Improving Multilingual Semantic Interoperation in Cross-Organizational Enterprise Systems Through Concept Disambiguation

Jingzhi Guo, *Member, IEEE*, Li Da Xu, *Senior Member, IEEE*, Guangyi Xiao, *Member, IEEE*, and Zhiguo Gong, *Member, IEEE*

***Abstract*—For the multilingual semantic interoperations in cross-organizational enterprise systems and e-commerce systems, semantic consistency is a research issue that has not been well resolved. This paper contributes to improving multilingual semantic interoperation by proposing a concept-connected near synonym (NSG) framework for concept disambiguation. NSG framework provides a vocabulary preprocessing process of collaborative vocabulary editing, which further ensures semantically consistent vocabulary for building semantically consistent business processes and documents between context-different information systems. The vocabulary preprocessing offered by NSG automates the process of finding potential near synonym sets and identifying collaboratively editable near synonym sets. The realization of NSG framework includes a probability model that computes concept values between concepts based on a newly introduced semantic relatedness method—SRCT. In this paper, SRCT-based methods are implemented and compared with some existing semantic relatedness methods. Experiments have shown that SRCT-based methods outperform the existing methods. This paper has made an improvement on the existing methods of semantic relatedness and reduces the collaboration cost of collaborative vocabulary editing.**

***Index Terms*—Data engineering, enterprise systems, industrial informatics, interenterprise multilingual interoperation, knowledge management, text analysis, understandability.**

I. Introduction

BUSINESS process is an important research topic in industrial informatics [23], [24], concerning the flows of business and manufacturing activities in cross-organizational enterprise systems [27], [43], [46]–[48] and e-commerce systems [18]. Its major task is to facilitate the business collaboration and cooperation between enterprises through their underlying information systems [16]. A technical challenge is the lack of business process interoperation across domains of involved information systems [1]. This is because cross-organizational business processes are often heterogeneous in the two aspects of maintaining syntactic consistency in structuring and modeling business processes [23], [24], and ensuring semantic consistency for delivering accurate activity meaning to recipients [15]. At present, syntactic consistency has been well researched (e.g., [23] and [24]) while semantic consistency is rarely explored except for a few such as [15] and [18]. This hinders business collaboration and cooperation.

This paper covers a specific research topic of ensuring semantic consistency between multilingual business processes. It aims at automating concept disambiguation process for preprocessing of dictionary entries for collaboratively developing multilingual vocabulary [15], [41], used for building semantically consistent cross-organizational information systems.

Currently, there are three alternative approaches for ensuring semantic consistency: mandatory standardization, intelligent mediation, and collaborative conceptualization [15]. The first two approaches have now been widely applied in industrial informatics, for example, adopting IEC standard 61131 and 61499 [5], [6] to develop control logic software of industrial applications, using various standards for new smart grid technology [13], and applying intelligent mediators with agent technology to mediate various industrial applications and different information resources [31], [44]. Yet, for cross-organizational systems, it is highly possible that users often adopt different standards for information creation and exchange according to their own contexts. Thus, standardization approach is not always useful. Intelligent mediation approach attempts to solve standardization problem based on agent technology. Nevertheless, since different industrial applications are mostly created and run in different contexts, rules of creating and using information by agents often have different semantic assumptions. This leads to semantic conflicts between cross-context users for information exchange.

To enable information exchange across heterogeneous information systems without misinterpretation, the research of [15] provides a collaborative conceptualization approach, which resolves semantic conflicts of information exchange based on a collaborative vocabulary editing mechanism. The main task of this approach is to create semantically consistent concepts across heterogeneous information systems. Such concepts can be used to build consistent business documents and processes for accurate information exchange and are both syntactically and semantically consistent for cross-organizational enterprise

Manuscript received December 19, 2011; revised January 29, 2012; accepted February 08, 2012. Date of publication February 28, 2012; date of current version July 23, 2012. This work was supported in part by the University of Macau Research Committee Grant RG055/08-09S/11R/GJZ/FST, in part by the National Natural Science Foundation of China (NSFC) Grant 71132008, and in part by the Changjiang Scholar Program of the Ministry of Education of China. Paper no. TII-11-1042.

J. Guo, G. Xiao and Z. Gong are with the University of Macau, Taipa, Macao (e-mail: jzguo@ umac.mo; ya97409@umac.mo; fstzgg@umac.mo).

L. D. Xu is with the Old Dominion University, Norfolk, VA 23529, USA, and also with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: LXu@odu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2012.2188899

systems. A disadvantage of this approach is that collaboration always implies labor cost, though collaboration is inevitable.

In this paper, we will improve collaborative conceptualization approach [15] by inserting a preprocessing mechanism for collaborative vocabulary development to maximum reduce the required human effort, thus to reduce labor cost. The main idea is: for all vocabulary entries that need to be collaboratively edited, they are preprocessed by a near synonym finding process, so that collaborative editors can resolve semantic conflicts between vocabulary entries using sets of near synonyms found in the preprocessing.

The importance of providing a preprocessing mechanism for finding near synonym sets for collaborative editing can be illustrated in the following example. Suppose that there are context-different enterprise systems of A, B, C, D, E, F, and G. Each of them has a product vocabulary to name their products such that $\{\text{refrigerator}\} \subset A$, $\{\text{fridge}\} \subset B$, $\{\text{rèfrigèrateur}\} \subset C$, $\{\text{refrigerator}\} \subset D$, $\{\text{cooler bag}\} \subset E$, $\{\text{冰箱}\}, \subset F$ and $\{\text{雪櫃}\} \subset G$. It is obvious that ambiguous interpretations of the product name may happen if a business process is across these systems. The detailed causes are: (1) A, C, and F cannot recognize the product name with each other if no multilingual translations; (2) A and D may interpret differently for "refrigerator"; (3) "refrigerator," "cooler bag," and " 雪櫃" may refer to the same; and (4) "雪櫃 " of G cannot be recognized by F. When collaborative conceptualization approach is applied as a solution to disambiguating the sense of these terms, in practice, there is high value to provide an automatic preprocessing mechanism to identify near synonym sets as many as possible as it can drastically reduce collaboration effort.

To implement the above-mentioned idea, we adopt an existing multilingual dictionary as an initial vocabulary for preprocessing, through which target vocabulary can be further built. This saves collaboration time by avoiding editing a brand new vocabulary from scratch. Nevertheless, an initial multilingual vocabulary (e.g., Multilingual American Heritage Dictionary adopted in this research) often contains undesirable semantic inconsistency between multilingual concepts, that is, synonyms and homonyms that confuse vocabulary editors. A key phenomenon is non-associated synonyms that might have same or similar meanings yet cannot be detected between different natural languages, for instance, "实现 (fruition)" and "realization" are synonymous but we cannot find their association in the initial vocabulary. Obviously, it is time-consuming for editors to pick up all associated multilingual synonyms in editing period even if they are knowledgeable lexical experts.

To solve this problem, this paper proposes a near synonym graph (NSG) framework based on WordNet ([32], [45]) for automating the process of multilingual concept disambiguation in order to find multilingual near synonyms, that is, the semantically equivalent and similar concepts in an initial multilingual vocabulary.

There are two major issues regarding finding near synonyms from an initial multilingual vocabulary. First, how to retrieve potential near synonym sets from a large multilingual dictionary. Second, how to identify near synonym sets from the retrieved potential near-synonym sets. To tackle these two issues, the framework we propose correspondingly consists of two stages: retrieving potential near synonym sets from a multilingual dictionary and identifying near synonym sets. In the first stage, potential near synonym sets of initial vocabulary are collected by using concept match technology using a given external synonym dictionary (i.e., WordNet [32] in this research). To generalize the process of concept association, we introduce a concept-connected graph to increase the quality of finding sets of potential near synonyms.

In the second stage, we proceed to identify near synonym sets from potential near synonym sets. Since a potential near synonym set might consist of some concepts that have low similarity or non-synonyms at all, concepts with low or no similarity in senses are discarded. Experimental results show that the proposed framework is effective for retrieving potential near synonym sets and identifying near synonym sets.

Unlike many studies that only consider semantic relatedness to only a thesaurus or to a corpus, our NSG framework improves semantic relatedness between concepts by associating concepts not only to a thesaurus but also to concepts of initial vocabulary itself using both methods of thesaurus-based and corpus-based.

The rest of this paper is organized as follows. Section II discusses the related work. Section III presents a near synonym graph framework to build a WordNet-based concept-connected graph. Section IV proposes a method of retrieving potential near synonym sets. Section V proposes an approach of how to identify near synonym sets. In Section VI, we discuss various methods of how to measure semantic relatedness. In Section VII, experiments are made to evaluate the accuracy and performance under different methods of measuring semantic relatedness. Finally, a conclusion is made.

II. Related Work

The issue of identifying near synonymous sets of an existing bilingual dictionary, in particular, finding concepts with same or similar meaning has long been studied in machine translation and computational linguistics [2], [4], [8], [10], [20], where near synonyms are extracted and selected for use. The study can be classified into two directions: retrieving the associated words for building potential near synonym sets using corpus/texts/Web [4], and identifying near synonyms from potential near synonym sets using thesaurus [10], [20]. The former direction involves a most important step, namely, a statistical method is applied to extract potential near synonym sets from corpus, text or Web pages. A limitation of this method is that the corpus may fail to provide sufficient information about relevant word/word relationships. To improve the word/word synonymous relevance, the latter study direction is often proposed, that is, a third-party synonym dictionary or electronic database (e.g., English WordNet [32] or Chinese Cilin [36]) is introduced to find the potential synonym sets of target dictionary. This category of methods helps quickly find potential near synonym sets that are relevant to the given third-party synonym dictionary or electronic database.

At present, WordNet [32], [45], an electronic English lexical database, has been widely adopted to solve a variety of problems, such as information extraction and information retrieval [7], ontology engineering [39], and word sense disambiguation [34]. Particularly, it is often used for: 1) finding and identifying pure English near synonyms and 2) finding and identifying

multilingual near synonyms such as building EuroWordNet [11], Italy WordNet [33], Portugese WordNet [9], and Chinese WordNet [7]. In this paper, WordNet is used for concept association and disambiguation between words of an English-Chinese Dictionary to find potential near synonym sets and identify near synonym sets for building precollaborative vocabulary.

Using English WordNet to find and identify near synonyms of another English dictionary is relatively easy. Its major application is to adopt WordNet to disambiguate word senses among a same set of near synonyms and select the best near synonym. However, using English WordNet to develop another multilingual WordNet or find/identify near synonym sets of non-English WordNet is a rather difficult task. Most existing approaches can be summarized as follows: finding a local dictionary and an English WordNet, translating the word definitions of the local dictionary into English to form a bilingual dictionary, comparing the English word definitions of local dictionary with the synset definitions of WordNet, and identifying the near synonyms appeared in local dictionary. This type of approaches is rather intuitive. Their advantage is that they can immediately identify which words are known to WordNet and which are not. For the known words, they can immediately find the unambiguous words and identify the similarity of ambiguous words. The disadvantages are that: 1) the identified similarity may be incorrect because the English translation of word definitions in local dictionary may be not accurate and 2) the unknown words outside of WordNet cannot be processed by WordNet.

A remedy approach is often proposed to process the unknown words by introducing a second synonym dictionary to handle unknown words. For example, in [7], CILIN [36] is functioned as the first synonym dictionary where known words are picked up and tagged by senses. For unknown words, English WordNet is adopted as the second synonym dictionary to relate to the first synonym dictionary based on synset of English WordNet. The advantage is that some unknown words, which are known by WordNet, can be disambiguated and synonymously identified. However, accurate translation still is a task that has not been resolved in this approach. This may lower the accuracy of identified synonyms. In addition, some unknown words still exist and cannot be identified because they are not covered by both WordNet and CILIN.

III. Overview of Near Synonym Graph Framework

To improve the existing solutions of finding and identifying near synonym sets, particularly avoiding translation ambiguity of a locally designed vocabulary, this section propose a novel framework, called *near synonym graph* (NSG). This framework assumes that a well-known multilingual dictionary is reputable and accurate in multilingual translations for its corresponding words and definitions. The criteria of selecting such a reputable and accurate multilingual dictionary are that: 1) a major dictionary edited by renowned lexical experts and internationally used. The list is American Heritage Dictionary, Longman Dictionary, Oxford Dictionary, Webster's Dictionary, etc., and 2) having natural language translation versions as many as possible. Based on these two criteria, we adopt English-Chinese American Heritage Dictionary (BD) (with 99 890 words and 175 238 senses) as a target local dictionary (as an initial vocabulary) where English and Chinese near synonym sets must be found and identified as our research target. In this framework, we adopt English WordNet (WN) [45] (with 147 306 words, 206 904 concepts and 117 659 synsets) as an external resource of English synonym dictionary.

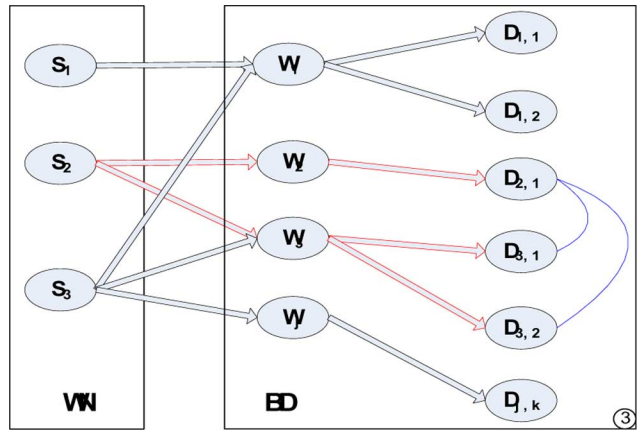


Fig. 1. Overview of NSG framework.

The key research problem of this paper is that how concepts of BD can be disambiguated to find their semantic similarity hence to identify all possible near synonym sets. To resolve this problem, our framework is *synonym dictionary driven* (i.e., finding potential near synonym sets starting from a synonym dictionary, that is, WN), which is different from the existing approach that is often *local dictionary driven* (i.e., finding and identifying near synonym sets starting from local dictionary). The benefit of synonym dictionary driven approach is that the corresponding English synonymous relationships can be easily found in local dictionary and the Chinese synonymous relationships can be detected and merged through the existing standard translation of BD. To implement the idea of synonym dictionary driven approach, NSG framework is designed as a concept-connected near synonym graph, where a concept of WN connects to one-to-many words of BD, of which each again connects to many concepts (i.e., word glosses) of BD. Each concept-connected graph constitutes a potential near synonym set, which is later disambiguated for identifying a final usable near synonym set.

The NSG framework can be illustrated in Fig. 1 as a graph with examples of $S_1 = \{\text{recognition}\}$, $S_2 = \{\text{actualization, realization}\}$ and $S_3 = \{\text{realization, realization, recognition}\}$; $W_1 =$ "recognition", $W_2 =$ "actualization", $W_3 =$ "realization" and $W_4 =$ "realization"; $D_{1,1} =$ gloss 1 of W_1 in BD, $D_{1,2} =$ gloss 2 of W_1 in BD, $D_{2,1} =$ gloss of W_2 in BD, and etc.

Technically, NSG framework is a 7-tuple of $< S, W, D, R_1, R_2, R_3, P >$ combining WN with BD, where:

— $S \in$ WN is a set of concepts and $s \in S$ is a concept;
— $W \in$ WN $\cap$ BD is a set words and $w \in W$ is a word;
— $D \in$ BD is a set of concepts and $d \in D$ is a concept;
— R_1: $S \rightarrow W$ is a one-to-many concept-to-word relation;

— R_2: $W \rightarrow D$ is a one-to-many word-to-concept relation;
— R_3: $D_1 \rightarrow D_2$ is a one-to-many concept-to-concept relation;
— $P : S \times W \times D$ is a triple constrained by R_1 and R_2, such that (1) for each $p \in P$, $p : s \rightarrow w \rightarrow d$, $p = s \times w \times d$, $p = (s, w, d)$ or $p = s \cdot w \cdot d$, p is called a single path s.w.d from s to d, and (2) for some elements $P' \subseteq P$, $P' : S' \rightarrow W' \rightarrow D'$, $P' = S' \times W' \times D'$ or $P' = (S', W', D')$, P' is called a set of connected paths or a potential near synonym set. It is obvious that $\{P'\} = P$.

In this framework, we have two operations *od* and *oi* on a *direct path* and an *indirect path*, such that:

— $\mathrm{od} : R_1 \rightarrow R_2$ on $p \in P$, notated by $_{\mathrm{od}}(s, w, d)$;
— $\mathrm{oi} : R_1 \rightarrow R_2 \wedge R_1' \rightarrow R_2' \rightarrow R_3$ across $p_1, p_{i \neq 1} \in P$ where $R_2(W) = R_3(D_1)$, notated by $_{\mathrm{oi}}(s, w, d)$.

The entire design of the framework is as follows:

1) Adopt WN as a synonym dictionary using its synsets.
2) Provide a unique tag for each synset of WN.
3) Adopt a reputable digitalized multilingual dictionary (i.e., BD) as an initial vocabulary for finding and identifying near synonym sets for target vocabulary. In this research, BD is Multilingual American Heritage Dictionary.
4) Provide a unique tag for each word and for each concept in BD. It implies that each word W_j in BD can have many tags to identify different word definitions such as $D_{3,1}$ and $D_{3,2}$, where W_j and $D_{j,k}$ are all bilingual in English and Chinese.
5) For some sense-connected synsets S_i of WN, build a sense-connected subgraph from WN to BD for describing a potential near synonym set.
6) For each potential near synonym set, compare concept similarity between all concepts of WN and all concepts of BD, and between all concepts within BD.
7) Identify each near synonym set from each potential near synonym set.

In NSG framework, the solution is clearly divided into two stages: finding potential near synonym sets and identifying final usable near synonym sets. In the later two sections, we will discuss them one by one.

IV. Extracting Potential Near Synonym Sets

The first stage of NSG framework is to find all potential near synonym sets $\{P'\}$ from WN to BD. Our method is to build WN synset-based concept connections through a set of common words appeared in both WN and BD. To accomplish it, we categorize P into many potential near synonym sets $\{P'\}$ in 22 categories, shown in Table I, in the following categorization approach.

Category 1 (C_1): a set of words W_j in both WN and BD has only one synset S_i^1 and one concept $D_{j,k}^1$ such that $C_1 = (S_i^1, W_j, D_{j,k}^1)$.

Category n (C_n): a set of words W_j in both WN and BD has n synsets S_i^n and m concepts $D_{j,k}^m$ such that $C_n = (S_i^n, W_j, D_{j,k}^m)$.

Apparently, C_1 directly derives a set of near synonym sets for BD, which exactly has one concept mapping onto a WN synset. For C_n, it derives a complex potential near synonym set associated by the shared words W_j of both WN and BD. When n becomes larger, the set becomes more complex.

TABLE I
Statistics of Potential Near Synonym Sets

Category	Synsets	Words	Potential sets	Words per synset	Words per potential set
C1	1	90522	57415	1.576626	1.576626
C2	2	15299	6837	1.118839	2.237677
C3	3	5421	1796	1.006125	3.018374
C4	4	3024	697	1.084648	4.338594
C5	5	1692	333	1.016216	5.081081
C6	6	1120	178	1.048689	6.292135
C7	7	648	93	0.995392	6.967742
C8	8	473	59	1.002119	8.016949
C9	9	441	46	1.065217	9.586957
C10	10	269	25	1.076000	10.760000
C11	11	265	23	1.047431	11.521740
C12	12	156	15	0.866667	10.400000
C13	13	77	8	0.740385	9.625000
C14	14	77	4	1.375000	19.250000
C15	15	84	4	1.400000	21.000000
C16	16	52	2	1.625000	26.000000
C17	17	39	3	0.764706	13.000000
C18	18	105	5	1.166667	21.000000
C19	19	13	1	0.684211	13.000000
C20	20	27	1	1.350000	27.000000
C21	21	226	9	0.865900	25.111111
C22+	22	27276	1	0.827272	27276.000

The categorization result for all potential near synonym sets can be shown in Table I.

Table I shows all categories of potential near synonym sets $\{P'\} = P$ starting from WN synsets. In this statistic, Category 1 occupies 61%, which implies a big reduction of computing effort for finding the final near synonym sets because the sets of C_1 can be directly recommended as the near synonym sets for later collaborative vocabulary editing. From C_2 to C_{21}, there is a general trend that the number of potential near synonym sets is decreasing. This also reduces the computing cost of identifying the final near synonym sets. For each higher number category, the number of words needs to be disambiguated is increasing. $C22+$ is a large mixed set including words associated together through more than 22 synsets.

The extracted result of potential near synonym sets $\{P'\}$ is the source for further identifying final near synonym sets. To better serve the process of identifying, the actual content of Table I is resorted into four files based on the part of speech types like Noun, Verb, Abject, and Adverb, using an ascending ordering of WN synset tag. Each resorted file, a partial example shown in Fig. 3, consists of a set of WN synset based entries $\{p\}$ initiated by $\backslash s$. Each entry $p \in P' \in P$ consists of a set of $W' \subseteq W$. Each $w \in W'$ has a set of BD concepts $D' \subseteq D$.

V. Identifying Near Synonym Sets

Given an entry p of a potential near synonym set P' fallen in Category 2 to 22, the final usable near synonym sets for collaborative vocabulary editing is computed in NSG framework aforementioned. The approach by this framework suggests a weight-based probability model such that a three-level graph is composed of nodes of concepts and edges of probable weights. In this approach, a subgraph of a potential near synonym set $p = (s_i, w_j, d_{j,k})$ is reformulated in Fig. 2.

In this subgraph, the first-level vertices $\{s_i\}$ are WN synsets, the second-level vertices $\{w_j\}$ are words appeared in both WN and BD, and the third-level vertices $\{d_k\}$ are BD concepts (as

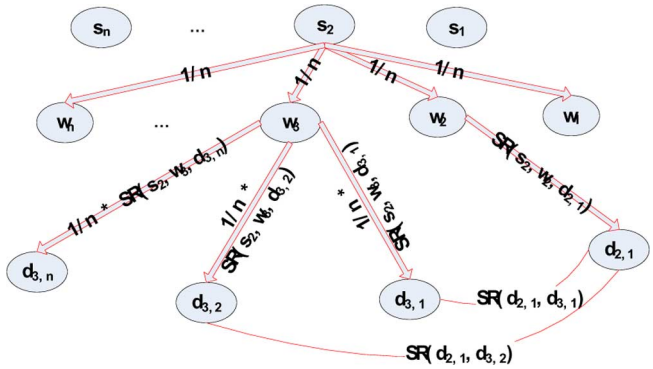


Fig. 2. A concept-connected NSG.

word definitions). Each edge has a weight. The weight of edge from a synset s_i to a word w_j is computed as follows:

$$w(s_i, w_j) = \frac{1}{|s_i \cdot words|} \tag{1}$$

where s_i.words means the number of words.

The weight of an edge from a word w_j to a concept $d_{j,k}$ is

$$w(w_j, d_{j,k}) = \frac{1}{|w_j \cdot concepts|} \times SR(s_i, w_j, d_{j,k}) \tag{2}$$

where w_j.concepts means the number of concepts and $\mathrm{SR}(s_i, w_j, d_{j,k})$ is a function measuring the semantic relatedness between two concepts that are concept-connected by their corresponding synset of WN and concept of BD. The details of the semantic relatedness measure will be discussed in Section VI.

The third-level vertices are also sense-connected for every two vertices $d_{j,1}$ and $d_{j',2}$ if and only if $d_{j,1}$.word is not equal to $d_{j',2}$.word. The weight of the edge between $d_{j,1}$ and $d_{j',2}$ is defined as follows:

$$\begin{aligned} &w(d_{j,1}, d_{j',2}) = SR(d_{j,1}, d_{j',2}) \\ &iff\ d_{j,1} \cdot word \neq d_{j',2} \cdot word. \end{aligned} \tag{3}$$

When $d_{j,1} \cdot \mathrm{word} = d_{j',2}$.word, it means that two concepts define the same word but a polysemous word.

A. Probability Model for Concept Value Identification

To more accurately identify near synonym sets within a potential near synonym set. This paper identifies not only the directly associated concepts of BD with WN synsets (called concept value identification by *direct path*) but also the seemingly non-associated concepts of BD, that is, between BD word definitions (called concept value identification by *indirect path*). For example, in Fig. 2, the direct path from s_2 to $d_{3,1}$ through d_3 is $s_2 \Rightarrow w_3 \Rightarrow d_{3,1}$. Differently, an indirect path from s_2 to $d_{3,1}$ through w_2 is $s_2 \Rightarrow w_2 \Rightarrow d_{2,1} \Rightarrow d_{3,1}$.

We state that there exists a *weight-based probability p* for identifying the concept values along both direct and indirect paths, such that

$$p_{\mathrm{od}}(s_1, w_j, d_k) = SR_{\mathrm{od}}(s_1, w_j, d_k) \tag{4}$$

$$\begin{aligned} p_{\mathrm{oi}}(s_1, w_j, d_k) &= SR_{\mathrm{oi}}(s_1, w_j, d_k) \\ &= SR_{\mathrm{od}}(s_1, w_j, d_k) + SR_{\mathrm{oi}}(s_1, w_x, d_k) \end{aligned} \tag{5}$$

where

$$SR_{\mathrm{od}}(s_1, w_j, d_k) = \frac{SR(T(s_1), T(d_k))}{|\{s_1 \times W\}|} \tag{6}$$

$$\begin{aligned} &SR_{\mathrm{oi}}(s_1, w_x, d_k) \\ &\quad = \sum_{x_w} \sum_{y_d} \frac{SR(s_1, w_x, d_y) \times SR(T(d_y), T(d_k))}{|\{w_x \times D\}|}. \end{aligned} \tag{7}$$

Formula (4) is a direct computation of semantic relatedness value disregarding the semantic impacts from others. Formula (5) is an indirect semantic relatedness computation that considers itself and other impacts. In (6) and (7), s_1, w_j, w_x, d_k and d_y are single elements, $x \neq j$, $y \neq k$, x experiences every word of s_1 except for w_j and y experiences every BD concept in $s_1 \times w_x \times d_y \subseteq P$ except for d_k for every w_x. For example, the probability of concept value from s_2 to w_3 to $d_{3,1}$ is $p(s_2, w_3, d_{3,1})$, that is, $SR(T(s_2), T(d_{3,1}))/n$ computed from direct path plus values from all indirect paths $s_2 \rightarrow (w_1, w_2, w_4, \ldots) \rightarrow (d_{1,y}, d_{2,y}, d_{4,y}, \ldots)$.

Based on the above formulas, it is obvious that the research of semantic relatedness problem can be converted to find semantic relatedness between two gloss texts $SR(T_1, T_2)$. In Section VI, we will devote to discussing how to compute semantic relatedness between two texts $SR(T_1, T_2)$.

B. Identifying Near Synonym Sets From Max Similarity

Given all semantic relatedness methods $SR(s_i, w_j, d_{j,k})$, this subsection proposes a Near Synonyms Finding algorithm (NSF) to identify near synonym sets from a set of max concept-similar values. The algorithm is designed to identify all near synonym sets of Noun, Verb, Abject, or Adverb. The NSF algorithm can be described in Fig. 4.

In this high-level algorithm, the $SR(s, s \cdot w, s \cdot w \cdot d)$ function is to compute semantic relatedness score between a concept s and a concept s.w.d. along the path from s. It is a kernel function in the entire algorithmic process. Two approaches can be alternatively applied to this function. One is the *direct approach* computing $SR_{\mathrm{od}}(s_1, s_1 \cdot w_j, s_1 \cdot w_j \cdot d_k)$ along a direct path and the other is the *indirect approach* computing $SR_{\mathrm{oi}}(s_1, s_1 \cdot w_j, s_1 \cdot w_j \cdot d_k)$ along both direct and indirect paths.

VI. Semantic Relatedness

Semantic relatedness is to find the semantic similarity between two objects. In this section, we first review the existing methods of computing semantic similarity between two concepts $\mathrm{SR2C}(s_p, s_q)$, between two words $\mathrm{SR2W}(w_p, w_q)$ and between two texts $\mathrm{SR2T}(t_p, t_q)$. Then, we introduce our own method, $\mathrm{SRCT}(T_1, T_2)$ that also measures the semantic relatedness between two texts. In the following, we will first discuss the existing methods of SR2C, SR2W, and SR2T, and then describe SRCT method in details.

A. SR2C Methods

SR2C is to find the semantic relatedness between any two concepts within a taxonomy, utilizing the hierarchical relationship presented by hypernyms and hyponyms. It is often called taxonomy-based or thesaurus-based.

\an annotation for the part of potential near synonym set uniquely tagged by 106567689 2
\s 106567689 2 \i software used in art and architecture and engineering and manufacturing to assist in precision drawing \w 28402 computer-aided design 0 1
\d 430604 n.Abbr. \ib The use of computer programs and systems to design detailed two- or three-dimensional models of physical objects, such as mechanical parts, buildings, and molecules. 4.07906e-302 \zh 计算机支持设计 计算机程序或系统的使用，用来设计详细的物体平面或三度空间的模型，如机械部分、建筑及分子 （名词）缩写 &$b\{CAD\}$
\w 19140 cad 0 2
\d 424191 abbr. \ib Computer-aided design. 0 \zh 计算机辅助设计 计算机辅助设计 （略语）
\d 424190 n. \ib An unprincipled, ungentlemanly man. 0 \zh 卑贱的人 不道德的，无教养的人 （名词）
\an annotation for the part of potential near synonym set uniquely tagged by 102941716 1
\s 102941716 1 \i a rotating disk shaped to convert circular into linear motion
\w 19741 cam 0 2
\d 424621 abbr.&I{Computer \ib Computer-aided manufacturing. 3.65322e-302 \zh 计算机辅助制造 计算机辅助制造 （略语）&$I\{$【计算机科学】}
\d 424620 n. \ib An eccentric or multiply curved wheel mounted on a rotating shaft, used to produce variable or reciprocating motion in another engaged or contacted part. 0 \zh 凸轮 安装于旋转轴上的偏心轮或复合曲线轮，用以使相连的或相接触的其它部件产生多变的或往复的运动 （名词）

Fig. 3. Overview of an Element in an Noun File of Joint Set between Synset in WN and concepts in BD.

NSF Algorithm. Identifying a set of near synonym sets.
Input: P_1 /* $P_1 \subseteq s \times w \times d \subseteq P \subseteq NSG$ *is a file, e.g. a* $p_1 \in P_1$ *is defined as a set of sets of sets in the form of* \s × \w × \d, *as illustrated in Figure 3* */
Output: $P_2 \leftarrow \varnothing$ /* P_2 *is a set of all identified near synonym sets as a file* */
1. for($P_1.p_i \in P_1$){ /* $P_1.p_i$ *is a synset in* P_1*/
2. for($s.w_i \in P_1.p_i$){ /* $s.w_i$ *is a polysemous word concept set in* $P_1.p_i$*/
3. for($s.w.d_i \in s.w_i$){ /* $s.w.d_i$ *is a concept in* $s.w_i$ */
4. $v_1 \leftarrow$ SR(s, s.w, $s.w.d_i$) /* *Determine concept value* v_1 */
5. if $v_1 > v_2$ then { /* v_2 *is the last chosen concept value of s.w.d* */
6. $d_2 \leftarrow s.w.d_i$ /* d_2 *is the last chosen path to s.w.d* */
7. $v_2 \leftarrow v_1$ }
8. } /* *Found a max concept-valued path* d_2 *within an s.w set* */
9. $p_2 \leftarrow p_2 \cup d_2$ /* p_2 *is a last identified near synonym set* */
10. } /* *Identified a near synonym set* p_2 */
11. $P_2 \leftarrow P_2 \cup p_2$ /* P_2 *is a last identified set of all near synonym sets* */
12. } /* *An identified set for all near synonym sets* P_2 *in* P_1 */

Fig. 4. NSF algorithm.

Early researches can be found in Leacock and Chodorow (LCA) [25], which measures the shortest path between two concepts s_1 and s_2 in hierarchy. Its formula is

$$\mathrm{SR2C}_{\mathrm{lca2}}(s_1, s_2) = -\frac{\log(\min \mathrm{len}(s_1, s_2))}{2D_{\max}} \tag{8}$$

where $D_{\max}$ is the maximum depth of concepts in hierarchy and $\min \mathrm{len}(s_1, s_2)$ is the shortest path between s_1 and s_2.

Differently, Wu and Palmer (WUP) [35] measures semantic relatedness by computing the depth of two concepts s_1 and s_2 and the depth of least common subsumer $(\mathrm{LCS}_{s1,s2})$ of s_1 and s_2

$$\mathrm{SR2C}_{\mathrm{wup}}(s_1, s_2) = \frac{2d(\mathrm{LCS}_{s1,s2})}{d(s_1) + d(s_2)} \tag{9}$$

where $d(s)$ is a depth function counting nodes from root node to the nodes of s_1, s_2 or $\mathrm{LCS}_{s1,s2}$.

In Resnik (RES) [37], information content (IC) is introduced to measure semantic similarity between two concepts s_1 and s_2 applying Least Common Subsumer (LCS) of two concepts s_1 and s_2 such that

$$\mathrm{SR2C}_{\mathrm{res}}(s_1, s_2) = \mathrm{IC}(\mathrm{LCS}_{s1,s2}). \tag{10}$$

RES believes that the more information two concepts share, the more similarity they are.

Following RES, LIN [30] measures the semantic relatedness between two concepts s_1 and s_2 in LCS and add a normalization function with the information contents of two given concepts. Its formula is

$$\mathrm{SR2C}_{\mathrm{lin}}(s_1, s_2) = \frac{2\mathrm{IC}(\mathrm{LCS}_{s1,s2})}{\mathrm{IC}(s_1) + \mathrm{IC}(s_2)}. \tag{11}$$

For both RES and LIN, information content is computed by $\mathrm{IC}(\mathrm{LCS}_{s1,s2}) = \mathrm{MAX}_{s \in S(s_1,s_2)}[-\log p(s)]$. An improved IC function for LCS value is later provided in SVH (Seco,Veale and Hayes) [38]) as follows:

$$\mathrm{IC}_{\mathrm{SVH}}(s_i) = 1 - \frac{\log(\mathrm{hypos}(s_i) + 1)}{\log(|S|)} \tag{12}$$

where hypos function counts hyponym links of concept s_i in WN, and $|S|$ is the number of total concept nodes of in WN. In this research, $\mathrm{IC}_{\mathrm{SVH}}(S_i)$ is adopted to replace the original IC functions used in the methods of $\mathrm{SR2C}_{\mathrm{wup}}(S_1, S_2)$, $\mathrm{SR2C}_{\mathrm{res}}(S_1, S_2)$, and $\mathrm{SR2C}_{\mathrm{lin}}(S_1, S_2)$.

B. SR2W Methods

SR2W is to find the semantic relatedness between any two words. Often, it has two approaches: one is to find the similarity employing a knowledge base such as a taxonomy like WN, and the other is to find the similarity employing a corpus or a large text by looking for the co-occurrence of the two words.

For the knowledge-based approach, the problem is often defined as finding the maximum similarity pair of two words from any two near-synonym sets with the formula as follows:

$$\mathrm{SR2W}(w_p, w_q) = \max_{w_p \in S_q, w_q \in S_q} \mathrm{SR2C}(w_p, w_q) \tag{13}$$

where SR2C is described in Section VI-A.

For the corpus-based approach, the research of information retrieval contributes a lot. For example, PMI-IR [42] measures semantic similarity of two words by finding the co-occurrence of two words with a normalization. Its formula is as follows:

$$\mathrm{SR2W}_{\mathrm{PMI-IR}}(w_1, w_2) = \log_2 \frac{\mathrm{hits}(w_1 \ \mathrm{AND} \ w_2) \times N}{\mathrm{hits}(w_1) \times \mathrm{hits}(w_2)} \tag{14}$$

where N is the total documents in search engine, and hits is the number of search hits of query words.

Alternatively, SOC-PMI [21] measures the semantic relatedness of two words by computing the important neighbor co-occurrence when co-occurrence of two words is very low but

with high co-occurrence of sorted common neighbors. However, SOC-PMI needs a complex preprocessing in counting each document in a large corpus by sliding a small window size. When text as a corpus is small, it is not recommended to use this method due to the efficiency consideration.

C. SR2T Methods

SR2T is to find the semantic similarity between any two texts. In literature, there are two basic methods, which are the following.

- Lesk [26]: it finds semantic similarity between two texts by counting the overlap of words between two texts.
- Term vector [3]: it finds semantic similarity between two texts by computing the cosine similarity of two term vector in the space model of TFIDF.

Besides these two basic methods, there are many other methods that are proposed in literature. In the following, we discuss some of them.

STASIS [28] measures semantic relatedness between two texts T_1 and T_2 through a cosine similarity by building a joint set $T = T_1 \cup T_2$ and then giving similarity score to each word of T_1 and T_2 both appeared and not appeared in T based on a word similarity measure.

Semantic text similarity (STS) [21] adopts a combined method based on SOC-PMI's word similarity, where semantic relatedness between two texts is computed by combining string similarity, semantic similarity, and common-word order similarity with normalization. In this method, its string similarity between two words is computed as follows:

$$\alpha = \omega_1\nu_1 + \omega_2\nu_2 + \omega_3\nu_3 \tag{15}$$

where ω is the weight and ν is the string match value in three different modes. The semantic similarity between two words adopts SOC-PMI's word similarity. By combining string similarity and semantic similarity, its combined semantic similarity is computed as follows:

$$\mathrm{SR2T}_{\mathrm{STS}}(X,Y) = \frac{(\delta + \sum_{1=1}^{|\rho|}\rho_i) \times (m+n)}{2mn} \tag{16}$$

where X and Y are two input texts with m number words in X and n number of words in Y, δ is number of common words in both texts, and ρ is iteratively extracted from the semantic relatedness matrix of the non-common words between two texts X and Y with extraction rule of $\rho_1 > \rho_2 > \cdots \rho_{m-\delta}$.

Tsatsaronis *et al.* [40] measures semantic relatedness between two texts based on implicit semantic links between the words of WordNet (WN). This method requires preprocessing and requires pretty much computing time because of the computing complexity.

Latent semantic analysis (LSA) [22] is a method where term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by singular value decomposition (SVD) on a term-by-document matrix (T) that represents the corpus.

ESA (explicit semantic analysis) [12] is a variation on the standard vectorial model in which the dimensions of the vector are directly equivalent to abstract concepts. When an article in Wikipedia represents a concept in the ESA vector, the method is called as WikiESA, where the relatedness of a term to a concept is defined by the $\mathrm{tf}^*\mathrm{idf}$ score. WikiESA uses the cosine of the two concept vectors in a high-dimensional space to measure semantic relatedness. Similarly, when the gloss of a WordNet synset represents a concept, method can be called as WordNetESA.

Salient semantic analysis (SSA) [19] measures semantic relatedness by analyzing the links among documents in Wikipedia. Similar to Term Vector and ESA, the cosine similarity for two concept vectors in space model is used. SSA built a corpus by analyzing Wikipedia articles yet it is hard to build required dimensions for documents.

D. SRCT Method: Our Proposal

SRCT method, introduced in this subsection, is to measure semantic relatedness $\mathrm{SRCT}(T_1, T_2)$ to identify the similarity score between two gloss texts T_1 and T_2 such that $\mathrm{SRCT}(T_1, T_2) = \mathrm{SR}(T_1, T_2)$ [*see* (10)]. It can be computed in two methods of knowledge-based and corpus-based.

1) SRCT_K*: A Knowledge-Based Method:* SRCT_K method computes semantic relatedness between and utilizing WN knowledge of hypernyms and hyponyms.

Given two texts T_1 and T_2, w_p and w_q are the non-common words of T_1 and T_2, respectively. δ is the number of common words appeared in both T_1 and T_2. The computation of $\mathrm{SRCT}_K(T_1, T_2)$ can be presented as follows:

$$\mathrm{SRCT}_K(T_1,T_2) = \frac{\delta + \sum_{j=0}^{n-\delta} \max \mathrm{SR2C}(w_p, w_q)}{n} \tag{17}$$

where $\mathrm{SR2C}_K(w_p, w_q)$ is knowledge-based and alternatively uses the existing methods of LCA, WUP, RES, and LIN of formulas (8)–(11). To differentiate these SRCT_K methods, we name them as SRCT_LCN, SRCT_WU, SRCT_RES, and SRCT_LIN, respectively.

2) SRCT_C*: A Corpus-Based Method:* SRCT_C method computes semantic relatedness between T_1 and T_2 utilizing the statistical information of a corpus.

Given two gloss texts as two word lists $T_1 = (w_1, w_2, \ldots, w_m)$ and $T_2 = (w_1, w_2, \ldots, w_n)$ where $m = |T_1|$, $n = |T_2|$, and $\{k\}$ are the common words of T_1 and T_2 with $\delta = |k|$. Then, we sequentially find the non-common word lists $\{w_p\} \in T_1$ and $\{w_q\} \in T_2$ but $w_p, w_q \notin k$.

Now, we construct three semantic relatedness matrixes of M_1, M_2, and M_3 in the form of $e_{ij}(m-\delta)(n-\delta)$ from non-common words $\{w_p\}$ and $\{w_q\}$ of T_1 and T_2, such that

$$e_{ij}M_1 = \alpha \tag{18}$$

$$e_{ij}M_2 = \mathrm{TMI}(s_i, w_p, w_q) \tag{19}$$

$$e_{ij}M_1 = \phi e_{ij}M_1 + \gamma e_{ij}M_2 \tag{20}$$

where α is the string similarity defined in formula (15), ϕ and γ are the weights of $e_{ij}M_1$ and $e_{ij}M_2$ with $\phi + \gamma = 1$, and $\mathrm{TMI}(s_i, w_p, w_q)$ is a co-occurrence function returning a semantic relatedness value between s_i, w_p, and w_q.

Given M_3, we find out the maximum e_{ij} value $\rho_1 = \max(e_{ij}M_3)$ and then cross out the row and column that contain ρ_1 and construct a new $M_3' = e_{ij}(m-\delta-1)(n-\delta-1)$ to further find the maximum e_{ij} value $\rho_2 = \max(e_{ij}M_3')$. This

process continues until $m-\delta-i=0$ or $n-\delta-i=0$ and a list $(\rho_1,\rho_2,\ldots,\rho_i)$ is obtained.

Given the list $(\rho_1,\rho_2,\ldots,\rho_i)$, compute $\mathrm{SRCT}_C(T_1,T_2)$ applying (19) of STS [21] such that

$$\mathrm{SRCT}_C(T_1,T_2)=\frac{(\delta+\sum_{1=1}^{|\rho|}\rho_i)\times(m+n)}{2mn}. \tag{21}$$

SRCT_C is a combined method based on PMI-IR [42] and STS [21]. For semantic relatedness between two non-common words w_p and w_q of T_1 and T_2, that is, $\mathrm{TMI}(s_i,w_p,w_q)$ shown in (22), it does not only consider semantic co-occurrence of w_p and w_q in a corpus but also semantically relates to a synset s_i in WN to see whether s_i, w_p and w_q have co-occurrence. Particularly, TMI is computed as follows:

$$\mathrm{TMI}(s_i,w_p,w_q)=\log_2\frac{p(s_i,w_p,w_q)}{p(s_i)p(w_q)p(w_q)} \tag{22}$$

where p is a probability function of co-occurrence and $p(s_i,w_p,w_q)$ is

$$p(s_i,w_p,w_q)=\frac{1+p(s_i,w_p)+p(s_i,w_q)+p(w_p,w_q)-p(s_i)-p(w_p)-p(w_q)}{2}. \tag{23}$$

By deduction, $\mathrm{TMI}(s_i,w_p,w_q)$ is further computed as follows:

$$\mathrm{TMI}(s_i,w_p,w_q)=\log\frac{N+\mathrm{co}(s_i,w_p)+\mathrm{co}(s_i,w_q)+\mathrm{co}(w_p,w_q)-\mathrm{fr}(s_i)-\mathrm{fr}(w_p)-\mathrm{fr}(w_q)}{2\times\mathrm{fr}(s_i)\times\mathrm{fr}(w_p)\times\frac{\mathrm{fr}(w_q)}{N^2}} \tag{24}$$

where co-function with two arguments is the two-word co-occurrence, and fr function with one argument is the frequency of each word. If s_i has three words w_1, w_2, and w_3, the $\mathrm{fr}(s_i)$ will use the query like (w_1 OR w_2 OR w_3) and $\mathrm{co}(s_i,w_t)$ will use the query like (w_1 OR w_2 OR w_3) AND w_t.

Since SRCT_C method has combined with the methods of PMI-IR and STS, we rename this method as SRCT_CPS.

VII. Experiments and Evaluation

A. Evaluation Setting

We adopt nine steps to create a larger English-Chinese bilingual vocabulary for making experiment on proposed SRCT methods using the NSG framework. Particularly, the steps are as follows.

1) Collected English WN dictionary from Internet and prepared an English-Chinese bilingual dictionary (BD).
2) Parsed the WN and BD dictionaries to structured data. Each concept/sense of BD has an identity associated with English gloss, Chinese gloss, part-of-speech, and zero or more samples. WN is organized by a set of synonyms (synsets). Each synset contains one or more English words, an English gloss, part of speech for each word, and some samples.
3) Implemented MSCA algorithm and semantic relatedness API for processing all elements of potential near-synonym sets from preprocessing.
4) Implemented different semantic relatedness methods of Lesk [26], TermVector [3], STASIS [28], WikiESA [12], WordNetESA, STS2008 [21], SRCT_LCN, SRCT_WUP, SRCT_RES, SRCT_JC, SRCT_LIN, and SRCT_CPS.
5) Implemented both direct and indirect approaches (discussed in Section V-B of identifying near synonym sets from potential near synonym sets for each semantic relatedness method mentioned in 4) above.
6) Provided a search engine with index of Wikipedia articles, WN glosses, BD glosses and samples based on CLucene Library (sourceforge.net/projects/clucene/).
7) Retrieved process data in file and store experiment results in file with different semantic relatedness methods in both direct and indirect approaches.
8) Use Intel Library TBB (threadingbuildingblocks.org/) for parallel computing.

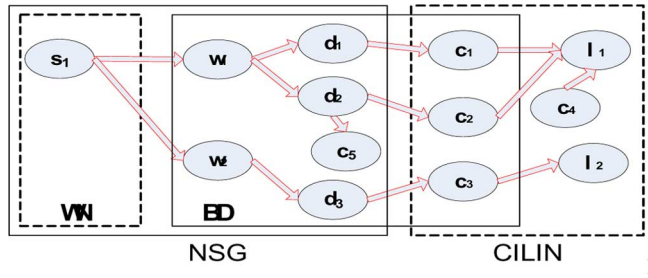


Fig. 5. Paths from NSG to CILIN.

B. Evaluation Method

To evaluate our experiment results, we have developed two evaluation methods: one is to evaluate the result of all identified near-synonym sets by checking whether they are falling into Tongyici CILIN [36]. If any pair of identified near synonymous words co-occur in CILIN, then this pair of near synonyms is deemed to be accurate. The second evaluation method is human evaluation using an implemented human verification interface.

1) Evaluation Through CILIN: The evaluation method of the first is computed on a common word set $C=\{c\}$, shown in Fig. 5, where $C\in\mathrm{BD}\cap\mathrm{CILIN}$. The method assumes there exist paths d.c.l. from $d\in D\in\mathrm{NSG}$ to $c\in C$ to $l\in L\in\mathrm{CILIN}$. Each $l\in L$ is a synonym set.

To fairly evaluate each method, we have experimented, we have built a Validation Set Finding (VSF) algorithm, shown in Fig. 6, to find validation set P_3 for all experimented methods.

Based on VSF algorithm, the evaluation matrix can be described by an Accuracy Ratio (AR) as follows:

$$\mathrm{AR}=\frac{|P_2\cap P_3|}{|P_3|} \tag{25}$$

where P_2 is computed from NSF algorithm shown in Fig. 4 and P_3 is computed from VSF algorithm shown in Fig. 6.

The AR assumes that the more identified near synonym sets of NSG intersect with the synonym sets of CILIN, the more accurate they are. It also implies that the higher of AR, the more accurate for the applied method.

Since there exists nonintersected sets between NSG and CILIN, the AR can only evaluate roughly about 1/10 of the

VSF Algorithm. Finding a validation set
Input: *L, C, a 1-to-1 relation* R_4*: D→C, a m-to-m relation* R_5*: C→L, and* $P \in$ *NSG as defined in Section III.*
Output: $P_3 \leftarrow \varnothing$
for $(P.s_i)\{$ /* $P.s_i \in P$ */
  for$(w_j.d_j.c_j \wedge w_k.d_k.c_k)\{$ /* $w_j.d_j, w_k.d_k \in s_i, j \neq k$* /
    $c_j \leftarrow R_4(w_j.d_j)$ /* $w_j.d_j$ *is a valued concept in* w_j */
    $c_k \leftarrow R_4(w_k.d_k)$ /* $w_k.d_k$ *is a valued concept in* w_k */
    if$(\neg\varnothing \leftarrow R_5(c_j) \cap R_5(c_k))$ **then**
      $P_3 \leftarrow P_3 \cup w_j.d$
      $P_3 \leftarrow P_3 \cup w_k.d$ }
} /* *building a set of validated sets* */

Fig. 6. VSF algorithm.

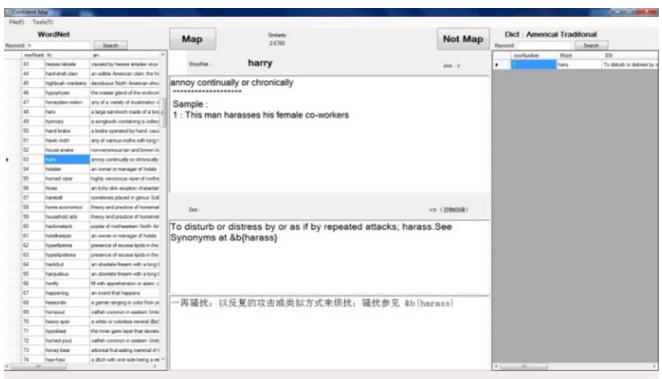


Fig. 7. Human verification interface.

total identified near-synonyms of NSG (i.e., 11260 pairs). Thus, we need additional evaluation method whenever possible.

2) Human Varification Interface: To complement the insufficiency of the evaluation through CILIN, we design a human verification interface for human to manually verify the correctness of near-synonyms, as shown in Fig. 7. This interface is also a preliminary tool for collaborative vocabulary editing that ensures the semantic consistency between words. The tool works by editors' evaluation based on the similarity scores that are automatically created by different methods that generate the identified near-synonym sets.

The design and implementation of the human verification interface is particularly discussed in the research field of collaborative vocabulary editing (*see* some of our research results in [14] and [17]), which is out of the research scope of this paper. However, the major technical design principles are serialization and lock (i.e., concurrency control for collaborative editing), multiversioning (i.e., disagreements reservation), arbitration (i.e., final decision making for conflicts), and sinks as history (i.e., for any arbitrated synonym set, it is no longer editable).

In Fig. 7, all identified near-synonym sets are displayed in the left of the table ordered by the similarity score and can be searched by keyword. The similarity score is shown in the middle of the top. The collaborative editor can evaluate the correctness of the identified near synonym sets according to English explanation, samples of WN, English-Chinese explanation, and samples of BD.

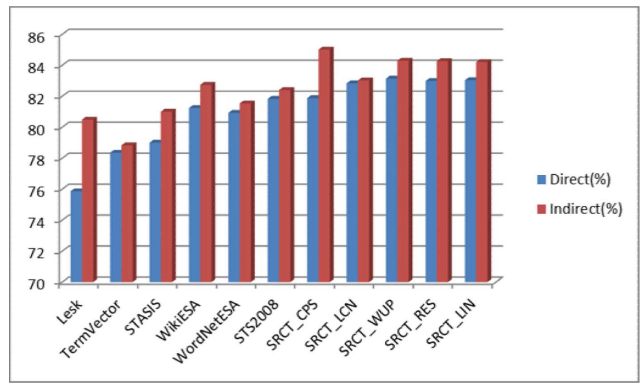


Fig. 8. Comparison Result of accuracy ratio for different approaches.

Based on this interface, experiments are set to compare the time costs of identifying synonym sets between software agents and human editors based on the following assumptions.

1) There are no concurrent editing on a same block. A *block* refers to a set of connected paths initiated by a WN synset, defined as P' in NSG framework of Section III or intuitively a hierarchical record rooted from a $\backslash s$ in Fig. 3.
2) Every editor is also an arbitrator for the edited content.
3) Human verification is the continuous work of software agent that has identified the near synonym sets.

Thus, experiments were made in two groups. One group as a control group directly identifies synonym sets based on given blocks without giving near synonym sets. The another group is the target group which verifies the near synonym sets identified by software agent earlier.

C. Result and Discussion

Methods of measuring semantic relatedness between texts include baseline methods (Lesk, TermVector), non-baseline methods (STASIS, WikiESA, WordNetESA, STS2008) and our proposed methods prefixed with SRCT (knowledge-based SRCT_LCN, SRCT_WUP, SRCT_RES and SRCT_LIN, and corpus-based SRCT_CPS). Since all these methods can be computed in our proposed NSG framework, we evaluate these methods in both direct approach and indirect approach (*see* Section V-B).

1) Accuracy Ratio: Fig. 8 shows the experimental results of semantic relatedness by both direct approach (shown in the left bar) and indirect approach (shown in the right bar), using different semantic relatedness methods described in Section VI. For the direct approach, methods of SRCT-prefixed outperform 1% or 2% comparing with the non-baseline methods using direct approach like WikiESA, WordNetESA, and STS2008. For the indirect method, methods of SRCT-prefixed outperform 2% more than non-baseline methods using indirect approach. The results of SRCT-prefixed methods in indirect approach outperform 3% more than the non-baseline methods using direct approach.

2) Time Cost: Table II shows the time cost of the control group consisting of ten human editors to directly identify the

TABLE II
EXPERIMENT RESULT OF CONTROL GROUP (MINUTES)

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Total
31.5	33.3	34.4	29.9	30.2	35.9	38.3	40.1	36.2	29.6	339.4

TABLE III
EXPERIMENT RESULT OF TARGET GROUP (MINUTES)

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Total
WikiESA	8.54	9.67	8.23	10.68	11.2	13.6	7.98	7.3	9.3	8.13	94.63
SRCT_CPS	6.38	7.96	6.84	6.12	5.77	8.12	6.12	6.1	7.21	5.42	66.09

TABLE IV
COMPARISON BETWEEN TARGET GROUP AND CONTROL GROUP (HOUR)

Control Group	**Target Group**				
All Human	**Part Human**	Direct Approach	Indirect Approach	Human Verification	Semantic Relatedness Method
63.69		0.158	0.258		Lesk
		0.004	0.004		TermVector
		2.230	3.000		STASIS
	21.444	0.702	4.537	16.907	WikiESA
		0.025	0.077		WordNetESA
		0.311	0.837		STS2008
	15.102	1.461	3.294	11.808	SRCT_CPS
		2.279	2.277		SRCT_LCN
		2.287	2.286		SRCT_WUP
		2.283	2.287		SRCT_RES
		0.820	0.825		SRCT_LIN

Note: *All Human* refers to that human editors identify synonym sets without near synonym sets identified by software agent; Part Human refers to that human editors identify synonym sets by verifying near synonym sets identified by software agents in indirect approach in early time.

synonym sets from 500 blocks. Each editor is assigned 50 blocks.

Table III shows the time cost of the target group consisting of ten human editors to verify the near synonym sets from 500 blocks, which have been identified by software agent in two semantic relatedness methods of WikiESA and SRCT_CPS.

Table IV shows the time cost comparison between the target group and the control group.

In Table IV, software agents have identified near synonym sets from 5630 blocks. To make it comparable between target group and control group, the time used by human editors of Tables II and III is normalized based on the block number 5630.

The experiments in comparison (*see* Table IV and Fig. 9) show that Part Human method of target group is much better than All Human method of control group because the automation of identifying near synonym sets drastically reduces the human editing time. The experiments also show that indirect approach has consumed more computing time than direct approach. This is because the indirect approach first need the iterative computing along direct path in the NSG graph. The baseline methods like Lesk and TermVector are the quickest. Wiki-based methods cost more time than baseline methods since the high dimensions of articles in explicit semantic vectors. SRCT_ prefixed methods are moderate in terms of time cost. Except for SRCT_CPS, other SRCT-prefixed methods spend less time than STASIS and WikiESA but more time than WordNetESA and STS2008.

In summary, the experiment results indicate that our proposed SRCT method is promising. By this method, the precision rate

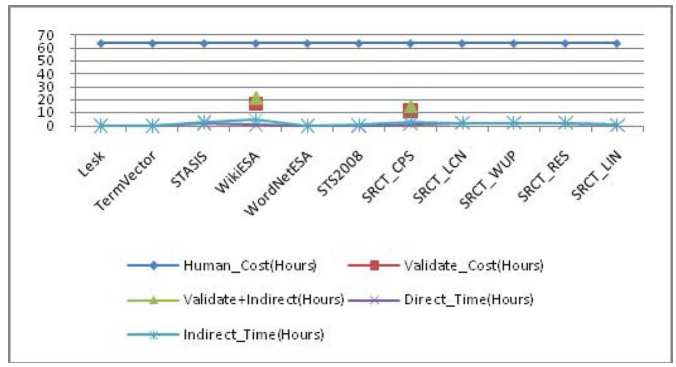


Fig. 9. Comparison result of time cost between all human method (human_cost legend) and part human method Validate + Indirect legend, and between direct approach (direct_time legend) and indirect approach (indirect_time legend).

has been increased comparing with both baseline methods and non-baseline methods. This improvement will help save a lot of time for collaborative vocabulary editors when they collaboratively find synonyms and incrementally create new concepts in Chinese-English vocabularies. Meanwhile, when collaborative editors supply some confident mapping information or provide certain standard definition patterns, the algorithms will work more efficiently and further improve the precision rate automatically. In this sense, a learning effect will be produced. Thus, we propose to simultaneously do collaborative work together with automatic computing work. This will greatly reduces the collaboration time of vocabulary editors and further improve the accuracy ratio, reduce computing time, and increase efficiency of collaborative work.

VIII. CONCLUSION

The contribution of this paper is the improvement of multilingual semantic interoperation in cross-organizational enterprise systems through best supporting collaborative vocabulary editing, which provides a semantically consistent vocabulary for building semantically consistent business processes and documents. It has proposed a NSG framework for automating the process of multilingual concept disambiguation to automatically find multilingual near synonym sets within a multilingual dictionary. The NSG framework is designed on a three-level concept-connected near synonym graph, where potential near synonym sets are categorized based on WordNet synsets, and usable near synonym sets are identified from the potential near synonym sets through a proposed probability model for concept value identification. This model is computed using the proposed SRCT semantic relatedness methods, which is experimented and compared with some existing methods of computing semantic relatedness.

This paper has several particular contributions as follows.

1) Proposed a concept-connected NSG framework based on WordNet synset. It makes easier to find all potential near synonym sets and reduces the overhead in finding potential near synonym sets.
2) Proposed a probability model for concept value identification. It not only considers the WordNet concept-connected words, word-connected definitions and the semantic relatedness between WordNet synsets and word definitions of

bilingual dictionary but also takes care of the semantic relatedness between word definitions of bilingual dictionary.

3) Provided a Near Synonym Finding (NSF) algorithm to identify all near-synonym sets from sets of max concept-similar values. It also implements two approaches of computing semantic relatedness along the direct and indirect paths of NSG.
4) Introduced a new SRCT method of computing semantic relatedness between two texts by not only incorporating the existing methods of PMI-IR [42] and STS [21] but also accounting of the knowledge of WordNet synsets.
5) Proposed two evaluation methods of CILIN-based [36] and human verification interface based.

In addition, this research has implemented the new SRCT based methods. Experiments have been made on these methods by comparing them with the existing methods of semantic relatedness. It shows that SRCT based methods outperform the existing methods.

In future, we plan to improve the human verification interface so that all identified near synonym sets can be quickly and conveniently verified as the final usable synonym sets.

REFERENCES

[1] L. Aldin and S. de Cesare, "A literature review on business process modelling: New frontiers of reusability," *Enterprise Inform. Syst.*, vol. 5, no. 3, pp. 359–383, Aug. 2011.

[2] R. Baayen and R. Lieber, "Word frequency distribution and lexical semantics," *Comput. Humanities*, vol. 30, pp. 281–291, 1997.

[3] R. Baeza-Yates and B. Ribeiro-Neto*, Modern Information Retrieval*. New York: ACM Press, 1999.

[4] M. Carpuat, P. Fung, and G. Ngai, "Aligning word sense using bilingual corpora," *ACM Trans. Asia Language Inform. Process.*, vol. 5, no. 2, pp. 89–120, Jun. 2006.

[5] G. Cengic and K. Åkesson, "On formal analysis of IEC 61499 applications, part a: Modeling," *IEEE Trans. Ind. Informat.*, vol. 6, no. 2, pp. 136–144, May 2010.

[6] G. Cengic and K. Åkesson, "On formal analysis of IEC 61499 applications, part B: Execution semantics," *IEEE Trans. Ind. Informat.*, vol. 6, no. 2, pp. 145–154, May 2010.

[7] H.-H. Chen, C.-C. Lin, and W.-C. Lin, "Building a Chinese-English WordNet for translingual applications," *ACM Trans. Asian Language Inform.n Process.*, vol. 1, no. 2, pp. 103–122, Jun. 2002.

[8] L.-L. Chang, K.-J. Chen, and C.-R. Huang, "A lexical-semantic analysis of Mandarin Chinese verbs: Representation and methodology," *Computational Linguistics and Chinese Language Processing*, vol. 5, no. 1, pp. 1–18, Feb. 2000.

[9] B. Dias-Da-Silva, "Brazilian Portugese WordNet: A computational linguistic exercise of encoding bilingual relational lexicons," *Int. J. Computat. Linguistics Appl.*, vol. 1, no. 1–2, pp. 137–150, Jan.–Dec. 2010.

[10] P. Edmonds and G. Hirst, "Near-Synonymy and lexical choice," *Computat. Linguistics*, vol. 28, no. 2, pp. 105–144, 2002.

[11] EuroWordNet., 2012. [Online]. Available: http://www.illc.uva.nl/EuroWordNet/

[12] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *J. Artif. Intell. Res.*, vol. 34, pp. 443–498, 2009.

[13] V. Güngör, D. Sahin, T. Kocak, S. Ergüt, C. Buccella, C. Cecati, and G. Hancke, "Smart grid technologies: Communication technologies and standards," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 529–539, Nov. 2011.

[14] J. Guo*, Collaborative Concept Exchange*. Saarbrücken, Germany: VDM Pub., 2008.

[15] J. Guo, "Collaborative conceptualization: Towards a conceptual foundation of interoperable electronic product catalogue system design," *Enterprise Inform. Syst.*, vol. 3, no. 1, pp. 59–94, 2009.

[16] J. Guo, "Collaboration role in semantic integration for electronic marketplace," *Int. J. Electron. Bus.*, vol. 8, no. 6, pp. 528–549, 2010.

[17] J. Guo, I.-H. Lam, C. Chan, and G. Xiao, "Collaboratively maintaining semantic consistency of heterogeneous concepts towards a common concept set," in *Proc. 2nd ACM SIGCHI Symp. Eng. Interactive Comput. Syst. (EICS 2010)*, 2010, pp. 213–218.

[18] J. Guo, L. Xu, Z. Gong, C.-P. Che, and S. Chaudhry, "Semantic inference on heterogeneous E-marketplace activities," *IEEE Trans. Syst., Man, Cybern.—Part A: Syst. Humans*, vol. 42, no. 2, pp. 316–330, Mar. 2012.

[19] S. Hassan and R. Mihalcea, "Semantic relatedness using salient semantic analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 884–889.

[20] D. Inkpen, "A statistical model for near-synonym choice," *ACM Trans. Speech and Language Process.*, vol. 4, no. 1, article 2, pp. 2.1–2.17, Jan. 2007.

[21] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discovery from Data*, vol. 2, no. 2, pp. 1–25, 2008.

[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*, vol. 104, no. 2, pp. 211–240, Apr. 1997.

[23] M. L. Rosa, A. ter Hofstede, P. Wohed, H. Reijers, J. Mendling, and W. van der Aalst, "Managing process model complexity via concrete syntax modifications," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 2, pp. 255–265, May 2011.

[24] M. L. Rosa, P. Wohed, J. Mendling, A. ter Hofstede, H. Reijers, and W. van der Aalst, "Managing process model complexity via abstract syntax modifications," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 614–629, Nov. 2011.

[25] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998, pp. 265–283.

[26] M. Lesk, "Automatic sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proc. SIGDOC'86*, 1986, pp. 24–26.

[27] S. Li, L. Xu, X. Wang, and J. Wang, "Integration of hybrid wireless networks in cloud services oriented enterprise information systems," *Enterprise Inform. Syst.*, vol. 6, no. 2, pp. 165–187, 2012.

[28] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[29] X. Li, S. Szpakowicz, and S. Matwin, "A WordNet-based algorithm for word sense disambiguation," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, vol. 2, pp. 1368–1374.

[30] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.

[31] M. Metzger and G. Polakow, "A survey on applications of agent technology in industrial process control," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 570–581, Nov. 2011.

[32] G. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[33] MultiWordNet, 2012. [Online]. Available: http://multiwordnet.fbk.eu/english/home.php

[34] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput, Surveys*, vol. 41, no. 2, article 10, pp. 10.1–10.69, Feb. 2009.

[35] M. Palmer and Z. Wu, "Verb semantics for English-Chinese translation," *Mach. Translation*, vol. 10, pp. 59–92, 1995.

[36] Research Center for Social Computing and Information Retrieval, Cilin for Synonyms, 2012. [Online]. Available: http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162

[37] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, Jul. 1999.

[38] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," in *Proc. ECAI-04*, 2004, pp. 1089–1090.

[39] F. Suchanek, G. kasneci, and G. Weikum, "YAGO: A large ontology from Wikipedia and WordNet," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, pp. 203–217, 2008.

[40] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," *Artif. Intell. Res.*, vol. 37, pp. 1–39, 2010.

[41] T. Tudorache, N. Noy, S. Tu, and M. Musen, "Supporting collaborative ontology development in protégé," in *Semantic Web—ISWC 2008*, 2008, vol. 5318, pp. 17–32.

[42] P. Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Trans. Inform. Syst.*, vol. 21, no. 4, pp. 315–346, Oct. 2003.

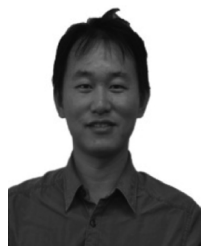
[43] K. Wang, X. Bai, J. Li, and C. Ding, "A service-based framework for pharmacogenomics data integration," *Enterprise Inform. Syst.*, vol. 4, no. 3, pp. 225–245, 2010.
[44] G. Wiederhold and M. Genesereth, "The conceptual basis for mediation services," *IEEE Expert*, vol. 12, no. 5, pp. 38–47, Sep./Oct. 1997.
[45] WordNet, 2012. [Online]. Available: http://wordnet.princeton.edu/
[46] L. Xu, "Enterprise systems: State-of-the-Art and future trends," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 630–640, Nov. 2011.
[47] L. Xu, H. Liu, S. Wang, and K. Wang, "Modeling and analysis techniques for cross-organizational workflow systems," *Syst. Res. Behav. Sci.*, vol. 26, no. 3, pp. 367–389, Mar. 2009.
[48] M. Zdravković, H. Panetto, M. Trajanovic, and A. Aubry, "An approach for formalizing the supply chain operations," *Enterprise Inform. Syst.*, vol. 5, no. 4, pp. 401–421, 2011.

Jingzhi Guo (M'05) received the B.Econ. degree in international business management from the University of International Business and Economics, Beijing, China, in 1988, the M.Sc. degree in computation from the University of Manchester, Manchester, U.K., in 2000 and the Ph.D. degree in Internet computing and e-commerce from Griffith University, Brisbane, Australia, in 2005.

He is currently an Assistant Professor in e-commerce technology with the University of Macau, Macao. His principal research is in the field of concept representation, semantic integration and collaboration systems, mainly applied to the fields of e-commerce, e-marketplace, e-banking, and virtual world.

Li Da Xu (M'86–SM'11) received the M.S. degree in information science and engineering from the University of Science and Technology of China, Hefei, in 1981, and the Ph.D. degree in systems science and engineering from Portland State University, Portland, OR, in 1986.

He serves as the Founding Chair of IFIP TC8 WG8.9 and the Founding Chair of the IEEE SMC Society Technical Committee on Enterprise Information Systems.



Guangyi Xiao (M'10) received the M.S. degree in software engineering from University of Macau, Macao, in 2009. He is currently working towards the Ph.D. degree in e-commerce technology at the Department of Computer and Information Science, University of Macau, Macao.

His principal researches are in the field of controlled vocabulary and business documents, mainly applied to the fields of e-commerce, e-marketplace and virtual world.

Zhiguo Gong (M'10) received the M.S. degree in mathematics from Peking University, Peking, China, in 1998, the B.S. degree in mathematics from Heibei Normal University, Heibei, China, in 1983, and the Ph.D. degree from the Department of Computer Science, Chinese Academy of Science, Beijing, China, in 1998.

He is currently an Associate Professor and Head of Computer Science at the Department of Computer and Information Science, University of Macau, Macao. His research interests are databases, digital library, web information retrieval and web mining.