A Thesaurus-Based Semantic Classification of English Collocations

Chung-chi Huang, Chiung-hui Tseng, Kate H. Kao, Jason S. Chang
ISA, National Tsing Hua University
{u901571, smilet, msgkate, jason.jschang}@gmail.com

Abstract

We propose a new method for organizing the numerous collocates into semantic thesaurus categories. The approach introduces a thesaurus-based semantic classification model automatically learning semantic relations for classifying adjective-noun (A-N) and verb-noun (V-N) collocations into different categories. Our model uses a random walk over weighted graph derived from WordNet semantic relation. We compute a semantic label stationary distribution via an iterative graphical algorithm. The performance for semantic cluster similarity and the conformity of semantic labels are both evaluated. The resulting semantic classification establishes as close consistency as human judgments. Moreover, our experimental results indicate that the thesaurus structure is successfully imposed to facilitate grasping concepts of collocations. It might improve the performance of the state-of-art collocation reference tools.

Keywords: Collocations, Semantic classification, Semantic relations, Random walk algorithm, Meaning access index.

1. Introduction

Submitting queries (e.g., a search keyword "*beach*" for a set of adjective collocates) to collocation reference tools typically return many collocates (e.g., collocate adjectives with a pivot word "*beach*": "*rocky*", "*golden*", "*beautiful*", "*pebbly*", "*splendid*", "*crowded*", "*superb*", etc.) extracted from a English corpus. Applications of automatic extraction of collocations such as *TANGO* (Jian, Chang & Chang, 2004) have been created to answer queries of collocation usage.

Unfortunately, existing collocation reference tools sometimes present too much information in a batch for a single screen. With web corpus sizes rapidly growing, it is not uncommon to find thousands collocates for a query word. An effective reference tool might strike a balance between quantity and accessibility of information. To satisfy the need for presenting a digestible amount of information, a promising approach is to automatically partition words into various categories to support meaning access to search results and thus give a thesaurus index.

Instead of generating a long list of collocates, a good, better presentation could be composed of clusters of collocates inserted into distinct semantic categories. We present a robust thesaurus-based classification model that automatically group collocates of a given pivot word focusing on: (1) the adjectives in adjective-noun pairs (A-N); (2) the verbs in verb-noun pairs (V-N); and (3) the nouns in verb-noun pairs (V-N) into semantically related classes.

Our model has determined collocation pairs that learn the semantic labels automatically during random walk algorithm by applying an iterative graphical approach and partitions collocates for each collocation types (A-N, V-N and V-N mentioned above). At runtime, we start with collocates in question with a pivot word, which is to be assigned under a set of semantically related labels for the semantic classification. An automatic classification model is developed for collocates from a set of A-N and V-N collocations. A random walk algorithm is proposed to disambiguate word senses, assign semantic labels and partition collocates into meaningful groups.

As part of our evaluation, two metrics are designed. We assess the performance of collocation clusters classified by a robust evaluation metric and evaluate the conformity of semantic labels by a three-point rubric test over collocation pairs chosen randomly from the results. Our results indicate that the thesaurus structure is successfully imposed to facilitate grasping concepts of collocations and to improve the functionality of the state-of-art collocation reference tools.

2. Related Work

2.1 Collocations

The past decade has seen an increasing interest in the studies on collocations. This has been evident not only from a collection of papers introducing different definitions of the term “collocation” (Firth, 1957; Benson, 1985; Lewis, 1997), but also from a number of research on collocation teaching/acquisition associating to language learning (Lewis, 2000; Nation, 2001). When analyzing Taiwanese EFL writing, Chen (2002) and Liu (2002) investigated that the common lexical collocational error patterns include verb-noun (V-N) and adjective-noun (A-N). Furthermore, with the technique progress of NLP, Word Sketch (Kilgarriff & Tugwell, 2001) or *TANGO* (Jian, Chang & Chang, 2004) became the novel applications as collocation reference tools.

2.2 Meaning Access Indexing

Some attention has been paid to the investigation of the dictionary needs and reference skills of language learners (Scholfield, 1982; Béjoint 1994), especially the structure for easy comprehending. According to Tono (1992 & 1997), menus that summarize or subdivide definitions into groups ahead of entries in dictionaries would help users with limited reference skills. The System “Signposts” of the *Longman Dictionary of Contemporary English*, 3rd edition, the index “Guide Word” of the *Cambridge International Dictionary of English*, as well as the “Menus” of the *Macmillan English Dictionary for Advanced Learners* all value the principle.

2.3 Similarity of Semantic Relations

The construction of practical, general word sense classification has been acknowledged to be one of the most ambitious and frustrating tasks in NLP (Nirenburg & Raskin, 1987), even *WordNet* with more significant contribution of a wide range of lexical-semantic resources (Fellbaum, 1998). Lin (1997) presented an algorithm for word similarity measure by its distributional similarity. Unlike most corpus-based word sense disambiguation (WSD) algorithms where different classifiers are trained for separate words, Lin used the same local context database as the knowledge sources for measuring all word similarities. Distributional similarity allows pair wise word similarity measure to deal with infrequent words or unknown proper nouns. However, compared to distributional similarity measure, our model by random walk algorithm has remarkable feature to deal with any kind of constraints, thus, not limited to pair-wise word similarities, and can be improved by adding any algorithm constraints available.

More specifically, the problem is focused on classifying semantic relations. Approaches presented to solve problems on recognizing synonyms in application have been studied (Lesk, 1986; Landauer and Dumais, 1997). However, measures of recognizing collocate similarity are not as well developed as measures of word similarity, the potential applications of semantic classification are not as well known. Nastase and Szpakowicz (2003) presented how to automatically classify a noun-modifier pair, such as “laser printer”, according to the semantic relation between the head noun (printer) and the modifier (laser). Turney (2006) proposed the semantic relations in noun pairs for automatically classifying. As for VerbOcean, a semi-automatic method was used to extract fine-grained semantic relations between verbs (Chklovski & Pantel, 2004). Hatzivassiloglou and McKeown (1993) presented a method towards the automatic identification of adjectival scales. More recently, Wanner et al. (2006) has sought to semi-automatically classify the collocation from corpora by using the lexical functions in dictionary as the semantic typology of collocation elements. Nevertheless, there is still a lack of fine-grained semantically-oriented organization for collocation.

3. Methodology

We focus on the preparation step of partitioning collocations into categories: providing each word with a semantic label and thus presenting collocates under thesaurus categories. The collocations with the same semantic attributes by the batch size are then returned as the output. Thus, it is crucial that the collocation categories be fairly assigned for users’ easy-access. Therefore, our goal is to provide a semantic-based collocation thesaurus that automatically adopts characterizing semantic attributes. Figure 1 shows a comprehensive framework for our unified approach.

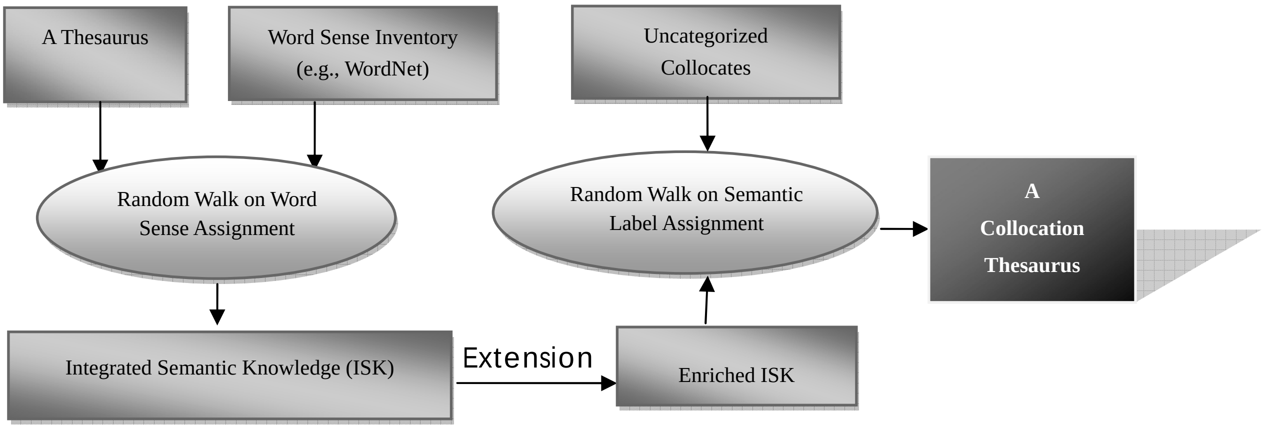
A Thesaurus
Word Sense Inventory (e.g., WordNet)
Uncategorized Collocates
Random Walk on Word Sense Assignment
Random Walk on Semantic Label Assignment
A Collocation Thesaurus
Integrated Semantic Knowledge (ISK)
Extension
Enriched ISK

Figure 1.A comprehensive framework for our classification model.

3.1 Problem Statement

We are given (1) a set of collocates *Col* = {C_1, C_2, …, C_n} (e.g., *sandy, beautiful, superb, rocky, etc*.) denoted with a set of part-of-speech tags *P*, {*P* ∈ *Pos* | *P* = adjective P_{adj}, verb P_v, or noun P_n} for a pivot word *X* (e.g., *beach*) extracted from a corpus of English texts (e.g., *British National Corpus*); (2) a combination of thesaurus categories (e.g., *Roget’s Thesaurus*), *TC* = {(*W*, *P*, *L*) | *W* ∈ *Voc*, *P* ∈ *Pos*, *L* ∈ *Cat*}, where *Voc* is the thesaurus vocabulary words *W*, ordered by general-purpose topics hereinafter called the semantic labels (e.g., feelings, materials, art, food, time, etc.), *Cat* = {L_1, L_2, …, L_m}, with conceptual-semantic attributes as the basis for organization; and (3) a lexical database (e.g., *WordNet*) as our word sense inventory *SI* for semantic relation population. *SI* is equipped with a measure of semantic relatedness of *W*, REL(*S*, *S*’) encoding semantic relations REL ∈ *SR* holding between word sense *S* and *S*’.

Our goal is to partition *Col* into subsets *Sub* of similar collocates, *Sub* ⊆ *Col*, by means of an integrated semantic knowledge crafted from the mapping of *TC* and *SI* that is likely to express closely related meanings of *Col* in the same context of *X* mentioned herein *beach*. For this, we use a graph-based algorithm to give collocations a thesaurus index by giving each collocate in *Col* a semantic label *L*.

3.2 Learning to Build a Semantic Knowledge by Iterative Graphical Algorithms

Recall that we attempt to provide each word with a semantic label and partition collocations into thesaurus categories. In order to partition a large-scale collocation input and reduce the out-of-vocabulary (OOV) words occurred, automating the task of building an integrated semantic knowledge base is a necessary step, but also imposes a huge effort on the side of knowledge integration and validation. An integrated semantic knowledge (*ISK*) is defined to interpret a word in triples (*W*, *L*, *S*), i.e., the given word, a semantic label representing one of thesaurus categories, and its corresponding word sense, as cognitive reference knowledge. At this first stage, interconnection is still between words and labels from the given thesaurus category *TC* and not between word senses and semantic labels. For interpreting words in triples (*W*, *L*, *S*) as an *ISK* and corresponding to the fact that there's a limited, almost scarcely found, resource that is intended for such semantic knowledge, we proceeded as follows to establish one comprehensive *ISK* allowing concentrating on our task of populating it with new semantic relations between words and labels, overcoming the problem of constructing a resource from scratch.

3.2.1 Word Sense Assignment for Integrated Semantic Knowledge

In the first stage of the learning process, we used a graph-based sense linking algorithm which automatically assigns senses to all words under a thesaurus category by exploiting semantic relations identified among word senses. It creates a graph of vertices representing a set of words and their admissible word senses in the context of a semantically consistent list. The pseudo code for the algorithm is shown as Figure 2.

By adding synonymous words through semantic relations, it can broaden the word coverage of *TC*, which may reduce significantly the number of OOV words in *TC* and cope with the problem of collocates that form a group by itself. This strategy relies on a set of general-purpose topics as semantic labels *L* in a thesaurus category *TC* and a word sense inventory *SI* encoding semantic relations. *TC* and *SI* are derived from separate lexicographical resources, such as *Longman Lexicon of Contemporary English* and *WordNet*.

The algorithm assumes the availability of a word sense inventory *SI* encoding a set of semantic relations as a measure of semantic relatedness. Given a set of words with corresponding admissible senses in *SI*, we build a weighted graph $G = (V, E)$ for *SI* such that there is a vertex *V* for each admissible sense, and a directed edge *E* for each semantic relation between a pair of senses (vertices).

The input to this stage is a word sense inventory *SI* encoding a set of semantic relations *SR* attributing the senses of *SI*, and a set of words $W = \{w_1, w_2, \ldots, w_n\}$ listed under L_i in a set of semantic labels *Cat* used in a thesaurus *TC*. The semantic relations *SR* comprise REL(*S*, *S'*) where *S* and *S*' are admissible senses in *SI*, and REL is a semantic relation (e.g., synonyms, hypernyms, and hyponyms holding between senses) existing between *S* and *S'* and explicitly encoded in *SI*. Notice that semantic relations typically hold between word senses but not necessarily between words. We apply semantic relations to identify the intended senses for each word in the list. Accordingly these intended senses will form a semantically consistent set with maximal interconnecting relations

We use random walk on the weighted graph *G* encoding admissible senses as vertices *V* and semantic relations *SR* as edges *E* with a view to discovering the most probable sense S^* for *W*. The edges will be stepped through by imaginary walkers during the random walk in a probabilistic fashion. Through the random walk on *G*, the probability of intended senses will converge to a higher than usual level because of the influx via incoming edges representing semantic relations. All vertices in the weighted graph *G* start with a uniform probability distribution. The probability is reinforced by edges that participate in a *SR* until the reinforcement of probability converges for the given sense consistency, leading to a stationary distribution over sense probability P_s, represented as scores Q_s attached to vertices in the graph. In all, the weights on *G* indicating the sense strength converge to arrive at the consistency of senses, which become the output of this learning stage. The procedure is repeated for all word lists in *TC*. Recall that these most probable senses are useful for extending the limited coverage of *TC* and reducing the number of OOV words effectively.

Algorithm 1. Graph-based Word Sense Assignment

Input: A word *W* from a set annotated with a semantic label *L* under a category *Cat* from a thesaurus *TC*;
A word sense inventory *SI* with a measure of semantic relatedness of *W*, REL (*S*, *S'*) encoding semantic relations REL ∈ *SR* holding between word meanings *S* and *S'*.
S is one of the admissible senses of *W* listed in *SI*, and so as *S'* of *W'*.

Output: A list of linked word sense pairs (*W*, *S**)

Notation: Graph $G = \{V, E\}$ is defined for admissible word senses and their semantic relations, where a vertex $v \in V$ is used to represent each sense *S* whereas an edge in *E* represents a semantic relation in *SR* between *S* and *S'*. Word sense inventory *SI* is organized by semantic relations *SR*, where REL (*S*, *S'*), REL ∈ *SR* is used to represent one of the *SR* holding between word sense *S* of *W* and *S'* of *W'*.

PROCEDURE AssignWordSense(*L*,*SI*)

Build weighted graph *G* of word senses and semantic relations

```
(1)   INITIALIZE $V$ and $E$ as two empty sets
      FOR each word $W$ in $L$
          FOR each of $n$ admissible word sense $S$ of $W$ in $SI$, $n = n(W)$
              ADD node $S$ to $V$
      FOR each node pair $(S,S')$ in $V \times V$
          IF $(S$  REL $S') \in SR$ and $S \neq S'$ THEN ADD edge $E(S,S')$ to $E$
      FOR each word $W$ AND each of its word senses $S$ in $V$
(2)       INITIALIZE $P_s = 1/n(W)$ as the initial probability
(2a)      ASSIGN weight $(1-d)$ to matrix element $M_{S,S}$
(2b)      COMPUTE $e(S)$ as the number of edges leaving $S$
          FOR each other word $W' \neq W$ in $L$ AND each of $W'$ senses $S'$
(3)           IF $E(S,S') \in E$ THEN ASSIGN Weight $d/e(S)$ to $M_{S,S'}$
              OTHERWISE ASSIGN 0 to $M_{S,S'}$
```

Score vertices in *G*

```
      REPEAT
          FOR each word $W$ AND each of its word senses $S$ in $V$
(4)           INTIALIZE $Q_S$ to $P_S * M_{S,S}$
              FOR each other word $W' \neq W$ in $L$ AND each of $W'$ senses $S'$
(4a)              INCREMENT $Q_S$ by $P_{S'} * M_{S',S}$
          FOR each word $W$ AND
              Sum $Q_S$ over $n(W)$ senses as $N_w$
          FOR each sense $S$ of $W$
(4b)          Replace $P_S$ by $Q_S/N_w$  so as normalize to sum to 1
      UNTIL probability $P_S$ converges
```

Assign word sense

```
(5)   INITIALIZE $List$ as NULL
      FOR each word $W$
(6)       APPEND $(W,S^*)$ to $List$ where $S^*$ maximizes $P_s$
(7)   OUTPUT $List$
```

Figure 2.Algorithm for graph-based word sense assignment.

The algorithm (referring to Figure 2) for the best sense assignment S^* for W consists of three main steps: (1) construction of a word sense graph; (2) sense scoring using graph-based probability ranking algorithm; and (3) word sense assignment.

In Step 1, the weighted graph $G = (V, E)$ is built by populating candidate $n(W)$ admissible senses S of each given word W as vertices from SI, such that for each word W and its sense S, there is a vertex V for every intended sense S. In addition, the edge $E(S, S')$ in E, a subset of $V \times V$, is built up by adding a link from vertex S to vertex S' for which a semantic relation REL(S, S') between the two vertices is derived, where S is one of the admissible senses of W and S' of W'.

In Step 2, we initialize the probability P_s to a uniform distribution over each vertex S. And we set the weight of self-loop edge as (*1-d*) (Step 2a), and the weights of other outbound edges as $d/e(S)$, calculated as $Q_s = Q_s + \frac{d \times p_{s'}}{e(S')}$

In our ranking algorithm for the weighted graph, the decision on what edge to follow during a random walk considers the weights of outbound edges. One with a higher probability follows an edge that has a larger weight. The ranking algorithm is particularly useful for sense assignment, since the semantic relations between pairs of senses (vertices) are intrinsically modeled through weights indicating their strength, rather than a decision on binary 0/1 values.

As described in Step 3, the weights are represented as a matrix M for which the weights of all outbound edges from S are normalized to sum to 1. Our random walk algorithm holds that an imaginary walker who is randomly stepping over edges will eventually stop walking. The probability, at any step, that the walker will continue is a damping factor, a parameter usually denoted by d. The d factor is defined as the vertex ratio of the outgoing edges and the self-loop edge as the result of dividing the vertex weight of the damping constant. The damping factor is subtracted from 1. The value for (*1-d*) introduced is the principal eigenvector for the matrix M. The value of the eigenvector is fast to approximate (a few iterations are needed) and in practice it yields fairly optimal results. In the original definition of a damping factor introduced by PageRank (Brin and Page, 1998), a link analysis algorithm, various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85 whereas we use variant value for d in our implementation.

In Step 4 of vertex scoring, we compute the probabilistic values of each vertex at every iteration. The set of probabilities Q_s of each sense S for the next iteration is computed by multiplying the current probability P_s with the matrix $M_{s,s}$. For instance (Step 4a), suppose a walker is to start at one vertex of the graph. The probability of Q_s is the probability of a walker stands at a vertex of S forming a self-loop plus the sum of the influx of $P_{s'}$ weighted by $M_{s',s}$. In Step 4b, we normalize Q_s for the probability of all admissible senses with each word to sum to 1 and replace P_s by Q_s.

The normalized weighted score is determined as: $P_s(W) = \frac{Q_s(W)}{\sum_{l \in senses(W)} Q_l(W)}$

Subsequently, in Step 5, we calculate the ranking score of maximum probability P_s that integrates the scores of its start node. And thus the resulting stationary distribution of probabilities can be used to decide on the most probable set of admissible senses for the given word. For instance, for the graph drawn in Figure 3, the vertex on the vertical axis represented as the *sense #3* of "*fine*" will be selected as the best sense for "*fine*" under the thesaurus category "*Goodness*" with other entry words, such as, "*lovely*", "*superb*", "*beautiful*", and "*splendid*". The output of this stage is a set of linked word sense pairs (W, S*) that can be used to extend the limited thesaurus coverage. The overall goal of ranking admissible senses is to weight highly the senses that tend to arrive at the consistency of word senses.

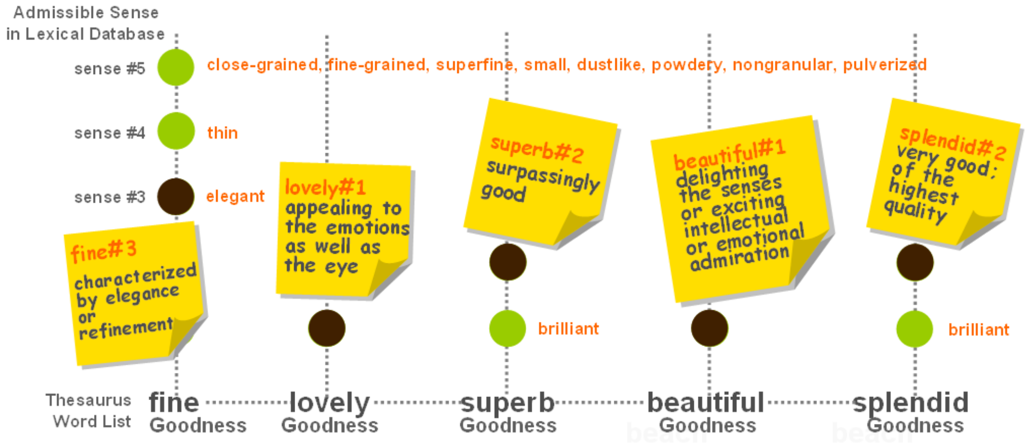
Admissible Sense
in Lexical Database
sense #5
sense #4
sense #3
close-grained, fine-grained, superfine, small, dustlike, powdery, nongranular, pulverized
thin
elegant
fine#3 characterized by elegance or refinement
lovely#1 appealing to the emotions as well as the eye
superb#2 surpassingly good
beautiful#1 delighting the senses or exciting intellectual or emotional admiration
splendid#2 very good; of the highest quality
brilliant
brilliant
Thesaurus
Word List
fine
Goodness
lovely
Goodness
superb
Goodness
beautiful
Goodness
splendid
Goodness

Figure 3.Highest scoring word sense under category "*Goodness*" assigned automatically by random walk.

Recall that our goal is to select the word senses for each specific collocate, categorized by the corresponding semantic label, for example, *sandy*, *rocky*, *pebbly beach* with label *Materials*; *beautiful*, *lovely*, *fine*, *splendid*, *superb* beach with *Goodness*. In order for the word coverage under thesaurus category to be comprehensive and useful, we need to expand the words listed under a label. This output dataset of the learning process is created by selecting the optimal linked word sense pairs (*W*, *S**) from each semantic relation in our word sense inventory where the specific semantic relation is explicitly defined.

Although alternative approaches can be used to identify word senses of given words, our iterative graphical approach has two distinctive advantages. First, it enables a principled combination of integrated similarity measure by modeling through a multiple types of semantic relations (edges). Secondly, it transitively merits local aggregated similarity statistics across the entire graph. To perform sense propagation, a weighted graph was constructed. On the graph, interconnection of edges is aggregated on a semantic relatedness level by random walk. The sense edge voltage is transitively propagated to the matching sense vertex. The effect depends on the reinforcement of the semantic relations (edges) and magnitude of the sense relations (vertices), creating a flexible amplitude-preserving playground like no other optional way of modeling a transcended graph propagation of senses. By doing so, our model is carved out to be a robust, more flexible solution with possible alternatives of combining additional resources or more sophisticated semantic knowledge. This approach is relatively computationally inexpensive for unsupervised approach to the WSD problem, targeting the annotation of all open-class words in lexical database using information derived exclusively from categories in a thesaurus. The approach also explicitly defines semantic relations between word senses, which are iteratively determined in our algorithm.

3.2.2 Extending the Coverage of Thesaurus

Automating the task of building a large-scale semantic knowledge base for semantic classification imposes a huge effort on the side of knowledge integration and validation. Starting from a widespread computational lexical database such as *WordNet* overcomes the difficulties of constructing a knowledge base from scratch. In the second stage of the learning process, we attempt to broaden the limited thesaurus coverage as the basis of our applied semantic knowledge that may induce to unknown words in collocation label assignment in Section 3.3. The sense-annotated word lists generated as a result of the previous step are useful for extending the thesaurus and reducing OOV words that may render words that form a group by itself.

In the previous learning process, "*fine*" with other adjective entries "*beautiful, lovely,* *splendid*, *superb*" under semantic label "*Goodness*" can be identified as belonging to the word sense *fine#3* "*characterized by elegance or refinement or accomplishment*" rather than other admissible senses (as shown in Table 1). Consider the task of adding similar word to the set of "*fine#3*" in the thesaurus category "*Goodness*". We apply semantic relation operators for novel word extension for "*fine#3*". Some semantic relations and semantic operators available in the word sense inventory are shown in Table 2.

In this case, "***similar_to***", the semantic relation operator of "*fine#3*" can be applied to derive similar word "*elegant#1*" as the extended word for "*fine#3*" identified with the sense definition "*characterized by elegance or refinement*".

Table 1.Admissible senses for adjective "*fine.*"

Sense Number	**Definition**	**Example**	**Synsets of Synonym**
fine #1	(being satisfactory or in satisfactory condition)	*"an all-right movie"; "everything's fine"; "the passengers were shaken up but are all right"; "things are okay"*	all ight#1, o.k.#1,ok#1, okay#1
fine #3	(characterized by elegance or refinement or accomplishment)	*"fine wine" ; "a fine gentleman";* *"fine china and crystal"; "a fine violinist"*	elegant#1
fine #4	(thin in thickness or diameter)	*"a fine film of oil"; "fine hairs";* *"read the fine print"*	thin#1

Table 2.Some semantic operators in word sense inventory.

***SR* Operators**	**Description**	**Relations Hold for**
syn operator	synonym sets for every word that are interchangeable in some context	all words
sim operator	adjective synsets contained in adjective clusters	adjectives

3.3 Giving Thesaurus Structure to Collocation by Iterative Graphical Algorithms

The stage takes full advantage of the foundation built in the prior learning process, established an extended semantic knowledge to build a thesaurus structure for online collocation reference tools. We aim to partition collocations in groups according to semantic relatedness by exploiting semantic labels in a thesaurus and assign each collocate to a thesaurus category.

In this stage of the process, we apply the previously stated random walk algorithm and automatically assign semantic labels to all collocations by exploiting semantic relatedness identified among collocates. By doing so, our approach for collocation label assignment can cluster collocations together in groups, which is helpful for dictionary look-up and learners to find their desired collocation or collocations under a semantic label.

We use a set of corresponding admissible semantic labels L to assign labels under thesaurus category $L \in Cat$ to each collocate $C \in Col$, such that the collocates annotated with L can be partitioned into a subset corresponding to a thesaurus category, $Sub = \{ (C, L) \mid C \in Col, L \in Cat \in TC \}$, which facilitate meaning-based access to the collocation reference for learners. We define a label graph $G = (V, E)$ such that there is a vertex $v \in V$ for every admissible label L of a given collocate C, and there is an edge $e \in E$ between two vertices where the two vertices have the same label. Edge reinforcement of the label (vertex) similarity distance between pairs of labels is represented as directed edges $e \in E$, defined over the set of vertex pairs $V \times V$. Such semantic label information typically lists in a thesaurus.

Given such a label graph G associated with a set of collocates Col, the probability of each label P_L can be iteratively determined using a graph-based ranking algorithm, which runs over the graph of labels and identifies the likelihood of each label (vertex) in the graph. The iterative algorithm is modeled as a random walk, leading to a stationary distribution over label probabilities P_L, represented as scores Q_L attached to vertices in the graph. These scores Q_L are then used to identify the most probable semantic label L_* for each collocate C, resulting in a list of annotations (C, L_*) for all collocates in the input set. The algorithm is quite similar to the one for graph-based word sense assignment shown in Figure 2. But note that the overall goal of ranking admissible labels is to weight highly the semantic labels that help arrange collocations in a thesaurus category and provide learners with a thesaurus index.

In other word, our goal is to assign corresponding semantic labels to each specific collocate, for example, "*sandy*, *rocky*, *pebbly* *beach* with label *Materials.*" In order for the semantic structure to be comprehensive and useful, we try to cover as much OOV words as possible by applying semantic relation operators (e.g., derivational relations). We propose the replacement of OOV words for their derivational words such as the replacement of "rocky" for "rock" and "dietary" for "diet". For a few number of derivationally substitutable OOV words occurred, such as *pebbly* *beach*, we apply the built-in vocabulary of words, i.e., *pebble*, as a substitution for *pebbly* by exploiting the derivational relations from the obtainable sense inventory as we will discuss in more detail in the section of experimental set-up.

The output of this stage is a list of linked label-annotated collocate pairs (C, L^*) that can be used to classify collocations in categories.

4. Experimental Settings

4.1 Experimental Data

In our experiments, we applied random walk algorithm to partitioning collocations into existing thesaurus categories, thus imposing a semantic structure on the raw data. In analysis of learners' collocation error patterns, the types of verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns (Liu, 2002; Chen, 2002). Hence, for our experiments and evaluation, we focused our attention particularly on V-N and A-N collocations.

Recall that our classification model starts with a thesaurus consisting of lists of semantic related words extended by a word sense inventory via random walk Algorithm. Then, the extended semantic knowledge provides collocates with topic labels for semantic classification of interest. Preparing the semantic knowledge base in our experiment consists of two main steps: (1) Integration, and (2) Extension. Two kinds of resources are applied as the input data of this learning process of semantic knowledge integration described below.

4.1.1 Input Data 1: A Thesaurus for Semantic Knowledge Integration

We selected the set of thesaurus categories from the dictionary of *Longman Lexicon of Contemporary English* (*LLOCE*). *LLOCE* contains 15,000 distinct entries for all open-class words, providing semantic fields of a pragmatic, everyday common sense index for easy reference. The words in *LLOCE* are organized into approximately 2,500 semantic word sets. These sets are divided into 129 semantic categories and further organized as 14 semantic fields. Thus the semantic field, category, and semantic set in *LLOCE* constitute a three-level hierarchy, in which each semantic field contains 7 to 12 categories and each category contains 10 to 50 sets of semantic related words. The *LLOCE* is based on coarse, topical semantic classes, making them more appropriate for WSD than other finer-grained lexicon.

4.1.2 Input Data 2: A Word Sense Inventory for Semantic Knowledge Extension

For our experiments, we need comprehensive coverage of word senses. Word senses can be easily obtained from any definitive records of the English language (e.g. an English dictionary, encyclopedia or thesaurus). In this case, we applied *WordNet* to broaden our word coverage from 15,000 to 39,000. *WordNet* is a broad-coverage machine-readable lexical database, publicly available in parsed form (Fellbaum, 1998). *WordNet* 3.0 lists 212,557 sense entries for open-class words, including nouns, verbs, adjectives, and adverbs. In order to extend the sense coverage, we applied random walk Algorithm to match a significant and manageable portion of the *WordNet* sense inventory to the *LLOCE* thesaurus.

WordNet can be considered a graph over synsets where the word senses are populated as vertices and the semantic relations edges. *WordNet* is organized by the sets of synsets; a synset is best thought of as a concept represented by a small set of synonymous senses: the adjective {*excellent*, *first-class*, *fantabulous*, *splendid*}, the noun {*enemy*, *foe*, *foeman*, *opposition*}, and the verb {*fight*, *contend*, *struggle*} form a synset.

4.2 Experimental Configurations

We acquired all materials of the input data (1) and (2) to train and run the proposed model, using the procedure and a number of parameters as follows:

4.2.1 Step 1: Integrating Semantic Knowledge

To facilitate the development of integrated semantic knowledge, we organize synsets of entries in the first input data, *LLOCE*, into several thesaurus categories, based on semantic coherence and semantic relations created by lexicographers from *WordNet*. The integrated semantic knowledge can help interpret a word by providing information on its word sense and its corresponding semantic label, (i.e., “*fine*” tagged with “*Materials*”).

Recall that our model for integrating word senses and semantic labels is based on random walk algorithm on a weighted directed graph whose vertices (word senses) and edges (semantic relations) are extracted from *LLOCE* and *WordNet* 3.0. All edges are drawn as semantic relatedness among words and senses, derived using the semantic relation operators (Table 3).

Table 3.The semantic relation operators used to link the lexical connection between word senses.

Relation Operators	Semantic Relations for Word Meanings	Relations Hold for
Syn operator	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
hyp operator	hypernym/hyponym (superordinate/subordinate) relations between synonym sets	nouns verbs
vgp operator	verb synsets that are similar in meaning and should be grouped together when displayed in response to a grouped synset search.	verbs
Sim operator	adjective synsets contained in adjective clusters	adjectives
der operator	words that have the same root form and are semantically related	all words

In particular for all semantic relation operators, we construct a maximum allowable edge distance *MaxED*, informing a constraint over the edge path between words for which the word sense likelihood is sought. For our experiments, the *MaxED* is set to 4.

4.2.2 Step 2: Extending Semantic Knowledge

Once we have mapped the sense-label from the stationary distribution in the random walk graph, another step is taken to take advantage of the mapped semantic knowledge by adding more novel words to the thesaurus categories. The word coverage in question is extended by more than twice as many *LLOCE* thesaurus entries. For the extension of our semantic knowledge, we need information on joint word sense and semantic label pairs, and semantic relation among words from the previous step. Various kinds of the above-mentioned semantic relation operators can be derived, depending on the type of semantic operators available for the word class at hand. In experiments, we focus on the synset operation provided in *WordNet*.

4.3 Test Data

We used a collection of 859 V-N and A-N collocation pairs for testing, obtained from the website, *JustTheWord* (http://193.133.140.102/JustTheWord/). *JustTheWord* clusters collocates into sets without understandable label. As a result, we will compare the performance of our model with *JustTheWord* in Section 5

We evaluated semantic classification of three types of collocation pairs, focusing on **A**-N, **V**-N and V-**N**. We selected five pivot words for each type of collocation pairs for their varying level of abstractness and extracted a subset of their respective collocates from the *JustTheWord*. Among 859 testing pairs, 307 collocates were extracted for **A**-N, 184 for **V**-N, and 368 for V-**N**.

To make the most appropriate selection from testing data in *JustTheWord*, we have been guided here by research into language learners' and dictionary users' needs and skills for second language learning, taking account especially of the meanings of complex words with many collocates (Tono, 1992; Rundell, 2002). The pivot words we selected for testing are words that have many respective collocations and are shown in boxes around each entry in *Macmillan English Dictionary for Advance Learners.*

5. Results and Discussions

Two pertinent sides were addressed for the evaluation of our results. The first was whether such a model for a thesaurus-based semantic classification could generate collocation clusters based on human-like word meaning similarities to a significant extent. Second, supposing it did, would its success of semantic label assignment also strongly excel in language learner collocation production? We propose innovative evaluation metrics to examine our results respectively in these two respects and assess whether our classification model can reliably cluster collocates and assign a helpful label in terms of language learning. In the first subsection, first we explain why we propose a new evaluation metrics in order to explore how the method results in simple, robust designs yet influences each facet of the question for lexicographic and pedagogical purposes. In the following subsections, the evaluation metrics are presented individually in two regards, for assessing the performance of collocation clusters, and for the conformity of assigned semantic labels.

5.1 Performance Evaluation for Semantic Cluster Similarity

The collection of the traditional evaluation (Salton, 1989) of clustering works best for certain type of *clustering* method but might not be well suited to evaluate our *classification* model, where we aim to facilitate collocation referencing and help learners improve their collocation production. In that case, for assessing collocation clusters, we propose a robust evaluation method by setting up the items to be evaluated as a test for semantic similarity to judge the performance of clustering results. For semantic labeling results, we developed a grading rubric with performance descriptions for the conformity of labels as a reference guide. Two human judges were asked to give performance assessment by scoring each item. The evaluation methodology is aimed at fostering the development of innovative evaluation designs as well as encouraging discussion regarding language learning by means of the proposed method.

Landauer and Dumais (1997) were first proposed using the synonym test items of the Test of English as a Foreign Language (TOEFL) as an evaluation method for semantic similarity. Fewer fully automatic methods of a knowledge acquisition evaluation, one that does not depend on knowledge being entered by a human, have been capable of performing well on a full scale test used for measuring semantic similarity. An example provided by Landauer (1997) is shown below where "*crossroads*" is the real synonym for "*intersection*".

You will find the office at the main ***intersection***.

(a) place (b) crossroads (c) roundabout (d) building

For this experiment, we conducted the task of evaluating the semantic relatedness among collocation clusters according to the above-mentioned TOEFL benchmark to measure semantic similarity and set up target items out of our test data as sheet of clustering performance test. Our human judges performed a decision task similar to TOEFL test takers: They had to decide which one of the four alternatives was synonymous with the target word. A sample question is shown below where grouping "***sandy*"** and "***rocky*"** together with the target word "*beach*" because they belong to the same category of concept as the collocation is more appropriate than clustering "***sandy*"** and any of others together.

sand**y** beach

(a) long (b) rocky (c)super (4)narrow

There are 150 multiple choice questions randomly constructed to test the cluster validation, 50 questions for each 3 testing collocation types and therein 10 for each of **A**-N, **V**-N, and V-**N** testing collocation pairs. In order to judge how much degree our model ultimately has achieved in producing good clusters, two judges were asked to primarily choose the one most nearly correct answer. If the judges find one of the distracters to be also the plausible answer, giving collective answer options is allowed for our evaluation in order to test the cluster validation thoroughly from grey area among options given inadvertently. If the judges think no single correct answer is plausible enough, 0 point can be given for no satisfactory option considered. Table 4 shows the performance figures of collocation clusters generated by the two systems. As is evidence from the table, our model showed significant improvements on the precision and recall in comparison with *JustTheWord.*

Table 4.Precision and recall of our classification model and those of *JustTheWord*

Results / System	**Judge 1**		**Judge 2**		**Inter-Judge Agreement**
	Precision	**Recall**	**Precision**	**Recall**	
Ours	.79	.71	.73	.67	.82
JustTheWord	.57	.58	.57	.59	

Without doubt, subjectivity of human judgments interferes with the performance evaluation of collocation clusters, for inter-judge agreement is just above 80 %. The closer our precision (79% and 73%) is to the discrimination ratio, the more effectively that an automatic method distinguishes subjects in accordance with human judgment.

5.2 Conformity of Semantic Labels

The second evaluation task here focuses on whether the semantic labels facilitate users to scan the entry quickly and find the desired concept of the collocations. From the experiments, we show that the present online collocation learning tools may not be an appropriate place to seek guidance on fine discrimination between near synonyms. This problem could be alleviated if the alphabetical frequency ordering of the learning tool could be supplemented by thematic treatment in our thesaurus-based semantic classification model. Our evaluation result will indicate the extent to which semantic labels are useful, to what degree of reliability. Only to the extent that evaluation scores are reliable and the test items are solidly grounded in its practical viewpoint can they be useful and fair to the assessment.

Two human informants were asked to grade collocation with label, half of them randomly selected from our output results. The assessment was obtainable through different judges that participated in evaluating all of the collocation clusters as described above. One native American graduate and a non-native PhD researcher specializing in English collocation reference tools for language learners were requested to help with the evaluation. We set up a three-point rubric score to evaluate the conformity of semantic labels. When earning two points on a three-point rubric, a label has performed well in terms of guiding a user finding a desired collocation in a collocation reference tool. If the assigned label is somewhat helpful in collocation look-up, a score of one is shown that labels are achieving at an acceptable level. To assign judgments fairly and to calculate a fair reflection of the conformity of the labels, a zero score can be given if the labels can be considerably misleading to what is more indicative of the concepts. We set up an evaluation guide to present judges with the description for each rubric point, and allow the judges to grade each question as “0”, “0.5” or “1” for the item.

Table 5 shows that 77% of the semantic labels assigned as a reference guide has been judged as adequate in terms of guiding a user finding a desired collocation in a collocation learning tool, and that our classification model provably yields productive performance of semantic labeling of collocates to be used to assist language learners. The results justify the move towards semantic classification of collocations is of probative value.

Table 5.Performance evaluation for assigning semantic labels as a reference guide

	Judge 1	**Judge 2**
Ours	.79	.75
JustTheWord	Not available	Not available

6. Conclusion

The research sought to create a thesaurus-based semantic classifier within a collocation reference tool limited to the collocates occurring without meaning access indexes. We describe a thesaurus-based semantic classification for a semantic grouping of collocates with a pivot word and the construction of a collocation thesaurus that is used by learners to enhance collocation production. The thesaurus-based semantic classification classifies objects into semantically related groups that can participate in the same semantic relation with a given word. Rather than relying on a distributional analysis, our model is resourced from an integrated semantic knowledge, which is then generalized to combat sparsity. The evaluation shows that this robustly designed classification model facilitates the existing computational collocation reference tools and provides users with the collocations they desire to make semantically valid choices. The thesaurus structure is successfully imposed to facilitate grasping concepts of collocations.

Given that there is very little precedent review for us to follow, this research offers insights into how such a collocation thesaurus could be structured and useful. The semantic labeling described here improves collocation reference tools and has given us a tool for studies of collocation acquisition. The final results convincingly motivate the move towards semantic classification of collocations.

Many avenues exist for future research and improvement of our classification model. Another possibility would be to train more set of variables, each of which may take one among several different semantic relations for each collocation types. There is also a set of constraints which state compatibility or incompatibility of a combination of variable semantic relations.

To top it all off, existing methods for extracting the best collocation pairs from a corpus of text could be implemented. Domain knowledge, heuristics, and WSD techniques could be used to improve the identification of semantic label types. Semantic relations could be routed to classification model that performs best for more types of collocation pair (such as adverb-adjective pairs).

References

Béjoint, H. 1994. Tradition and Innovation in Modern English Dictionaries. Oxford: Clarendon Press.

Benson, M. 1985. Collocations and Idioms. In *Dictionaries, Lexicography and Language Learning*, R. Ilson (Ed.), 61-68. Oxford: Pergamon Press.

Brin, S. & Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.

Chen, P. C. 2002. A Corpus-Based Study of the Collocational Errors in the Writings of the EFL Learners in Taiwan. Unpublished master’s thesis, National Taiwan Normal University, Taipei.

Chklovski, T. & Pantel, P. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of EMNLP*, 33-40.

Fellbaum, C.(Ed.) 1998. WordNet: An Electronic Lexical Database. MA: MIT Press.

Firth, J. R. 1957. The Semantics of Linguistics Science. Papers in linguistics, 1934-1951. London: Oxford University Press.

Hatzivassiloglou, V. & McKeown, K. R. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering adjectives according to meaning. In *Proceedings of ACL*, 172–182.

Hindle, D. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of ACL*,268-275.

Jian, J. Y., Chang, Y. C. & Chang, J. S. 2004. TANGO: Bilingual Collocational Concordancer. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Kilgarriff, A. & Tugwell, D. 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of ACL Workshop on Collocations*, 32-38.

Landauer, T. K. & Dumais, S. T. 1997. A Solution to Plato’s Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2):211-240.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC*, 24–26.

Lewis, M. 1997. Implementing the Lexical Approach. Hove : Language Teaching Publications.

Lewis, M. 2000. Teaching Collocation: Further Development in the Lexica1 Approach. Hove: Language Teaching Publications.

Lin, D. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In *Proceedings of ACL*, 64-71.

Liu, L. E. 2002. A Corpus-Based Lexical Semantic Investigation of Verb-Noun Miscollocations in Taiwan Learners' English. Unpublished master's thesis, Tamkang University, Taipei.

Rundell, M. (Editor-in-Chief). 2002. The Macmillan English Dictionary for Advanced Learners. Oxford: Macmillan Publishers Limited.

Nastase, V. & Szpakowicz, S. 2003. Exploring Noun–Modifier Semantic Relations. In *Proceedings of International Workshop on Computational Semantics*, 285–301.

Nation, I. S. P. 2001. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nirenburg, S. & Raskin, V. 1987. The Subworld Concept Lexicon and the Lexicon Management System. Computational Linguistics, 13(3/4): 276-289.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley Publishing.

Scholfield, P. 1982. Using the English Dictionary for Comprehension. TESOL Quarterly, 16(2):185-194.

Tono, Y. 1992. The Effect of Menus on EFL Learners' Look-up Processes. LEXIKOS 2: 229-253.

Tono, Y. 1997. Guide Word or Signpost? An Experimental Study on the Effect of Meaning Access Indexes in EFL Learners' Dictionaries. English Studies, 28: 55-77.

Turney, P. D. 2006. Similarity of Semantic Relations. Computational Linguistics, 32(3):379–416.

Wanner, L., Bohnet, B. and Giereth, M. 2006. What is beyond Collocations? Insights from Machine Learning Experiments. *EURALEX*.

Summers, D. (Director) 1995. Longman Dictionary of Contemporary English (3rd edition). Harlow:Longman.

Procter, P. (ed.) 1995. Cambridge International Dictionary of English. Cambridge: Cambridge University Press.