,2})$ are calculated using the following expression:

$$\mathrm{pcc}(Sim_k(I_{i,1},I_{i,2}), J(I_{i,1},I_{i,2})) = \frac{\mathrm{cov}(\boldsymbol{U},\boldsymbol{V})}{\sqrt{\mathrm{cov}(\boldsymbol{U},\boldsymbol{U})}\sqrt{\mathrm{cov}(\boldsymbol{V},\boldsymbol{V})}} \tag{16}$$

where cov is a covariance function. Now the n local similarity measures in one of the $m_1m_2\ldots m_n$ linearly combined similarity measures that has the greatest Pearson correlation coefficient are selected. With these n local similarity measures and the assigned weights, a linearly combined similarity measure for I_1 and I_2 with high accuracy is constructed.

C. Measure Construction Algorithm

Based on the above explanations of how to assign weights and how to select local similarity measures, an algorithm for constructing a linearly combined similarity measure with high accuracy is designed as follow:

Linearly combined measure construction algorithm

Input: The number of the contribution components n
N randomly selected samples $(I_{i,1}, I_{i,2})$ $(i = 1, 2,\ldots, N)$
The judged similarities of these N samples $J(I_{i,1}, I_{i,2})$
m_1 measures $Sim_{1,1}, Sim_{1,2},\ldots, Sim_{1,m_1}$ for $Sim(f_1(I_1), f_1(I_2))$
m_2 measures $Sim_{2,1}, Sim_{2,2},\ldots, Sim_{2,m_2}$ for $Sim(f_2(I_1), f_2(I_2))$
……
m_n measures $Sim_{n,1}, Sim_{n,2},\ldots, Sim_{n,m_n}$ for $Sim(f_n(I_1), f_n(I_2))$
Output: A linearly combined measure with high accuracy for (I_1, I_2)

1 **for integer** $i \leftarrow 1$ **to** N **do**
Compute $Sim(f_1(I_{i,1}), f_1(I_{i,2}))$ using $Sim_{1,1}, Sim_{1,2},\ldots, Sim_{1,m_1}$
Compute $Sim(f_2(I_{i,1}), f_2(I_{i,2}))$ using $Sim_{2,1}, Sim_{2,2},\ldots, Sim_{2,m_2}$
……
Compute $Sim(f_n(I_{i,1}), f_n(I_{i,2}))$ using $Sim_{n,1}, Sim_{n,2},\ldots, Sim_{n,m_n}$
end for
2 **for integer** $i \leftarrow 1$ **to** N **do**
Compute all the elements of the matrix $\Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}$
end for
3 Solve the vector $\boldsymbol{\omega} = [\omega_1, \omega_2,\ldots, \omega_n]^{\mathrm{T}}$
for integer $j \leftarrow 1$ **to** n **do**
if $\omega_j < 0$ **then**
$w_j \leftarrow 0$ and return to 2
end if
end for
4 $\boldsymbol{w} \leftarrow \boldsymbol{\omega}/(\omega_1+\omega_2+\ldots+\omega_n)$
for integer $k \leftarrow 1$ **to** $m_1m_2\ldots m_n$ **do**
for integer $i \leftarrow 1$ **to** N **do**
Compute $Sim_k(I_{i,1}, I_{i,2})$ using $Sim_k(I_1, I_2)$
end for
end for
5 **double** $pcc\_max \leftarrow 0$
integer $p \leftarrow 0$
for integer $k \leftarrow 1$ **to** $m_1m_2\ldots m_n$ **do**
Compute $\mathrm{pcc}(Sim_k(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2}))$
if $\mathrm{pcc}(Sim_k(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2})) > pcc\_max$ **then**
$pcc\_max \leftarrow \mathrm{pcc}(Sim_k(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2}))$
$p \leftarrow k$
end if
end for
6 Output the linearly combined measure is $Sim_p(I_1, I_2)$

The designed algorithm takes as input N randomly selected samples and their judged similarities and a certain number of candidate measures that can be chose to compute the local similarities $Sim(f_j(I_{i,1}), f_j(I_{i,2}))$ $(i = 1, 2,\ldots, N; j = 1, 2,\ldots, n)$. It returns as output a linearly combined measure with high accuracy for the two linguistic items (I_1, I_2). The main ideas behind the algorithm are informally described as follows. The algorithm firstly uses the input measures to assess the local similarities $Sim(f_j(I_{i,1}), f_j(I_{i,2}))$. Then it successively computes the values of the elements of the matrix $\Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{YY}}^{-1}\Sigma_{\boldsymbol{YX}}$, the intermediate vector $\boldsymbol{\omega}$, and the real weight vector $\boldsymbol{w}$. After that, the algorithm uses each one of the $m_1m_2\ldots m_n$ measures to assess the similarities of the N samples. Finally, the Pearson correlation coefficients between the assessed similarities $Sim_k(I_{i,1}, I_{i,2})$ $(k = 1, 2,\ldots, m_1m_2\ldots m_n)$ and the judged similarities $J(I_{i,1}, I_{i,2})$ are computed and the measure $Sim_p(I_1, I_2)$, where $\mathrm{pcc}(Sim_p(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2}))$ is the greatest one among the k correlation coefficients, is output as the constructed linearly combined measure with high accuracy.

The time complexity of the designed algorithm is analyzed as follow. It appears that step 4 needs the largest computation amount among all six steps. Thus, the time complexity of the algorithm is $O(m_1m_2\ldots m_nN)$, which contains three cases: (1) If $m_1m_2\ldots m_n$ is far greater than N, the time complexity is $O(m_1m_2\ldots m_n)$. (2) If $m_1m_2\ldots m_n$ and N are in the same magnitude, the time complexity is $O(N^2)$. (3) If $m_1m_2\ldots m_n$ is far less than N, the time complexity is $O(N)$.

IV. Evaluation

This section firstly provides a theoretical proof of the effectiveness of the proposed approach. It then verifies the proof result by a set of experiments. Finally, the results of the theoretical proof and experiments are analyzed.

A. Theoretical Proof

The following theoretical proof proves that the accuracy of the linearly combined similarity measure constructed by the designed algorithm is the highest one among the accuracies of

all possible linear combinations of the n local similarity measures in each of the $m_1m_2\ldots m_n$ linearly combined measures in Expression (15). That is, the inequality $\text{pcc}(Sim_p(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2})) \geq \text{pcc}(Sim_{k,q,r}(I_{i,1}, I_{i,2}), J(I_{i,1}, I_{i,2}))$ holds for all $k = 1, 2,\ldots, m_1m_2\ldots m_n$ and q, $r = 1, 2,\ldots, n$, where q stands for the number of the contribution components in a linear combination of the n local measures in one of the $m_1m_2\ldots m_n$ linearly combined measures in Expression (15), and r denotes the sequence number of the linear combinations with the same number of contribution components. As an example, all possible linear combinations of the n local measures in the linearly combined measure $Sim_1(I_1, I_2)$ in Expression (15) are as follows:

$Sim_{1,1,1}(I_{i,1}, I_{i,2}) = Sim_{1,1}(f_1(I_{i,1}), f_1(I_{i,2}))$
$Sim_{1,1,2}(I_{i,1}, I_{i,2}) = Sim_{2,1}(f_2(I_{i,1}), f_2(I_{i,2}))$
.....
$Sim_{1,1,(n!)/[1!(n-1)!]}(I_{i,1}, I_{i,2}) = Sim_{n,1}(f_n(I_{i,1}), f_n(I_{i,2}))$
$Sim_{1,2,1}(I_{i,1}, I_{i,2}) = w_{1,2,1,1}Sim_{1,1}(f_1(I_{i,1}), f_1(I_{i,2})) + w_{1,2,1,2}Sim_{2,1}(f_2(I_{i,1}), f_2(I_{i,2}))$
$Sim_{1,2,2}(I_{i,1}, I_{i,2}) = w_{1,2,2,1}Sim_{1,1}(f_1(I_{i,1}), f_1(I_{i,2})) + w_{1,2,2,2}Sim_{3,1}(f_3(I_{i,1}), f_3(I_{i,2}))$
.....
$Sim_{1,2,(n!)/[2!(n-2)!]}(I_{i,1}, I_{i,2}) = w_{1,2,(n!)/[2!(n-2)!],1}Sim_{n-1,1}(f_{n-1}(I_{i,1}), f_{n-1}(I_{i,2})) + w_{1,2,(n!)/[2!(n-2)!],2}Sim_{n,1}(f_n(I_{i,1}), f_n(I_{i,2}))$
.....
$Sim_{1,n,(n!)/[n!(n-n)!]}(I_{i,1}, I_{i,2}) = w_{1,n,(n!)/[n!(n-n)!],1}Sim(f_1(I_{i,1}), f_1(I_{i,2})) + w_{1,n,(n!)/[n!(n-n)!],2}Sim(f_2(I_{i,1}), f_2(I_{i,2})) + \ldots + w_{1,n,(n!)/[n!(n-n)!],n}Sim(f_n(I_{i,1}), f_n(I_{i,2}))$

Proof. Let vectors $\boldsymbol{J} = [J(I_{i,1}, I_{i,2})]^{\mathrm{T}}$, $\boldsymbol{X} = [Sim_p(I_{i,1}, I_{i,2})]^{\mathrm{T}}$, $\boldsymbol{Y_1} = [Sim_1(I_{i,1}, I_{i,2})]^{\mathrm{T}}$, $\boldsymbol{Y_2} = [Sim_2(I_{i,1}, I_{i,2})]^{\mathrm{T}},\ldots$, $\boldsymbol{Y_t} = [Sim_t(I_{i,1}, I_{i,2})]^{\mathrm{T}}$ ($t = m_1m_2\ldots m_n$). According to the designed algorithm, $\text{pcc}(\boldsymbol{X}, \boldsymbol{J})$ is the greatest Pearson correlation coefficient among all $\text{pcc}(\boldsymbol{Y_k}, \boldsymbol{J})$ ($k = 1, 2,\ldots, m_1m_2\ldots m_n$). Therefore, the inequality "$\text{pcc}(\boldsymbol{X}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Y_k}, \boldsymbol{J})$" holds.

For the linearly combined measure $Sim_1(I_1, I_2)$ in Expression (15), let vectors:

$\boldsymbol{Z_{1,1,1}} = [Sim_{1,1,1}(I_{i,1}, I_{i,2})]^{\mathrm{T}}$, $\boldsymbol{Z_{1,1,2}} = [Sim_{1,1,2}(I_{i,1}, I_{i,2})]^{\mathrm{T}},\ldots$,
$\boldsymbol{Z_{1,1,(n!)/[1!(n-1)!]}} = [Sim_{1,1,(n!)/[1!(n-1)!]}(I_{i,1}, I_{i,2})]^{\mathrm{T}}$,
$\boldsymbol{Z_{1,2,1}} = [Sim_{1,2,1}(I_{i,1}, I_{i,2})]^{\mathrm{T}}$, $\boldsymbol{Z_{1,2,2}} = [Sim_{1,2,2}(I_{i,1}, I_{i,2})]^{\mathrm{T}},\ldots$,
$\boldsymbol{Z_{1,2,(n!)/[2!(n-2)!]}} = [Sim_{1,2,(n!)/[2!(n-2)!]}(I_{i,1}, I_{i,2})]^{\mathrm{T}},\ldots$,
$\boldsymbol{Z_{1,n,(n!)/[n!(n-n)!]}} = [Sim_{1,n,(n!)/[n!(n-n)!]}(I_{i,1}, I_{i,2})]^{\mathrm{T}}$.

Then $\boldsymbol{Z_{1,n,(n!)/[n!(n-n)!]}} = \boldsymbol{Y_1}$ because the linear combination is unique when the number of the contribution components is n. According to the designed algorithm, the weight vector $\boldsymbol{w_1} = [w_{1,1}, w_{1,2},\ldots, w_{1,n}]^{\mathrm{T}}$ in $Sim_1(I_1, I_2)$ is solved by maximizing $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$. So $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ is the greatest Pearson correlation coefficient among all $\text{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$ ($u = 1, 2,\ldots, n$; $v = 1, 2,\ldots, (n!)/[1!(n\text{-}1)!]$), which contains the following cases:

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $w_{1,1} = 1$ and $w_{1,2} = w_{1,3} = \ldots = w_{1,n} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,1,1}}, \boldsymbol{J})$";

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $w_{1,2} = 1$ and $w_{1,1} = w_{1,3} = \ldots = w_{1,n} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,1,2}}, \boldsymbol{J})$";

.....

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $w_{1,n} = 1$ and $w_{1,1} = w_{1,2} = \ldots = w_{1,n\text{-}1} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,1,(n!)/[1!(n-1)!]}}, \boldsymbol{J})$";

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $0 < w_{1,1}, w_{1,2} < 1$ and $w_{1,3} = w_{1,4} = \ldots = w_{1,n} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,2,1}}, \boldsymbol{J})$";

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $0 < w_{1,1}, w_{1,3} < 1$ and $w_{1,2} = w_{1,4} = \ldots = w_{1,n} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,2,2}}, \boldsymbol{J})$";

.....

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $0 < w_{1,n\text{-}1}, w_{1,n} < 1$ and $w_{1,1} = w_{1,2} = \ldots = w_{1,n\text{-}2} = 0$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,2,(n!)/[2!(n-2)!]}}, \boldsymbol{J})$";

.....

If $\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J})$ obtains the greatest value when $0 < w_{1,1}, w_{1,2}, \ldots, w_{1,n} < 1$, then "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) = \text{pcc}(\boldsymbol{Z_{1,n,(n!)/[n!(n-n)!]}}, \boldsymbol{J}) > \text{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$".

It can be concluded from the above cases that "$\text{pcc}(\boldsymbol{Y_1}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$". For the remaining linearly combined measures in Expression (15), i.e. $Sim_2(I_1, I_2)$, $Sim_3(I_1, I_2),\ldots, Sim_t(I_1, I_2)$ ($t = m_1m_2\ldots m_n$), it can be proved that "$\text{pcc}(\boldsymbol{Y_2}, \boldsymbol{J})$, $\text{pcc}(\boldsymbol{Y_3}, \boldsymbol{J}),\ldots$, $\text{pcc}(\boldsymbol{Y_t}, \boldsymbol{J})$ are greater than or equal to the Pearson correlation coefficients between the similarities computed by all possible linear combinations of their respective local similarity measures and the judged similarities".

Based on the proved conclusions "$\text{pcc}(\boldsymbol{X}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Y_k}, \boldsymbol{J})$" and "$\text{pcc}(\boldsymbol{Y_k}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$", the inequality "$\text{pcc}(\boldsymbol{X}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Y_k}, \boldsymbol{J}) \geq \text{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$" holds. □

B. Experimental Verification

Generally, an experiment for evaluating the effectiveness of a similarity measure can be carried out using an identical benchmark consisting of a certain number of sense pairs, word pairs, or text pairs and their judged similarities. During the past few decades, various benchmarks at different linguistic levels (i.e. sense, word, and text levels) have been designed, where the most widely used benchmarks are SENSEVAL-2 (sense-level) [33], OntoNotes (sense-level) [34], RG-65 (word-level) [35], YP-130 (word-level) [36], WordSimilarity-353 (word-level) [37], and MSRvid, OnWN, MSRpar, SMTeuroparl, and SMTnews (text-level) in the SemEval-2012 task 6 [38]. Among these benchmarks, the benchmarks that contain the same number of linguistic item pairs are MSRvid, OnWN, and MSRpar (each of them contains 750 text pairs). Since OnWN is not provided with any training data, MSRpar is not generic (belongs to the newswire genre), and MSRvid is provided with training data and generic, MSRvid will be used to verify the theoretical proof result in the following 8 experiments.

Based on the selected benchmark, it is assumed that the purpose of the 8 experiments is to construct a linearly combined measure $Sim(I_1, I_2)$, which totally has three contribution components $Sim(f_1(I_1), f_1(I_2))$, $Sim(f_2(I_1), f_2(I_2))$, and $Sim(f_3(I_1), f_3(I_2))$, to assess the similarities of text pairs. Meanwhile, it is also assumed that Carrillo et al.'s BUAPRUN-1 measure ($Sim_{\mathbf{CAR}}$) [39] and Yeh and Agirre's System 2 measure ($Sim_{\mathbf{YEH}}$) [40] are the candidate local measures that can be selected to compute $Sim(f_1(I_1), f_1(I_2))$, Croce et al.'s Sys_2 measure ($Sim_{\mathbf{CRO}}$) [41] and Malandrakis et al.'s Hierarchical measure ($Sim_{\mathbf{MAL}}$) [42] are the candidate local measures that can be selected to compute $Sim(f_2(I_1), f_2(I_2))$, and Caputo et al.'s UNIBA-LSARI measure ($Sim_{\mathbf{CAP}}$) [43] and Banea et al.'s IndividualRegression measure ($Sim_{\mathbf{BAN}}$) [44] are the candidate local measures that can be selected to compute $Sim(f_3(I_1), f_3(I_2))$. According to these conditions and Expression (15), 2×2×2 = 8 linearly combined measures can be constructed as follows:

$$\begin{cases} Sim_1(I_1,I_2) = w_{1,1}Sim_{\mathbf{CAR}} + w_{1,2}Sim_{\mathbf{CRO}} + w_{1,3}Sim_{\mathbf{CAP}} \\ Sim_2(I_1,I_2) = w_{2,1}Sim_{\mathbf{CAR}} + w_{2,2}Sim_{\mathbf{CRO}} + w_{2,3}Sim_{\mathbf{BAN}} \\ Sim_3(I_1,I_2) = w_{3,1}Sim_{\mathbf{CAR}} + w_{3,2}Sim_{\mathbf{MAL}} + w_{3,3}Sim_{\mathbf{CAP}} \\ Sim_4(I_1,I_2) = w_{4,1}Sim_{\mathbf{CAR}} + w_{4,2}Sim_{\mathbf{MAL}} + w_{4,3}Sim_{\mathbf{BAN}} \\ Sim_5(I_1,I_2) = w_{5,1}Sim_{\mathbf{YEH}} + w_{5,2}Sim_{\mathbf{CRO}} + w_{5,3}Sim_{\mathbf{CAP}} \\ Sim_6(I_1,I_2) = w_{6,1}Sim_{\mathbf{YEH}} + w_{6,2}Sim_{\mathbf{CRO}} + w_{6,3}Sim_{\mathbf{BAN}} \\ Sim_7(I_1,I_2) = w_{7,1}Sim_{\mathbf{YEH}} + w_{7,2}Sim_{\mathbf{MAL}} + w_{7,3}Sim_{\mathbf{CAP}} \\ Sim_8(I_1,I_2) = w_{8,1}Sim_{\mathbf{YEH}} + w_{8,2}Sim_{\mathbf{MAL}} + w_{8,3}Sim_{\mathbf{BAN}} \end{cases} \quad (17)$$

In each of the 8 experiments, the weights in each linearly combined measure in Expression (17) and in all possible linear combinations of the three local measures in this linearly combined measure and the Pearson correlation coefficient between the similarities of the 750 text pairs computed by each linearly combined measure or each linear combination and the judged similarities of the 750 text pairs are calculated and listed in Table 1 according to the designed algorithm. As an example, the first experiment (Experiment 1) calculates the weights in $Sim_1(I_1, I_2)$ and in all possible linear combinations of $Sim_{\mathbf{CAR}}$, $Sim_{\mathbf{CRO}}$, and $Sim_{\mathbf{CAP}}$ (please see Table 2) and the Pearson correlation coefficient between the similarities of the 750 text pairs computed by $Sim_1(I_1, I_2)$ or each linear combination and the judged similarities of the 750 text pairs (please see Table 1).

Table 1. The calculated weights and Pearson correlation coefficients in the 8 experiments.

Experiment	$Sim_{k,u,v}$	$w_{k,u,v,1}$	$w_{k,u,v,2}$	$w_{k,u,v,3}$	$pcc_{k,u,v}$
Experiment 1	$Sim_{1,1,1}$	1.0000	—	—	0.6532
	$Sim_{1,1,2}$	1.0000	—	—	0.8217
	$Sim_{1,1,3}$	1.0000	—	—	0.7908
	$Sim_{1,2,1}$	0.2093	0.7907	—	0.8258
	$Sim_{1,2,2}$	0.1275	0.8725	—	0.7918
	$Sim_{1,2,3}$	0.6515	0.3485	—	0.8335
	$Sim_{1,3,1}$	0.0544	0.6301	0.3155	0.8337
	$Sim_1(I_1, I_2)$	0.0544	0.6301	0.3155	**0.8337**
Experiment 2	$Sim_{2,1,1}$	1.0000	—	—	0.6532
	$Sim_{2,1,2}$	1.0000	—	—	0.8217
	$Sim_{2,1,3}$	1.0000	—	—	0.8750
	$Sim_{2,2,1}$	0.2093	0.7907	—	0.8258
	$Sim_{2,2,2}$	0.0931	0.9069	—	0.8757
	$Sim_{2,2,3}$	0.2459	0.7541	—	0.8803
	$Sim_{2,3,1}$	0.0237	0.2367	0.7396	0.8803
	$Sim_2(I_1, I_2)$	0.0237	0.2367	0.7396	**0.8803**
Experiment 3	$Sim_{3,1,1}$	1.0000	—	—	0.6532
	$Sim_{3,1,2}$	1.0000	—	—	0.7717
	$Sim_{3,1,3}$	1.0000	—	—	0.7908
	$Sim_{3,2,1}$	0.1724	0.8276	—	0.7740
	$Sim_{3,2,2}$	0.1275	0.8725	—	0.7918
	$Sim_{3,2,3}$	0.3981	0.6019	—	0.8015
	$Sim_{3,3,1}$	0.0000	0.3981	0.6019	0.8015
	$Sim_3(I_1, I_2)$	0.0000	0.3981	0.6019	**0.8015**
Experiment 4	$Sim_{4,1,1}$	1.0000	—	—	0.6532
	$Sim_{4,1,2}$	1.0000	—	—	0.7717
	$Sim_{4,1,3}$	1.0000	—	—	0.8750
	$Sim_{4,2,1}$	0.1724	0.8276	—	0.7740
	$Sim_{4,2,2}$	0.0931	0.9069	—	0.8757
	$Sim_{4,2,3}$	0.1285	0.8715	—	0.8765
	$Sim_{4,3,1}$	0.0282	0.1137	0.8581	0.8766
	$Sim_4(I_1, I_2)$	0.0282	0.1137	0.8581	**0.8766**
Experiment 5	$Sim_{5,1,1}$	1.0000	—	—	0.7939
	$Sim_{5,1,2}$	1.0000	—	—	0.8217
	$Sim_{5,1,3}$	1.0000	—	—	0.7908
	$Sim_{5,2,1}$	0.4015	0.5985	—	0.8437
	$Sim_{5,2,2}$	0.5102	0.4898	—	0.8230
	$Sim_{5,2,3}$	0.6515	0.3485	—	0.8335
	$Sim_{5,3,1}$	0.3346	0.5002	0.1652	0.8460
	$Sim_5(I_1, I_2)$	0.3346	0.5002	0.1652	**0.8460**
Experiment 6	$Sim_{6,1,1}$	1.0000	—	—	0.7939
	$Sim_{6,1,2}$	1.0000	—	—	0.8217
	$Sim_{6,1,3}$	1.0000	—	—	0.8750
	$Sim_{6,2,1}$	0.4015	0.5985	—	0.8437
	$Sim_{6,2,2}$	0.1800	0.8200	—	0.8784
	$Sim_{6,2,3}$	0.2459	0.7541	—	0.8803
	$Sim_{6,3,1}$	0.1186	0.2009	0.6805	0.8817
	$Sim_6(I_1, I_2)$	**0.1186**	**0.2009**	**0.6805**	**0.8817**
Experiment 7	$Sim_{7,1,1}$	1.0000	—	—	0.7939
	$Sim_{7,1,2}$	1.0000	—	—	0.7717
	$Sim_{7,1,3}$	1.0000	—	—	0.7908
	$Sim_{7,2,1}$	0.5488	0.4512	—	0.8206
	$Sim_{7,2,2}$	0.5102	0.4898	—	0.8230
	$Sim_{7,2,3}$	0.3981	0.6019	—	0.8015
	$Sim_{7,3,1}$	0.4508	0.2424	0.3068	0.8270
	$Sim_7(I_1, I_2)$	0.4508	0.2424	0.3068	**0.8270**
Experiment 8	$Sim_{8,1,1}$	1.0000	—	—	0.7939
	$Sim_{8,1,2}$	1.0000	—	—	0.7717
	$Sim_{8,1,3}$	1.0000	—	—	0.8750
	$Sim_{8,2,1}$	0.5488	0.4512	—	0.8206
	$Sim_{8,2,2}$	0.1800	0.8200	—	0.8784
	$Sim_{8,2,3}$	0.1285	0.8715	—	0.8765
	$Sim_{8,3,1}$	0.1546	0.0711	0.7743	0.8788
	$Sim_8(I_1, I_2)$	0.1546	0.0711	0.7743	**0.8788**

Table 2. All possible linear combinations of the three local measures in $Sim_1(I_1, I_2)$.

Components	Linear combination
One	$Sim_{1,1,1}(I_1, I_2) = Sim_{\mathbf{CAR}}$;
	$Sim_{1,1,2}(I_1, I_2) = Sim_{\mathbf{CRO}}$;
	$Sim_{1,1,3}(I_1, I_2) = Sim_{\mathbf{CAP}}$
Two	$Sim_{1,2,1}(I_1, I_2) = w_{1,2,1,1}Sim_{\mathbf{CAR}} + w_{1,2,1,2}Sim_{\mathbf{CRO}}$
	$Sim_{1,2,2}(I_1, I_2) = w_{1,2,2,1}Sim_{\mathbf{CAR}} + w_{1,2,2,2}Sim_{\mathbf{CAP}}$
	$Sim_{1,2,3}(I_1, I_2) = w_{1,2,3,1}Sim_{\mathbf{CRO}} + w_{1,2,3,2}Sim_{\mathbf{CAP}}$
Three	$Sim_{1,3,1}(I_1, I_2) = w_{1,3,1,1}Sim_{\mathbf{CAR}} + w_{1,3,1,2}Sim_{\mathbf{CRO}} + w_{1,3,1,3}Sim_{\mathbf{CAP}}$

As shown in Table 1, the highest Pearson correlation coefficient is the one of $Sim_6(I_1, I_2)$. Thus, a linearly combined similarity measure for text pairs with high accuracy is constructed as:

$$Sim(I_1, I_2) = 0.1186Sim_{\mathbf{YEH}} + 0.2009Sim_{\mathbf{CRO}} + 0.6805Sim_{\mathbf{BAN}} \quad (18)$$

C. Evaluation Results Analysis

The theoretical proof has proved that the measure constructed the algorithm has the highest accuracy among all possible linear combinations of its local measures. It is actually a proof of the effectiveness of the weight assignment method in the algorithm since the objective of weight assignment is to maximize the accuracy of the measure. In the existing approaches for constructing a linearly combined similarity measure [16−19, 21−23, 25, 26], weights are fixed or adjusted to achieve certain purposes (which do not include

maximizing the accuracy of the measure). This does not necessarily obtain the highest accuracy. While in the weight assignment method, weights are adjusted to maximize the accuracy of the measure. Thus compared to the existing approaches [16−19, 21−23, 25, 26], the proposed approach has an advantage of constantly achieving the highest accuracy.

In the experimental verification, 8 experiments have been carried out to verify the result of the theoretical proof. As can be seen from the results of these experiments (Table 1), the inequality "$\mathrm{pcc}(\boldsymbol{X}, \boldsymbol{J}) \geq \mathrm{pcc}(\boldsymbol{Y_k}, \boldsymbol{J}) \geq \mathrm{pcc}(\boldsymbol{Z_{1,u,v}}, \boldsymbol{J})$" holds in every experiment (e.g. in Experiment 1, $0.8817 \geq \{0.8337, 0.8803, 0.8015, 0.8766, 0.8460, 0.8817, 0.8270, 0.8788\} \geq \{0.6532, 0.8217, 0.7908, 0.8258, 0.7918, 0.8335, 0.8337\}$), which experimentally verifies the correctness of the theoretical proof and the effectiveness of the designed algorithm. Besides, careful readers may find that there are slight differences among all $\mathrm{pcc}(\boldsymbol{Y_k}, \boldsymbol{J})$ ($k = 1, 2, \ldots, 8$). This is because $\mathrm{pcc}(\boldsymbol{Y_k}, \boldsymbol{J})$ are all the optimal values computed using the same method to optimize the Pearson correlation coefficients of the 8 linearly combined measures in Expression (17) that differentiate with each other in only one local measure. As can also be seen from the results, the accuracy of a linear combination with more contribution components is higher than the accuracy of a linear combination with less contribution components. As an example, $\mathrm{pcc}_{1,3,1} > \{\mathrm{pcc}_{1,2,1}, \mathrm{pcc}_{1,2,2}, \mathrm{pcc}_{1,2,3}\} > \{\mathrm{pcc}_{1,1,1}, \mathrm{pcc}_{1,1,2}, \mathrm{pcc}_{1,1,3}\}$ holds in Experiment 1. This laterally demonstrates the effectiveness of the weight assignment method in the algorithm. In addition, it can be seen from the results that the Pearson correlation coefficient of the constructed linearly combined measure with respect to the MSRvid benchmark [38] (i.e. 0.8817) is higher than the Pearson correlation coefficients of the six local measures $Sim_{\mathbf{CAR}}$ [39], $Sim_{\mathbf{YEH}}$ [40], $Sim_{\mathbf{CRO}}$ [41], $Sim_{\mathbf{MAL}}$ [42], $Sim_{\mathbf{CAP}}$ [43], and $Sim_{\mathbf{BAN}}$ [44] with respect to the same benchmark (i.e. 0.6532, 0.7939, 0.8217, 0.7717, 0.7908, and 0.8750). This further signifies that the proposed linearly combined measure construction approach is indeed helpful to improve the accuracy of the linearly combined measure.

In summary, the findings of the 8 experiments are: (1) The accuracy of the linearly combined measure constructed by the designed algorithm is higher than the accuracies of all possible linear combinations of its local measures. (2) The weight assignment and local measure selection methods in the algorithm are helpful to improve the accuracy of the linearly combined measure.

V. Conclusions

In this paper, an approach for constructing a linearly combined similarity measure with high accuracy for assessing the similarity between linguistic items has been proposed. This approach mainly consists of a weight assignment method and a local measure selection method. The two methods respectively assigned different weights and chose different local measures to construct a linearly combined measure for different cases through maximizing the accuracy of this linearly combined measure. This can ensure the accuracy of the constructed linearly combined measure always be high in any cases. The paper has also presented the theoretical and experimental evaluation of the approach. The evaluation results show that the linearly combined measure constructed by the approach has high accuracy and the weight assignment and local measure selection methods are effective to improve the accuracy of the linearly combined measure.

One future study will aim especially at using the proposed approach in practical applications. In some applications such as data integration, information retrieval, and knowledge extraction, a linearly combined measure with high accuracy is needed to be constructed sometimes and the approach could be directly applied in such construction. Another future study will focus on overcoming a major limitation of the proposed approach. The approach mainly shows how to construct a linearly combined measure with high accuracy for assessing the similarity between linguistic items and does not address the construction of a nonlinearly combined measure with high accuracy. Because a nonlinearly combined measure may be of use in some applications, it would be desirable to study how to construct such a measure with high accuracy.

Acknowledgments

The authors would like to appreciate the financial supports by the National Natural Science Foundation of PR China (No. 71371081) and the Specialized Research Foundation for the Doctoral Program of Higher Education of PR China (No. 20130142110044).

References

[1] Jurgens D, Pilehvar MT, Navigli R. Cross level semantic similarity: An evaluation framework for universal measures of similarity. Lang Resour Eval 2016; 50(1): 5–33.

[2] Wu JB, Wu ZL. Comprehensive approach to semantic similarity for rapid data integration. Int J Control Autom Syst 2014; 12(3): 680–687.

[3] Mori U, Mendiburu A, Lozano JA. Similarity Measure Selection for Clustering Time Series Databases. IEEE Trans Knowl Data Eng 2016; 28(1): 181–195.

[4] Otegi A, Arregi X, Ansa O, Agirre E. Using knowledge-based relatedness for information retrieval. Knowl Inf Syst 2015; 44(3): 689–718.

[5] Cui X, Cai S, Qin Y. Similarity-based approach for accurately retrieving similar cases to intelligently handle online complaints. Kybernetes 2017; 46(7): 1223–1244.

[6] Moschopoulos T, Iosif E, Demetropoulou L, Potamianos A, Narayanan SS. Toward the automatic extraction of policy networks using web links and documents. IEEE Trans Knowl Data Eng 2013; 25(10): 2404–2417.

[7] Li S, Zhou L, Li Y. Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. Inf Process Manage 2015; 51(1): 58–67.

[8] Shvaiko P, Euzenat J. Ontology matching: State of the art and future challenges. IEEE Trans Knowl Data Eng 2013; 25(1): 158–176.

[9] Kim H, Kang S, Oh S. Ontology-based quantitative similarity metric for event matching in publish/subscribe system. Neurocomputing 2015; 152(3): 77–84.

[10] Pilehvar MT, Navigli R. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. Artif Intell 2015; 228(11): 95–128.

[11] Fellbaum C (Ed.). WordNet: An Electronic Database. MIT Press, 1998.

[12] Navigli R, Ponzetto SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 2012; 193(12): 217–250.

[13] Meng L, Huang R, Gu J. A review of semantic similarity measures in wordnet. Int J Hybrid Inf Technol 2013; 6(1): 1–12.

[14] Taieb MAH, Aouicha MB, Hamadou AB. Ontology-based approach for measuring semantic similarity. Eng Appl Artif Intell 2014; 36(11): 238–261.

[15] Gomaa WH, Fahmy AA. A survey of text similarity approaches. Int J Comput Appl 2013; 68(13): 13–18.

[16] Rodríguez MA, Egenhofer MJ. Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowl Data Eng 2003; 15(2): 442–456.

[17] Li Y, McLean D, Bandar ZA, O'Shea JD, Crockett K. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Trans Knowl Data Eng 2006; 18(8): 1138–1150.

[18] Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 2003; 15(4): 871–882.
[19] Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans Knowl Disc Data 2008; 2(2): 1–25.
[20] Tversky A. Features of similarity. Psychol Rev 1977; 84(4): 327–352.
[21] Petrakis EGM, Varelas G, Hliaoutakis A, Raftopoulou P. X-similarity: Computing semantic similarity between concepts from different ontologies. J Digit Inf Manage 2006; 4(4): 233–237.
[22] Furlan B, Batanović V, Nikolić B. Semantic similarity of short texts in languages with a deficient natural language processing support. Decis Support Syst 2013; 55(3): 710–719.
[23] Li P, Wang H, Zhu KQ, Wang Z, Hu X, Wu X. A large probabilistic semantic network based approach to compute term similarity. IEEE Trans Knowl Data Eng 2015; 27(10): 2604–2617.
[24] Maedche A, Staab S. Comparing ontologies-similarity measures and a comparison study. Internal Report No. 408, Institute AIFB, University of Karlsruhe, Karlsruhe, Germany, Mar. 2001.
[25] Jiang Y, Zhang X, Tang Y, Nie R. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Inf Process Manage 2015; 51(3): 215–234.
[26] Akmal S, Shih LH, Batres R. Ontology-based similarity for product information retrieval. Comput Ind 2014; 65(1): 91–107.
[27] Wu Z, Palmer M. Verbs semantics and lexical selection. Proc 32nd Ann Meeting Assoc Comput Linguist, 1994; pp. 133–138.
[28] Lin D. An information-theoretic definition of similarity. Proc 15th Int Conf Machine Learning, 1998; pp. 296–304.
[29] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945; 26(3): 297–302.
[30] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers, 2011.
[31] van der Weken D, Nachtegael M, Kerre EE. Using similarity measures and homogeneity for the comparison of images. Image Vision Comput 2004; 22(9): 695–702.
[32] Lu W, Qin Y, Qi Q, Zeng W, Zhong Y, Liu X, Jiang X. Selecting a semantic similarity measure for concepts in two different CAD model data ontologies. Adv Eng Inform 2016; 30(3): 449–466.
[33] Kilgarriff A. English lexical sample task description. Proc 2nd Int Workshop Eval Word Sense Disamb Syst, 2001; pp. 17–20.
[34] Weischedel R, Palmer M, Marcus M, Hovy E, Pradhan S, Ramshaw L, Xue N, Taylor A, Kaufman J, Franchini M, El-Bachouti M, Belvin R, Houston A. OntoNotes Release Version 5.0 LDC2013T19. Linguistic Data Consortium, 2013.
[35] Rubenstein H, Goodenough JB. Contextual correlates of synonymy. Commun ACM 1965; 8(10): 627–633.
[36] Yang D, Powers DMW. Verb similarity on the taxonomy of WordNet. Proc 3rd Int WordNet Conf, 2006; pp. 121–128.
[37] Finkelstein L, Evgenly G, Yossi M, Ehud R, Zach S, Gadi W, Eytan R. Placing search in context: The concept revisited. ACM Trans Inf Syst 2002; 20(1): 116–131.
[38] Agirre E, Cer D, Diab M, Gonzalez-Agirre A. SemEval-2012 task 6: A pilot on semantic textual similarity. Proc SemEval-2012, 2012; pp. 385–393.
[39] Carrillo M, Vilarino D, Pinto D, Tovar M, León S, Castillo E. Buap: Three approaches for semantic textual similarity. Proc 1st Joint Conf Lexical Comput Semantics, 2012; pp. 631–634.
[40] Yeh E, Agirre E. SRIUBC: Simple similarity features for semantic textual similarity Proc 1st Joint Conf Lexical Comput Semantics, 2012; pp. 617–623.
[41] Croce D, Annesi P, Storch V, Basili R. UNITOR: Combining semantic text similarity functions through SV regression. Proc 1st Joint Conf Lexical Comput Semantics, 2012; pp. 597–602.
[42] Malandrakis N, Iosif E, Potamianos A. DeepPurple: Estimating sentence semantic similarity using N-gram regression models and web snippets. Proc 1st Joint Conf Lexical Comput Semantics, 2012; pp. 565–570.
[43] Caputo A, Basile P, Semeraro G. UNIBA: Distributional semantics for textual similarity. Proc 1st Joint Conf Lexical Comput Semantics 2012; pp. 591–596.
[44] Banea C, Hassan S, Mohler M, Mihalcea R. UNT: A supervised synergistic approach to semantic text similarity. Proc 1st Joint Conf Lexical Comput Semantics, 2012; pp. 635–642.