Journal of Web Engineering, Vol. 16, No. 1&2 (2017) 039–062

ASSOCIATION LINK NETWORK BASED SEMANTIC COHERENCE MEASUREMENT FOR SHORT TEXTS OF WEB EVENTS

WEIDONG LIU XIANGFENG LUO JUNYU XUAN DANDAN JIANG
School of Computer Engineering and Science, Shanghai University
Shanghai, China
liuwd@shu.edu.cn luoxf@shu.edu.cn xuanjunyu@shu.edu.cn Emily_dan@shu.edu.cn

ZHENG XU
The Third Research Institute of Ministry of Public Security
Shanghai, China
xuzheng@shu.edu.cn

Received July 25, 2015
Revised April 4, 2016

As novel web social Media emerges on the web, large-scale short texts are springing up. Although these massive short texts contain rich information, their disorder nature makes users difficult to obtain the desired knowledge from them, especially the semantic coherent knowledge. Different orders of these short texts often express different semantic coherence states. Therefore, how to automatically measure semantic coherence of short texts is a fundamental and significant problem for web knowledge services. Existing related works on the semantic coherence measurement of different orders of short texts/sentences seldom focus on graph structure of semantic link network for reflecting coherence change, measuring coherence by these graph-based features and discovering some interesting coherence patterns. In this paper, we propose an association link network based semantic coherence measurement for short texts of web events. Our method firstly construct an association link network from which some graph-based features are then extracted to measure semantic coherence of different orders and lastly some coherence patterns are discovered for guiding automatically text ordering/generation. To validate correctness of our method, we conduct a series of experiments including sentence order permutation, sentence removal and adding/replacing sentence and compare with other two methods. The results show that our method can measure semantic coherence with higher accuracy and outperforms other methods in some experiments. Such method can be widely applied in web text automatic generation, web short text organization and web event summarization etc.

Keywords: association link network, semantic coherence measurement, short text of web events
Communicated by: M. Gaedke & Q. Li

1. Introduction

Coherence is defined as a "continuity of senses" and "the mutual access and relevance within a configuration of concepts and relations" [1]. Semantic coherence is related with semantic association and sound organization structure of these concepts and their association. These specific concepts are distributed in conceptual space in semantic link network[2]. Different

configurations of concepts and relations will cause different coherence states on semantic link network, which exhibit measurable by graph-based features on semantic link network.

In the human discourse process, semantic coherence is a key problem since readers/writers routinely attempt to construct coherent meanings and connections among text constituents [3]. When facing massive unordered sentences, we will try to pursue semantic coherence from them as shown in Fig. 1. The user first acquires the meaning in keywords/sentences level as steps 1-2. Then, he/she tries to understand the semantic association between keywords/sentences. If the semantic association implied by these sentences are semantic incoherent, the user will reorder sentences to generate different links between keywords/concepts for maximizing the semantic coherence as loops in 3-4. Lastly, the user obtains semantic coherence as steps 5-6.

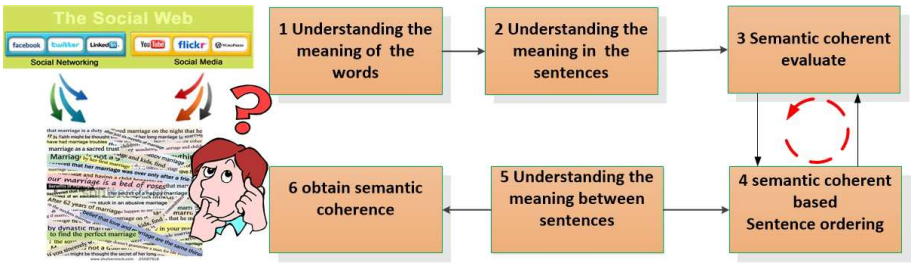


Fig. 1. human beings' semantics coherence processing

As the increasing growth of microblog usage, short texts become the main message passing forms and massive unordered short texts are emerging on the web. How to discovery semantic coherence from these large scale short texts is a practical and challenging problem since large scale data set, loose associations and unordered distribution are beyond human beings' ability to construct semantic coherence on them. Although some methods are proposed to summarize core semantics and automatically organize unordered sentences, these existing methods, however, do not have a fundamental method for measuring semantic coherence. In fact, a reasonable and efficient coherence measurement is a foundational problem for web text processing and semantic computation, such as web event summary and web text organization etc. In this paper, our research mainly focus on these unordered short texts on web. For simplicity, we use sentences to refer short texts in the following parts since the short texts and sentences are alike in length.

To solve the above fundamental problem, we face two challenging issues:

- what features can reflect semantic coherence.
- how to use the coherence features to measure semantic coherence states of different sentence orders.

To solve the above issues, we propose association link network based semantic coherent measurement for short texts on web. Such method constructs different association link networks to represent different sentence semantic coherence state under different sentence orders, extracts graph-based features from association link network and discoveries interesting coherence patterns by combination of coherence features,which has the following merits:

1. Different from previous work, our method is based on semantic link network to measure semantic coherence. The semantic relations and structure can reflect the text coherence since the semantic process and knowledge storage are on semantic net and different coherence is related with different configuration of concepts and relations.

2. Compared with other methods, our method can discover easily understood coherence patterns which conform to the human cognitive. Such simple patterns are easily measurable for sentence coherence. So our method satisfies human cognitive theory more than other methods.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 mainly gives the framework of our method; Section 4 mainly computes semantic coherent elements, including the sentence representation and the coherence features computation. Section 5 uses these coherent features to compute semantic coherence and discover coherence patterns by combination different coherence features. Section 6 reports experiments. Section 7 makes conclusion.

2. Related Work

The researches of discourse semantic coherence measurement methods are summarized as follows.

Sentence similarity is often used to measure coherence of text, which assumes that sentences of coherent text have most overlaps between sentence in keyword semantic space[4]. The Word Net, Wikipedia and Google search engine are used for topic coherence evaluation by measuring sentence similarity in concept semantic space[7]. Coh-metrix is developed to analyze text in point view of cohesion and readability[8]. Entity distribution between sentence exhibits some patterns for reflecting semantic coherence. For example, keyword pair pattern[12] and entity grids[4, 5] for which syntactic role transition between sentences are used for measuring semantic coherent. Discourse rhetorical structure relation transition between sentences provide some clues for textual coherence, so distribution of discourse relations are used in coherence anaysis[6]. Some other works study influence of temporal feature on semantic coherence in News[9].

Some semantic coherence assessment methods are proposed from a sentence ordering point view. The structure of text is assumed as a tree. The tree-node are linked rhetorical or discourse relation by Marcu[10], where sentence ordering is changed into searching an optimization tree structure. Sentence grouping and similarity computation are often used in sentence ordering method which groups sentences by similarity and orders sentences by maximizing similarity between sentence groups and similarity between sentences in each group[11]. Domain-dependent methods are often used in sentence ordering. For example, the entity dependent relations and probabilistic methods are learned from domain to guide sentence ordering[12]; HHM methods are used to represent clustered topics and topic shift relations, the best sentence order are obtained when possibility is maximal[13]. Entity-based and HMM-based models are combined to be complementary to each other in coherence assessment[15, 16]. Publication date and Temporal cues are used in sentence ordering[14]. Textual complexity changed with different sentence orders[17].

However, the above methods have the following limitations or unsolved issues as follows:

1. latent structure changes are not discovered from semantic link network. The discourse knowledge is organized and stored in semantic link network[18] whose structure changes are guiding signification for semantic coherence analysis. However, the above methods do not reveal structure changes in semantic link network.

2. less attentions are paid on coherence pattern analysis. For a coherent text, these are some coherence patterns used in composition. Automatically and accurately analyzing some patterns can guide web text generation and summarization. however, little related work provides coherence patterns.

3. Preliminary Work and Problem Statement

3.1. *Preliminary Work*

Given a sentence order $S_{(1:T)} = (s_{(l)}|l = 1, 2, \cdots, T)$, where $s_{(l)}$ denotes the l^{th} sentence, two types of knowledge support semantic coherence of the sentence order, including 1) explicit knowledge from sentence context knowledge; 2) implicit knowledge from domain knowledge.

Herein, we propose explicit association link network and implicit association link network, which respectively correspond to explicit knowledge and implicit knowledge.

Definition 1: Explicit Association Link Network, $EALN_{S_{(1:T)}}$
EALN explicitly represent keywords and relations given a sentence order. The relations in EALN are within each sentences, but not include that between sentences. It is denoted by

$$EALN_{S_{(1:n)}} =< N, E > \tag{1}$$

where $S_{(1:n)} = (s_{(l)}|l = 1, 2, \cdots, T)$;$N = \{k_1, k_2, \ldots, k_k\}$; k_i denotes keyword index i; $E = \{E_{i,j}\}$,$(i \neq j)$ denotes association strength between keywords k_i and k_j; $E_{i,j}$ is calculated by

$$E_{i,j} = \frac{\sum_{l=1}^{|T|-w} I(k_i, k_j|s_{(l)} \square s_{(l+w)})}{\sum_{l=1}^{|T|-w} I(s_{(l)}|s_{(l)})} \tag{2}$$

where $s_{(l)} \square s_{(l+w)}$ denotes a sliding window across from $s_{(l)}$ to $s_{(l+w)}$ in a given sentence order $S_{(1:n)}$; w denotes the length of sliding window.

Definition 2: Implicit Association Link Network, $IALN$
$IALN$ is used to represent keywords and relations in domain knowledge. The relations in IALN not only includes that in sentences, but also includes that between sentences. It is denoted by

$$IALN =< N, I > \tag{3}$$

where $N = \{k_1, k_2, \ldots, k_k\}$; $I = \{I_{i,j}\}$,$(i \neq j)$ denotes association strength between keywords k_i and k_j; $I_{i,j}$ is calculated by

$$I_{i,j} = \frac{\sum_{S_{(1:T_i)} \in D} \sum_{l=1}^{|T_i|-w} I(k_i, k_j|s_{(l)} \square s_{(l+w)})}{\sum_{S_{(1:T_i)} \in D} \sum_{k=1}^{|T_i|-w} I(s_{(l)}|s_{(l)})}; \tag{4}$$

where $D = \{S_{(1:T_i)}|i = 1, 2, \cdot, m\}$ denotes domain knowledge which consist of m sentence orders; $s_{(l)}\square s_{(l+w)}$ denotes a sliding window across from $s_{(l)}$ to $s_{(l+w)}$ in $S_{(1:T_i)}$; w denotes the length of sliding window.

The semantic coherence link network of a sentence order is constructed by combining the explicit association link network and implicit association link network.

Definition 3: Semantic Coherence Link Network, $SCN_{S_{(1:T)}}$

SCN is used to represent keywords and relations included by a given sentence order $S_{(1:n)} = (s_{(l)}|l = 1, 2, \cdots, T)$. It is denoted by,

$$SCN_{S_{(1:T)}} =< N, R > \tag{5}$$

where $N = \{k_1, k_2, \ldots, k_k\}$; k_i denotes a keyword with index i; $R = \{R_{i,j}\}$, $(i \neq j)$ denotes association strength between k_i and k_j; $R_{i,j}$ is calculated by,

$$R_{i,j} = \alpha \times I_{i,j} + (1 - \alpha) \times E_{i,j} \tag{6}$$

where α denotes a tuning parameter which weights the association strengths of implicit association relation and that of explicit association relation.

3.2. *Problem Statement*

Before we introduce the problem statement, we introduce some definitions which are used in following sections.

Definition 4: coherence features of a sentence order, $CFs(S_{(1:T)})$

The coherence state of the sentence order $S_{(1:T)}$ is represented by $CFs(S_{(1:T)})$,

$$CFs(S_{(1:T)}) = (f_i|i = 1, 2, \cdots, n) \tag{7}$$

where f_i denotes the i^{th} coherence feature; n denotes the total number of semantic coherence feature.

Definition 5: coherence value of a sentence order $S_{(1:T)}$**,** $CV(CFs(S_{(1:T)}))$

the coherence state $CFs(S_{(1:T)}$ is measured by CV which satisfies:

if $S_{(1:T)}$ performs higher semantic coherence than $S^{'}_{(1:T)}$,

then

$$CV(CFs(S_{(1:T)})) > CV(CFs(S^{'}_{(1:T)})) \tag{8}$$

where $S_{\{1:T\}} = \{s_l|l = 1, 2, \cdots, T\}$ denotes a sentence set and s_l denotes a sentence in the set; $S_{(1:T)} = (s_{(l)}|l = 1, 2, \cdots, T)$ is one sentence order of $S_{\{1:T\}}$ and $s_{(l)}$ denotes the l^{th} sentence in $S_{(1:T)}$; $S^{'}_{(1:T)} = (s^{'}_{(l)}|l = 1, 2, \cdots, T)$ is another one sentence order of $S_{\{1:T\}}$ and $s^{'}_{(l)}$ denotes the l^{th} sentence in $S^{'}_{(1:T)}$.

Definition 6: Coherence pattern, P

P is a coherence pattern which occurs more in coherent sentence orders then incoherence ones.

$$P = (b_i|i = 1, 2, \cdots, n) \tag{9}$$

where b_i is binary element; each b_i corresponds to the semantic coherence feature f_i.

Give a sentence order $S_{(1:T)} = (s_{(l)}|l = 1, 2, \cdots, T)$, the following issues to be solve by this paper:

1. how to select coherence features $CFs(S_{(1:T)})$ as coherence state of $S_{(1:T)}$;
2. how to quantize the coherence sate by semantic coherence value $CV(CFs(S_{(1:T)}))$;
3. how to discover more coherence patterns $\{P_i\}$ which are widely used in coherent sentence orders;

To solve the above issues, Fig. 2 shows the framework of association link network based semantic coherence measurement for short texts of web events. We briefly describe the framework and each module.

Event discover methods mainly discover different events from physical space, where each event consists of short texts from Tweet, Weibo and Commend etc. Event discover methods are not our focus in this paper. We just adopt some existing event discover algorithms to obtain some core short texts of web event[20, 21, 22, 24].

Given a sentence order $S_{(1:T)} = (s_{(l)}|l = 1, 2, \cdots, T)$, our semantic coherence measurement is divided into 4 steps as follows:

1. To precisely represent semantic coherence, the 1^{st} step constructs semantic coherence link network(SCN) which exhibits different graph structures with different orders of short texts for each event.
2. To obtain some distinguished measurable features for semantic coherence, the 2^{nd} step extracts some measurable graph-features from association link network which show significant difference when comparing coherent sentence orders with unordered ones for each event.
3. To quantize different coherence state of different sentence orders, the 3^{th} step uses a coherence measurement which use different coherence features to measure semantic coherence.
4. To discover interesting coherence patterns from coherent sentence orders, the 4^{th} step discovers different coherence patterns which are useful for guiding generation of semantic coherence.

4. Basic Model and Method

4.1. *Basic Model*

The discourse processing is conducted on association link network, where semantics is represented as keywords and its relations. When different sentence orders are processed, different association link networks are generated since the keywords across the sliding window are different and are linked. Some graph-based features are extracted from the association link network for reflecting different coherence. To compute semantic coherence, 3 tasks are unfolded by: 1) semantic coherence link network construction; 2) coherence feature computation; 3) Semantic coherence computation and coherence pattern discovery.

Semantic coherence link network construction Semantic coherence link network is used to reflect coherence states changed with different sentence orders. Given a

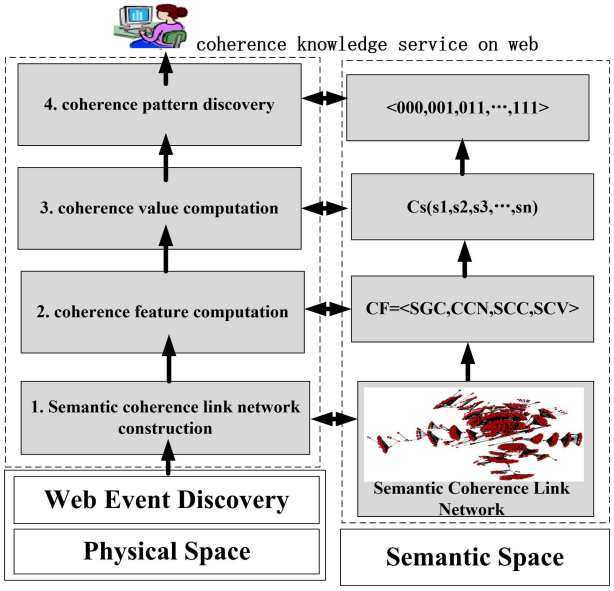
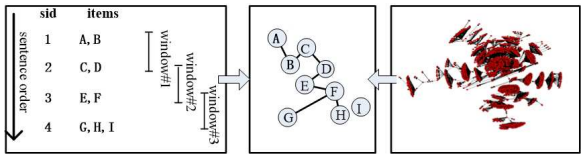
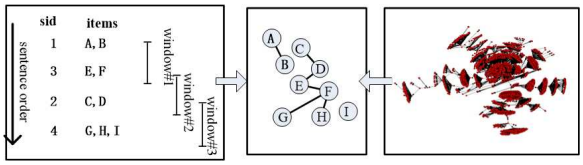


Fig. 2. Framework of association link network based semantic coherence measurement for short texts of web events

sentence order, whether an association relation exist between two keywords is conditioned on the association strength between the keywords in explicit association link network and that in implicit association link network by Eq. 6.



(a) Semantic coherence link network of sentence order 1



(b) Semantic coherence link network of sentence order 2

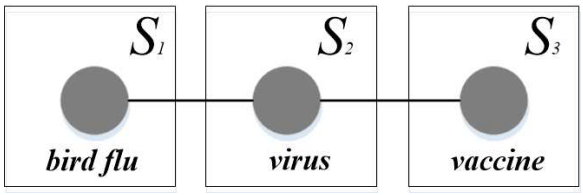
Fig. 3. construction process of semantic coherence link network

Fig. 3 shows that semantic coherence link network is constructed by collaborating with explicit association link network mined from a given sentence order and implicit association link network. In Fig 3, sentence order 1 is $S^1_{(1:4)}$={AB, CD, EF, GHI} and sentence order 2 is $S^2_{(1:4)}$={AB, EF, CD, GHI}, where A, B, C, D, E, F, G, H, I denote different keywords which distribute on 4 sentences: AB, CD, EF, GHI. In Fig. 3, different configurations of keywords occur in a sliding window when sentence orders are different. ABCD is a transaction for

$S^1_{(1:4)}$ when sliding window begin with A, while ABEF is a transaction for $S^2_{(1:4)}$. Whether an association relation exists is determined by explicit association link network and implicit association link network as Eq. 6.

Semantic coherence features computation Given a semantic coherence link network of a sentence order, we can extract some graph-based features from the semantic coherence link network. By comparing these features of coherent sentence orders with that of incoherent ones, we can discover some distinguished features which have significantly difference in feature values between coherent sentence orders and incoherent ones.

(a) weak global connectivity



(b) strong global connectivity

Fig. 4. Coherent change with global connectivity

Heuristic 1: global connectivity. the keywords in coherent sentence orders are interconnected more other than that of incoherence ones.

A coherent sentence order often has high global connectivity between keywords while incoherence sentence order includes isolated keywords or fragments which break coherence of sentences.

For example, the semantic coherence link network in Fig.4(b) is more coherent than that of Fig.4(a). In Fig.4(a), many isolated keywords have no relations with other keywords. We hardly obtain the association relations between keywords and thus can not build semantic coherence among the keywords. In Fig.4(b), the keywords 'bird flu', 'virus' and 'vaccine'are linked into a connective graph and thus we can easily obtain coherent semantics by accessing the connective graph.

Definition 7: Semantic global connectivity, *SGC*

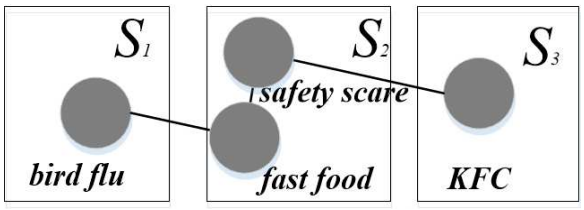
SGC represents global interconnectivity of keywords in semantic coherence link network. It can be defined as:

$$SGC = 1 - \frac{N_c}{N} \tag{10}$$

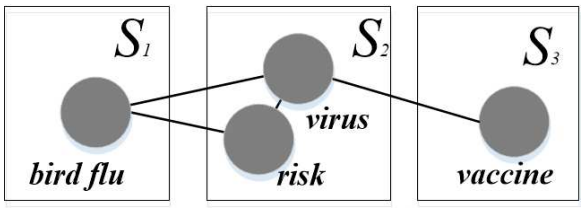
where N_c is the number of connected sub-graphs in semantic coherence link network; N is the number of keywords in semantic coherence link network. Low *SGC* shows that the semantic coherence link network lacks a global connectivity since there are many isolated keywords or connective sub-graphs in semantic coherence link network. Massive isolated keywords make

sentence difficult to be understood because of lacking semantic association between keywords. A coherent sentence order should have higher semantic global connectivity between keywords.

For example, the SGC of Fig.4(a) is calculated by 1-3/3=0; the SGC of Fig.4(b) is calculated by 1-1/3=2/3.



(a) weak coherent compactness



(b) strong coherent compactness

Fig. 5. Coherent change with coherent compactness

Heuristic 2: semantic compactness. A coherent sentence order often has high semantic compactness.

Semantic compactness measures how compact keywords in semantic coherence link network between each other. Compared with loose semantic coherence link network, compact semantic coherence network enables reader to access keywords more directly because of compact relations between keywords in the coherent sentence order.

For example, semantic coherence link network in Fig.5(b) is more coherent than that in Fig.5(a) since the keywords in Fig.5(b) obtains more direct explanation from others. In Fig.5(a), obtaining relations of 'bird flu' and 'KFC' needs 3 steps and obtaining all relations between keywords needs 10 steps. While, in Fig.5(b), obtaining the relation between 'bird flu' and 'vaccine' only needs 2 steps and grasping all relations between keywords needs 8 steps.

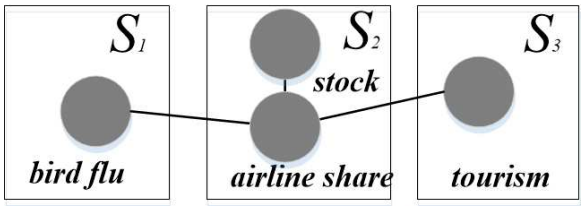
Definition 8: Coherent compactness, CCN

CCN represents the extent of cross referencing of keywords in SCN. The measure is defined as

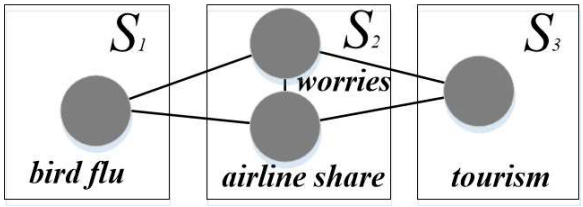
$$CNN = \frac{Max - \Sigma_{u \in N} \Sigma_{v \in N} d(u,v)}{Max - Min} \tag{11}$$

where $Max = N \times (N \times (N-1))$, $Min = N \times (N-1)$; $d(u,v)$ is the shortest path length of keyword u and keyword v; N is the number of keywords in semantic coherence link network. When CCN is high, keywords have more straightforward explanations from other keywords with fewer steps. Low CCN shows that the semantic association relations between keywords are loose where keywords are hardly directly explained by others.

For example, in Fig.4(a) $Max = 4 \times (4 \times 3) = 48$, $Min = 1 \times (4 \times 3) = 12$ and $\Sigma_{u \in N} \Sigma_{v \in N} d(u,v) = 20$ which is substituted into Eq.11, resulting in $CNN = \frac{48-20}{48-12} = 7/9$. In Fig.4(b), $Max = 4 \times (4 \times 3) = 48$, $Min = 1 \times (4 \times 3) = 12$, $\Sigma_{u \in N} \Sigma_{v \in N} d(u,v) = 16$ which is substituted into Eq.11, resulting in $CNN = \frac{48-16}{48-12} = 8/9$.



(a) weak clustered semantics



(b) strong clustered semantics

Fig. 6. Coherent change with clustered semantics

Heuristic 3: clustered semantics. A coherent sentence order often has highly clustered semantics

Clustered semantics are keyword clusters mined from sentences to represent semantics. Single keyword often does not express substantial semantics since it does not provide some association relations of this keyword as context.

For example, SCN in Fig.6(b) shows more semantic coherent than that in Fig.6(a) since it supplements some context information. In Fig.6(a), although 'bird flu' affects 'airline share' is given, why it can affect 'airlines' is still unknown. In Fig.6(b), we know that such 'bird flu' affects 'airline share' because 'tourists worry about bird flu and give up their air travel'. These association relations are understood after understanding their context information.

Definition 9: Semantic coherence clustering, SCC

SCC reflects the degree to which the keywords involve the clustered semantics. The measure is defined as

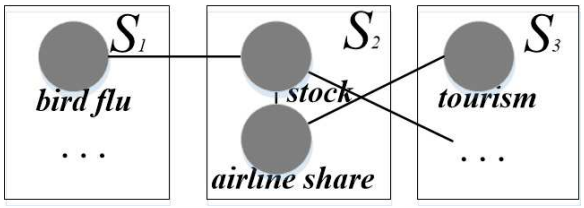
$$SCC = \frac{1}{|D|}\Sigma_{d_i \in D|} AC_{d_i} \tag{12}$$

where $D = \{d_i | i = 1, 2, \cdots\}$ denotes a set of keyword degrees in SCN; d_i denotes; AC_{d_i} denotes the average clustering coefficient of keyword with degree d_i; N is the number of keywords in SCN. When SCC is high, most keywords in semantic coherence link network are linked into some clustered semantics. Low SCC shows the few keywords involve clustered semantics. Coherent sentence orders should keep higher SCC, which enables these clustered keywords to express the main points and gives more context information (e.g. topic keywords).

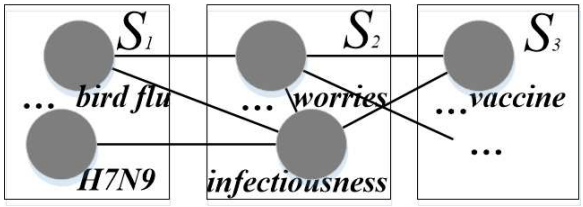
For example, in Fig.6(a), $D = \{1, 3\}$, $|D| = 2$, $AC_1 = 0$ and $AC_3 = 0$ which are substituted into Eq.12, resulting in SCC=0. In Fig.6(b), $D = \{2, 3\}$, $|D| = 2$, $AC_2 = 1$ and $AC_3 = 2/3$ which are substituted into Eq.12, resulting in SCC=5/6.

Heuristic 4: stable semantic variance. A semantic coherent sentence order often has relative stable semantic variance.

Semantic variance often means that the overall clustered semantic change between keywords



(a) weak stable semantics



(b) strong stable semantics

Fig. 7. Coherent change with semantic stability

of semantic coherence link network. A semantic coherent sentence order often has more stable change than incoherent one.

For example, SCN in Fig.7(b) is more semantic coherent than SCN in Fig.7(a) since the Fig.7(b) has more stable semantic changes between sentences than Fig.7(a). In Fig.7(a), there are unstable semantic changes from sentence 1 to sentence 2 since the keywords in sentence 2 are weakly explained by sentence 1. In Fig.7(b), there are stable changes from sentence 1 to sentence 2 since keywords in these sentences are explained by each other.

Definition 10: Semantic coherence variance, SCV

SCV has a positive correlation with standard deviation of coefficient distribution on different degree of keywords, which represents the variation of semantic distribution around average clustering coefficient. It can be defined by

$$SCV = 1 - \exp((\frac{1}{|D|-1} \times \Sigma_{d_i \in D}(AC_{d_i} - AC^{'})^2)^{\frac{1}{2}}) \tag{13}$$

where AC_{d_i} means average clustering coefficient of keywords with degree i; $AC^{'} = \frac{1}{|D|}\sum_{d_i \in D} AC_{d_i}$ means the average clustering coefficient of all keywords. High SCV suggests that clustering coefficient unevenly distributes on different keyword degrees.

For example, If $D = \{1, 2, 3\}$, $|D| = 3$, $AC_1 = 0$, $AC_3 = 2/3$, $AC_3 = 1/6$ and $AC^{'} = \frac{0+2/3+1/6}{3} = 1/12$, then $SCV = \frac{((1/12)^2+(7/12)^2+(1/12)^2)^{1/2}}{2}$ by Eq.13.

4.2. *Method for semantic coherence computation and coherence pattern discovery*

In previous section, some graph-based features are extracted from semantic coherence link network. Here, we try to study how to use these features to measure sentence semantic coherence and discover the coherence patterns in coherent sentence orders.

4.2.1. *Semantic coherence computation*

For a sentence order $S_{(1:T)}$, coherence features $CFs(S_{(1:T)})$ in Eq.7 is refined by Eq.14.

$$CFs(S_{(1:T)}) = (SGC, CCN, SCC, SCV) \tag{14}$$

where semantic global connectivity SGC is referred to Eq. 10; coherent compactness CCN is referred to Eq. 11; semantic coherence clustering SCC is referred to Eq. 12 and Semantic coherence variance SCV is referred to Eq. 13.
$CV(S_{(1:T)})$ in Eq.8 is calculated by,

$$CV_{\vec{W}}(S_{(1:T)}) = \vec{W} \times CFs(S_{(1:T)}) \tag{15}$$

where $\vec{W}$ denotes weight vectors for weighting influence of each coherence feature on coherence value; $CFs(S_{(1:T)})$ denotes the coherence features of sentence order $S_{(1:T)}$.

Through analyzing these coherence features, it is found that different features have different influence on the semantic coherence.

To learn the weight vector, we minimizes the number of violations of Eq. 8 by Eq.16,

$$\vec{W} = \arg\max_{\vec{W}} \Sigma_{(S_{(1:T)}, S'_{(1:T)}) \in CO \times ICO} I(CV_{\vec{W}}(S_{(1:T)}) > CV_{\vec{W}}(S'_{(1:T)})) \tag{16}$$

where $I(CV_{\vec{W}}(S_{(1:T)}) > CV_{\vec{W}}(S'_{(1:T)}))$ denotes an indicator function whose outcome is 1 when $CV_{\vec{W}}(S_{(1:T)}) > CV_{\vec{W}}(S'_{(1:T)})$,0 otherwise; CO denotes a set of coherent sentence order; ICO denotes a set of incoherent sentence order.

We use the SVMlight Package to learn the weight vector by Eq.16, more details in [19].

4.2.2. *Semantic coherence pattern discovery*

Given a sentence set, an intuitive idea is that coherent sentence order has higher coherent compactness, higher semantic coherent clustering and stable semantic coherent variant. What is other coherence patterns which can distinguish semantic coherent sentence from incoherence ones? In this section, we manly solve the above issue.

The coherence pattern P in Eq.9 is refined by Eq17.

$$P = (b_{f_i} | f_i \in CFs) \tag{17}$$

where f_i denotes a coherence features in CFs; b_{f_i} is binary element which corresponds to f_i.
Definition 11: Coherence pattern space, CPS
CPS contains all possible patterns of coherent features when two sentence orders are compared. It can be denoted as

$$CPS = \{P_l | l = 0, 1, \cdots, 2^{|CFs|} - 1\} \tag{18}$$

where $P_l = (b^l_{f_i} | f_i \in CFs)$ with coherence support $Sup(P_l)$ and coherence confidence $Conf(P_l)$, which respectively as:

1) $Sup(P_l)$ denotes the percentage of this coherence pattern occurs in the coherence pattern space, which is calculated by,

$$Sup(P_l) = \frac{n(P_i)}{\Sigma_i n(P_i)} \tag{19}$$

2) $Conf(P_l)$ represent the distinguish ability of coherence patterns P_l in predicting whether a sentence order is semantic coherent. It is calculated by,

$$Conf(P_l) = \frac{n(P_l)}{n(P_l) + n(\widetilde{P_l})} \tag{20}$$

where $n(P_l)$ is the frequency of pattern P_l; $\widetilde{P}_l$ denotes a couple pattern of P_l by,

$$\widetilde{P}_l = (\widetilde{b}^l_{f_i} | f_i \in CFs) \tag{21}$$

where $\widetilde{b}_{f_i} = 1 - b_{f_i}$. High confidence suggests that coherence pattern P_l have strong power to distinguish semantic coherence.

To discover coherence pattern useful semantic coherence, we conduct algorithm 8 as follows:

Fig. 8. coherence pattern discovery algorithm

Input: pairs of a original sentence order and one of its permutation
Output: coherence patterns space with support and confidence

1. calculating the coherence features for each pair by Eq. 10 to Eq. 13;
2. encoding the coherence pattern $P_l = (b_{f_i} | f_i \in CFs)$ where by: $b^l_{f_i} = 1$ if the f_i of the original sentence order are higher than that of its permutation, $b^l_{f_i} = 0$ otherwise;
3. calculating the support and confidence by Eq. 19 and Eq. 20;
4. **return** CPS with support and confidence

5. Experiment

In this section, we conduct some experiments to validate the correctness of our semantic coherence measurement method.

5.1. *Data Set*

Table 1. Description of datasets

Data Set	Dataset 1	Dataset 2
Source	Health news in routers	Environment news in routers
# News(benchmark data)	10000	10000
Avg.# sentences	15.67	13.29

To validate the correctness of our sentence semantic coherence measurement method based on association link network, we downloaded 20000 news about health and environment respectively 10000 on reuters website (http://www.reuters.com) from March 2009 to August 2009 as data set. Table 1 gives a description of the data set. It contains 10000 texts and each text has average 15.67 sentences for health news and 13.29 sentences for environment news. We shuffle sentences of the texts and use these unordered sentence sets as experimental datasets in the following experiments.

5.2. *Validating the correctness of semantic coherent feature computation*

To evaluate the performance of the proposed coherence features on reflecting semantic coherence, we observe the changes of coherent features when different numbers of sentence are

removed and the sentences are disordered.

5.2.1. *Coherent feature changes when removing sentences*

To verify that semantic coherence features can dynamically reflect semantic coherence, we observe changes of semantic coherence features changes caused by removing different number of sentences. For a sentence set, removing some sentences usually breaks its semantic coherence. The full sentence set is regarded as more semantic coherence than the remaining sentences after removing some sentences. Better coherent features should be sensitive to these changes caused by removing sentences.

We randomly select 1000 texts from data set 1 in table 1. For each text, N sentences are removed ($0 \leq N \leq 7$). We observe changes of coherent features after removing N sentences. The above process is conducted 10 times and then we calculate its average values, following by carrying out a one-way analysis of variance (ANOVA) to examine the effect on coherent features when removing N sentences. we list the results in table 2 and show them by Fig. 9.

Table 2. Average sematic coherent features change when removing different number of sentences

# removed sentences	Avg(SGC)	Avg(CCN)	Avg(SCC)	Avg(SCV)
0	0.7055	0.5754	0.4356	0.0956
1	0.6718	0.5255	0.4263	0.1033
2	0.6437	0.4822	0.4057	0.1122
3	0.6131	0.4384	0.3893	0.1133
4	0.5804	0.3954	0.3701	0.1182
5	0.5550	0.3628	0.3639	0.1208
6	0.5270	0.3304	0.3453	0.1243
7	0.4985	0.2992	0.3255	0.1232
Sig.	0.000	0.000	0.000	0.000

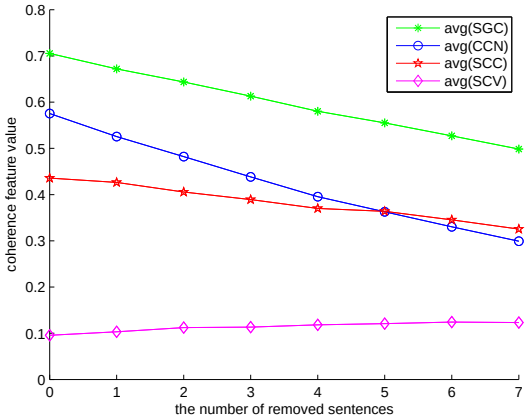


Fig. 9. The coherent feature changes when removing different number of sentences

Table 2 and Fig. 9 show the average values of semantic global connectivity (SGC), coherent compactness (CCN), semantic coherence clustering (SCC), and semantic coherence variance

(SCV) change when removing N sentences ($0 \leq N \leq 7$). ANOVA assumes that if coherent features have obvious signification difference between different groups, then level of significance $sig. \leq 0.05$. If a feature has $sig. < 0.05$, then such feather is statistically sentential to semantic coherent change. ANOVA analysis shows that SGC, CCN, SCC and SCV have significant changes($(Sig < 0.05)$) when N sentences are removed. So these coherent features are sensitive to the coherent changes caused by removing sentences.

5.2.2. *Coherent features changes when disordering sentences order*

To further verify that the semantic coherence features can reflect changes of semantic coherence, we observe the change of semantic coherence caused by permuting sentences. For a text, the original sentence order is thought to be more semantic coherent than its permutations. A better coherent feature can reflect such changes when comparing the coherence features of original sentence order with that of its permutations.

We randomly select 1000 texts from data set 1 and data set 2 in table 1. For each text, we disorder sentences and observe the difference of coherence feathers between the ordinal sentence order and its permutations. The above process is conducted 10 times and then we list their average values in table 3. We carry out a one-way analysis of variance (ANOVA) to examine the effect on coherent features caused by disordering sentences.

Table 3 and Fig. 10 show that 4 coherent features have significant difference between original sentence orders and disordered ones. The ANOVA reveals that SGC, CCN, SCC and SCV have significant changes when the sentence order is permutated, since $sig \leq 0.05$. So these coherence features are sensitive to coherence change caused by permuting sentences.

Table 3. Average sematic coherent features change when disordering sentences

# removed sentences	Avg(SGC)	Avg(CCN)	Avg(SCC)	Avg(SCV)
original sentence order	0.7047	0.5770	0.4360	0.0947
disordered sentences order	0.6699	0.5253	0.4107	0.1089
Sig.	0.003	0.00	0.040	0.023

5.2.3. *Coherence features selection by analysis of variance*

To discover which combinations of coherence feature are more efficient to measure semantic coherence, we evaluate different combinations of coherence features to distinguish coherent sentence order from incoherence ones. Table 4 lists all possible combinations of the coherence features. When a feature is used in experiments, it is marked by 1; otherwise 0. We think that effective feature combinations can distinguish coherent sentence order more precisely.

We randomly select 1000 texts from data set 1 and data set 2 in table 1. For each text, we mix the original sentence order with one of its permutations. The difference features combinations is listed in table 4, we use classification algorithms (ID3) [23] to distinguish coherent sentence orders from others. We think that a better combination exhibits higher accuracy. The above process is conducted 10 times and we calculate the average values listed in table 5.

Table 5 shows the accuracy of classification under different feature combinations in table 4. It shows that the combinations 2, 6, 14 have higher accuracy than others since their $mean >$

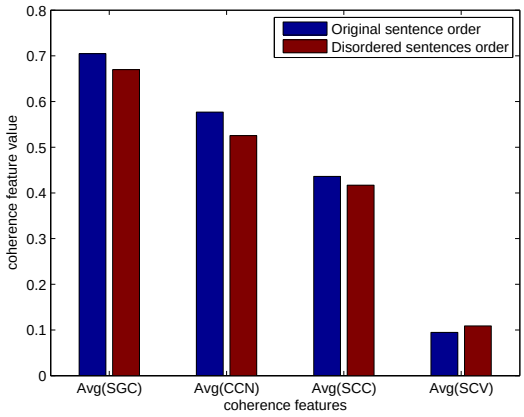


Fig. 10. The difference of coherent features between original sentence order and disordered one

Table 4. combinations of coherent features

Coherence patten	SCV	SCC	CCN	SGC
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
...	...	...	...	...
15	1	1	1	1

0.8. Among feature combinations, combination 14 has the most significant influence on the semantic coherence, which has minimum number of coherence features including coherence variance SCV, semantic coherence clustering SCC and coherent compactness CCN. So we keep only SCV, SCC and CCN as coherent features in the following experiments.

Table 5. Accuracy from different combinations of coherent features

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	10	0.6378	0.0039	0.0012	0.6350	0.6406	0.6296	0.6436
2	10	0.8076	0.0042	0.0013	0.8046	0.8106	0.8004	0.8153
3	10	0.6360	0.0044	0.0014	0.6329	0.6391	0.6304	0.6431
4	10	0.7377	0.0047	0.0015	0.7344	0.7411	0.7276	0.7458
5	10	0.6345	0.0047	0.0015	0.6312	0.6378	0.6284	0.6436
6	10	0.8150	0.0050	0.0016	0.8114	0.8186	0.8071	0.8218
7	10	0.6345	0.0065	0.0020	0.6298	0.6391	0.6224	0.6451
8	10	0.6362	0.0038	0.0012	0.6335	0.6389	0.6307	0.6416
9	10	0.6363	0.0076	0.0024	0.6308	0.6417	0.6256	0.6469
10	10	0.8344	0.0057	0.0018	0.8303	0.8384	0.8273	0.8433
11	10	0.6348	0.0049	0.0015	0.6313	0.6383	0.6289	0.6442
12	10	0.7405	0.0052	0.0016	0.7368	0.7443	0.7333	0.7522
13	10	0.6325	0.0041	0.0013	0.6295	0.6354	0.6269	0.6387
14	10	0.8360	0.0027	0.0009	0.8341	0.8380	0.8324	0.8404
15	10	0.6363	0.0029	0.0009	0.6342	0.6384	0.6302	0.6404
Total	150	0.7027	0.0877	0.0072	0.6885	0.7168	0.6224	0.8453

5.3. *Validating the correctness of the semantic coherent computation*

To verify the association link network based semantic coherence measurement method can precisely measure semantic coherence, we design coherent sentence choosing and coherent sentence order ranking experiments. Besides, we compare our method with other two methods for validating the correctness of our method.

5.3.1. *Coherent sentence choosing experiment*

In this experiment, we randomly select 1000 texts from data set 1 and data set 2 in table 1. For each text, one sentence is removed and the remaining sentences can be regard as a question with one blank. The removed sentence is the only right answer to the question. The only right answer is mixed with other N-1 sentences which are randomly selected from other texts to form N options. Just as the writer of text always chooses semantic coherent sentences as the best choice, a better semantic coherence measurement always gives higher coherence value of the original sentence removed from text. Our coherence measurement method chooses the right answer from N options for the 1000 questions by measuring semantic coherence value as Eq. 15 and choosing coherence sentence with the highest coherence value. The above process is conducted 10 times and then we calculate accuracy by Eq. 22.

$$accuracy = \frac{n(questions \;\; are \;\; answered \;\; correctly)}{n(questions)} \tag{22}$$

Table 6 shows the average accuracy of the coherent sentence choosing experiment. For 2 options, the accuracy is 89.5%. For 5 options, the accuracy is 88.7%. When sentences

increase to 50, our method chooses the semantic coherent sentence with 83.5% accuracy. It confirms that our method can choose semantic coherent sentences with high accuracy. Since the options are randomly selected from other texts, most options are irrelevant sentences with the questions. However, some semantic association sentences have high possibility to involve these options as the options increase. Since sentence options do not only include the right answer but also its association sentences, the average accuracy decreases as the sentence options increase in table 6.

Table 6. Average accuracy in choosing semantic coherent sentence

#Sentence options	Accuracy (%)
2	89.5
5	88.7
10	88.1
20	85.6
50	83.5

5.3.2. *Coherent sentence order ranking experiment*

To further validate correctness of our association link network based sentence semantic coherence measurement method, we conduct coherence sentence order ranking experiment. We randomly select 500 texts as questions from data set 1 and data set 2 in table 1. For each text, its sentences are permuted up to N-1 times. The N-1 permutations are mixed with original sentences set to form N sentence order options. We use our method to measure different sentence orders and rank these options by coherence values in descending order. We make an assumption that the original sentence order of a text is always more semantic coherent than its permutations. A better coherence measurement method can rank the original sentence order higher than its permutations. Our method ranks the N sentence order options and records the rank of original sentence order for the 500 questions. The experiment is conducted 10 times, and then we calculate the proportion of the original sentence set in different rank range as accuracy by Eq. 23.

$$accuracy = \frac{n(questions\ \ ranking\ \ original\ \ sentence\ \ order\ \ in\ \ rank\ \ range)}{n(questions)} \tag{23}$$

Table 7 shows the rank of the original sentence order and its proportion in all the questions including results from data set 1 and data set 2. The result of data set 1 is as follows. For 2 options, the proportion of the original sentence set in rank 1 is 82.5%. For 5 options, the original sentence set in rank 1 with 50.3% and rank 1-2 with 78.0%. For 50 options, the original sentence set is rank 1 with 10.2%, rank 1-2 with 16.6% and rank 1-5 with 30.2%. As the number of the options increases, the rank of the original sentence set keeps 1-5 with higher proportion than its random form. It confirms that our semantic coherence measurement can assign semantic coherent sentence orders with higher accuracy. The result of set 2 is similar with the results of data set 1. It suggests that our method can be used in two different domains and performed higher accuracy. Compared with experiments of coherent sentence choosing in table 6. Table 7 shows lower accuracy since the sentences in the same text have high association relation between each other. If two sentences have the same semantic association with others, then swapping the two sentences will not change semantic coherence.

Table 7. Average proportion of sematic coherent sentence order in different rank range in different data source

Data sources		Data Set 1	Data Set 2
#Sentence order options	Rank range	Accuracy (%)	Accuracy (%)
1 from 2	1	82.5	75.8
1 from 10	1	36.2	30.3
	1-2	54.3	54.7
	1-5	84.4	80.1
1 from 20	1	20.1	11.8
	1-2	33.5	32.3
	1-5	52.3	50.8
	1-10	83.1	81.2
1 from 50	1	10.2	9.5
	1-2	16.6	21.2
	1-5	30.2	30.5
	1-10	53.1	48.2
	1-25	76.5	78.1

5.3.3. *Coherence measurement method comparison experiment*

To evaluate our coherence measurement method, we compare our method with other two widely used coherence measurement methods such as LSA and entity grid based methods. We use these methods to choose coherent sentence and rank the sentence order. Better methods will assign the original sentence and rank the original sentence order with higher coherence value.

In experiment of sentence choosing, we randomly select 500 texts from data set 1 and data set 2 in table 1. For each text, one sentence is removed. The remaining sentences can be regard as a question with one blank. The removed sentence is the only one right answer to the question. The right answer is mixed with other N-1 sentences which are randomly selected from other texts to form N options. Our association link network based coherence measurement method chooses the right answer from N options for the 500 questions. The above process is conducted 10 times and then we calculate accuracy by equation 22. Table 8 shows the average accuracy of choosing a coherent sentence by three methods. Table 8 shows that our method has higher accuracy than other methods in coherent choosing.

Table 8. Average accuracy in choosing semantic coherent sentence in different rank range in different methods

#Sentence options	our method Accuracy (%)	LSA Accuracy (%)	Entity grid based method Accuracy(%)	sig.
2	87.2	73.2	78.5	0.001
5	85.1	75.7	77.5	0.000
10	83.7	69.3	72.1	0.003
20	80.6	61.2	68.6	0.000
50	77.5	57.1	63.5	0.000

In experiment of sentence ordering ranking, we randomly select 500 texts from data set in table 1. For each text, its sentences are permuted up to N-1 times and then the N-1 permutations are mixed with the original sentence order to consist N sentence set options.

We use different ranking methods to rank the N sentence order options for each question. The experiments are conducted 10 times, and then we calculate the average proportion of the

original sentence set in different rank range as table 8. For 2 options, our method has accuracy with 87.2%, LSA with 73.2% and Entity grid method with 78.5%. For 5 options, our method has accuracy with 85.1%, LSA with 75.7%, Entity grid method with 78.5%. Our method has higher accuracy than other methods from 10 to 50 options. From the comparison of the three methods, it suggests that our method has better performance on sentence choosing than entity grid methods and it shows statistical significant ($sig \leq 0.05$).

Table 9. Average proportion of semantic coherent sentence set in different rank range in different methods

Methods		Our method	LSA	Entity grid based method
#Sentence set options	Rank range	Accuracy (%)	Accuracy (%)	Accuracy (%)
1 from 2	1	82.5	67.5	78.7
1 from 5	1	50.3	43.3	51.1
	1-2	78.0	61.2	77.1
1 from 10	1	36.2	31.5	35.1
	1-2	54.3	53.9	55.6
	1-5	84.4	60.3	86.0
1 from 20	1	20.1	15.2	19.9
	1-2	33.5	26.5	30.3
	1-5	52.3	51.6	56.5
	1-10	83.1	61.5	82.6
1 from 50	1	10.2	7.0	10.7
	1-2	16.6	15.1	18.9
	1-5	30.2	30.4	33.5
	1-10	53.1	41.2	51.5
	1-25	76.5	58.0	82.9

Table 9 shows the rank of the original sentence set and its proportion in all the questions among our method, LSA and entity grid based method. For 2 options, our method ranks original sentence set in rank 1 with percentage 82.5%; LSA method with 67.5%; entity grid method with 78.7%. For 5 options, our method ranks the original sentence set in rank 1 with 50.3% and rank 1-2 with 78.0%; LSA method ranks the original sentence set in rank 1 with 43.3% and rank 1-2 with 61.2 %; Entity grid based method ranks the original sentence set in rank 1 with 55.1% and rank 1-2 with 71.2% . For 50 options, our method ranks the original sentence set in rank 1 with 10.2%, rank 1-2 with 16.6% and rank 1-5 with 30.2%; LSA method ranks the original sentence set in rank 1 with 7.0 rank 1-2 with 15.1% and rank 1-5 with 30.4%; Entity grid method ranks the original sentence set in rank 1 with 10.7%, rank 1-2 with 18.9% and rank 1-5 with 33.5%. The results show that our method has similarly performance with entity grid base method and has better performance than LSA method in measuring semantic coherence.

To give a comprehensive comparison of our method with other two methods, we compare the three methods in four prospects: 1) coherent elements; 2) syntactic demand; 3) distinguish ability in choosing coherent sentence; 4) distinguish ability to coherent change caused by ordering sentences.

Table 10 shows the comparison of these methods. Our coherence measurement method mainly uses association link network to dynamically reflect coherence from which some coherent features are extracted and used to measure coherence. LSA calculates the similarity between sentences. Entity grid based method analyzes transitions of syntactic roles of each entity, which assumes that some coherent texts always repeat important nouns. Three methods

Table 10. The comparison of different methods

	Our methods	LSA	Entity Grids based methods
Coherent elements	coherent features	similarity	transition of syntactic roles
Syntactic demand	weak	weak	strong
Semantic association ability	strong	weak	no
Distinguish ability in coherent order	strong	weak	strong
Distinguish ability in coherent sentence	strong	weak	weak

are common in all extracting nouns, but entity grid based method needs additional syntactic parsing to give syntactic role of each keyword per sentence. Our method has strong semantic association ability between keywords by mining association relations. LSA method has weak semantic associate ability by mapping keywords in concept space. Entity grids based method has weak semantic association ability since their keywords are independence. When these methods are used in 1) distinguishing coherent sentence and 2) distinguishing coherent sentence order, our method outperforms other methods in sentence choosing since its strong semantic association ability between sentences and has similar performance on distinguishing coherent sentence order compared with other two methods.

5.4. *Coherence patterns discovery and analysis*

To discover what coherence patterns work well on coherent sentence orders, we randomly select respectively 1000 texts from data set 1 and set 2 in table 1. For each text, its sentences are randomly permuted up to 10 times. We first obtain a pairs of the original sentence order so and its permutations so' and list the coherence patterns as coherence pattern space in table 11 which includes 8 patterns from 0 to 7. We execute algorithm 8 and list the statistics about support and confidence of different coherence patterns as table 12.

Table 11. coherence pattern space

pattern	SCV	SCC	CCN
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

Table 12. coherence patterns with support and confidence

coherence pattern	Coherent support(%)	Coherent confidence (%)
0	0.0418	0.0861
1	0.0532	0.6482
2	0.0292	0.2032
3	0.2314	0.8005
4	0.0577	0.1995
5	0.1146	0.7968
6	0.0289	0.3518
7	0.4432	0.9139

We calculate coherent support and coherent confidence for each coherence pattern and list the results in talbe 12. Table 12 shows that pattern 3, 5, 7 have higher support since their respective $sup \geq 10\%$ and higher coherent confidences since their respective $conf \geq 75\%$.

To further discover the effect of coherence pattern on coherence, we list coherent support and confidence of coherence pairs which consist of coherence pattern and its couple pattern. Table 13 shows some statistic information about support and confidence of these coherence pattern pairs and gives cognitive explanation about them.

Table 13. cognitive explanation of coherence pattern and its couple pattern

pattern pairs	pattern	SCV	SCC	CCN	Coherent confidence	Support
0_7	0	0	0	0	0.0861	0.4850
	7	1	1	1	0.9139	
1_6	1	0	0	1	0.6482	0.0821
	6	1	1	0	0.3518	
2_5	2	0	1	0	0.2032	0.1438
	5	1	0	1	0.7968	
3_4	3	0	1	1	0.8005	0.2891
	4	1	0	0	0.1995	
Cognitive Explain		Is stable transition	Is deeply discussed	Is strongly associated		

In table 13, coherence pattern pairs are distributed as:

Support of pair 0_7 accounts for 48.50 %; support of pair 1_6 account for 8.21%; support of pair 2_5 account for 14.38%; pair 3_4 reaches 28.91%. Obviously, the support of pairs 0_7, 2_5, 3_4 totally account for more than 91.79% in all the patterns.

For pair 0_7, pattern 7 is a typical coherence pattern since its higher confidence($conf = 0.9139$). It suggests coherent sentence orders usually have higher compactness, higher coherence clustering and low semantic variance. Higher compactness means keywords are easier to associate with others; higher coherence clustering denotes that the key points in sentence order have been deeply discussed or have rich context; lower semantic variance denotes all semantics of the sentence set have stable transition. Pattern 0 is a typical pattern for incoherence. For pair 2_5, pattern 5 is a distinguished coherence pattern whose $conf = 0.7968$. Pattern 2 is a distinguished pattern for incoherence. For pair 3_4, pattern 3 is a distinguish coherence pattern whose $conf = 0.8005$. Higher global compactness suggests that writer always make semantic association between keywords when discourse processing. Coherent sentence order can make strong association keywords occur adjacent sentences; higher clustering and higher semantic variance suggest that the sentence set may include many sub-topics and only the minority of topics have been deeply discussed. Pattern 4 is an pattern for semantic incoherence. Loose relations and weak semantic clustering have decided the incoherent semantics.

Although most coherent sentence orders have higher compactness, only higher semantic compactness is not enough for coherence. For example, although pattern 1 has higher compactness, it is not a distinguished pattern for coherence. We should observe other two features. Pattern pair 1_6 has not shown obvious preference for coherence. These pairs only account for support 8.2% in all the pairs.

6. Conclusion

As various novel web social Media appear, a large volume of short messages are transmitted by sentences such as Twitter, Facebook, Microblogs, etc. The massive and unordered short

texts/sentences are rich in semantic information, but hard to express semantic coherent information. To help users obtain semantic coherence, the precondition for developing automatic coherence organization method of short text is to solve the fundamental and practical problem that how to measure semantic coherence under different sentence orders. To solve the above problem, we propose a novel association link network based method to measure semantic coherence. Since human being discourse processing and knowledge storage occur in semantic link network , association link network are sensitive to any semantic coherence changes caused by changing sentence order, such as removing, adding, replacing and permuting sentence order. Our contributions are as follows.

1. To provide the computable elements of semantic coherence, we build association link network and extract its graph features as coherence features since semantic information is stored and processed on semantic association network and semantic association link network is changed by these graph features when the sentence order is changed.

2. To give a precise sentence semantic coherence measurement method, we investigate how the coherence features work on semantic coherence and use the combination of the coherence feature to measure semantic coherence.

3. To discover what's coherence pattern usually occur in coherent sentence order for guiding the further automatic sentence organization, we propose coherence pattern support and confidence for discovering some distinguished and interesting coherence patterns for semantic coherence.

To validate the correctness of our method, we conduct experiments to test whether coherence feathers are sensitive to sentence removal, sentence permutation; which combination of coherence feather is more efficient for measuring semantic coherence; how efficient of our semantic coherence measurement is to choose coherent sentence and rank coherent sentence order. Besides, we have compared our method to other two semantic measurement methods by experiments and theory analysis. The experimental and analysis results exhibits that our method outperforms the other two method and show great potential in some basic applications. We expect such methods can be extended into on-going works on online automatic short text organization, question-answering system and automatic text generation etc.

Acknowledgement

The Research in this paper was supported by the National Science Foundation of China (grant No. 61471232), the Key Innovation Program of Shanghai Municipal Education Commission (grant No.13ZZ064).

References

1. R. Beaugrande, "DE, DRESSLER W.(1981)," Introduction to text linguistics.
2. A. M. Collins and E. F. Loftus (1975), *A spreading-activation theory of semantic processing* Psychological review, vol. 82, pp. 407-428.
3. A. C. Graesser, M. Singer, and T. Trabasso (1994), *Constructing inferences during narrative text comprehension* Psychological review, vol. 101, p. 371.
4. M. Lapata and R. Barzilay(2005), *Automatic evaluation of text coherence: Models and representations* in International Joint Conference On Artificial Intelligence, p. 1085.
5. R. Barzilay and M. Lapata (2008), *Modeling local coherence: An entity-based approach* Computational Linguistics, vol. 34, pp. 1-34.

6. Z. Lin, H. T. Ng, and M.-Y. Kan(2011), *Automatically evaluating text coherence using discourse relations* in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 997-1006.
7. D. Newman, J. H. Lau, K. Grieser, and T. Baldwin(2010), *Automatic evaluation of topic coherence* in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108.
8. A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai(2004), *Coh-Metrix: Analysis of text on cohesion and language* Behavior Research Methods, Instruments, & Computers, vol. 36, pp. 193-202.
9. T. Nahnsen(2009), *Domain-independent shallow sentence ordering*in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium, 2009, pp. 78-83.
10. D. Marcu(1997), *From local to global coherence: A bottom-up approach to text planning* in Proceedings of the National Conference on Artificial Intelligence, pp. 629-636.
11. R. Zhang(2011), *Sentence ordering driven by local and global coherence for summary generation* in Proceedings of the ACL, pp. 6-11.
12. M. Lapata(2003), *Probabilistic text structuring: Experiments with sentence ordering* in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pp. 545-552.
13. R. Barzilay and L. Lee(2004), *Catching the drift: Probabilistic content models, with applications to generation and summarization* in Proceedings of HLT-NAACL.
14. R. Barzilay(2003), *Information fusion for multidocument summarization: paraphrasing and generation* Columbia University.
15. R. Soricut and D. Marcu(2006), *Discourse generation using utility-trained coherence models* in Proceedings of the COLING/ACL on Main conference poster sessions, pp. 803-810.
16. M. Elsner, J. Austerweil, and E. Charniak(2007), *A unified local and global model for discourse coherence* in Proceedings of NAACL/HLT.
17. N. Fang, X. Luo, and W. Xu(2009), *Measuring textual context based on cognitive principles* International Journal of Software Science and Computational Intelligence (IJSSCI), vol. 1, pp. 61-89.
18. J.L. Austerweil, J.T. Abbott and T.L. Griffiths(2012), *Human memory search as a random walk in a semantic network*//Advances in neural information processing systems.: 3041-3049.
19. T. Joachims(2002), *Optimizing search engines using clickthrough data.* In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133-142.
20. F. C. T. Chua and S. Asur (2013), *Automatic summarization of events from social media*, in: ICWSM.
21. L. Shou, Z. Wang, K. Chen and G. Chen(2013), *Sumblr: continuous summarization 490 of evolving tweet streams*, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 533-542.
22. J. Leskovec, L. Backstrom and J. Kleinberg(2009), *Meme-tracking and the dynamics of the news cycle* in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 495 ACM, pp. 497-506.
23. J. R. Quinlan(1986), *Induction of Decision Trees.* Mach. Learn. 1, 1, 81-106.
24. X. Luo, J. Xuan and H. Liu(2014), it Web event state prediction model: combining prior knowledge with real time data. Journal of Web Engineering 13.5-6. 483-506.