Language, Information and Computation(PACLIC 11), 1996, 357-362

Extracting Topics from Texts Based on Situations

Ma Zhiyi,Zhan Xuegong,Yao Tianshun
Department of Computer Science, Northeastern University
Shenyang, 110006 China

Abstract

To understand text, we must relate it with specified situations. This paper, on the basis of such an idea, discusses how the things that a text describes and the situation that the text relates to are expressed in a computer and how the topic of a text is extracted.

1.Introduction

The topic of a piece of writing is the central idea of it. People not only care for what it expresses, but also care for the most important idea of it. Therefore, if each piece of writing has a marked topic, then retrieval rate and retrieval precision of writings can be increased to a great extent. In computers, a piece of writing is a text. Only when a text relates to certain situations, can its topic be rightly extracted. On the basis of situations matching, the paper presents the method of extracting topics from texts.

A situation is something's world state in a specified moment and environment. So, when a thing's attributes are described, the thing's situations must be referred to, and at every moment, reasoning only can be executed in a situation. This is why situations are built.

A situation is a function of time and space. For example, things and their behaviors can constitute different situations at the same time and in difference space, or in the same space and at difference time by increasing or decreasing a little information.

The situation knowledge of something is the sum of perceptual experience and rational knowledge about a topic. The degree of people understanding a text often relies on their knowledge level of the field to which the text refers. That is, same or similar situation knowledge is the foundation of people's understanding each other.

Situations may be divided into broad sense situations and narrow sense situations. A broad sense situation includes all dynamic state events and static state events in the objective world. It is consecutive tableaus in consecutive time and space. A narrow sense situation is dynamic state or static state events that relates to a specified discourse, that is, it includes the environments in which the discourse happens and the events to which the discourse relates.

The task of situation theory is extracting common inner structure from large amounts of real situations in the objective world, probing into the restrained relations among situations, revealing meaning of the language expressions, and presenting a calculable mathematical model for natural language understanding based on situations(Sun Bo 1992). A real situation is a set of specific dynamic or static events. An abstract situation is a set that isconstructed by some elements

with common inner structure extracted from large amounts of real situations, and it is a mathematical abstraction. Objects, relations among objects, time and space are all basic elements that establish an abstract situation, and objects and relations among objects are of recurrent in time and space.

2.Text Understanding Based on Situations

Natural language understanding is mainly a process of the close on analyzing of a text information that reflects its natural features and attributes and judging of narrated facts, events and behaviors.

A text can be understood under a specific situation. The analysis of natural language based on situations is as follows:

Let r_n be an n-gram relation and p be a space-time location. We define $< r_n, x1, x2, ...xn, p >$ as an event state.

An event procedure is a sequence of event states which conforms to a certain time order.

An event class is an event procedure that contains undecided elements.

For an event class E, if a function f can determine the values of undecided elements of E, then f is called a fixed function of E.

If E[f] is a subprocedure of an event procedure e, it is called an event of E.

If there exists an event procedure e that makes all the variables unique, then e is a context of E.

A situation pattern is a causal set of event classes, and its organizational form is a network-like structure, which uses event classes as its basic nodes.

Let $[\![...]\!]$ be a causal linkage set, Δ be space-time variables, Xi be object variables, where $i \in [1,n]$. The following are some definitions:

<situation pattern>::= $[\![$<event class>,...,<event class>$]\!]$

<event class>::=$< r_n, X1, ..., Xn, > | [< r_n, X1, ..., Xn, >]$

An event state relates to time and space. Now, we analyze the words which express time or space.

Location words include region names, orgaization names and compound location words. The first two can immediately be extracted from a text, but the last can not, it only conjoins the first two and can give a real space meaning.

Time words may express moment or temporal intervals.

Direction words can not be used alone and must be used together with location words and time words.

Depicting an object still involves concepts, and concepts' meaning depends on situations. In another word, to express an object, those conceptual attributes, situations that a concept depends on and relations among concepts must be depicted. A conceptual structure is described as follows:

<conceptual class>::<conceptual name>

<name of member>::<content of member>

...

<member operation mode>::<member operation depicting>

...

<name of the situation depended>

where,

<conceptual class> may be general concepts that describe common attributes of concepts with the same attributes, operative functions, and situations depended on. It may also be special concepts.

<name of member> describes names of individual attributes and names of individual features.

<content of member> describes the value of members or restrictions to the value. It may also be another concept.

<member's operation model>::<member's operation depicting> describes operations that relate to the members in the form of rules.

A part of situation structure on concepts is as follows:

...
Entry
Modifier semantics
Pragmatic constraint
...

When reasoning, the conceptual organization not only restricts the range of the searching space, but also benefits reasoning and policy-making under circumstances that information is of uncertainty or insufficiency.

Conceptual situation knowledge is stored in the system collocation dictionary.

Sentences in a text describe relations among objects. According to the nature of a sentence that expresses information, sentences are divided into narrative sentences, descriptive sentences and determining sentences.

Narrative sentences express dynamic information. Its situation meaning is an event state or an event process.

Descriptive sentences express static information. Its situation meaning is modifying semantics of a real frameworks.

Determine sentences' predicates are copulatives. Determine sentences are similar to descriptive sentence.

Given a text, we combine situations, utilize the intertransmittal language of our system CETRAN(Yiao Tianshun 1995),constitute the event chains, object chains and time chains, establish situation patterns and understand it.

3.Extracting topics

The process of extracting the topic from a text based on situations is that the situation of the text matches the situation patterns that exist in the system. The situation pattern is a model of knowledge representation. In accordance with a thing happening and developing, the thing's event processes constitute an event network. If a text's event network can match a situation pattern, the situation that the text describes is a concrete instance of the situation pattern. Therefore, the system can ascertain that the text's topic is consistent with the situation pattern's topic.

In practice, there is little fully matching between situations. What the paper discusses is similarities between situations.

3.1 Similar Relations

If there are some similarities between a new situation and the situation

pattern that has existed, then other similarities can be inferred, and some natures can be processed by the new situation. This is a reasoning process.

The process pays attention to similarities between situations, and may reason under the circumstances that information is of uncertainty or insufficiency, and can get new information.

The process that determines whether two situations are similar can be divided into two steps:

Step1. The system judges whether the elements which composite a situation are parallelism in the situations' structure, i.e., whether two event processes are similar in the situations' topology.

Step2. An event process is an event chain, and each node of the chain is an event state. If each pair of corresponding event states in two event chains is similar, then the two situations are similar.

Because applied fields are extensive, the paper can not present a general outline about the causality among event processes. However, it is very clear that event states are basic units of a situation network.

A knowledge composition of similar reasoning between event states is as follows:

a. set of objects obj-set
   set of relations among objects rela-set
   set of event states es-set
b. similar relations among objects
   similar relations among the relations

Compositions of similar relations among event states are as follows:

a. corresponding relations among structures of elements which composite event states.

b. the similarities among elements.

The elements include objects and relations among the objects which composite event states.

If two event states s1 and s2 are similar, then we write $s1 \approx s2$.

$s1 \approx s2$ holds, if and only if there is an isomorphism w between s1 and s2.

a. Given an object a, if $a \in$ obj-set, then $w(a) \in$ obj-set, and $a \approx w(a)$.

b. Given a relation r, if $r \in$ rela-set, then $w(r) \in$ rela-set and $r \approx w(r)$.

c. If $s1 \in$ es-set, then $w(s1) \in$ es-set and

$$w(r(a1,...,an)) = (r)(w(a1,...,an))$$

$$w(\llbracket s1,...,sn \rrbracket) = \llbracket w(s1),...,w(sn) \rrbracket$$

Given two element e1 and e2, if e1 is a part of e2, then e2 includes e1.

Similar relations among objects are symmetric and reflexive, so are the similar relations among relations and similar relations among event states. Inclusion relations are reflexive and transitive.

3.2 Calculating similarity between situations

The similarity among situations is a measure of relation strength. It may be measured in two respects:

One is qualitative analysis, and it reflects specific relations' sudden change. The other is quantitative analysis, and it describes specific relations' gradual change. For example, attributes' values may be compared with one another.

Given two event states, it is not possible that isomorphism among their elements and difference among their attributes are accurately calculated, only similarity can be calculated. The following is an algorithm that calculates degrees of similarity between event states.

Given a set S, it includes n objects whose structures are the same, and each has m attributes. These constitute a relation table O(n,m).

Definition 1 Let u be a new object, if its semantic attributes and value fields entirely correspond to those in O, then u and t(t belong to S) have similarity, denoted So, So $\in$ [0,1].

Definition 2 Let a value field of an attribute Xi be A, if u.Xi $\in$ A and t.Xi $\in$ A, then u.Xi and t.Xi have similarity, denoted Sa, Sa $\in$ [0,1]. So is a function of Sa.

Objects that have existed are divided into K groups. The number of members in each group is N1,N2,...,Nk, respectively. Where,

$$\sum_{i=1}^{k} Ni=n$$

Supposed an attribute X has I difference values x1, x2,...,xl, then the frequency that each happens is Gx1,Gx2,...,Gxl, Nixj is the number of the group whose value is Xj in the ith group. where,

$$\sum_{i=1}^{k} Nixj=Gxj \qquad \sum_{j=1}^{l} Nixj=Nxi$$

Supposed an attribute X happens, and an attribute Y may happen in the same group, then the probability is denoted Sa(X,Y)=P(Y|X).

Supposed an attribute X happens in the ith group only, then the probability of the event P(i|X)=Nix/Gx.

Supposed an attribute Y happens in the ith group, then the probability of the event P(Y|i)=Niy/Ni.

Supposed X and Y are independent of each other, and X happens in the ith group, and Y may happen in the ith group, then the probability of the event p(Y|X,i)=P(i|X)*P(Y|i)=NixNij/GxNi.

$$Sa(Xi,Xj)=\sum_{K=1}^{k} P(Xj|Xi,K)=\sum_{K=1}^{k} \frac{NkxjNkxi}{GxiNk}=\frac{1}{Gxi}\sum_{K=1}^{k} \frac{NkxjNkxi}{Nk}$$

In practice, there are some different focal points. If the focal point is an entire match between the values of attributes, then the similarity of attributes X and Y is Sa(X,Y)/b, where b is an integer and b>1.

Given two objects u and v, then the similarity between them is S0(u,v) =(W1*Sa(u.X1,v.X1)+W2*Sa(u.X2,v.X2)+...+Wk*Sa(u.Xk,v.Xk))/(W1+W2+...+ Wk)

Where, Wi expresses importance degrees of attributes, and its value relates to practical applications. In our system, it is defined as Wi=1-(the number of u.Ai in the training set)/(the number of objects in training set)

Structures of the objects that the algorithm processes are the same. For the calculation of the similarity of objects whose structures are similar, a fuzzy number shall be introduced. The idea of its algorithm is similar as above.

The relations among objects are composed of expended case

relations, event relations and semantic relations and so on. The whole relation set includes 49 kinds of relations(Yiao Tianshun 1995). The similarity comparison between relations among objects is a kind of inclusion relation.

Calculating similarity between event states is based on results of comparison between objects and between relations among objects. Hence, it is easy to calculate similarity between event states in concrete application fields.

4. Conclusion and Future Work

Understanding a text, extracting information from it, building a situation, comparing it with other situations, deciding its topic are our works that we have been engaging in these years. We have gotten some satisfactory achievements.

A text must be understood in a specific situation. The paper discusses how situation theory is applied in understanding texts, describes the method of extracting topics from texts and presents an effective algorithm that may calculate the similarity between event states. We are ready for further research on causal relations among event processes and the readjustment of corresponding parameters in the process of calculating similarity.

References

Yiao Tianshun.1995.*Natural Language Understanding.*Tsinghua University.

Sun Bo, He Kekang. 1992. A Model Chinese Understanding Based on the Situation Theory.*Computer Search and Devolpment.1992(4).*

Gerard Salton. 1994. Automatic Analysis. Theme Generation, and Summarization of Machine-Readable Texts.*Science*, 264,3.

Luo yulong, Li Puo, Zhao qinping. 1995. An Analogical Knowledge Representation and its Logical Description. *CAI95 China.*

Han Ke, Li Deyi. 1995. A Resemblance Evaluation Based on Statistical Knowledge, *LDC/IIA95.* Tsinghua University Press.

Tang Hongying,Yao Tianshun. 1995. the Chinese Full-Text Analysis Based on a Sematic Network Computing Algorithm.*PNLPPRS'95.*