Sentence Similarity Measure Based on Events and Content Words

SHAN Jian-fang, LIU Zong-tian, ZHOU Wen
School of Computer Engineering and Science, Shanghai University, Shanghai, 200072 China
E-mail: sjfshan@163.com

Abstract

Similarity between sentences plays an important role in a variety of applications of natural language process. The paper proposes a novel method to calculate sentence similarity, which takes into account both semantic and syntax: measuring semantic similarity based on events, measuring syntactic similarity based on content words sequence. Our experimental results show that the method is reasonable and effective.

Keywords: Sentence similarity; Event; Event similarity; Content Words Sequence

1. Introduction

Similarity between sentences become increasing important in a variety of applications of natural language process such as text summarization, example-based machine translation, automatic question-answering, information extraction and text clustering.

Enormous achievements have been made over the past years in research on sentences similarity calculation, which can be classified into three categories: semantic-based, syntax-based and both. The first category usually base on co-occurring words and word statistics, basically include: Probabilistic Model [1], Vector Space Models (VSM) [2], Edit Distance Method [3], N-gram Model [4], etc. The method of co-occurring-word-based fits to long text not fits to sentences, especially headline and tabloid of news, because the short texts may very similar while co-occurring words are rare or even null. The VSM bases on TFIDF is one of the most effective models in statistic-based methods, neither fits to short text. For example, two sentences, "I love pop song" and "Mom likes classic music", describe the keenness for music, there must be a high similarity between them. However, they do not share words, the method based on co-occurring words and the VSM both compute the similarity as 0. The second category usually use WordNet [5,6] or HowNet [7–9] as semantic knowledge base, the semantic similarity between sentences is calculated through integrating the semantic similarity between words. In the last category, Peng Jing [10] proposes a similarity calculation method based on words similarity, the method converts text to words vector space model, and then splits words into a set of concepts. Through computing the inner products between concepts, it obtains the similarity between words. Xiao-Ying Liu [11] proposed a method to calculate sentence similarity, semantic similarity is derived from a lexical knowledge base and a corpus syntactic similarity is correlation coefficient between original and relative index vectors of sentences, the sentence similarity is then defined as a combination of semantic similarity and syntactic similarity.

Every considering semantic method treats texts as a bag of words and lose sight of semantic relation among words in the context. Obviously, Texts with same words can have different meaning if the words are put together into them in different way. In other words, a word can convey different meaning if it is different sentence element. For example, two sentences "Tom gave Mary a book as birthday gift" and "Mary gave Tom a book as birthday gift", traditional methods compute the similarity as 1. Obviously, the result does not accord with the consensus of people that the similarity is 1 if and only if one compare with oneself. So, word itself and its sentence element, such as subject, predicate, accusative, etc, are what we must take into consideration.

The paper proposes a novel method to calculate sentence similarity, which takes into account both events and content words sequence. Measuring semantic similarity based on events. With the knowledge granularity and Semantic-Based knowledge processing arrival, many scholars raise the research on events. Computational linguists think that the sentence is not only a bag of words, but also a narrator of events in the view of semantic understanding at higher granularity. The same bag of words can form different sentences if these words are in different morpheme sequence, the different sentences describe different events, every word plays a role in event, the same word can convey different meaning in an event if it plays another role. In order to address the problem of "a bag of words", the paper takes the word’s role into consideration through the event and event elements. The number of co-occurring content words and
the length of LCS(Longest Common Subsequence) contribute to measuring syntactic similarity between two sentences. At last, the method calculate the sentences overall similarity through the weighted sum of syntactic similarity and semantic similarity.

2. Event

2.1. The definition and expression of event

Event originated from Cognitive Science. It has been widely used in the computational linguistics literature as well as in information retrieval and various NLP applications, although with significant variance in what exactly an event is. Events and their semantic structure have been analyzed by several linguists. Wang Shen [12] defined event as "For some purposesomebody does something for some-one with some meannosometimes and somewhere". For the applicability of NLP, we have the following definition (we call it 5H elements event model), an event is "who do what to whom when and where", the formal expression is \( e = (H_s, H_p, H_a, H_l, H_t) \) The relations between 5H and sentence elements are: \( H_s \leftarrow \) subject, \( H_p \leftarrow \) predicate, \( H_a \leftarrow \) accusative, \( H_l \leftarrow \) time, \( H_t \leftarrow \) location.

2.2. Event extraction

Given a sentence \( s \), algorithm \( \text{extract}_\text{event}(s) \), events are extracted from \( s \) by using the following steps:

\( \bullet \) Mark sentence with POS tags and recognize named entities, such as \( \{ \text{Person}, \text{Organization}, \text{Location}, \text{Time} \} \), etc, by using Gate [13].

\( \bullet \) Parse the sentence by using Stanford Parser [14].

\( \bullet \) Analysis the results from the above steps, procuring event formal express of the sentence.

3. The proposed method

3.1. Semantic similarity between words

The way to measure the semantic similarity between two words is to treat words as an undirected graph and measure the distance between them in WordNet. Human usually consider that the shorter the path from one node to another, the more similar they are. Note that the path length is measured in words nodes rather than in links edges. As shown in Figure 1, we observe that the length between girl and miss is 1, apple and banana is 3, girl and apple is 12. A shared parent of two words is called a subsumer. The least subsumer (LS) of two words is the subsumer that does not have any children that are also the subsumer of two words. In other words, the LS of two words is the most specific subsumer of the two words. Back to the Figure 1, the LS of (girl, miss, young lady) and (boy, male child) is (person, individual). The path length is a simple way to compute the semantic distance between two words. This formula is proposed by Wu & Palmer, the measure takes into account both path length and depth of the least subsumer. Given two words \( w_1 \) and \( w_2 \), the semantic similarity \( S_{\text{word-sem}}(w_1, w_2) \) between them as follows:

\[
S_{\text{word-sem}}(w_1, w_2) = \frac{2 \times \text{depth}(s)}{\text{depth}(w_1) + \text{depth}(w_2)}
\]

where \( \text{depth}(s) \) is the shortest distance from root node/word to node/word \( s \) on the WordNet-style hierarchical graph. \( \text{ls} \) denotes the least subsumer of \( w_1 \) and \( w_2 \).

We classify the words that are not embodied by WordNet into three categories - named entity (basically include person name, location name, organization name and other proper nouns), numeric string (date, time, percentage, etc.), the remainder, and we set the following rules:

\( \bullet \) The similarity between two words is 0 if they belong to different category.

\( \bullet \) The similarity between two named entities is 1 if they refer to the same object (abbreviative notation, alias name, etc).

\( \bullet \) If the two named entities don’t refer to the same object, meanwhile, they belong to the same concept, such as two person-names, and we’ll assign a appropriate value to the similarity according to the application.

\( \bullet \) The similarity is 1 if the two numeric strings have the same actual value, such as 0.5 and 50%, otherwise the similarity between them is 0.

\( \bullet \) Statistics show that the last category plays only a very small part, the similarity between them is considered as 0.
3.2. Semantic similarity between sentences

Given two sentences $s_1$ and $s_2$ the event-based semantic similarity calculation between them involves the following steps: extracting events from sentences, event elements similarity measure based on words similarity, event elements similarity measure based on elements similarity, sentences semantic similarity measure based on events similarity. The specification of the steps is following:

- **Calling algorithm extract events($s$)**, events are extracted from $s_1$ and $s_2$, reaching the following events’ expresses:
  
  \[ e_1 = (H_{11}, H_{1p}, H_{1a}, H_{1t}, H_{1t}) \]
  \[ e_2 = (H_{11}, H_{1p}, H_{1a}, H_{1t}, H_{1t}) \]

- **Measuring similarity between event elements:**

  Let $i = (s, p, a, t, l)$, which denotes the subscript of event elements $H_i$, $H_{1i}$ represents the element $H_i$ of $e_1$, $H_{2i}$ represents the element $H_i$ of $e_2$. For each word $H_{1i}$ in $H_{11}$, we firstly identify the word $h_{2i}$ in $H_{2i}$ that has the highest semantic similarity between them according to the formula (1). Then for each word $h_{2i}$ in $H_{2i}$, we identify the word $h_{1i}$ in $H_{1i}$ that has the highest semantic similarity between them. Finally, we obtain the similarity between $H_{1i}$ and $H_{2i}$, the formulas as following:

  \[ S_H(H_{1i}, H_{2i}) = \max_{h_{1i} \in H_{1i}} \frac{\sum_{j \in H_{2i}} S_{\text{word-sem}}(h_{1i}, h_{2j})}{|H_{1i}|} \]

  \[ S_H(H_{2i}, H_{1i}) = \max_{h_{2i} \in H_{2i}} \frac{\sum_{j \in H_{1i}} S_{\text{word-sem}}(h_{2i}, h_{1j})}{|H_{2i}|} \]

  \[ \text{Sim}_H(H_{1i}, H_{2i}) = \frac{S_H(H_{1i}, H_{2i}) + S_H(H_{2i}, H_{1i})}{2} \]  (2)

where $|H_{1i}|$ is the number of words in $H_{1i}$, and $|H_{2i}|$ is the number of words in $H_{2i}$. If there is only one word in $H_{1i}$, and do $H_{2i}$, formula (2) is simplified as $\text{Sim}_H(H_{1i}, H_{2i}) = S_{\text{word-sem}}(H_{1i}, H_{2i})$, i.e., formula (1). Since some event element may null (has no words), we set two roles: the similarity between two null elements is 1, the similarity between null and non-null is 0.

- **Measuring similarity between $e_1$ and $e_2$** through the following formula:

  \[ \text{Sim}_e(e_1, e_2) = \sum_i w_i \text{Sim}_H(H_{1i}, H_{2i}) \]  (3)

where $i = (s, p, a, t, l), w_i$ is the weight of element $H_i$ when measuring the similarity between events, subject to $\sum_i w_i = 1$.

- The similarity between events $e_1$ and $e_2$ is also the semantic similarity between sentences $s_1$ and $s_2$ as following,

  \[ \text{sim}_{\text{sem}}(s_1, s_2) := \text{Sim}_e(e_1, e_2) \]  (4)

3.3. Syntactic similarity between sentences

Firstly, the sentences need to be preprocessed, including filter stop words and stemming words. Content words, such as nouns, verbs, adjectives, etc, are preserved. Given two post-preprocessing sentences $s_1$ and $s_2$, the following metrics should be considered:

- **Measurement based on the number of co-occurring content words**

  \[ \text{Sim}_{\text{syn-co}} = \frac{2 * C}{|s_1| + |s_2|} \]  (5)

where $C$ is the number of co-occurrence words between $s_1$ and $s_2$, $|s|$ the length of sentence $s$.

- **Measurement based on LCS (Longest Common Subsequence):**

  A sequence $Y = [y_1, y_2, y_3, ..., y_n]$ is a subsequence of another sequence $X = [x_1, x_2, x_3, ..., x_m]$, if there exists a strict increasing sequence $[i_1, i_2, i_3, ..., i_n]$ of indices of $X$ such that for all $j = 1, 2, 3, ..., k$, we have $x_{i_j} = y_j$. Given two sequences $X$ and $Y$, the longest common subsequence (LCS) of $X$ and $Y$ is a common subsequence with maximum length [15]. According the definition, we define the following formula:

  \[ \text{Sim}_{\text{syn-LCS}} = \frac{2 * |\text{LCS}(s_1, s_2)|}{|s_1| + |s_2|} \]  (6)

Where $|\text{LCS}((s_1, s_2))|$ is the length of longest subsequence of $s_1$ and $s_2$.

- Taking the above two metrics into consideration, the formula for overall sentence syntactic similarity is given:

  \[ \text{Sim}_{\text{syn}}(s_1, s_2) = \alpha * \text{Sim}_{\text{syn-co}} + (1 - \alpha) * \text{Sim}_{\text{syn-LCS}} \]  (7)

where $\alpha$ is weight ratio of the metric for calculating the syntactic similarity.

3.4. Sentence similarity

The overall sentence similarity is calculated, which consists of two parts: the semantic similarity and syntactic similarity weighted by a smoothing factor. Given two sentences $s_1$ and $s_2$, the total sentence similarity is calculated according to the following formula:

\[ \text{Sim}(s_1, s_2) = \lambda * \text{Sim}_{\text{sem}}(s_1, s_2) + (1 - \lambda) * \text{Sim}_{\text{syn}}(s_1, s_2) \]  (8)

4. Experiments and evaluation

The sentences used in the sentence retrieval based on similarity experiments are the model summarizations of
DUC 2004(task 1), which are very short manual summarization (not more than 75 bytes) of each document in corpus. The corpus contained on 50 TDT topics documents and each topic contained on average 10 documents. The number of sentences (model manual summarization) is 2000.

For each source sentence, five sentences are returned ordered by the similarity between the source sentence and the retrieval sentence. Partial experiment result is shown in table 1. Our experiment results show that the retrieved sentences based on similarity are logical and accord with people’s practical experience.

Though the method used more weight ratios, we can adjust relevance weight ratio according to the objective retrieval results. For example, if we only care about "somebody is at somewhere", we can set the weight ratios of $H_p, H_a, H_t$ to zero. In the paper’s experiment, we take all the elements into consideration. Despite that, the weight ratios determination is still a key problem. In the future, we will devote to parameter optimization to improve the accuracy.

5. Conclusion

The paper proposed a novel method to calculate the similarity between sentences, which combines semantic and syntax. Firstly, semantic similarity is calculated based on the events that are extracted from the sentences. Secondly, syntactic similarity is calculated based on the content words that are preserved in the sentences after pre-processing. The overall sentence similarity is then calculated as a combination of semantic similarity and syntactic similarity.

The whole calculation process is illustrated as figure 2.

Our experimental results of sentence retrieval based on similarity show that the proposed method is reasonable and effective.

In the future, we will perform parameter modifications to obtain better approximation by a large mount of experiments, and improve the method for Chinese sentences.

Acknowledgment

This work is supported by four projects of National Science Foundation of China (NSFC No.60575035), Shanghai Scientific Special Funds for Cultivation and Selection of Excellent Young Teachers (shu-07027), Innovation Fund of Shanghai University (A10-0108-08-004) and Shanghai Leading Academic Discipline Project(J50103).

References


<table>
<thead>
<tr>
<th>NO.</th>
<th>Source Sentences</th>
<th>Retrieved Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Former Hurri-cane Mitch leaves 231 dead in Hon-duras,357 in region.</td>
<td>- Hurricane Mitch’s death toll is 357, including at least 231 in Honduras&lt;br&gt;- At least 231 killed by Hurricane Mitch in Honduras; makes region toll 357.&lt;br&gt;- Honduras confirms 231 people dead in the wake of Hurricane Mitch.&lt;br&gt;- Hurricane Mitch punished Honduras with 120mph winds and killing 32 people.&lt;br&gt;- Hurricane Mitch hits Honduras with 120 mph winds, 32 people killed.</td>
</tr>
<tr>
<td>2.</td>
<td>Low petroleum prices and high exploration costs causing oil company mergers.</td>
<td>- Low petroleum prices, high cost of exploration motives for possible merger.&lt;br&gt;- Low petroleum prices, high costs, drive companies to consider mergers.&lt;br&gt;- Low oil prices, high exploration costs demand deep-pocket companies.&lt;br&gt;- Low petroleum prices causing oil companies to consider merger talks.&lt;br&gt;- Low petroleum prices causing oil companies to consider merger talks.</td>
</tr>
<tr>
<td>3.</td>
<td>New Cambodian government in limbo as Hun Sen rejects talks out of country.</td>
<td>- Cambodian leader Hun Sen rejects opposition demands for talks in Beijing.&lt;br&gt;- Hun Sen rejects out of country talks, Sihanouk asked to host summit.&lt;br&gt;- Cambodian government rejects opposition’s call for talks abroad.&lt;br&gt;- Political deadlock broken, Hun Sen to be Cambodia’s prime minister.&lt;br&gt;- Cambodian King announces coalition government with Hun Sen as sole Premier.</td>
</tr>
<tr>
<td>4.</td>
<td>European Commis-sion President hopes Pinochet will stand trial for genocide and terrorism.</td>
<td>- European Commission President endorses extradition and trial of Pinoche.&lt;br&gt;- European Commission president urges extradition, punishment of Pinochet for his crimes.&lt;br&gt;- European Commission President calls for Pinochet’s extradition to Spain.&lt;br&gt;- British and Spanish governments leave extradition of Pinochet to courts.&lt;br&gt;- Spanish and British governments say Pinochet arrest strictly a legal issue.</td>
</tr>
<tr>
<td>5.</td>
<td>Pope calls for aid to Central America where Hurricane Mitch killed 9,000.</td>
<td>- Pope appeals for aid for victims of Hurricane Mitch. Estimated 9,000 dead.&lt;br&gt;- Pope John Paul II appeals for aid for hurricane victims in Central America.&lt;br&gt;- Pope John appealed for aid to Central American victims of Hurricane Mitch.&lt;br&gt;- European Union votes funds to aid hurricane victims in Central America.&lt;br&gt;- European Union approves $8.18 million in aid to Hurricane Mitch victims.</td>
</tr>
</tbody>
</table>


