The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

HAN: Hierarchical Association Network for Computing Semantic Relatedness

Xiaolong Gong, Hao Xu, Linpeng Huang
Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China
{gxl121438, insanelun, lphuang}@sjtu.edu.cn

Abstract

Measuring semantic relatedness between two words is a significant problem in many areas such as natural language processing. Existing approaches to the semantic relatedness problem mainly adopt the co-occurrence principle and regard two words as highly related if they appear in the same sentence frequently. However, such solutions suffer from low coverage and low precision because i) the two highly related words may not appear close to each other in the sentences, e.g., the synonyms; and ii) the co-occurrence of words may happen by chance rather than implying the closeness in their semantics. In this paper, we explore the latent semantics (i.e., concepts) of the words to identify highly related word pairs. We propose a hierarchical association network to specify the complex relationships among the words and the concepts, and quantify each relationship with appropriate measurements. Extensive experiments are conducted on real datasets and the results show that our proposed method improves correlation precision compared with the state-of-the-art approaches.

1 Introduction

Measuring semantic relatedness between words is a fundamental problem in the areas of natural language processing, artificial intelligence and information retrieval. For example, document summarization and question answering systems leverage semantic relatedness scores to align sentences (Mogren 2015; Wen-tau et al. 2013); Information retrieval systems perform query expansion (Budanitsky and Hirst 2006) based on word relatedness scores; Informally, semantic relatedness reflects a *free association process*[1] by human brains. That is, when mentioning a cue word, the first few words coming into most people's mind exhibit high relatedness to the cue word. For instance, when a cue word "tea" is mentioned, the words that are highly related to "tea" could be "cup", "drink", "lemon", "leaf", etc.

Various solutions have been proposed to measure semantic relatedness between words. Most of them (Dagan, Lee, and Pereira 1999; Miller and Charles 1991; Keyang, Kenny, and Seung-won 2015) adopt the principle of co-occurrence. That is, the two words are regarded as semantically related if they co-occur in many sentences. However, such methods suffer from low precision and low coverage because i) the words that appear in the same sentence may not necessarily be closely related in their semantics but co-occur by chance and ii) it is very likely that two highly related words appear far away from each other, e.g., two synonyms seldomly appear in one sentence.

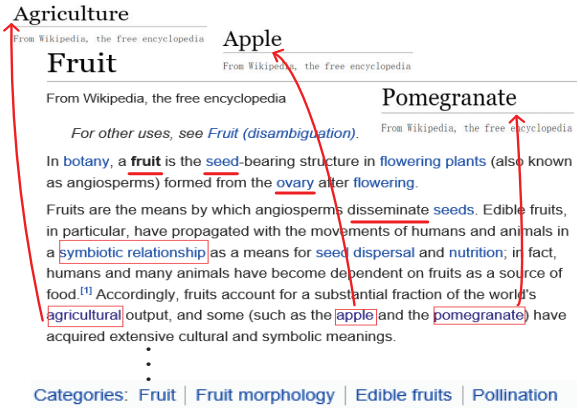

Figure 1: A sample Wikipedia page

To address the problems, several approaches (Agirre et al. 2010; Rada et al. 1989; Resnik 1995; Peipei, Haixun, and Kenny 2013) leverage linguistic resources such as WordNet (Miller 1995) and Roget's Thesaurus (Roget 1852) and identify the belonging categories of the words following the *isA* relationship. These categories are also referred to as the *concepts* of the words. The semantic relatedness between two words is then measured by the closeness of their concepts. Some approaches exploit concepts of every occurrence of the word as a weighted vector (Gabrilovich and S.Markovitch 2007), or model the semantic meaning of a word by the salient concepts that frequently co-occur in the immediate context of the word (Hassan and Mihalcea 2011). While these approaches explore latent semantics for the words, they suffer from some drawbacks such as word ambiguity and only focus on identifying concepts for the noun. This leads to the low coverage problem due to the fact that many semantically related words are adjectives and verbs, e.g., "blue" and "unhappy", "pardon" and "sorry".

In this paper, we leverage Wikipedia pages as our data

[1] http://web.usf.edu/FreeAssociation

Symbol	Description
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	a graph with vertices $\mathcal{V}$ and edges $\mathcal{E}$
n	size of vocabulary
m	the total number of concepts
$\mathcal{W}$	a set of n words
$\mathcal{C}$	a set of m concepts
w_i	a word in $\mathcal{W}$
c_i	a concept in $\mathcal{C}$
$\mathcal{P}$	set of wiki pages
$\mathcal{E}_{\{w,wc,c\}}$	three types of edge sets
$\mathcal{F}_{\{w,wc,c\}}$	relatedness measurement

Table 1: Symbols and Their Meanings

sources and aim to construct a comprehensive set of word pairs that are highly related in semantics. Intuitively, the title of a wiki page presents a concept and a hyperlink towards this page contains words with the corresponding concept. The words associated with the hyperlink is referred to as an anchor. Figure 1 shows an example of anchor link where word "apple" to its concept "Apple" via the hyperlink. Hence we can obtain a complete concept set and a high-quality word-to-concept mapping in a natural way. To guarantee the precision and coverage of word semantic relatedness results, we propose to develop a hierarchical association network, named HAN, to capture three kinds of relationships among words and concepts, namely word-to-word, word-to-concept and concept-to-concept relationships.

We summarize the contributions of this paper as follows.

- We propose a *hierarchical association network* (HAN) to capture the complex relationships among words and concepts, and compute word relatedness by considering all the relationships into account.
- We introduce various score functions to quantify the word-to-word, word-to-concept and concept-to-concept relationships and provide a novel word relatedness score function which encapsulates the closeness of the words as well as their latent concepts.
- We conduct extensive experiments on a real dataset to evaluate the effectiveness of our proposed solution using HAN. The results show that our solution achieves improvement in correlation precision, compared with other state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 provides model definition and introduces the hierarchical association network; Section 3 provides various semantic relatedness measurements and Section 4 defines the final relatedness score function. Section 5 demonstrates the experimental results. Finally we conclude in Section 6.

2 Hierarchical Association Network

Definition

We consider a set $\mathcal{P} = \{p_1, \cdots, p_k\}$ of web pages from the Wikipedia website[2]. We refer to the title of p_i as a *concept* denoted by c_i, all categories of p_i denoted by T_i. And the text of a hyperlink towards page p_i as an *anchor* for the concept c_i of p_i. Further, we denote by $\mathcal{C}$ the set of concepts involved in $\mathcal{P}$, i.e., $\mathcal{C} = \bigcup_{p_i \in \mathcal{P}} c_i$, and denote by $\mathcal{W}$ the vocabulary for $\mathcal{P}$. Table 2 lists the symbols and their meanings used throughout this paper.

[2]https://en.wikipedia.org/wiki/Main_Page

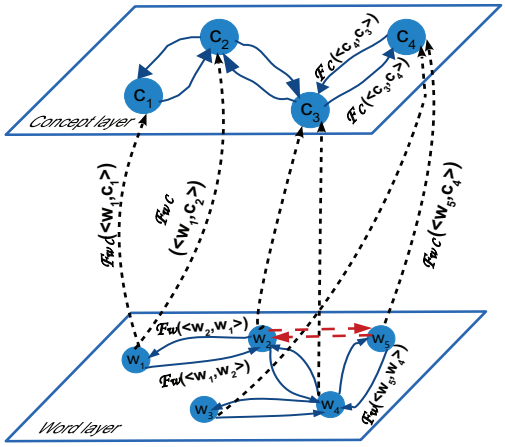

Figure 2: Hierarchical Association Network (HAN)

Figure 1 provides an example wiki page including concept $c =$ "Fruit". This concept belongs to several categories $T =$ {"Fruit", "Fruit morphology", "Edible fruits Pollination"}. We highlight three anchors "agricultural", "apple" and "pomegranate" for concepts "Agriculture", "Apple" and "Pomegranate" following the hyperlinks, respectively. Intuitively, two words are very likely to be semantically related if they refer to two concepts that are closely related. Following our intuition, first we define $\mathcal{F} : \mathcal{W} \times \mathcal{W} \to [0, 1]$ as final relatedness score function, and then we decompose $\mathcal{F}$ into two parts: word-level relatedness $\mathcal{F}_w$ and concept-level relatedness $\mathcal{F}_c$. We use $\mathcal{F}_c$ to enhance the relationships for truly related words and penalize the words that are co-recurrent by chance, thus improving the coverage and precision.

Network Construction

We introduce a hierarchical association network (HAN) to represent the words in $\mathcal{W}$, the concepts in $\mathcal{C}$ and their relationships. There is a directed edge from word w to concept c (i.e., $\langle w, c \rangle \in \mathcal{E}_{wc}$) *iff* w is an anchor for c. However, we notice that many anchors contain multiple words. For example, "symbiotic relationship" in Figure 1 is an anchor for concept "Symbiosis". HAN also includes a directed edge from those multiple words to the concept vertex c if the corresponding multiple word is an anchor for c. Note that we leverage those relevant concept vertices to identify concept relatedness and we do not include those multiple words in word layer.

Definition 1 (Hierarchical Association Network) *Consider a set $\mathcal{P}$ of wiki pages with the concept set $\mathcal{C}$ and vocabulary set $\mathcal{W}$. Let $\mathcal{D}$ denote the all the anchors. The* hierarchical association network *is a weighted directed graph $\mathcal{G}(\mathcal{P}, \mathcal{C}, \mathcal{W}, \mathcal{D}) = (\mathcal{V}, \mathcal{E}, \mathcal{F}_c, \mathcal{F}_{wc}, \mathcal{F}_w)$ where the vertex set contains all the words and concepts, i.e.,*

$\mathcal{V} = \mathcal{C} \cup \mathcal{W} \cup \mathcal{D}$, *and the edge set* $\mathcal{E}$ *includes three categories* $\mathcal{E}_c, \mathcal{E}_w, \mathcal{E}_{wc}$ *of edges defined as follows.*
(1) $\mathcal{E}_c = \{\langle c_i, c_j \rangle \mid c_i, c_j \in \mathcal{C}\}$;
(2) $\mathcal{E}_w = \{\langle w_i, w_j \rangle \mid w_i, w_j \in \mathcal{W} \cup \mathcal{D}\}$;
(3) $\mathcal{E}_{wc} = \{\langle w, c \rangle \mid w \in \mathcal{W} \cup \mathcal{D} \wedge c \in \mathcal{C} \wedge$ w is anchor for c $\}$.

We define $\mathcal{E} = \mathcal{E}_c \cup \mathcal{E}_{wc} \cup \mathcal{E}_w$. Every edge in $\mathcal{E}$ is associated with a weight, which indicates the strength of the relation. We denote by $\mathcal{F}_c : \mathcal{E}_c \to [0, 1]$, $\mathcal{F}_{wc} : \mathcal{E}_{wc} \to [0, 1]$ and $\mathcal{F}_w : \mathcal{E}_w \to [0, 1]$ the weighting functions for the edges in $\mathcal{E}_c$, $\mathcal{E}_{wc}$ and $\mathcal{E}_w$, respectively.

Figure 2 shows an example of hierarchical association network $\mathcal{G}$. $\mathcal{G}$ conceptually organizes 9 vertices using two layers. The concept layer has 4 concept vertices, i.e., $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$ and the word layer has 5 vertices, i.e., $\mathcal{W} = \{w_1, w_2, w_3, w_4, w_5\}$. Every pair of distinct concept (resp., word) vertices should be connected by two directed edges, but we omit the edges with zero weight such as $\langle c_1, c_4 \rangle$ and $\langle w_1, w_3 \rangle$ for simplicity. Every word in $\mathcal{W}$ is an anchor for certain concept, e.g., w_1 is an anchor for two concepts c_1, c_2.

3 Semantic Relatedness Computation

Word-level Relatedness $\mathcal{F}_w$

In this paper, we propose two strategies for word-level relatedness: a) word2vec (Mikolov et al. 2013) & GloVe (Pennington, Socher, and Christopher 2014). b) word co-occurrence. Word representation vector can be utilized to calculate their relatedness based on cosine function. Furthermore, we consider two kinds of proximity closeness in the word co-occurrence strategy: (1) within the same sentence and (2) within a fixed window size K_1. Figure 1 illustrates the idea of the closeness. The underlying words "fruit" and "seed" appear in the same sentence; "ovary" and "disseminate" co-occur within a window size of 30 while they are not included in one sentence. Let $f(\cdot, \cdot)$ denote the co-occurrence frequency for any two words. Formally, for any word-to-word edge $\langle w_i, w_i \rangle \in \mathcal{E}_w$, we have:

$$\mathcal{F}_w(\langle w_i, w_j \rangle) = \frac{f(w_i, w_j)}{\sum_{k=1}^{n} f(w_i, w_k)} \quad (1)$$

where n is the size of the vocabulary, i.e., $n = |\mathcal{W}|$. We compute f in two ways, namely sentence-sized and fixed window-sized, and compare their effectiveness experimentally.

Word-to-concept Relatedness $\mathcal{F}_{wc}$

Recall that there exists a directed edge from word w to concept c in HAN if w is an anchor for c. It is important to note that the relationship between words and concepts is a one-to-many mapping. That is, a word can be an anchor for different concepts due to the word ambiguity. Therefore, to compute $\mathcal{F}_{wc}$, our main insight is that w and c are closely related if c is the only semantic meaning for word w. To capture our intuition, let $l(w, c)$ be the total number of times when w is an anchor of c given all the Wikipedia pages $\mathcal{P}$. We formally define P_{link} as follows.

$$P_{link}(w, c) = \frac{l(w, c)}{\sum_{c' \in \mathcal{C}} l(w, c')} \quad (2)$$

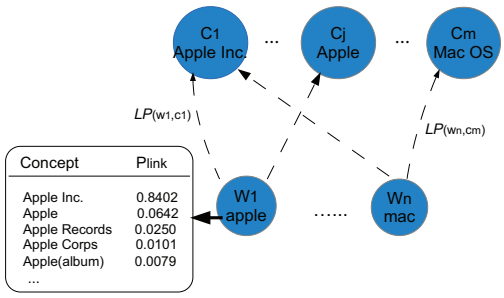

Figure 3: Mapping from anchors to concepts

Intuitively, P_{link} indicates whether concept c is the major semantic meaning for w. Here, we observe a special case in Figure 3, the score $P_{link}(w, c)$ indicates that the relatedness of word "apple" and the concept "Apple" is very low. There are two primary reasons for this situation. First is the "semantic drift", where many texts are beginning to mention "apple" as concept "Apple Inc.". The second is that some words "apple" in document are not marked as an anchor, which means $l(w, c)$ is insufficient in dataset. Therefore, we have to consider all contextual words surround w, not just the only anchor w. We propose *link popularity* (LP) that reflects the degree of strong connections between anchor words and their link concepts which defined as follows.

$$LP(w, c) = \sum_{\mathcal{P}} \sum_{w \in S} \frac{\sum_{w' \in S} tf\_idf(w', c)}{\sum_{c' \in c(w)} \sum_{w' \in S} tf\_idf(w', c')} \quad (3)$$

where $\mathcal{P}$ denote a wiki page, S represents one sentence in a wiki page which contain the word w, and w' means every contextual word in S. $c(w)$ is a set of concepts which are linked from anchor word w.

Finally, for any directed edge in the word-to-concept edge set $\mathcal{E}_{wc}$, our $\mathcal{F}_{wc}$ is defined as follows.

$$\mathcal{F}_{wc}(\langle w, c \rangle) = \frac{LP(w, c)}{\sum_{c' \in \mathcal{C}} LP(w, c')} \quad (4)$$

Concept-level Relatedness $\mathcal{F}_c$

We propose three score functions, namely the co-occurrence based score function M_1, the category based score function M_2 and the link-detection based score function M_3, respectively. The concept-level relatedness $\mathcal{F}_c$ is then computed by combining three score functions linearly. Formally, for any edge $\langle c_i, c_j \rangle$ in $\mathcal{E}_c$, we have:

$$\mathcal{F}_c(\langle c_i, c_j \rangle) = \alpha_1 M_1(c_i, c_j) + \alpha_2 M_2(c_i, c_j) + \alpha_3 M_3(c_i, c_j) \quad (5)$$

We also require $\alpha_1 + \alpha_2 + \alpha_3 = 1$ to guarantee that the concept-level relatedness values are normalized, i.e., in the range of $[0, 1]$.

Co-occurrence based Score The co-occurrence based score function M_1 considers two kinds of co-occurrences: (1) global co-occurrence: the number of wiki pages where two concepts co-occur; (2) local co-occurrence: co-occurrence frequencies of two concepts that appear closely

in proximity. We adopt the normalized pointwise mutual information (NPMI) (Bouma 2009) to measure both the global and the local associations of the concepts.

Global co-occurrence $NPMI_g$. Consider a set $\mathcal{P}$ of wiki pages and two concepts $c_i, c_j \in \mathcal{C}$. Let $df(c_i, c_j)$ denote the number of wiki pages in which c_i, c_j co-occur, and $df(c_i)$ denote the number of wiki pages which include c_i. The pointwise mutual information for the global associations between concepts is the following.

$$PMI_g(c_i, c_j) = \log(\frac{df(c_i, c_j) \times |\mathcal{P}|}{df(c_i) \times df(c_j)})$$

The global co-occurrence score $NPMI_g$ for two concepts c_i, c_j is the normalized PMI_g value.

$$NPMI_g(c_i, c_j) = \frac{PMI_g(c_i, c_j)}{-\log(df(c_i, c_j) \times |\mathcal{P}|)} \quad (6)$$

Local co-occurrence $NPMI_l$. Let $f(c_i, p)$ denote the occurrence frequencies of concept c_i in wiki page p, and $Co(c_i, c_j, p)$ denote the number of times that the two concepts appear within a fixed window size (K_2) or in the same sentence in wiki page p. Similar to the global co-occurrence, we compute PMI for the local associations between concepts as follows.

$$PMI_l(c_i, c_j) = \log \frac{\sum_{p \in \mathcal{P}} Co(c_i, c_j, p) \times |\mathcal{P}|}{\sum_{p \in \mathcal{P}} f(c_i, p) \times \sum_{p \in \mathcal{P}} f(c_j, p)}$$

The local co-occurrence $NPMI_l$ is the normalized PMI_l value:

$$NPMI_l(c_i, c_j) = \frac{PMI_l(c_i, c_j)}{-\log(\sum_{p \in \mathcal{P}} Co(c_i, c_j, p) \times |\mathcal{P}|)} \quad (7)$$

We define our co-occurrence based score function M_1 as a mixed normalized PMI value that combines $NPMI_g$ and $NPMI_l$. Specifically, for any two concepts $c_i, c_j \in \mathcal{C}$, we have:

$$M_1 = \max\{0, (1-\beta)NPMI_g(c_i, c_j) + \beta NPMI_l(c_i, c_j)\} \quad (8)$$

where $\beta \in [0, 1]$. Note that both $NPMI_g$ and $NPMI_l$ are within the range of $[-1, 1]$. Since we are only interested in related concept pairs, we retain the non-negative mixed normalized PMI values in M_1. Intuitively, a large value of β favors local co-occurrence and vice versa. We evaluate the effect of different values of β in Section 5.

Category based Score In our implementation, we clean the graph by removing useless categories (i.e., name of category contains "page", "error", "redirects"). Specifically, for every concept pair c_i and c_j, we compute Jaccard coefficient as its category based score. Let $cat(c_i)$ denote a set of categories to which the concept c_i belongs. For any two concepts $c_i, c_j \in \mathcal{C}$, we define the category based score M_2 as follows.

$$M_2(c_i, c_j) = \frac{|cat(c_i) \cap cat(c_j)|}{|cat(c_i) \cup cat(c_j)| - |cat(c_i) \cap cat(c_j)| + 1} \quad (9)$$

Link-structure based Score In addition to the co-occurrence, two concepts can also be related via hyperlinks. The rationale behind is that some common concepts are used to explain c_i and c_j; hence, c_i, c_j are related. We leverage a popular relatedness measure *Normalized Wikipedia Distance* (NWD) suggested by Milne and Witten (Milne and Witten 2008) to compute our link-structured based score function. Specifically, let I_c be the set of outgoing links from concept c. For any two concepts $c_i, c_j \in \mathcal{C}$, $M_3(c_i, c_j)$ is formally defined as follows.

$$M_3(c_i, c_j) = 1 - \frac{\log_2(\max\{|I_{c_i}|, |I_{c_j}|\} - \log_2 |I_{c_i} \cap I_{c_j}|}{\log_2 |\mathcal{P}| - \log_2 \min\{|I_{c_i}|, |I_{c_j}|\}} \quad (10)$$

4 Deriving Word relatedness $\mathcal{F}$

Finally, our word relatedness function $\mathcal{F}$ consists of two parts: (1) $\mathcal{F}_w$ measures the word-level relatedness and (2) $\textsc{Crel}(w_i, w_j)$ captures the semantic relatedness of the word pairs with respect to their relevant concepts. Formally, $\textsc{Crel}(w_i, w_j)$ is defined as follows.

$$\sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}} \mathcal{F}_{wc}(\langle w_i, c_i \rangle) \mathcal{F}_c(\langle c_i, c_j \rangle) \mathcal{F}_{wc}(\langle w_j, c_j \rangle) \quad (11)$$

We use $\lambda \in [0, 1]$ trades off the importance of word-level relatedness $\mathcal{F}_w$ against that of word relatedness w.r.t the relevant concepts $\textsc{Crel}(w_i, w_j)$. For any two words $w_i, w_j \in \mathcal{W}$, we have:

$$\mathcal{F}(\langle w_i, w_j \rangle) = \lambda \mathcal{F}_w(\langle w_i, w_j \rangle) + (1 - \lambda) \textsc{Crel}(w_i, w_j) \quad (12)$$

Training procedure

Let $\Theta = (\alpha_1, \alpha_2, \alpha_3)$ denote the set of three unknown parameters in our model, and we compare different values for λ, β, K_1, K_2 experimentally. For simplicity, we only use $\mathcal{F}_{wc}, \mathcal{F}_c$ and Equation 5 to represent the relevance items in Equation 12, which can be written into the following format.

$$\begin{aligned} \textsc{Crel} &= \sum_{c_i} \sum_{c_j} \mathcal{F}_{wc} \mathcal{F}_c \mathcal{F}_{wc} \\ &= \sum_{c_i} \sum_{c_j} \mathcal{F}_{wc} (\alpha_1 M_1 + \alpha_2 M_2 + \alpha_3 M_3) \mathcal{F}_{wc} \\ &= \alpha_1 {\mathcal{F}_w}^1 + \alpha_2 {\mathcal{F}_w}^2 + \alpha_3 {\mathcal{F}_w}^3 \end{aligned}$$

where ${\mathcal{F}_w}^1$, ${\mathcal{F}_w}^2$ and ${\mathcal{F}_w}^3$ represent $\sum_{c_i} \sum_{c_j} \mathcal{F}_{wc_i} M_1 \mathcal{F}_{wc_j}$, $\sum_{c_i} \sum_{c_j} \mathcal{F}_{wc_i} M_2 \mathcal{F}_{wc_j}$ and $\sum_{c_i} \sum_{c_j} \mathcal{F}_{wc_i} M_3 \mathcal{F}_{wc_j}$, respectively.

It is easy to see that, we integrate the four types of features into a single strength score $\mathcal{F}$. The objective is to learn the model $\Theta = (\alpha_1, \alpha_2, \alpha_3)$. Specifically, we choose linear regression algorithm to learn unknown parameter Θ on the training set Florida Norms (denoted by $\mathcal{FN}(\langle w_i, w_j \rangle)$). This training set is generated by a well-studied psychological process called *free association*. Table 2 shows a fragment of the free association norms collected by University of South Florida (Nelson, McEvoy, and Schreiber 2004).

Cue	Target	Forward	Backward
basket	weave	30/143	9/138
basket	ball	19/143	--
basket	fruit	12/143	2/184
basket	picnic	5/143	22/143

Table 2: Relatedness of Cue-to-Target pairs about a cue word *basket*

The first column in data file presents the normed words or cues, and the second field presents their responses or Targets. The cues and their targets are presented as pairs, the 3th & 4th field are called Forward/Backward strength respectively what has sometimes been called cue-to-target strength (a fraction that the number of participants who responded with this pair in experiment, '- -' means without this pair when "basket" as a cue word. The general idea is to find an optimal Θ that can minimize the expected loss.

$$\Theta = argmin \sum_{w_i} \sum_{w_j} \mathcal{L}(\mathcal{F}(\langle w_i, w_j \rangle), \mathcal{FN}(\langle w_i, w_j \rangle)) \tag{13}$$

5 Experiments

Datasets

Wikipedia & Florida norms In this paper, we constructed two hierarchical association networks HAN_{wiki} and HAN_{free}, HAN_{wiki} is based on the Wikipedia dump on October 2, 2015.[3] After parsing the Wikipedia XML dump, we obtained 15.5GB hypertexts with 4,950,533 articles. Not all of these articles are useful to generate our vocabulary and features. Several preprocessing steps are conducted as follows:

1. Tokenization, stop words removal. Lemmatization of tokens, which makes each token turn into its morphological stem. Remove those words that is contained in less than 5 articles or occur less than 50 times in all the wiki pages.
2. Discard pages that have less than 150 nonstop words, and discard navigation pages without introducing any concept.

After these steps, we collected 2,231,468 articles, 158,071, 728 sentences and 793,486 words. In our experiment, we only took 30,000 most frequent words as our dictionary. The vertices of HAN_{free} are based on the Florida norms(see Table 2). The original Florida free association norms data contains 5019 cue words and a total of 72,176 cue-target pairs. Some target words are also cue words, so we eventually got 63,619 pairs as directed edges. We used those cue words to construct our small hierarchical association networks.

Conceptual test set: ConceptRel-250 In our work, we proposed to use three measurements M_1, M_2, M_3 to evaluate the relatedness between two concepts. Especially in M_1, we want to determine parameter K_2 and β in order to achieve the highest accuracy among concepts. So we need a conceptual test set to take care of this problem. However, many previous datasets are built by human judgments on the similarity only for pair of words other than concepts, such as WordSim-353, MC and RG. In this paper, we constructed the ConceptRel-250, a new conceptual test set to address the above situation by evaluating correlation on ConceptRel-250. Our procedure of constructing the dataset consists of two steps:

1. Firstly, we extracted a set of the concept pairs that they co-occur in wiki pages. Then, we randomly drew 250 concept pairs from this set to construct our conceptRel-250.
2. Concept relatedness scores were evaluated by volunteers, with an average of 10 ratings for each concept pair. Ratings were collected on a 1-5 scale, where 5 stands for "highly related" and 1 stands for "not related".

[3] You can download here: https://dumps.wikimedia.org/

	Florida norms	WS353
sentence	0.380	0.462
$K_1 = 10$	0.395	0.446
$K_1 = 20$	0.405	0.457
$K_1 = 30$	**0.408**	**0.458**
$K_1 = 40$	0.407	0.454

Table 3: Impact of context size: Correlation test in two datasets for various values of K_1 in traditional relatedness on words (measured by μ)

Evaluation procedure

We evaluated our hierarchy association networks for relatedness measures on three standard datasets:

- Miller & Charles (1991) list of 30 noun pairs, using a scale from 0 to 4. (MC)
- Rubenstein & Goodenough (1965) is a 65 word synonymity list also scoring from 0 to 4. (RG)
- WordSimilarity-353 Test Collection (Finkelstein et al. 2002), with 353 word pair annotated on a scale from 0 to 10. (WS353)

We adopted two important measures, Pearson correlation metric γ and Spearman correlation metric ρ, to evaluate the semantic relatedness results. In our experiments, we also followed Hassan and Mihalcea (2011) by computing the harmonic mean of Pearson and Spearman metrics $\mu = \frac{2\gamma\rho}{\gamma+\rho}$.

Baselines

To evaluate the performance of the proposed method, we compared our method to the six methods that showed good results, namely, LSA (Deerwester et al. 1990), ESA (Gabrilovich and S.Markovitch 2007), SSA (Hassan and Mihalcea 2011), W2V[4] (Mikolov et al. 2013), GloVe[5] (Pennington, Socher, and Christopher 2014) and SaSA (Zhaohui and Giles 2015). We trained W2V& GloVe

[4] https://radimrehurek.com/gensim/models/word2vec.html

[5] https://nlp.stanford.edu/projects/glove/c

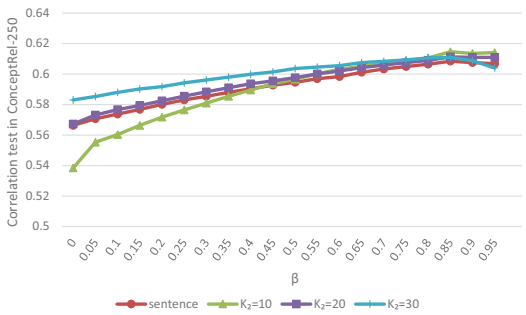

Figure 4: Impact of context size with different β in co-occurrence based score function (Correlation test in *ConceptRel-250* dataset)

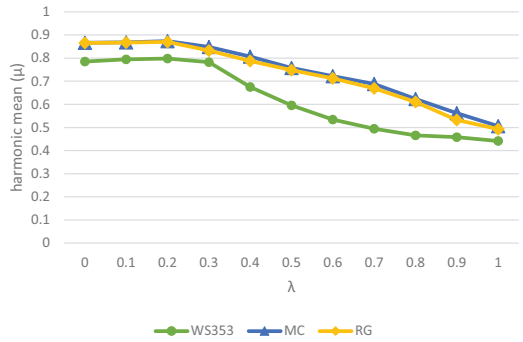

Figure 5: Performance with various value of λ

in wiki set and we selected 100 as an vector dimensionality. But the primary comparison is a recent algorithm AN_{wiki} (Keyang, Kenny, and Seung-won 2015).

In order to judge a well semantic relatedness for a pair of words, AN_{wiki} proposed five types of co-occurrences extracted from the rich structures of Wikipedia so as to determine the edge set in association network and the weight of each edge. Specifically, 1) sentence level co-occurrences(slc), 2) title link co-occurrences(tlc), 3) title gloss co-occurrences(tgc), 4) title body co-occurrences(tbc), 5) category level co-occurrences(clc). All these co-occurrence types are measurements in single level, which means concepts in this algorithm are treated as a common word.

Experimental results

Parameters tuning We now evaluate the performance of HAN performance by varying the following parameters:

• The parameter K_1 and K_2 controls the context size in word co-occurrence $\mathcal{F}_w$ and concept relatedness $\mathcal{F}_c$ respectively. (Section 3)

• The parameter β trades off the importance of two related items. (Section 3)

• The parameter λ adjusts the contribution between $\mathcal{F}_w$ and $\textsc{Crel}(w_i, w_j)$.

Table 3 analyzes the effect of different context sizes in word co-occurrence relatedness ($\mathcal{F}_w$). We see a slight variation of the harmonic mean correlation as K_1 changes, with the best results around $K_1 = 30$. Here we let Florida norms as test set, and we found that the results on the Florida norms also show the best results with parameter $K_1 = 30$.

Metric	**WS227**	**WS353**
AN^0_{free}	0.645	0.476
AN^+_{free}	0.752	0.512
HAN^a_{free}	**0.781**	**0.586**

Table 4: Spearman correlation(ρ) on WS353 dataset

Figure 4 shows the correlation between our score M_1 and test set ConceptRel-250 with different contextual size and variation of β. We found the optimal correlation is obtained when relatedness of concept pairs are taken with the fixed-size window $K_2 = 10$ and $\beta = 0.85$.

Figure 5 shows the results w.r.t the different λ on HAN_{wiki}. λ actually adjusts the contribution of $\mathcal{F}_w$ against the contribution of $\textsc{Crel}(w_i, w_j)$. The optimal result is obtained when $\lambda = 0.2$, which means our proposed $\textsc{Crel}(w_i, w_j)$ has a better contribution to final result $\mathcal{F}$ than $\mathcal{F}_w$. However, the correlation result with $\lambda = 0$ is not am optimal value which means that $\mathcal{F}_w$ has certain contribution to final effectiveness.

Comparison results First of all, we illustrate that our HAN based on word embedding strategies is more effective in finding strong relatedness between words. We compared the performance of HAN^a_{free} with AN_{free} on WS-353 dataset. We also made comparison on WS227, a subset of WS353 in which all words belong to some vertices in HAN^a_{free}. Our HAN^a_{free} performed better than AN^0_{free} and AN^+_{free} (Keyang, Kenny, and Seung-won 2015) on WS353 and WS227 measured by Spearman correlation (ρ) (see Table 4), which indicates that our hierarchical association network can be useful in computing relatedness between words after aggregates the conceptual information. The results show that conceptual information can strengthen the relationship in Florida norms and illustrate its usefulness in computing semantic relatedness. We also observed that HAN_{free} has better performance on WS227 than WS353, which is a common case when applying AN^0_{free} and AN^+_{free}. This means that degradation in performance is mainly due to the limited vocabulary.

Considering the limitation in vocabulary of HAN_{free}, we created HAN_{wiki}, which has a large lexical coverage. HAN^b_{wiki} based on co-occurrence strategy and HAN^a_{wiki} based on word embedding strategy. The results show that HAN^a_{wiki} has a better performance than other algorithms on three standard datasets. Note that HAN_{wiki} and $SaSA$ outperformed other co-occurrence based algorithm in MC&RG, which indicates the necessity to consider semantic relatedness between concepts behind words. Furthermore, our HAN^a_{wiki} combining both free association and concepts relatedness achieved a higher precision with μ improving by 8.4% compared to $SaSA$, and by 3% compared to AN_{wiki}.

In order to demonstrate the usefulness of the co-

	γ			ρ			μ		
Metric	**MC**	**RG**	**WS353**	**MC**	**RG**	**WS353**	**MC**	**RG**	**WS353**
LSA	0.725	0.644	0.563	0.662	0.609	0.581	0.692	0.626	0.572
ESA	0.588	--	0.503	0.727	--	0.748	0.650	--	0.602
SSA_S	0.871	0.847	0.622	0.810	0.830	0.629	0.839	0.838	0.626
SSA_C	0.879	0.861	0.590	0.843	0.833	0.604	0.861	0.847	0.597
$W2V$	0.852	0.834	0.633	0.836	0.812	0.645	0.844	0.823	0.639
$GloVe$	0.837	0.828	0.603	0.809	0.781	0.620	0.823	0.804	0.611
$SaSA_t$	0.883	0.870	0.721	0.849	0.841	0.733	0.866	0.855	0.727
$SaSA$	0.886	0.882	0.733	0.855	0.851	0.739	0.870	0.866	0.736
AN_{wiki}	0.865	0.858	0.740	0.848	0.843	0.813	0.856	0.850	0.775
HAN^b_{wiki}	0.869	0.861	0.744	0.851	0.849	0.814	0.860	0.855	0.778
HAN^a_{wiki}	0.886	0.884	0.772	0.860	0.857	0.826	0.873	0.870	0.798

Table 5: Pearson(γ) , Spearman(ρ) and harmonic mean(μ) on the word relatedness datasets

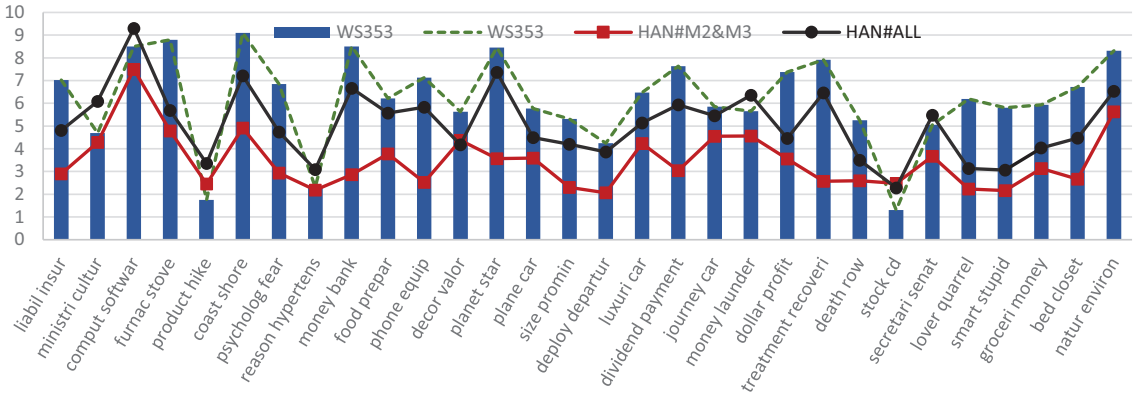

Figure 6: Correlation with human ratings using different methods

occurrence based score function (M_1), we conducted experiment with both category based score function and link-structure based score function in HAN (denote by $HAN\#M2\&M3$), and compared it with standard HAN with Equation 5 ($HAN\#All$) in WS353 dataset. Figure 6 shows a chart with the relatedness value obtained by the previous human ratings and the values obtained by HAN. The output values obtained by HAN in the range [0, 1] were mapped into [0, 10] for convenient comparison. For an intuitive understanding in chart, we sampled 30 word pairs from the WS353 dataset, then used both histogram and dashed line to represent it. The overall trend indicates that our method exhibited high correlation with the human ratings. For example, from "(phone,equip)" to "(plane,car)" , the $HAN\#All$ curve fits better than $HAN\#M2\&M3$ curve with the human ratings, which means our approach combined with co-occurrence based score functions outperformed the approach that does not use it. However, from "(money,launder)" to "(dollar,profit)", two curves have contrary trend with human ratings, we believe that this situation is due to the limitation of Wikipedia corpus.

6 Conclusion and Future Work

This paper presents a novel two-layers association network named HAN, to capture three kinds of relationships among words and concepts, namely word-to-word, word-to-concept and concept-to-concept relationships. We provide a holistic view to model complex relationships among words and concepts, and fully utilize three relationships to identify highly related word pairs. Our empirical evaluation confirms that using HAN leads to well improvements in computing words relatedness over Florida Norms and Wikipedia corpus. In future work, computing semantic relatedness is not restricted to word. We need to move from words to phrases, sentences, and much larger pieces of texts. We need to design a new algorithm for better capturing contextual information hidden in the encyclopedia knowledge database.

Acknowledgments

We thank Yanyan Shen from Shanghai Jiao Tong University and Ning Li from the University of Iowa for useful discussions and feedback.

References

Agirre, E.; Cuadros, M.; Rigau, G.; and A.Soroa. 2010. Exploring knowledge bases for similarity. *In Proceedings of LREC'10* 373–377.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL'09 Conference*.

Budanitsky, A., and Hirst, G. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1):13–47.

Dagan, I.; Lee, L.; and Pereira, F. 1999. Similarity-based models of word co-occurrence probabilities. *Machine Learning* 34(1-3):43–69.

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Harshman, R. 1990. Indexing by latent semantic analysis. *JASIS* 41(6):391–407.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems* 20(1):116–131.

Gabrilovich, E., and S.Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In Proceedings of IJCAI'07*.

Hassan, S., and Mihalcea, R. 2011. Semantic relatedness using salient semantic analysis. *In Proceedings of AAAI'11* 884–889.

Keyang, Z.; Kenny, Q.; and Seung-won, H. 2015. An association network for computing semantic relatedness. *In Proceedings of AAAI'15*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space.

Miller, G., and Charles, W. 1991. Contextual correlates of semantic similarity. *Language & Cognitive Processes* 6(1):1–28.

Miller, G. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.

Milne, D., and Witten, I. 2008. Learning to link with wikipedia. *In Proceedings of CIKM'08* 509–518.

Mogren, O. 2015. *Multi-Document Summarization and Semantic Relatedness*. Gothenburg, Sweden: Department of Computer Science and Engineering.

Nelson, D.; McEvoy, C.; and Schreiber, T. 2004. *The university of south florida free association, rhyme, and word fragment norms*, volume 36. Behavior Res. Methods, Instruments & Computers.

Peipei, L.; Haixun, W.; and Kenny, Q. 2013. Computing term similarity by large probabilistic isa knowledge. *In Proceedings of CIKM'13*.

Pennington, J.; Socher, R.; and Christopher, M. 2014. Glove: Global vectors for word representation. 1532–1543.

Rada, R.; Mili, H.; Bichnell, E.; and Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics* 9:17–30.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of IJCAI'95* 448–453.

Roget, P. 1852. Roget's thesaurus of english words and phrases. *Longman Group Ltd*.

Rubenstein, H., and Goodenough, J. 1965. Contextual correlates of synonymy. *Commun. ACM* 8(10):627–633.

Wen-tau, Y.; Ming-Wei, C.; Christopher, M.; and Andrzej, P. 2013. Quertion answering using enhanced lexical semantic models. *Proceedings of the 51st Annual Meeting of the Association for Computational linguistics* 1744–1753.

Zhaohui, W., and Giles, C. L. 2015. Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. *In Proceedings of AAAI'15*.