Expert Systems with Applications 38 (2011) 6392–6399

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Short communication

A novel sentence similarity measure for semantic-based expert systems

Ming Che Lee

Department of Computer and Communication Engineering, Ming Chuan University, Taoyuan, Taiwan

ARTICLE INFO

Keywords:
Sentence
Similarity
Semantic web
Ontology

ABSTRACT

A novel sentence similarity measure for semantic based expert systems is presented. The well-known problem in the fields of semantic processing, such as QA systems, is to evaluate the semantic similarity between irregular sentences. This paper takes advantage of corpus-based ontology to overcome this problem. A transformed vector space model is introduced in this article. The proposed two-phase algorithm evaluates the semantic similarity for two or more sentences via a semantic vector space. The first phase built part-of-speech (POS) based subspaces by the raw data, and the latter carried out a cosine evaluation and adopted the WordNet ontology to construct the semantic vectors. Unlike other related researches that focused only on short sentences, our algorithm is applicable to short (4–5 words), medium (8–12 words), and even long sentences (over 12 words). The experiment demonstrates that the proposed algorithm has outstanding performance in handling long sentences with complex syntax. The significance of this research lies in the semantic similarity extraction of sentences, with arbitrary structures.

1. Introduction

A semantic expert system is software that attempts to processing and understanding natural language, which refers to irregular, complex, and diverse philosophic meaning and context of human language. Recently the field of natural language processing (NLP) presents a need for efficient algorithms and methodologies to evaluate the similarity between short texts and sentences (Michie, 2001), such as QA systems, FAQ matching systems, machine translation, news and article summarization systems, and information retrieval systems. The widely used technology is the vector space model (VSM). In VSM, the words, phrases, sentences, or articles are represented by a high dimensional vector, and the base space was constructed by all the non-stopwords presented in the system. The elements are correlated to each other according with geometric distance of their associated vectors in this space. However, the traditional vector-based models have a deficiency in semantic-awareness capability.

The issue of semantic aware among texts is increasingly pointing towards Semantic Web technologies in general and ontology in particular as a solution. Ontology has being a philosophical theory about the nature of being. A typical ontology has a taxonomy defining the concepts and their relationships of a domain, and a set of inference rules that powers its reasoning functions (Lee, Hendler, & Lassila, 2001). In the knowledge representation community, the most commonly used or cited ontology definition is from Gruber (1993). The Semantic Web is an evolving extension of the World Wide Web in which web content can be expressed in natural languages, and in a form that can be understood, interpreted, and used by software agents. Elements of the Semantic Web are expressed in formal specifications (Davies, Fensel, & Van Harmelen, 2003), which including Resource Description Framework (RDF), a variety of data interchange formats (such as RDF/XML, N3, Turtle, N-Triples) (XML; xmlschema), and notations such as RDF Schema and the Web Ontology Language (OWL-REF). In recent years, the WordNet has become the most widely used general-purpose ontology of English. Verbs, nouns, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), and each synonym expressing a distinct concept. Following the vigorous development of semantic web and ontology, there are many semantic-based systems been presents in the literature. Chiang, Ho, and Wang (2008) proposed a gene–gene relation extracting model that based on the gene ontology (GO). A business process integration model is presented in Jung (2009), which is based on the ontology alignment. Jeong et al. proposed a methodology for measuring the semantic similarity of XML schemas (Jeong, Lee, Cho, & Lee, 2008). Ontology technologies are also widely adopted in other semantic processing fields, such as text summarization (Aliguliyev, 2009; Zhan, Loh, & Liu, 2009), web service and agent computation (Garcı'a-Sa'nchez, Valencia-Garcı'a, Martı'nez-Be'jar, & Ferna'ndez-Breis, 2009; Guzman Arenas & Olivares Ceja, 2006; Liu, Shen, Hao, & Yan, 2009), and QA systems (Chu, Chen, & Chen, 2009; Guo & Zhang, 2009).

This paper presents a novel sentence similarity computation algorithm for English natural language processing. In the proposed algorithm, the semantic space is separated into two subspaces – the *noun space* and the *verb space*, and in each subspace, the value of the vector is determined via a WordNet similarity measure

E-mail address: leemc@mail.mcu.edu.tw

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.10.043

instead of the frequency or probability that the appearance of a word in the sentences. The rest of this paper is organized as follows. Section 2 outlines the system framework and the core functions. Section 3 introduces the proposed sentence similarity evaluation algorithm and gives some examples. Section 4 shows the experimental results, and the final gives the conclusion.

2. System framework

As shown in Fig. 1, the proposed framework is divided into two sub-systems – the semantic quantification and the semantic inferring core functions, which are described as follows:

2.1. The semantic quantification

This subsystem quantifies the input sentences and builds the POS-based semantic space. There are three major components:

I. *Sentence Formalization* – The inputted sentence pair is formalized before processing. This procedure includes tokenization, lower-casing, stemming, and stop-word removing. A stop words list is referred while removing useless terms. Stemming reduces inflected (or sometimes derived) words to their stem, base or root form. For example, the words ended with "*ed*", "*ing*", or "*ly*", are removed. The Porter's stemming algorithm (Porter) is adopted in this research.
II. *Part-of-Speech* – The words in each sentence that after pre-processing are categorized into two sets – the *Noun* and the *Verb* sets. The semantic space is defined to the union of words of the same POS sets.
III. *WordNet Similarity Measure* – Compared to other traditional vector space models and cosine evaluating algorithms, this research uses the WordNet semantic tree to determine the value of the vector instead of the time and probability of appearances. The concept is described in the following section in more details.

2.2. The semantic extracting core functions

This subsystem extracting the correlation of the two sentences via the semantic distance evaluated in part *A*. There are three components in this subsystem:

I. *POS based semantic coordinate* – This procedure computes the cosine angle for each vector of the same semantic space.
II. *Combined Sentence Similarity* – This procedure combines the two scores of the previous step into one integrated score.
III. *Optimization* – The optimized weight of the two score described above will be adjusted in the experiment.

3. The sentence similarity evaluation algorithm

3.1. The algorithm

This section describes the proposed algorithm and the corresponding formulas in details. This algorithm accepts two sentences as the input and outputs the similarity score of the two sentences. The main steps are described as follows:

Step 1. *Pre-processing.*
This step formalized the input sentences as described in Section 2.1-I.
Step 2. *Words Similarity*
Each word in the sentences that after pre-processing, is categorizes into two sets – the *Verb* set and the *Noun* set, as well as Definition 1.

Definition 1. The word sets of sentences A and B

$$SEN_A = \{S\_V_A, S\_N_A\},$$
$$SEN_B = \{S\_V_B, S\_N_B\}.$$

Definition 2. The Noun Vector(*NV*), Verb Vector(*VV*), Noun Semantic Space, and the Verb Semantic Space.

Noun Vector is the vector of nouns corresponding to the base space ($S\_it\ N_A \cup S\_N_B$), and *Verb Vector* is the vector of verbs corresponding to the base space ($S\_V_A \cup S\_V_B$), among which

$$|NV_{SEN\_A}| = |NV_{SEN\_B}| = |S\_N_A \cup S\_N_B|, \quad \text{and}$$
$$|VV_{SEN\_A}| = |VV_{SEN\_B}| = |S\_V_A \cup S\_V_B|.$$

In Definition 1, SEN_A and SEN_B are the sets of words after pre-processing. S_V_A and S_N_A are the sets of verbs and nouns in SEN_A, respectively. NV_{SEN_A} and VV_{SEN_A} are the vector space of verbs and nouns in sentence *A*. In Definition 2, the Noun Semantic Space (base space), and the Verb Semantic Space (base space) are defined as the

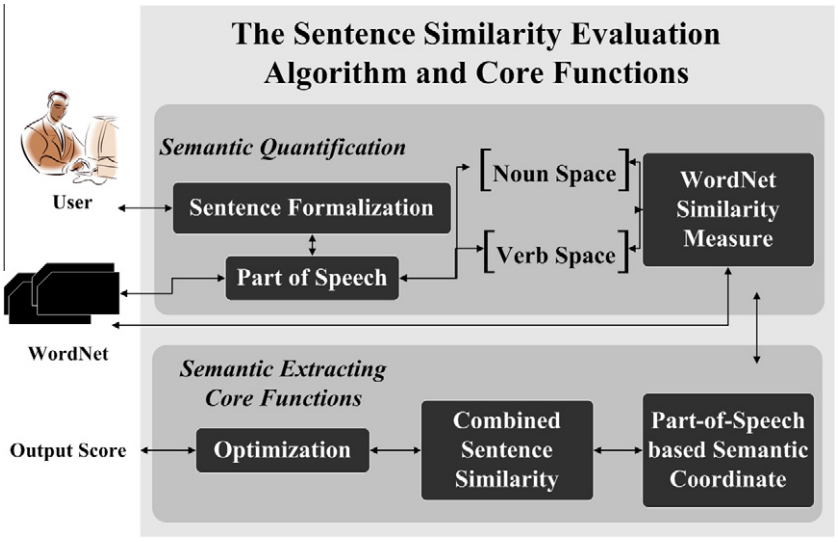

Fig. 1. The sentence similarity evaluation framework and the core functions.

union of nouns in SEN_A and SEN_B, and the union of verbs in SEN_A and SEN_B, respectively. The Wu & Palmer similarity measure (Wu & Palmer, 1994) has become somewhat of a standard for measuring similarity in lexical taxonomies. This research adopts the Wu & Palmer similarity measurement to determine the similarity between two nouns or verbs. The formula is listed as follows:

$$\begin{aligned} Similarity_{(WORD\_A,WORD\_B)} &= 2 \times DEPTH(H_l) \\ &\times (D_{Path\_Length}(WORD\_A, H_l) \\ &+ D_{Path\_Length}(WORD\_B, H_l) + 2 \\ &\times DEPTH(H_l))^{-1}. \end{aligned} \quad (1)$$

In Formula (1), H_l is the depth of the lowest shared hypernym of *WORD_A* and *WORD_B*. $DEPTH(H_i)$ is the level of H_l in the WordNet semantic tree. $D_{Path\_Length}(WORD\_A,H_l)$ is the semantic distance (number of hops) form H_l to *WORD_A*. $D_{Path\_Length}(WORD\_B,H_l)$ is the semantic distance (number of hops) form H_l to *WORD_B*. Each word is compared to the base space to obtain the value of each field via formula (1).

Step 3. Noun Vector and Verb Vector

This step determines the value of *NV* and *VV* of each sentence. In step 2, each word in *NV* or *NN* is computed to the whole corresponding semantic space. And the largest value is chosen as the final value of each field in the vector. The formal formulas are listed as follows:

$$NV_{SEN\_A_i} = \max_{k=1}^{|S\_N_A \cup S\_N_B|} (Similarity(WORD\_A, NOUN\_BASE_k)), \quad (2)$$

$$VV_{SEN\_A_i} = \max_{k=1}^{|S\_V_A \cup S\_V_B|} (Similarity(WORD\_A, VERB\_BASE_k)). \quad (3)$$

In formulas (2) and (3), NV_{SEN_Ai} denotes the value of *NV* of *SEN_A* in field *i*, and VV_{SEN_Ai} denotes the value of *VV* of *SEN_A* in field *i*.

Step 4. Cosine Measurement

This step computes the cosine angle of the *VV* and *NV* of the sentences, which are called *Verb Cosine* (*VC*) and *Noun Cosine* (*NC*). In this algorithm, the traditional cosine measurement was improved to meet our design. The formulas are listed as follows:

$$\begin{aligned} NC_{A,B} &= \left(\frac{\vec{NV}_{SEN\_A} \bullet \vec{NV}_{SEN\_B}}{\left|\vec{NV}_{SEN\_A}\right| \times \left|\vec{NV}_{SEN\_B}\right|} \right)^2 \\ &= \left(\frac{\sum_{i=1}^{|S\_V_A \cup S\_V_B|} NV_{SEN\_A_i} \times NV_{SEN\_B_i}}{\sqrt{NV^2_{SEN\_A_i}} \times \sqrt{NV^2_{SEN\_B_i}}} \right)^2, \end{aligned} \quad (4)$$

$$\begin{aligned} VC_{A,B} &= \left(\frac{\vec{VV}_{SEN\_A} \bullet \vec{VV}_{SEN\_B}}{\left|\vec{VV}_{SEN\_A}\right| \times \left|\vec{VV}_{SEN\_B}\right|} \right)^2 \\ &= \left(\frac{\sum_{i=1}^{|S\_N_A \cup S\_N_B|} VV_{SEN\_A_i} \times VV_{SEN\_B_i}}{\sqrt{VV^2_{SEN\_A_i}} \times \sqrt{VV^2_{SEN\_B_i}}} \right)^2. \end{aligned} \quad (5)$$

In formulas (4) and (5), the square is to reduce the degree of self comparison.

Step 5. The Integrated Sentence Similarity

This step combines the *VC* and *NC* into an integrated score. The weights of *VC* and *NC* are adjusted by a balance coefficient ζ, which is determined either via the experiment or by the users manually.

$$Similarity_{A,B} = \zeta \times (NC_{A,B}) + (1 - \zeta) \times (VC_{A,B}). \quad (6)$$

3.2. Examples

This subsection gives an example to illustrate the overall sentence similarity of the proposed algorithm. The notation of this example is listed in Table 1. Table 2-A lists the example sentences *A*, *B*, and *C* with POS tags. In this example, the similarity score is evaluated triple for pairs *A-B*, *A-C*, and *B-C*, which are denoted as case I, II, and III, and balance coefficient ζ is set to 0.65.

Case I. Pair A-B. In this case after pre-processing, the noun set of sentence *A* (S_N_A) is {*food*,*price*,*accommodation*}, and the noun set of sentence *B*(S_N_B) is {*price*,*accommodation*,*food*}, the noun semantic space is thus {*food*,*price*,*accommodation*}. The verb sets of sentences $A(S\_V_A)$ and $B(S\_V_B)$ are {*be*,*include*} and {*include*}, respectively, and the verb semantic space is {*be*,*include*}. According to the formulas (1)–(3), the final vectors of nouns and verbs of sentences *A* and *B* are: $NV_{SEN\_A} = [1,1,1] = NV_{SEN\_B}$, $VV_{SEN\_A} = [1,1]$, and $VV_{SEN\_B} = [0,1]$,

Table 1
Notation used in Section 3.2.

Notation	Meaning
V	Verbs
N	Nouns
J	Adjectives
A	Adverbs
S	Stop-words

Table 2-A
Sentence samples with POS notations.

Sentences	Raw sentences with POS notations
Sentence A	Food[*N*] is[*V*] also[*S*] included[*V*] in[*S*] the[*S*] price[*N*] of[*C*] the[*C*] accommodation[*N*]
Sentence B	The[*C*] price[*N*] of[*C*] the[*C*] accommodation[*N*] also[*C*] includes[*V*] the[*C*] food[*C*]
Sentence C	He[*N*] introduced[*V*] better[*J*] methods[*N*] of[*C*] management[*N*] in[*C*] this[*C*] company[*N*]

Table 2-B
Case I. NV_{SEN_A} v.s. noun base space.

NV_{SEN_A}/*A-B* noun base space	Food	Price	Accommodation
Food	1	0.43	0.59
Price	0.43	1	0.59
Accommodation	0.59	0.59	1
Vector	**1**	**1**	**1**

Table 2-C
Case I. VV_{SEN_A} v.s. verb base space.

VV_{SEN_A}/*A-B* verb base space	Be	Include
Be	1	0
Include	0	1
Vector	**1**	**1**

Table 3
Test data sets and experimental results.

	Raw sentences		
Triple A			
Sentence A-1	If she can be more considerate to others, she will be more popular		
Sentence A-2	She is not considerate enough to be more popular to others		
Sentence A-3	You are not supposed to touch any of the art works in this exhibition		
Similarity	A-1 v.s. A-2 = 0.9125	A-1 v.s. A-3 = 0.01956859	A-2 v.s. A-3 = 0.02903207
Triple B			
Sentence B-1	I won't give you a second chance unless you promise to be careful this time		
Sentence B-2	If you could promise to be careful, I would consider to give you a second chance		
Sentence B-3	The obscurity of the language means that few people are able to understand the new legislation		
Similarity	B-1 v.s. B-2 = 0.9384236	B-1 v.s. B-3 = 0.4190409	B-2 v.s. B-3 = 0.3293912
Triple C			
Sentence C-1	About 100 officers in riot gear were needed to break up the fight		
Sentence C-2	The army entered in the forest to stop the fight with weapon		
Sentence C-3	He thus avoided a pack of journalists eager to question him		
Similarity	C-1 v.s. C-2 = 0.6952305	C-1 v.s. C-3 = 0.4072169	C-2 v.s. C-3 = 0.5830132
Triple D			
Sentence D-1	Your digestive system is the organs in your body that digest the food you eat		
Sentence D-2	Stomach is one of organs in human body to digest the food you eat		
Sentence D-3	We had better wait to see what our competitors do before we make a move		
Similarity	D-1 v.s. D-2 = 0.9187595	D-1 v.s. D-3 = 0.2684233	D-2 v.s. D-3 = 0.2639506
Triple E			
Sentence E-1	I don't think it is a clever idea to use an illegal means to get what you want		
Sentence E-2	It is an illegal way to get what you want, you should stop and think carefully		
Sentence E-3	There is something wrong with the steel supporting member of the device		
Similarity	E-1 v.s. E-2 = 0.5911233	E-1 v.s. E-3 = 0.2679752	E-2 v.s. E-3 = 0.1166667
Triple F			
Sentence F-1	The powerful authority is partial to the members in the same party with it		
Sentence F-2	Political person sometimes abuse their authority that it is unfair to the citizen		
Sentence F-3	He reasoned that we could be there by noon if we started at dawn		
Similarity	F-1 v.s. F-2 = 0.872057	F-1 v.s. F-3 = 0.1842038	F-2 v.s. F-3 = 0.1540446
Triple G			
Sentence G-1	The fire department is an organization which has the job of putting out fires		
Sentence G-2	An organization which has the job of putting out fires is the fire department		
Sentence G-3	The man wore a bathrobe and had evidently just come from the bathroom		
Similarity	G-1 v.s. G-2 = 1	G-1 v.s. G-3 = 0.5586169	G-2 v.s. G-3 = 0.5586169

which are partly listed in Tables 2-B and 2-C. The next step is to compute the cosine angle via formulas (4) and (5). The results are:

$$NC_{A,B} = \left(\frac{\vec{VV}_{SEN\_A} \bullet \vec{VV}_{SEN\_B}}{\left|\vec{VV}_{SEN\_A}\right| \times \left|\vec{VV}_{SEN\_B}\right|} \right)^2 = \left(\frac{1\times 1 + 1\times 1 + 1\times 1}{\sqrt{1^2+1^2+1^2} \times \sqrt{1^2+1^2+1^2}} \right)^2 = 1$$

and

$$VC_{A,B} = \left(\frac{\vec{NV}_{SEN\_A} \bullet \vec{NV}_{SEN\_B}}{\left|\vec{NV}_{SEN\_A}\right| \times \left|\vec{NV}_{SEN\_B}\right|} \right)^2 = \left(\frac{1\times 0 + 1\times 1}{\sqrt{1^2+1^2} \times \sqrt{0^2+1^2}} \right)^2 = 0.666.$$

The final similarity of sentences A and B is

$$Similarity_{A,B} = \zeta \times (NC_{A,B}) + (1-\zeta) \times (VC_{A,B}) = 0.65 \times 1 + 0.35 \times 0.666 = 0.8831.$$

Case II. Pair A-C. In this case after pre-processing, the noun set of sentence A (S_N_A) is the same as case I, and the noun set of sentence C (S_N_C) is {*method*, *management*, *company*}, the noun semantic space is thus {*food*, *price*, *accommodation*, *method*, *management*, *company*}. The verb sets of sentences A (S_V_A) and B (S_V_B) are {*be*, *include*} and {*introduce*}, respectively, and the verb semantic space is {*be*, *include*, *introduce*}. According to the formulas (1)–(3), the final vectors of nouns and verbs of sentences A and C are: $NV_{SEN\_A} = [1, 1, 1, 0.78, 0.67, 0.53]$, $NV_{SEN\_C} = [0.67, 0.53, 0.78, 1, 1, 1]$, $VV_{SEN\_A} = [1, 1, 0.5]$, and $VV_{SEN\_C} = [0, 0.5, 1]$ The next step is to compute the cosine angle via formulas (4) and (5). The results are:

$$NC_{A,C} = \left(\frac{\vec{VV}_{SEN\_A} \bullet \vec{VV}_{SEN\_C}}{\left|\vec{VV}_{SEN\_A}\right| \times \left|\vec{VV}_{SEN\_C}\right|} \right)^2 = \left(\begin{bmatrix} (1\times 0.67) + (1\times 0.53) + (1\times 0.78) + \\ (0.78\times 1) + (0.67\times 1) + (0.53\times 1) \end{bmatrix} \times \begin{bmatrix} \sqrt{1^2+1^2+1^2+(0.78)^2+(0.67)^2+(0.53)^2} \times \\ \sqrt{(0.67)^2+(0.53)^2+(0.78)^2+1^2+1^2+1^2} \end{bmatrix}^{-1} \right)^2 = 0.68$$

and

Table 4
The similarity scores of human judgments, Yuhua Li, and ours with EBC = 1 ∼ 10.

R&G No.	Human Sim.	Exponential Balance Coefficient										Yuhua
		1	2	3	4	5	6	7	8	9	10	
1	0.01	0.66	0.44	0.28	0.18	0.12	0.08	0.05	0.03	0.02	0.02	0.33
5	0.01	0.77	0.59	0.45	0.34	0.26	0.20	0.15	0.12	0.09	0.07	0.29
9	0.01	0.68	0.47	0.34	0.25	0.19	0.14	0.11	0.08	0.06	0.05	0.21
13	0.11	0.82	0.69	0.60	0.52	0.47	0.42	0.38	0.35	0.32	0.29	0.53
17	0.13	0.94	0.86	0.80	0.74	0.69	0.65	0.60	0.57	0.53	0.50	0.36
21	0.04	0.82	0.68	0.57	0.48	0.40	0.34	0.30	0.25	0.22	0.19	0.51
25	0.07	0.83	0.70	0.60	0.52	0.46	0.40	0.36	0.32	0.29	0.26	0.55
29	0.01	0.80	0.65	0.54	0.45	0.39	0.33	0.29	0.25	0.22	0.19	0.33
33	0.15	0.90	0.82	0.74	0.67	0.61	0.55	0.50	0.46	0.42	0.38	0.59
37	0.13	0.69	0.59	0.55	0.51	0.48	0.45	0.43	0.40	0.38	0.36	0.44
41	0.28	0.73	0.55	0.42	0.34	0.27	0.22	0.18	0.15	0.12	0.1	0.43
47	0.35	0.95	0.90	0.85	0.81	0.76	0.72	0.69	0.65	0.62	0.59	0.72
48	0.36	0.82	0.70	0.62	0.56	0.51	0.48	0.45	0.42	0.39	0.37	0.65
49	0.29	0.90	0.80	0.72	0.64	0.58	0.52	0.47	0.42	0.38	0.34	0.74
50	0.47	0.84	0.74	0.68	0.64	0.61	0.59	0.57	0.56	0.55	0.53	0.68
51	0.14	0.87	0.77	0.68	0.61	0.54	0.49	0.44	0.40	0.37	0.33	0.65
52	0.49	0.70	0.49	0.35	0.25	0.18	0.13	0.09	0.07	0.05	0.04	0.49
53	0.48	0.75	0.61	0.52	0.46	0.41	0.37	0.34	0.30	0.28	0.25	0.39
54	0.36	0.52	0.42	0.34	0.27	0.22	0.18	0.14	0.12	0.09	0.08	0.52
55	0.41	0.82	0.68	0.57	0.49	0.43	0.38	0.34	0.31	0.34	0.27	0.55
56	0.59	0.97	0.93	0.90	0.87	0.85	0.82	0.80	0.77	0.75	0.73	0.76
57	0.63	0.93	0.88	0.83	0.80	0.76	0.74	0.72	0.70	0.68	0.67	0.70
58	0.59	0.83	0.71	0.62	0.55	0.50	0.46	0.42	0.39	0.36	0.33	0.75
59	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
60	0.58	0.87	0.79	0.74	0.70	0.68	0.66	0.66	0.65	0.65	0.64	0.66
61	0.52	0.92	0.85	0.78	0.73	0.67	0.63	0.58	0.54	0.51	0.48	0.66
62	0.77	0.91	0.83	0.76	0.69	0.63	0.58	0.53	0.48	0.44	0.40	0.73
63	0.56	0.97	0.94	0.92	0.89	0.87	0.84	0.84	0.40	0.78	0.76	0.64
64	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
65	0.65	0.89	0.80	0.72	0.65	0.58	0.52	0.47	0.43	0.39	0.35	0.83

$$VC_{A,C} = \left(\frac{\vec{NV}_{SEN\_A} \bullet \vec{NV}_{SEN\_C}}{\left|\vec{NV}_{SEN\_A}\right| \times \left|\vec{NV}_{SEN\_C}\right|} \right)^2 = \left(\frac{(1\times 0)+(1\times 0.5)+(0.5\times 1)}{\sqrt{1^2+1^2+0.5^2}\times\sqrt{0^2+0.5^2+1^2}} \right)^2 = 0.36.$$

The final similarity of sentences A and B is

$$Similarity_{A,C} = \zeta \times (NC_{A,C}) + (1-\zeta)\times(VC_{A,C}) = 0.65\times 0.68 + 0.35\times 0.36 = 0.568.$$

Case II. Pair BC. This case computes the similarity between sentences B and C. After pre-processing, the noun semantic space is {*price*, *accommodation*, *food*, *method*, *management*, *company*}, and the verb semantic space is {*include*, *introduce*}. According to the formulas (1)–(3), the vectors of nouns and verbs of sentences B and C are: NV_{SEN_B} = [1, 1, 1, 0.78, 0.67, 0.53] and NV_{SEN_C} = [0.53, 0.78, 0.67, 1, 1, 1], VV_{SEN_B} = [1, 0.5], and VV_{SEN_C} = [0.5, 1]. The cosine angles of sentences B and C are:

$$NC_{B,C} = \left(\frac{\vec{VV}_{SEN\_B} \bullet \vec{VV}_{SEN\_C}}{\left|\vec{VV}_{SEN\_B}\right| \times \left|\vec{VV}_{SEN\_C}\right|} \right)^2 = \left(\begin{bmatrix} (1\times 0.53)+(1\times 0.78)+(1\times 0.67)+ \\ (0.78\times 1)+(0.67\times 1)+(0.53\times 1) \end{bmatrix} \times \begin{bmatrix} \sqrt{1^2+1^2+1^2+(0.78)^2+(0.67)^2+(0.53)^2}\times \\ \sqrt{(0.53)^2+(0.78)^2+(0.67)^2+1^2+1^2+1^2} \end{bmatrix}^{-1} \right)^2 = 0.68,$$

$$VC_{A,C} = \left(\frac{\vec{NV}_{SEN\_A} \bullet \vec{NV}_{SEN\_C}}{\left|\vec{NV}_{SEN\_A}\right| \times \left|\vec{NV}_{SEN\_C}\right|} \right)^2 = \left(\frac{(1\times 0.5)+(0.5\times 1)}{\sqrt{1^2+0.5^2}\times\sqrt{0.5^2+1^2}} \right)^2 = 0.48.$$

The final similarity of sentences B and C is

$$Similarity_{B,C} = \zeta \times (NC_{B,C}) + (1-\zeta)\times(VC_{B,C}) = 0.65\times 0.68 + 0.35\times 0.48 = 0.61.$$

The results show that sentences pair A and B have the highest similarity with score 0.88.

4. Experiments

4.1. The WordNet ontology

WordNet is an online lexical database for the English language. WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of a group led by Miller. The version used in this study is WordNet 1.6, which contains over 120,000 words organized into over 99,000 synonym sets. In WordNet, words were partitioned into four categories – nouns, verbs, adjectives, and adverbs, and each concept was organized into a synonym set, called synset. A synset represents a concept in which all words have a similar or the same meaning. This research only uses the noun and verb senses since there are only noun and verbs be organized into hierarchies that based on the hypernymy/hyponymy relation between synsets.

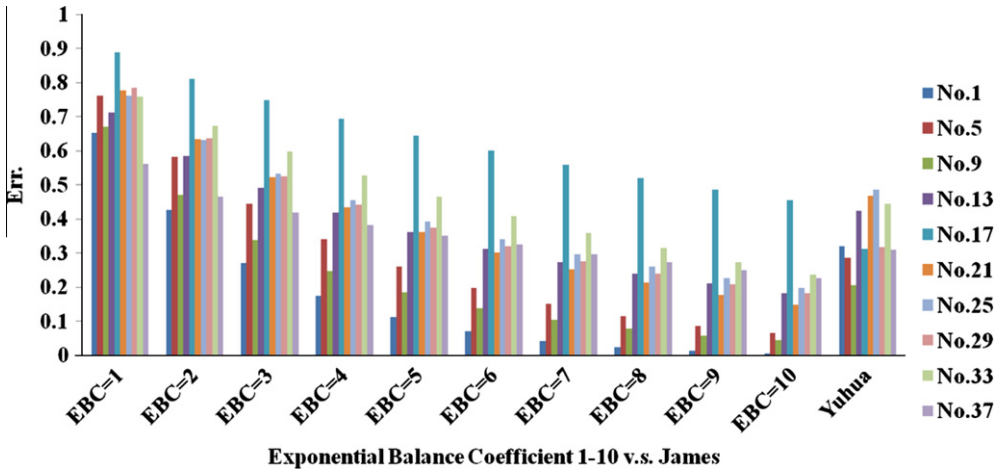

Fig. 2. Deviations from human judgments in test data no. 1–37.

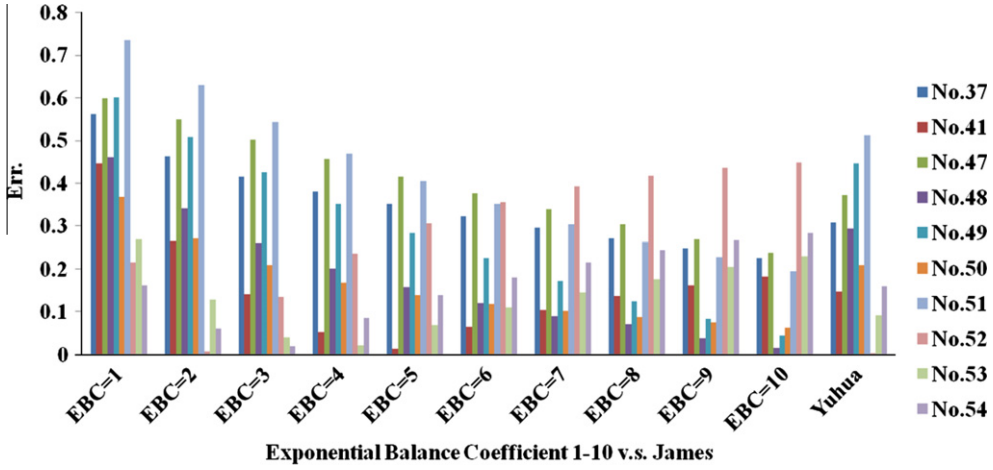

Fig. 3. Deviations from human judgments in test data no. 37–54.

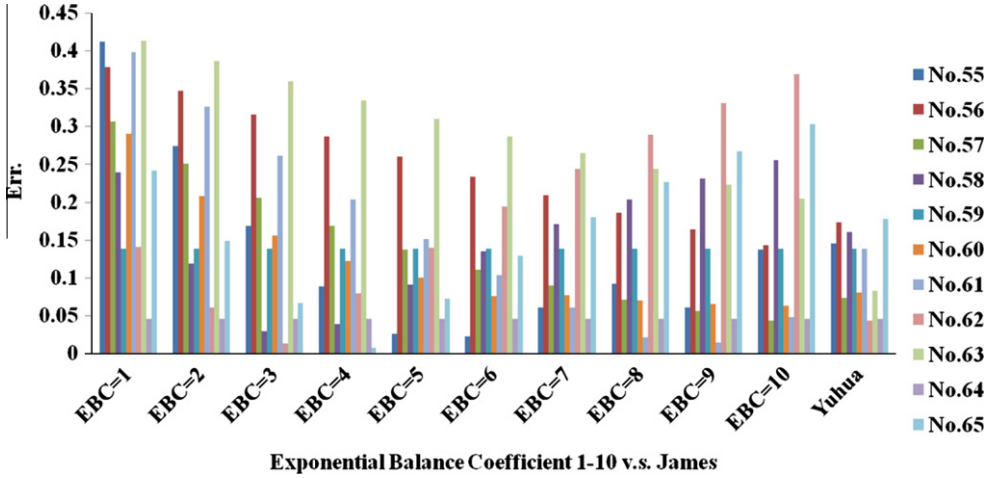

Fig. 4. Deviations from human judgments in test data no. 55–65.

4.2. Experimental results

Although a few related studies have been published, there are currently barely suitable data sets for evaluating the proposed algorithm since the performance of other approaches was mostly shown in very short sentences. To demonstrate that our approach can be applied to not only very short sentences but also long sentences, this research designed seven long sentence sets in this experiment, and each of them has three sentences. This research did not provide human similarity for the test data and the performance shown in this paper is left to the reader to judge. The test data sets and results are shown in Table 3. In this experiment, the balance coefficient ζ is the same as Section 2.2.

4.3. Compared to Yuhua Li et al.

Based on the notion of semantic and syntactic information contributed to the understanding of a sentence, Li et al. (Li, McLean, Bandar, O'Shea, & Crockett, 2006) defined a sentence similarity measure as a linear combination that based on the similarity of semantic vector and word order. A preliminary data set was constructed by Li et al. with human similarity scores provided by 32 volunteers who are all native speakers of English. The data set used 65 noun word pairs whose semantic similarities were originally measured by Rubenstein and Goodenough (1965) and were replaced with the definitions from the Collins Cobuild dictionary (Sinclair, 2001). The dictionary was constructed from a large corpus and the data set contains more than 400 million words. This experiment uses the same data set as Li et al. The complete sentence data set used in this experiment is available at http://www.docm.mmu.ac.uk/STAFF/D.McLean/SentenceResults.htm. In this experiment we use an exponential variable named the "Exponential Balance Coefficient (EBC)" to replace the square operation in formulas (4) and (5) to improve the flexibility of our method. Table 4 shows human similarity scores along with Li et al. and our semantic measure under EBCs 1-10. Human similarity scores are provided as the mean score for each pair and both scores were normalized into 0–1. Figs. 2–4 present the deviations from human judgments. The distribution of the scores and the average errors between human judgments and different EBCs were shown in Figs. 5 and 6, respectively. Our algorithm's similarity measure achieved a reasonably good performance while EBC > 3, the observation is that our approach will try to identify and quantify the potential semantic relation among words, although the common syntaxes or words of the compared sentence pairs are few or even none.

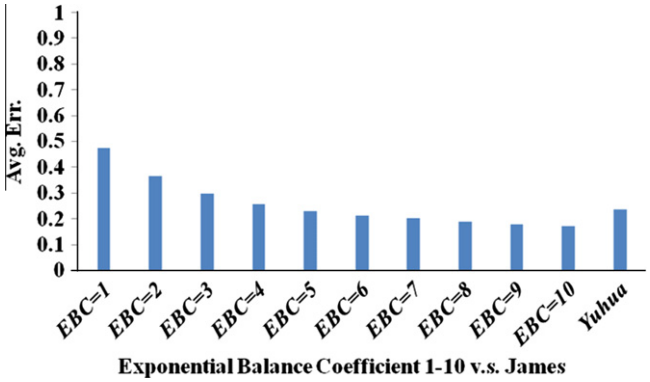

Fig. 6. Average error chart under EBC 1–10 and Li et al.

5. Conclusions

This paper presents a practical sentence similarity evaluation algorithm that based on part-of-speech and the WordNet lexical database. The similarity of sentences is difficult to evaluate since the structure of sentences may be complicated and there has no extra information to characterize the words of the sentences. Some approaches deal with this problem via determining the order of words; however, they are hard to be applied to compare the sentences with complex syntax as well as long sentences. Our approach solves this problem by the specific designed semantic space. The experiment results also demonstrate the effectiveness and significance of our approach.

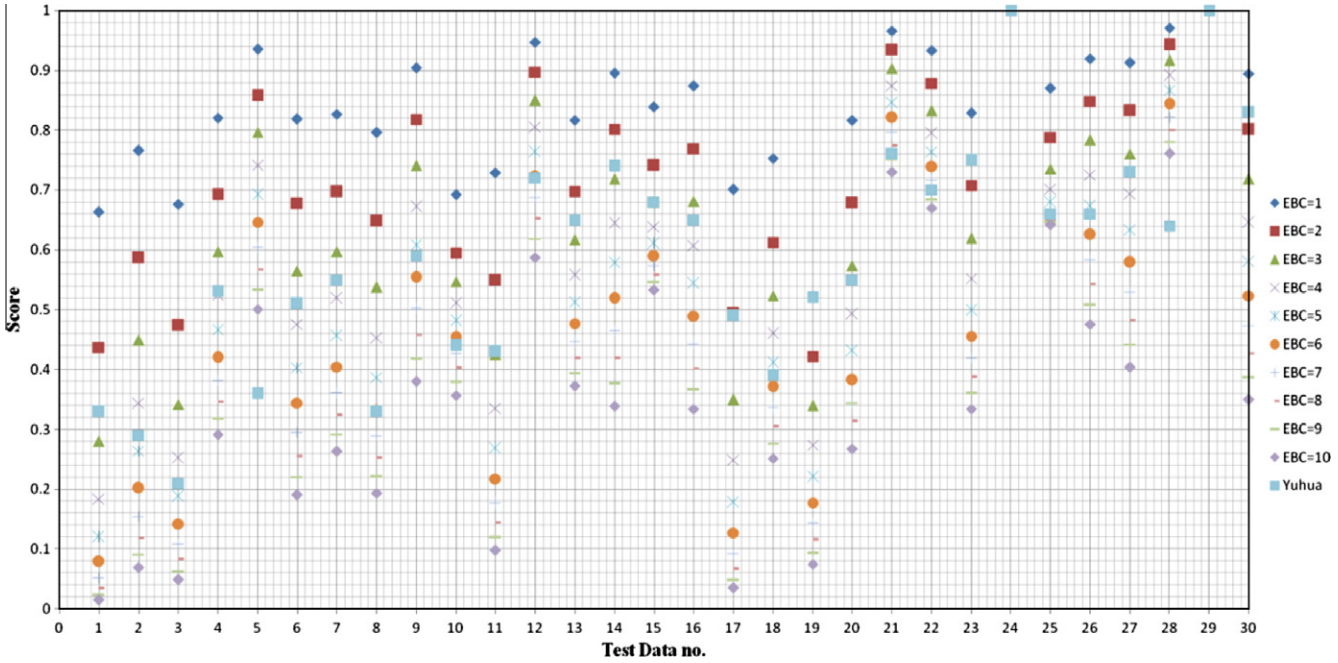

Fig. 5. The scores distribution of all test data.

References

Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications, 36*, 7764–7772.

Chiang, J. H., Ho, S. H., & Wang, W. H. (2008). Similar genes discovery system (SGDS): Application for predicting possible pathways by using GO semantic similarity measure. *Expert Systems with Applications, 35*(3), 1115–1121.

Chu, H. C., Chen, M. Y., & Chen, Y. M. (2009). A semantic-based approach to content abstraction and annotation for content management. *Expert Systems with Applications, 36*, 2360–2376.

Davies, J., Fensel, D., & Van Harmelen, F. (2003). *Towards the semantic web: Ontology-driven knowledge management*. John Wiley and Sons.

Garcı'a-Sa'nchez, F., Valencia-Garcı'a, R., Martı'nez-Be'jar, R., & Ferna'ndez-Breis, J. T. (2009). An ontology, intelligent agent-based framework for the provision of semantic web services. *Expert Systems with Applications, 36*, 3167–3187.

GO. Available from http://www.geneontology.org/GO.evidence.shtml.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199–220.

Guo, Q. l., & Zhang, M. (2009). Semantic information integration and question answering based on pervasive agent ontology. *Expert Systems with Applications, 36*, 10068–10077.

Guzman Arenas, A., & Olivares Ceja, J. M. (2006). Measuring the understanding between two agents through concept similarity. *Expert Systems with Applications, 30*(4), 577–591.

Jeong, B., Lee, D., Cho, H., & Lee, J. (2008). A novel method for measuring semantic similarity for XML schema matching. *Expert Systems with Applications, 34*, 1651–1658.

Jung, J. J. (2009). Semantic business process integration based on ontology alignment. *Expert Systems with Applications, 36*(8), 11013–11020.

Lee, T. B., Hendler, J., & Lassila, O. (2001). *The semantic web*. Scientific American.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering, 18*(8), 1138–1150.

Liu, M., Shen, W., Hao, Q., & Yan, J. (2009). An weighted ontology-based semantic similarity algorithm for web service. *Expert Systems with Applications, 36*, 12480–12490.

Michie, D. (2001). Return of the imitation game. *Electronic Transactions in Artificial Intelligence, 6*(2), 203–221.

OWL-REF. Available from http://www.w3.org/TR/owl-ref/.

Porter. Available from http://www.tartarus.org/martin/PorterStemmer/.

RDF. Available from http://www.w3.org/RDF/.

RDF-Schema. Available from http://www.w3.org/TR/rdf-schema/.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627–633.

Sinclair, J. (Ed.). (2001). *Collins Cobuild English Dictionary for Advanced Learners* (3rd ed.). Harper Collins.

WordNet. Available from http://wordnet.princeton.edu/.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. Paper presented at the proceedings of the 32nd annual meeting of the associations for computational linguistics.

XML. Available from http://www.w3.org/XML/.

xmlschema. Available from http://www.w3.org/TR/xmlschema-0/.

Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews – A text summarization approach. *Expert Systems with Applications, 36*, 2107–2115.