Journal of Computing and Information Technology - CIT 17, 2009, 2, 157–164
doi:10.2498/cit.1001104

Applying Dynamic Co-occurrence in Story Link Detection

Hua Zhao[1] and Tiejun Zhao[2]

[1]College of Information Science and Engineering, Shandong University of Science and Technology, Huangdao, Qingdao, Shandong Province, Heilongjiang Province, P. R. China
[2]School of Computer Science and Technology, Harbin Institute of Technology, P. R. China

Story link detection is part of a broader initiative called Topic Detection and Tracking, which is defined to be the task of determining whether two stories, such as news articles or radio broadcasts, are about the same event, or linked. In order to mine more information from the contents of the stories being compared and achieve a more high-powered system, motivated by the idea of the word co-occurrence analysis, we propose our dynamic co-occurrence, which is defined to be a pair of words that satisfy certain relation restriction. In this paper, relation restriction refers to a set of features. This paper evaluates three features: capital, location and distance. We use dynamic co-occurrence in the similarity computation when we apply it in the story link detection system. Experimental results show that the story link detection systems based on the dynamic co-occurrence perform very well, which testifies the great capabilities of the dynamic co-occurrence. At the same time, we also find that relation restriction is critical to the performance of dynamic co-occurrence.

Keywords: topic detection and tracking, story link detection, word co-occurrence, dynamic co-occurrence, relation restriction, detection cost

1. Introduction

Topic Detection and Tracking (TDT) is a new line of research pursued by the National Institute of Standards and Technology (NIST), which embraces a variety of technical challenges for information retrieval: Story Segmentation, Topic Tracking, Topic Detection, First Story Detection and Story Link Detection.

This paper focuses on the story link detection, which is defined to be the task of determining whether two stories, such as news articles or radio broadcasts, are about the same event, or linked. In TDT, an event is defined as “something that happens at some specific time and place" [1,2]. For example, a story about an earthquake in Japan in October and another story about an earthquake in Japan in May should not be classified as linked because they are about different events, although they both fall under the same general “topic" of natural disasters. Story link detection is thought of as the basis for other event-based topic analysis tasks, such as topic tracking, topic detection, and first story detection [1].

Because the story link detection task is focused on the streams of news stories where new events occur relatively frequently, and comparisons of interest are focused on events that are not known in advance. One consequence of this is that the prior knowledge we can use is poor, so in order to achieve a high-powered story link detection system, we must make the best of the content of the stories being tested, and mine more information from them. To do so, motivated by the idea of the word co-occurrence analysis, we propose our dynamic co-occurrence, which refers to a pair of words appearing in a story at the same time and satisfy certain relation restriction. In this paper, relation restriction is defined to be a set of features.

The structure of the paper is as follows. Section 2 is about a short overview of the current approaches used in story link detection. Section 3 covers our basic model for story link detection. Section 4 lays a strong emphasis on the dynamic co-occurrence. Section 5 focuses on the story link detection system based on the dynamic co-occurrence. Section 6 discusses the experimental results and analyses. Section 7 gives the conclusions inferred from our work.

2. History and Related Work

2.1. Concise History of TDT

The basic idea for TDT originated in 1996, when the Defense Advanced Research Projects Agency (DARPA) realized that it needed technology to determine the topical structure of news streams without human intervention. The domain of the TDT's interest is all broadcast news, i.e., written and spoken news stories in multiple languages (English, Chinese and Arabic).

The TDT tasks and evaluation approaches were developed by a joint effort between DARPA, the University of Massachusetts' Center for Intelligent Information Retrieval, Carnegie Mellon's Language Technology Institute, and Dragon Systems. In 1997, a pilot study laid the essential groundwork, producing a small corpus and establishing feasibility. Between 1998 and 2004, TDT research blossomed, with new and more challenging tasks, many more participating sites, and considerably larger multilingual corpora (adding Chinese data in 1999 and Arabic data in 2002).

LDC provided five corpora to support TDT research. These are namely the TDT Pilot corpus, the TDT2, TDT3, TDT4 and TDT5 corpora. These corpora are collections of news, including both text and speech, from a number of sources and languages. Each story in the TDT2, TDT3, and TDT4 corpora is tagged according to whether it discusses each of the defined topics.

2.2. Related Work

A number of research groups have developed story link detection systems. The best current technology for link detection relies on the use of Cosine similarity between document terms vectors with TF-IDF term weighting [3,4,5]. In a TF-IDF model, the frequency of a term in a document (TF) is weighted by the inverse document frequency (IDF), the inverse of the number of documents containing a term. UMass has examined a number of similarity measures in the link detection task [6,7], including weighted sum, language modeling and Kullback-Leibler divergence, and found that the cosine similarity produced the best results. Motivated by the performance improvement observed in the classifier combination [8], Francine Chen explored the combination of similarity measures for improving story link detection system [4]. In order to make the best of the content of stories, Ying-Ju Chen applied many NLP and IR approaches to monolingual and multilingual story link detection, which include story expansion, topic segmentation, and so on [9].

Because story link detection is the basis for other event-based TDT tasks, some other researchers research into the relations between story link detection task and other TDT tasks [10,11,12].

Word co-occurrence analysis has been widely used in various forms of research concerning the domains of content analysis, text mining, construction of thesauri and query expansion, etc. [13]. In general, its aim is to determine related word or terms and to find similarities of meaning between word pairs.

In order to mine more information from the contents of the stories, motivated by the performance improvements of the word co-occurrence analysis, we explored the dynamic co-occurrence in the story link detection.

3. The Basic Model for Story Link Detection

3.1. Basic Architecture

Our basic story link detection algorithm is shown as follows.

- Pre-processing to create a vector with TF-IDF weighting to represent each story in a given pair;
- Using the Cosine function to compute the similarity between two stories;
- Employing a predefined threshold to decide whether two stories discuss the same topic or not. That is, if the similarity is larger than the predefined threshold, then two stories are on the same topic, or else they are not linked.

3.2. Pre-processing and Story Model

Before a story is modeled, we apply pre-processing to it, which includes removing stopwords and lemmatizationb.

In the basic model, stories are expressed by Vector Space Model (VSM). Suppose S is a story which has been pre-processed, and $term_1$, $term_2$, ..., $term_k$ are distinct words that appear in S. Then S can be expressed by $S = (term_1, w_1; term_2, w_2; ...; term_k, w_k)$, where w_i is the weight of $term_i$ in S and is calculated by the incremental TF-IDF method showed in (1):

$$w_i = tf_i \times log(\frac{N}{n_i} + 0.01) \qquad (1)$$

Where tf_i is the frequency of $term_i$ in S, N is the total number of stories seen so far, and n_i is the document frequency of $term_i$ in all the stories seen so far.

3.3. Similarity Measure

We use classical Cosine function to compute the similarity between two stories in a pair. Assume $w_{s11}, w_{s12}, ..., w_{s1n}$ and $w_{s21}, w_{s22}, ..., w_{s2n}$ are weights of features $\delta_1, \delta_2, ..., \delta_n$ in story $S1$ and topic $S2$, respectively. Then the similarity between $S1$ and $S2$ is calculated as follows:

$$Cos(S1, S2) = \frac{\sum_{k=1}^{n} w_{s1k} \times w_{s2k}}{\sqrt{\sum_{k=1}^{n} w_{s1k}^2 \times \sum_{k=1}^{n} w_{s2k}^2}} \qquad (2)$$

4. Dynamic Co-occurrence

The motivation for proposing dynamic co-occurrence is to mine more information from the contents of the stories and make the best of the contents of the stories, as a result of which we will gain a more portable story link detection system with better performance. In this paper, dynamic co-occurrence is defined as follows:

Definition 1 A Dynamic Co-occurrence (DC) is a pair of words which appear in a story and satisfy certain relation restriction. A DC can be formalized as a three-tuples $\langle W1, W2, RR \rangle$, where W1 and W2 represent the words, and RR denotes the relation restriction satisfied by W1 and W2.

Definition 2 Relation Restriction is a set of features, and can be formalized as $\emptyset$ or $\{f1, f2, ...\}$, where $f1$ and $f2$ denote features.

When $RR = \emptyset$, any pair of words is a DC, as a result of which is that we get lots of DC, but many of these DC are redundant. On the other hand, from [14] we can conclude that a news story tends to focus on the important words and phrases to distinguish between different news events. So, in order to capture these useful dynamic co-occurrences successfully, we evaluate the performance of the following features in our experiments:

- **Capital**: In English news stories, the initials of names (people, location and organization) are usually capital, and these names are the key words to differentiate the similar topics [15,16]. Based on this, this paper chooses “capital" as a feature used in the relation restriction. In the experiments, we use C to denote “Capital".
- **Location**: News stories have the top-heavy, inverse-pyramidal structure, and usually put the important content in the former part. So we think that the words in the former part of the story are more important than those in the latter part. So we choose “location" as a feature, and use L to denote it in the experiments.
- **Distance**: Distance, or window size, is an important feature used in the word co-occurrence analysis, so it's reasonable for us to believe that distance is a useful feature in the dynamic co-occurrence. In our DC method, the distance between two words is defined to be the number of words appearing between these two words. For example, the story content after pre-processing is “*well United State celebrate Independence Day*", and the distance between *well* and *United* is 0, and the distance between *State* and *Day* is 2. We will use D to denote “Distance" in this paper.

The above features can combine with each other to produce composite features. As a result, we will evaluate the following RR in the experiment section all in all:

- $RR = \{C\}$: Each pair of words whose initials are capital forms a DC;
- $RR = \{D\}$: Each pair of words within certain distance forms a DC, and the distance can be specified by the parameter D. In our experiment, $D = \alpha$ means that the distance between two words is equal to α, where $1 \leq \alpha \leq \omega - 2$, and ω is the length of the story.
- $RR = \{L\}$: Each pair of words located in a certain former part of the story forms a DC, and the certain part can be specified by the parameter L. In the experiment, $L = \beta$ means that the words appear in the former β part of the story, where $1 \leq \beta \leq \omega$, and ω is the length of the story.
- $RR = \{C, D\}$: Each pair of words within certain distance, and the initials of which are capital, forms a DC.
- $RR = \{C, L\}$: Each pair of words located in a certain former part of the story, and the initials of which are capital, forms a DC.
- $RR = \{D, L\}$: Each pair of words located in a certain former part of the story, and within certain distance, forms a DC.
- $RR = \{C, D, L\}$: Each pair of words located in a certain former part of the story, within certain distance and the initials of which are capital, forms a DC.

In this paper, we remove stop-words before extracting dynamic co-occurrence.

In order to explain the definition in more detail, we give the following example: for the story content after pre-processing : *well United State celebrate Independence Day*, the dynamic co-occurrence under each relation restriction is listed in Table 1. In order to save space, we only give the words in the dynamic co-occurrences, and the relation restrictions are listed in the first column all together.

5. Story Link Detection Based on Dynamic Co-occurrence

5.1. Using Dynamic Co-occurrence

We use dynamic co-occurrence information in the similarity computation. Suppose $DC(S1)$ and $DC(S2)$ are the numbers of the dynamic co-occurrences in the story $S1$ and story $S2$ respectively, and $SameDC(S1, S2)$ is the number of the mutual dynamic co-occurrences between story $S1$ and story $S2$. The similarity based on the dynamic co-occurrence between story $S1$

RR	**Dynamic Co-occurrence**
{C}	⟨United,State⟩ ; ⟨United,Independence⟩ ; ⟨United,Day⟩ ; ⟨State,Independence⟩ ; ⟨State,Day⟩ ; ⟨Independence,Day⟩
{D=0}	⟨well,United⟩ ; ⟨United,State⟩ ; ⟨State,celebrate⟩ ; ⟨celebrate,Independence⟩ ; ⟨Independence,Day⟩
{L=$\frac{1}{2}$}	⟨well,United⟩ ; ⟨well,State⟩ ; ⟨United,State⟩
{C, D=0}	⟨United,State⟩ ; ⟨Independence,Day⟩
{C, L=$\frac{1}{2}$}	⟨United,State⟩
{D=0, L=$\frac{1}{2}$}	⟨well,United⟩ ; ⟨United,State⟩
{C, D=0, L=$\frac{1}{2}$}	⟨United,State⟩

Table 1. Examples of dynamic co-occurrence under different relation restriction.

and story $S2$, $DCSim(S1,S2)$, is as follows:

$$DCSim(S1,S2) = \frac{2 * SameDC(S1,S2)}{DC(S1) + DC(S2)} \quad (3)$$

5.2. Link Detection System Based on Dynamic Co-occurrence

The architecture of story link detection system based on the dynamic co-occurrence is shown as follows:

- Pre-processing to create a vector with TF-IDF weighting to represent each story ($S1$ and $S2$)in a given pair;
- Extracting the dynamic co-occurrences appearing in story $S1$ and story $S2$, respectively.
- Measuring the similarity between $S1$ and $S2$ based on Cosine and dynamic co-occurrence, respectively.
- Computing the final similarity between $S1$ and $S2$ as follows:

$$\begin{aligned} FinalSim(S1,S2) = {} & Cos(S1,S2) \\ & + DCSim(S1,S2) \end{aligned} \quad (4)$$

- Employing a predefined threshold to decide whether $S1$ and $S2$ discuss the same topic or not.

6. Experiments

6.1. Corpora

We manually created the training corpora and the test corpora from the TDT Pilot Corpus.

The TDT Pilot Corpus, covering the period from July 1, 1994 to June 30, 1995, was the first benchmark evaluation corpus for TDT research. This corpus contains 15,863 English stories which are represented as a stream of text. A set of 25 target topics are defined, which span a variety of topic types and cover a subset of all the topics discussed in the corpus stories. There are about 43 stories per topic on the average. Each story is assigned a label of YES, NO or BRIEF for each of the 25 topics.

We use only the topic 1 to topic 8 to create the corpora for the story link detection evaluation, where topic 1 to topic 5 is adopted to create the training set, and topic 6 to topic 8 is used to create the test set. The creation process of the corpora is as follows: (1) Every two stories which are related (the label is YES) to the same topic are a linked pair; (2) Every two stories which are related (the label is YES) to different topics are a un-linked pair. As a result, the training set and the test set are listed in Table 2.

6.2. Evaluation Measures

We adopt the evaluation methodology defined in TDT to evaluate our system performance. The cost function for the task defined by TDT is shown as follows. The better story link detection, the lower detection cost. In this paper, all experimental results are evaluated by this metric.

$$\begin{aligned} C_{Det} = {} & C_{Miss} \times P_{Miss} \times P_{target} \\ & + C_{FA} \times P_{FA} \times P_{non-target} \end{aligned} \quad (5)$$

Where P_{Miss} is the probability of missing a story; P_{FA} is the probability of a false alarm; P_{target} is the probability of seeing a new story in the stream; C_{Miss} is the cost of missing a new story; $P_{non-target}$ is the probability of seeing an old story, $P_{non-target} = 1 - P_{target}$; C_{FA} is the cost of a false alarm. A miss occurs when a linked story pair is not identified as linked by

	Training Set	Test Set
Number of the Linked Pairs	1545	2644
Number of the Un-Linked Pairs	4126	2816
Total	5671	5460

Table 2. Corpora used in the story link detection evaluation.

the system. A false alarm occurs when a pair of stories that are not linked are identified as linked by the system. In our experiments, we followed TDT's tradion, C_{Miss}, C_{FA}, and P_{target} are set to 1.0, 0.1, and 0.02, respectively.

The cost for each topic is equally weighted and normalized so that for a given system,"$(C_{Det})_{Norm}$ can be no less than one without extracting information from the source data"[1].

$$(C_{Det})_{Norm} = \frac{C_{Det}}{min(C_{Miss} \times P_{target}, C_{FA} \times P_{non-target})} \qquad (6)$$

6.3. Experimental Results and Analysis

We firstly evaluate the basic model of the story link detection, which gains the results: $P_{Miss} = 0.0783$, $P_{FA} = 0.0071$, $(C_{Det})_{Norm} = 0.1131$. Then, in order to evaluate the performance of dynamic co-occurrence, we do several experiments on the link detection system based on dynamic co-occurrence with different relation restriction. All the parameters are trained using the training corpora. Table 3 listed the experimental results.

All the performances of the story link detection based on the dynamic co-occurrence are better than those of the baseline, which proves the validity of dynamic co-occurrence.

In order to compare the performance of the features used in this paper in more detail, we give the following two figures: Figure 1 and Figure 2, where D and L are set to different values.

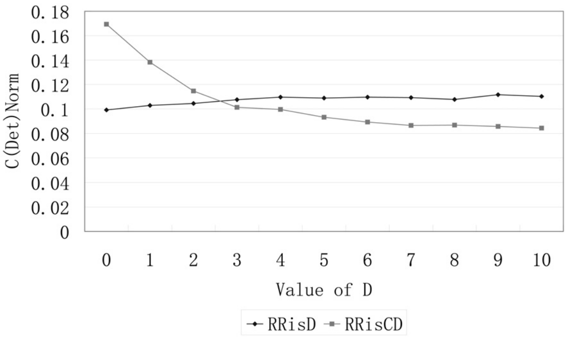

Figure 1. Performance of RRisD and RRisCD under different D.

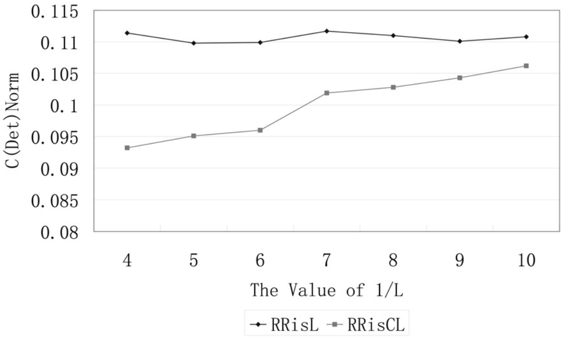

Figure 2. Performance of RRisL and RRisCL under different L.

From all the above results, we can see:

- From the performances of the **RRisC**, **RRisD** and **RRisL**, we can see that the features used in the relation restriction are very successful;

RR	P_{Miss}	P_{FA}	$(C_{Det})_{Norm}$	Experiment Name
{C}	0.0386	0.0089	0.0821	**RRisC**
{D=0}	0.0609	0.0078	0.0992	**RRisD**
{L=1/9}	0.0753	0.0071	0.1101	**RRisL**
{C,D=20}	0.0257	0.0103	0.0762	**RRisCD**
{C,L=1/3}	0.0480	0.0082	0.0881	**RRisCL**
{D=0,L=1/4}	0.0643	0.0075	0.1008	**RRisDL**
{C,D=10,L=1/3}	0.0556	0.0075	0.0921	**RRisCDL**
—	0.0783	0.0071	0.1131	**BaseLine**

Table 3. Comparison between performance of DC-based Story Link Detection System and of Baseline.

- From contrast between **RRisD** and **RRisCD**, as well as **RRisL** and **RRisCL**, we can see that the capital feature is more useful. This is because the initials of names (people, locations and organizations) are usually capital, these names are the key words to differentiate similar topics;
- From the results, we can conclude that the location feature is not as good as the capital feature and the distance feature. We think that the reason is the dynamic evolution of the topic [17].

7. Conclusions and Future Work

In order to mine more information from tested stories, we propose a new technology, dynamic co-occurrence, which is defined to be a pair of words satisfying some relation restrictions.

Experimental results indicate that dynamic co-occurrence is a useful method in the story link detection. We also find that the features used in the relation restriction is very important. This paper uses capital, location and distance as the features which are tested to be successful.

An obvious extension to the current work would find more helpful features in the creation of relation restriction, and research into more effective usage of the dynamic co-occurrence.

References

[1] The 2003 Topic Detection and Tracking Task Definition and Evaluation Plan, April, 2003. `http://www.nist.gov/speech/tests/tdt/tdt2003/-evalplan.htm`

[2] CHARLES L. WAYNE, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Presented in the *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Athens,Greece, pp. 1487–1494, 2000.

[3] RALF D. BROWN, THOMAS PIERCE, YIMING YANG, JAIME G. CARBONELL, Link Detection – Results and Analysis, 2000. `http://www.nist.gov/speech/tests/tdt/tdt99/-papers/cmu_sld/`

[4] FRANCINE CHEN, AYMAN FARAHAT, THORSTEN BRANTS, Multiple similarity measures and source-pair information in story link detection. Presented in the *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, Boston, Massachusetts, pp. 313–320, 2004.

[5] RALF D. BROWN, Dynamic Stopwording for Story Link Detection. Presented in the *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California, pp. 190–193, 2002.

[6] JAMES ALLAN, VICTOR LAVRENKO, DAVID FREY, VIKAS KHANDELWAL, UMass at TDT2000. Presented in the *Proceedings of Topic Detection and Tracking Workshop*, 2000.

[7] JAMES ALLAN, VICTOR LAVERNKO, RAMESH NALLAPTI, UMass at TDT2002. Presented in the *Proceedings of the Topic Detection and Tracking Workshop*, 2002.

[8] JOSEF KITTLER, MOHAMAD HATEF, ROBERT P. W. DUIN, JIRI MATAS, On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), pp. 226–239, 1998.

[9] YING-JU CHEN, HSIN-HIS CHEN, NLP and IR Approaches to Monolingual and Multilingual Link Detection. Presented in the *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 2002.

[10] OLIVIER FERRET, Using collocations for topic segmentation and link detection. Presented in the *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, pp. 1–7, 2002.

[11] FRANCINE CHEN, AYMAN FARAHAT, THORSTEN BRANTS, Story Link Detection and New Even Detection are Asymmetic. Presented in the *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, pp. 13–15, 2003.

[12] AYMAN FARAHAT, FRANCINE CHEN, THORSTEN BRANTS, Optimizing Story Link Detection is not Equivalent to Optimizing New Event Detection. Presented in the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics(ACL03)*, Sapporo, Japan, pp. 232–239, 2003.

[13] JING BAI, JIAN-YUN NIE, GUIHONG CAO, Content-Dependent Term Relations for Information Retrieval. Presented in the *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006)*, pp. 551–559, 2006.

[14] JAMES ALLAN, VICTOR LAVRENKO, RON PAPKA, Event Tracking. University of Massachusetts, Computer Science Department CIIR Technical Report IR-128, January 1998.

[15] JUHA MAKKONEN, HELENA AHONEN-MYKA, MARKO SALMENKIVI, Applying Semantic Classes in Event Detection and Tracking. Presented in the *Proceedings of International Conference on Natural Language Processing*, Mumbai, India, pp. 175–183, 2002.

[16] ZHAO HUA, ZHAO TIEJUN, ZHANG SHU, WANG HAOCHANG, Topic detection research based on content analysis. *Journal of Harbin Institute of Technology*, 38(10), pp. 1740-1743, 2006.

[17] ZHAO HUA, ZHAO TIEJUN, YU HAO, ZHANG SHU, Dynamic evolvement-oriented topic detection research. *Chinese high technology letters*, 16(12), pp. 1230–1235, 2006.

Received: July, 2007
Accepted: March, 2009

Contact addresses:

Ms. Hua Zhao
College of Information Science and Engineering
Shandong University of Science and Technology
579 Qianwangang Road
Huangdao, Qingdao 266510
Shandong Province, P. R. China
e-mail: doctorhuazhao@163.com

Prof. Tiejun Zhao
School of Computer Science and Technology
Harbin Institute of Technology
Harbin Province, P. R. China

MS. HUA ZHAO received her Bachelor's degree in computer science and technology from Liao Cheng University, China, in 2001, the Master's Degree in computer science and technology from Harbin Institute of Technology (HIT), China, in 2003, and the Doctor's degree in HIT in 2008. She currently is a lector at the College of Information Science and Engineering, Shandong University of Science and Technology, China. Her research interests include topic detection and tracking, natural language processing, machine learning.

DR. TIEJUN ZHAO received his Ph.D. degree in computer science and technology from Harbin Institute of Technology (HIT), China, in 1996. He is now a professor (doctorial supervisor) at the Research Center of Language Technology and the vice director of MOE-MS Key Laboratory of NLP & Speech in HIT. He is a member of NLP subject committee of Chinese Information Society, a member of editorial board of Journal of Chinese Information Processing, a member of China Language Data Consortium, a member of Harbin Expert Group on Information Security, a senior member of China Computer Federation. His research fields include: natural language processing, machine translation, content-based web information processing, applied artificial intelligence. In the last 5 years he has done and is doing about 10 projects from NSFC, 863 High-Tech Program, MOST etc. He has won 3 prizes of Ministry Science & Technology Award and has published over 60 academic papers and 2 books.