


A semantic frame-based intelligent agent for topic detection

Yung-Chun Chang^{1,2}  · Yu-Lun Hsieh^{1,3} · Cen-Chieh Chen^{1,3} · Wen-Lian Hsu¹

© Springer-Verlag Berlin Heidelberg 2015

Abstract Detecting the topic of documents can help readers construct the background of the topic and facilitate document comprehension. In this paper, we propose a semantic frame-based topic detection (SFTD) that simulates such process in human perception. We take advantage of multiple knowledge sources and extracted discriminative patterns from documents through a highly automated, knowledge-supported frame generation and matching mechanisms. Using a Chinese news corpus containing over 111,000 news articles, we provide a comprehensive performance evaluation which demonstrates that our novel approach can effectively detect the topic of a document by exploiting the syntactic structures, semantic association, and the context within the text. Experimental results show that SFTD is comparable to other well-known topic detection methods.

Keywords Topic detection · Semantic frame · Semantic class · Partial matching

Communicated by C.-S. Lee.

This research was supported by the National Science Council of Taiwan under Grant NSC102-3111-Y-001-012, NSC102-3113-P-001-006 and NSC 102-3114-Y-307-026.

✉ Yung-Chun Chang
changyc@iis.sinica.edu.tw

¹ Institute of Information Science, Academia Sinica, Taipei, Taiwan

² Department of Information Management, National Taiwan University, Taipei, Taiwan

³ Department of Computer Science, National Chengchi University, Taipei, Taiwan

1 Introduction

Due to recent technological advances, we are overwhelmed by the sheer number of documents. While keyword search systems nowadays can efficiently retrieve documents, users still have difficulty assimilating knowledge of interest from them. To promote research on this subject, the Defense Advanced Research Projects Agency (DARPA) initiated the Topic Detection and Tracking (TDT) project, with a goal to automatically detect topics and track-related documents from several document streams such as on-line news feeds. In essence, a topic is associated with specific times, places, and persons (Nallapati et al. 2004). Thus, detecting the topic of a document can help readers construct the background of the topic and facilitate document comprehension, and it is an active research area in information retrieval (IR). Nevertheless, current machine learning models applied to natural language processing have encountered various bottlenecks. The original purpose of machine learning is to learn text patterns that are expectedly general enough to be applied to other unseen texts. However, these patterns can only achieve a mediocre score. This is especially obvious when we compare the similarity of two sentences (Hsu et al. 1998). One can easily find two sentences that are literally different but convey similar semantic knowledge, which confuse most machine learning models.

To detect topic of documents effectively, we model topic detection as a classification problem. Our proposed method is different in that we took advantage of multiple knowledge sources, and implemented a frame generation algorithm to generate semantic frames that represent discriminative patterns in documents. Furthermore, to identify the topic of documents, we developed a frame matching algorithm to find the most relevant frames for each topic. The results

demonstrated that the semantic frame-based method is effective in topic detection. In addition, the proposed semantic frame generation and matching mechanism successfully exploits the syntactic structures, semantic association, and the content within the text. Consequently, our method outperforms others including the word vector model-based method (Li et al. 2010) and the latent Dirichlet allocation (LDA) method (Blei et al. 2003), which is a Bayesian networks-based topic model widely used to identify topics.

2 Related works

Much work has been done on automatic text categorization. Most of the topic detection methods are concerned with the assignment of texts onto a set of given categories. The original methods on topic detection rely on some measures of importance of the keywords. The weights of the features in these models are usually computed with the traditional methods such as *tf*idf* weighing, conditional probability and generation probability. For instance, Bun and Ishizuka (2002) present the *TF*PDF* algorithm which extends the well-known *tf*idf* to avoid the collapse of important terms when they appear in many text documents. Indeed, the *IDF* component decreases the frequency value for a keyword when it is frequently used. Considering different newswire sources or channels, the weight of a term from a single channel is linearly proportional to the term's frequency within it, while exponentially proportional to the ratio of documents that contain the term in the channel itself.

Several researchers have adopted machine learning approaches to recognize discriminative features for topic mining. For instance, Nallapati et al. (2004) attempted to find characteristics of topics by clustering keywords using a statistical similarity measure for grouping documents into clusters, each of which represents a topic. The clusters are then connected chronologically to form a timeline of the topic. Wu et al. (2010) use the tolerance rough set model to enrich the set of feature words into an approximated latent semantic space from which they extract hot topics by a complete-link clustering. Furthermore, many previous methods treated topic detection as a supervised classification problem (Blei et al. 2003; Zhang and Wang 2010). Given a training corpus containing a set of manually tagged examples of predefined topics, a supervised classification algorithm is employed to train a topic detection model to assign (i.e., classify) a topic to a document. These approaches can achieve substantial performance without much human involvement. However, to manifest topic-associated features, one often needs to annotate the features in documents, which is rarely done in most machine learning models Scott and Matwin

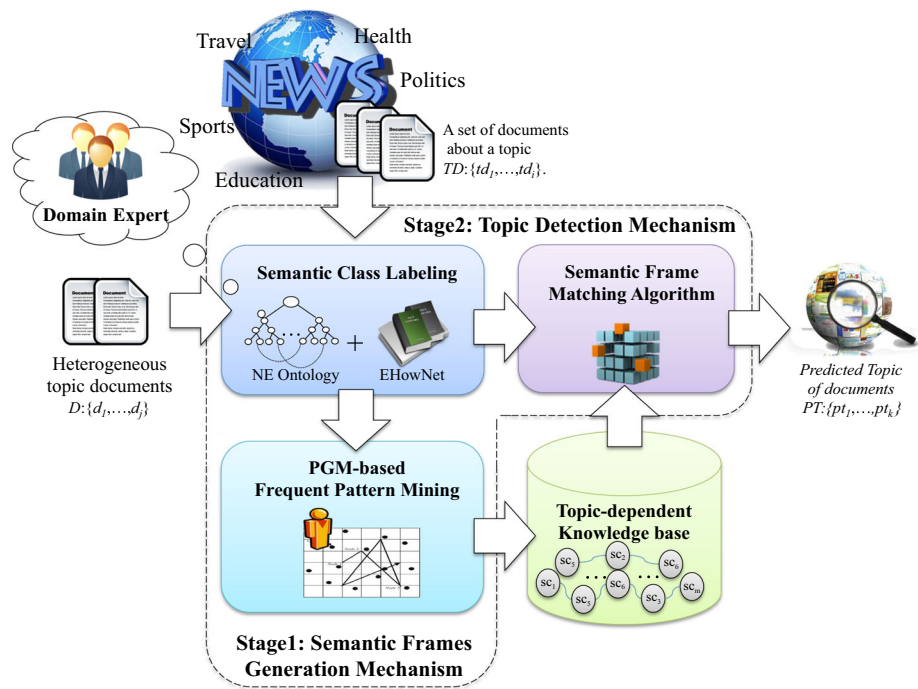
(1999). The shortage of knowledge, the data sparseness problem, and the lack of the ability to make generalizations are some common caveats of such an approach. Once the domain is changed, the models need to be retrained to obtain satisfactory results. Besides, fine-grained linguistic knowledge that is crucial in human understanding cannot be easily modeled, resulting in less desirable performance. One can easily find two sentences that are literally different but convey similar semantic knowledge, which could confuse most machine learning models. On the other hand, the main shortcoming of template-based or knowledge-based methods is the need of human effort to craft precise templates or rules.

Ontology is a conceptualization of a domain into a human understandable, machine-readable format consisting of entities, attributes, relationships, and axioms (Tho et al. 2006). It is also very reusable, which makes it very powerful for representing domain knowledge. The related applications of the ontology involve in many research fields. For instance, Alani et al. (2003) proposed the Artequakt that attempts to identify entity relationships using ontology relation declarations and lexical information to automatically extract knowledge about artists from the web. Some document detection methods made use of ontology and utilized the structured information in Wikipedia to enhance the performance (Grineva et al. 2009). García-Sánchez et al. (2006) proposed an ontology-based recruitment system to provide intelligent matching between employer advertisements and the curriculum vitae of the candidates. Moreover, Lee et al. (2009) used ontology to construct the knowledge of travel information in Tainan City, and further integrated fuzzy inference with ant colony optimization to recommend to the tourist a personalized travel route to enjoy Tainan City according to the tourist's requirements effectively.

Our method differs from existing topic detection approaches in a number of aspects. First, we proposed a semantic frame-based approach that mimics the perceptual behavior of humans in understanding. Second, the generated semantic frames can be represented as the domain knowledge required for detecting topics. In addition to syntactic features, we further consider the surrounding context and semantic associations to efficiently recognize topics. Finally, our research differs from other Chinese researches that rely on word segmentation for preprocessing by utilizing ontology for semantic class labeling.

The rest of the paper is organized as follows. In Sect. 2, we review previous works regarding different topic detection methods. Then, detailed description about the architecture of our topic detection system is given in Sect. 3. Section 4 presents the experimental results. We discuss the implications of the experimental results in Sect. 5. Finally, some conclusions are drawn in Sect. 6.

Fig. 1 Architecture of our semantic frame-based topic detection system



3 System architecture

Our system mainly consists of two mechanisms, the Semantic Frame Generation Mechanism (SFGM) and the Topic Detection Mechanism (TDM), as shown in Fig. 1. The SFGM first uses prior knowledge of each topic to mark the semantic classes of words in the corpus. Then it collects frequently co-occurring tuples, and generates frames for each topic by a Probability Graphical Model, or PGM. These frames are stored in the Topic-dependent Knowledge Base to provide domain-specific knowledge for our topic detection. In the TDM, an article is first labeled with semantic classes using the same knowledge as mentioned above. Then we apply a Semantic Frame Matching algorithm which utilizes our topic-dependent knowledge base to calculate the similarity between each topic and the article to determine the main topic of this article. Details of the two mechanisms will be explained in the following sections.

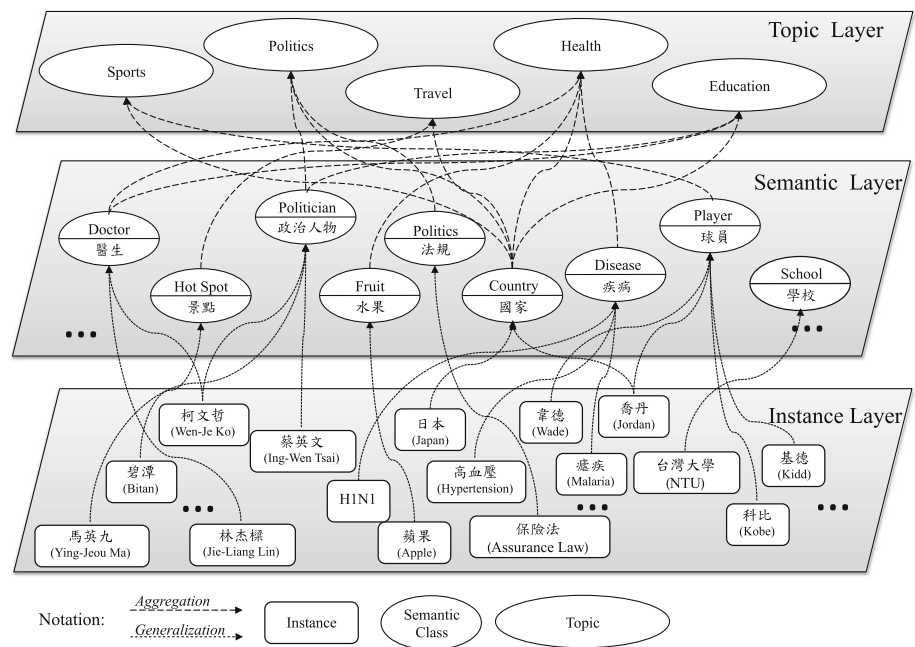
3.1 Semantic frame generation mechanism, SFGM

As depicted in Fig. 1, this mechanism first labels words in an article with semantic classes to increase the frequency of these classes, and enables us to extract distinctive semantic features of a certain topic. Then a graph-based frequent pattern mining algorithm is used to generate semantic frames. These frames form the basis of the topic detection mechanism that follows. In contrast, most Chinese topic detection researches rely on the word segmentation process, which is error-prone and can affect the accuracy of a system. In light

of this issue, our semantic labeling process uses the following two knowledge bases instead:

Named Entity Ontology (NEO) Ontology is a computational model for knowledge representation about the whole or some portions of the world. Recently, we have acknowledged an increasing interest in utilizing ontologies as artifacts to represent human knowledge and critical components in knowledge management, which can be observed in the Semantic Web, business-to-business applications, and several other application areas. Based on the levels of organization mentioned in Lee et al. (2005), Wang et al. (2010), this paper adopts a novel structure to construct the NE ontology for semantic labeling. Figure 2 depicts the architecture of the NE ontology, which includes a topic layer, a semantic layer, and an instance layer. There are five topics in the topic layer, namely “Sports”, “Politics”, “Travel”, “Health”, and “Education”. Moreover, there are 40 semantic classes in the semantic layer, including “doctor”, “hot spot” and others. Each semantic class denotes a general semantic meaning of named entities that can be aggregated from many topics. The instance layer represents 6323 named entities extracted from documents across five topics by the Stanford NER.¹ Domain experts further annotated each named entity by their corresponding semantic classes for the purpose of generalization. Each instance in the instance layer can connect to multiple semantic classes according to the generalized relations. For example, named entity “Jordan” can be generalized to “Player” and “Country”. And, the semantic class “Player”

¹ <http://www-nlp.stanford.edu/ner/>.

Fig. 2 Architecture of named entity ontology

is mentioned in documents with a topic of both health and sports.

Extended HowNet (EHowNet) Extended HowNet, or EHowNet, is an extension of the HowNet (Dong et al. 2010) for a structured representation of knowledge and semantics. It connected approximately 90 thousand words of the CKIP Chinese Lexical Knowledge Base and HowNet, and included extra frequent words that are specific in Traditional Chinese, resulting in 98,900 words. It also contains a different formulation of each word to better fit its semantic representation, as well as distinct definition of function and content words. The detailed specification can be found in CKIP (2009). A total of four basic semantic classes are applied, namely object, act, attribute, and value. Furthermore, compared to the HowNet, EHowNet possesses a layered definition scheme and complex relationship formulation, and uses simpler concepts to replace sememes as the basic element when defining a more complex concept or relationship. To illustrate the content of the EHowNet, let us take “dog food” for example. It is defined as the following:

Definition 1 (Dog food)

Simple Definition:

{food:telic={feed:target={dog},patient={~}}}

Expanded Definition:

{food:telic={feed:target={livestock:telic={TakeCare:patient={family},agent={~}}}, patient={~}}}

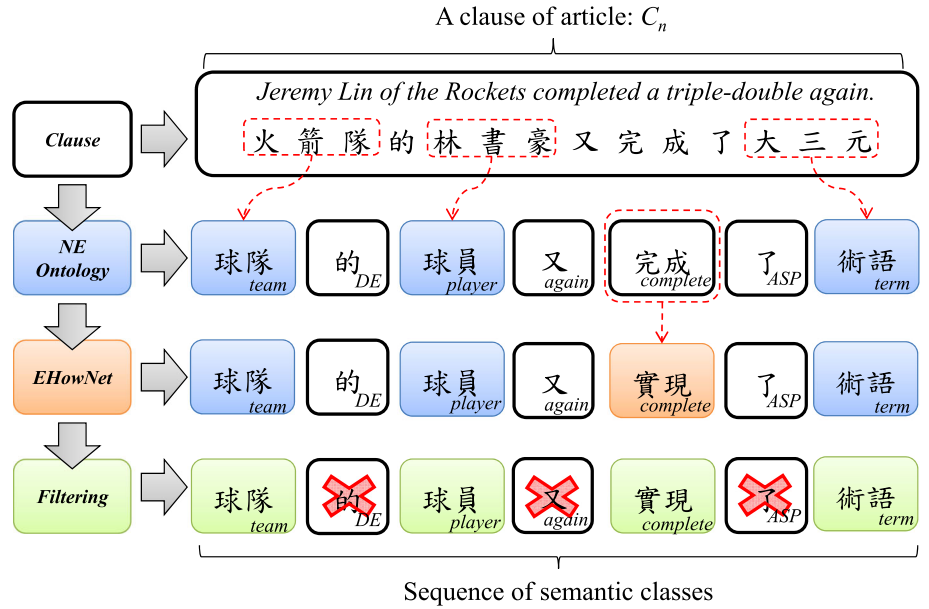
We can see that the EHowNet not only contains semantic representation of a word, but also its relations to other words or entities. The definition follows a specific pattern, as defined in CKIP (2009). This enables us to combine or dissect the meaning of words using its semantic components. Follow-

ing the method in Shih et al. (2014), we extracted the main definition of each word as the semantic class label.

With the above resources, the SFGM can transform words in the original documents into their corresponding semantic labels. Our research assigns clause as the unit for semantic labeling. To illustrate the labeling process, consider the sentence C_n = “火箭隊林書豪又完成了大三元 (Jeremy Lin of the Rockets completed a triple-double again)”, as shown in Fig. 3. First, the NEO converts all NEs to their corresponding semantic classes, and the clause becomes “[球隊_{team}] [球員_{player}] 又完成了大三元 ([player] of [team] completed a triple-double again)”. Then it is further labeled by the EHowNet to tag the main definition of all remaining words, as in, “[球隊_{team}] [球員_{player}] 又 [實現_{complete}] 了 [術語_{term}] ([player] of [team] [complete] a [term] again)”. Lastly, all the non-labeled words are removed, resulting in the frame “[球隊_{team}] [球員_{player}] [實現_{complete}] [術語_{term}]”. The semantic class labeling process cannot only eliminate the errors caused by Chinese word segmentation, but also group the synonyms of a word together by the same label to find distinctive and prominent semantic classes for a certain topic.

We formulate semantic frame generation as a frequent pattern mining problem. Based on the co-occurrence of semantic classes, we can construct a graph to describe the strength of relations between them. Since semantic classes are of an ordered nature, the graph is directed and can be made with association rules. To avoid the generation of frames with insufficient length, we set the minimum support of a semantic class as 50 and minimum confidence as 0.3 in our association rules. This is because we observed that the rank-frequency distribution of semantic classes followed Zipf’s law (Manning and Schütze 1999), and so does the normal-

Fig. 3 Semantic class labeling process



ized frequency of semantic frames. Low-frequency semantic classes usually identify semantic that are irrelevant to the topic. Hence, for each topic, we selected the first frequent semantic classes that accumulated frequencies that reached 80 % of the total semantic class frequency count in the topic documents. Thus, an association rule can be represented as (1).

$$\text{confidence}(SC_i \Rightarrow SC_j) = \frac{P(SC_j|SC_i)}{\frac{\text{support}(SC_i \cup SC_j)}{\text{support}(SC_i)}} \quad (1)$$

where $\text{support}_{\min} = 50$, $\text{confidence}_{\min} = 0.3$.

Figure 4 is an illustration of a semantic graph. In this graph, vertices (SC_x) represent semantic classes, and edges represent the co-occurrence of two classes, SC_i and SC_j , where SC_i precedes SC_j . The number on the edge denotes the confidence of two connecting vertices. After constructing all of the semantic graphs, we then generate semantic frames by applying the random walk theory (Lovász 1993) in search of high frequency and representative classes for each topic. Let a semantic graph G be defined as $G = (V, E)$, where $|V| = p$, $|E| = k$, a random walk process consisting of a series of random selections on the graph. Every edge (SC_n, SC_m) has its own weight M_{nm} , which denotes the probability of a semantic class SC_n , followed by another class SC_m . For each class, the sum of weight to all neighboring classes $N(SC_n)$ is defined as (2), and the probability matrix of the entire graph is defined as (3). As a result, a series of a random walk process becomes a Markov Chain. According to Li et al. (2010), the cover time of a random walk process on a normal graph is $\forall SC_n, C_{SC_n} \leq 4k^2$. We can conclude that using random walk to find frequent pat-

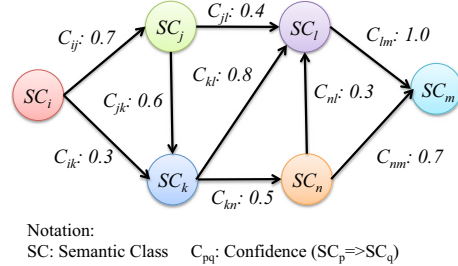


Fig. 4 A semantic graph for frame generation

terns on semantic graphs would help us capture even the low probability combinations and shorten the processing time.

$$\forall SC_n \sum_{m \in N(SC_n)} M_{nm} = 1 \quad (2)$$

$$Pr = \left[\begin{array}{c} X_t = SC_n \\ X_{t-1} = SC_k \\ \dots \\ X_0 = SC_i \end{array} \right] = Pr[X_{t+1} = SC_m | X_t = SC_n] = M_{nm} \quad (3)$$

Although the random walk process can help us generate frames from frequent patterns in semantic graphs, it can also create some redundancy. Hence, a merging procedure is required to eliminate the redundant results by only retaining the frames, with the longest length and highest coverage, and dispose off those that are completely covered by another frame. For example, the frame "[Country]-[Team]-[Player]" is completely covered by another frame "[Country]-[League]-[Team]-[Player]-[Match]-[Lost]". Thus,

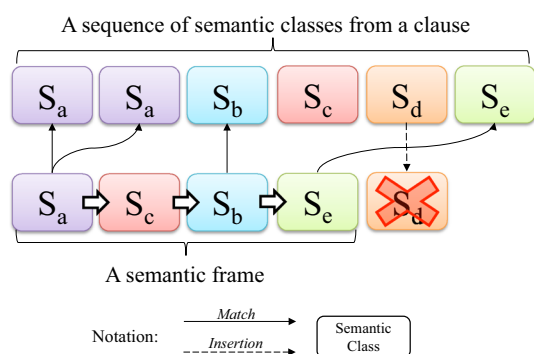


Fig. 5 Illustration of a semantic frame matching process

the former frame is removed. Moreover, the reduction of the semantic class space provided by frame selection is critical. It allows the execution of more sophisticated text classification algorithms, which lead to improved results. Those algorithms cannot be executed on the original semantic class space because their execution time would be excessively high, making them impractical (Baeza-Yates and Ribeiro-Neto 2011). Therefore, to select semantic frames closely associated with a topic would improve the performance of topic detection. We use the log likelihood ratio (LLR) (Manning and Schütze 1999), an effective feature selection method, to discriminate semantic classes for each topic. Given a training dataset comprised of different topics, the LLR calculates the likelihood of the occurrence of a semantic class in the topic. A semantic class with a large LLR value is thought to be closely associated with the topic. Lastly, we rank the semantic frames in the training dataset based on a sum of semantic classes LLR values and retain the top 100 for this topic.

3.2 Topic detection mechanism, TDM

Most of the previous machine learning-based topic detection studies focus on feature engineering to enhance the performance. However, once the range of topics is changed, the effort must be repeated to find another optimal set of features.

Unlike human knowledge, it is difficult to accumulate and share the knowledge collected from different topics. Moreover, the human perception of a topic is obtained through the recognition of important events or semantic contents to rapidly narrow down the scope of possible candidates. For example, when an article contains strongly correlated words like “Kobe Bryant (basketball player)”, “LA Lakers (basketball team)” and “NBA (basketball league)” simultaneously, it is natural to conclude that this is a sport-related article, with a less likelihood of an education-related one. This phenomenon can explain how humans can skim through an article to quickly capture the main topic. In light of this rationale, we proposed a novel approach for topic detection that simulates such process in human perception.

We use semantic frames derived from the frame generation mechanism as a basic knowledge for topic detection. During detection, an unknown article is first labeled with semantic classes, and a matching algorithm is applied to determine the topic of this article. The matching algorithm compares the sequence of semantic classes $C = \{s_1, \dots, s_n\}$ in each clause of an article to every frame $F = \{S_1, \dots, S_m\}$ in each topic. An illustration of the matching process of a sequence of semantic classes to a semantic frame is shown in Fig. 5. The matched and unmatched contents between the two sequences were given different scores according to their type. A match between the two sequences is given a positive score obtained from the LLR score of the semantic class in a topic. On the other hand, an insertion is defined as a label that is present in the article but not in the frame, and is given a negative score computed from the entropy of this label, which can be thought of as the *uniqueness* or *generality* of this label. Finally, the sum of scores of each topic was computed, and the topic with the highest score is considered as the winner. In this way, each individual semantic class label is given a different weight according to its characteristics. Thus, the order of these labels is not the only determining factor in matching. The detailed algorithm is described in Algorithm 1.

Algorithm 1 Semantic frame matching algorithm**Input:** A semantic frame $F = \{S_1, \dots, S_m\}$; A sequence of semantic classes from a clause $C = \{s_1, \dots, s_n\}$ **Output:** Matched score s

```

1:  $pos \leftarrow 0$ ;
2:  $s \leftarrow 0.0$ ;
3: for  $i = 1$  to  $m$  do
4:    $isMatched \leftarrow false$ ;
5:   for  $j = 1$  to  $n$  do
6:      $pos \leftarrow$  current matched position in  $C$  after  $pos$ ;
7:     if found  $s_j = S_i$  in  $p$  after  $pos$  then
8:        $s \leftarrow s + LLR$  value of matched  $S_i$ ;
9:        $isMatched \leftarrow true$ ;
10:    end if
11:  end for
12:  if  $isMatched = false$  then
13:     $s \leftarrow s - LLR$  value of matched  $S_i$ ;
14:  end if
15: end for

```

4 Experimental results

4.1 Dataset and experimental settings

To the best of our knowledge, there is no official corpus for Chinese topic detection. Therefore, we compiled a news corpus for the evaluations from a news agency database between the years 2010 and 2014. It contains five topics with the number of documents included in parentheses, i.e., “Sports” (28,920), “Politics” (29,024), “Travel” (22,257), “Health” (15,845), and “Education” (15,024). A more detailed statistics of this corpus is in Fig. 6. We use half of the documents of each topic as our training data, and the remaining half for testing. The evaluation metrics are the precision, recall, and F1-measure expressed as percentages (%), presented by micro-average to ensure the fairness of our evaluations. Other than our proposed method, SFTD, four additional methods were implemented and evaluated, including one random baseline and three widely used methods for topic detection. The first method is a word-based statistical model, namely, Naïve Bayes. Second, we implemented a vector-based model which chooses keywords based on the cosine distance to measure the similarity (denoted as VSM). Lastly, a probabilistic graphical model which uses the LDA model as the document representation to train an SVM to classify the documents as either topic relevant or irrelevant (Blei et al. 2003) (denoted as LDA).

4.2 Performance evaluation and comparison

Figure 7 depicts the performance of our system on five topics. Our system performs the best on the topic “Sports”, with precision, recall, and F1-measure of 75.66, 85.68, and 80.36 %, respectively. On the other hand, high precision and low recall were found in topics “Health” and “Travel”. The precision of “Health” is the highest among all topics, i.e., 89.52 %. Nevertheless, the topic “Politics” has a lower precision of 61.96 % and the highest recall of 90.63 %. Finally, the topic “Education” has a relatively lower performance of around 50 % for all three metrics.

To evaluate performance with different data splits, we use fivefold cross-validation to examine the effects of SFTD. For each evaluation run, one fifth of the documents are selected as test data, and the remaining are used to train SFTD. The results of the five evaluation runs are averaged to obtain the global performance. The result is in the format of mean \pm standard deviation. As shown in Fig. 8, our method achieves nearly identical performance, with precision, recall, and F1-measure of 72.04, 64.19, and 64.50 %, respectively. It is worthy to note that, the topic “Health” gets a slight improvement on precision with significant recall drop, while the recall of the topic “Travel” is improved along with a slight precision decrease. This may be due to the fact that the NEO can cover most terms in the topic “Travel” (mostly country or place names), but not so much for “Health” (prevalently disease

Notation:	(1) T5-SF: top 5 semantic frames with highest sum of semantic classes LLR values; (2) T10-SC: top 10 most frequent semantic classes; (3) T10-TW: top 10 topic words of LDA with highest probability; (4) AvLen: average length of documents in this topic.
Topics	Statistics
Sport (28920) AvLen: 314	<p>T5-SF: [運動員 athlete]->[運動隊伍 sport team]->[人 human]->[運動員 athlete]->[新聞 news]->[速度 speed]; [運動員 athlete]->[運動隊伍 sport team]->[人 human]->[運動員 athlete]->[人 human]; [運動員 athlete]->[書刊 books]->[運動員 athlete]->[3C 名詞 3C nouns]->[歡迎 welcome]->[網站 website]->[公司 company]->[國家 country]; [運動員 athlete]->[運動員 athlete]->[佔領 occupation]; [運動員 athlete]->[運動員 athlete]->[賺錢 make money]</p> <p>T10-SC: 運動員 athlete, 人 human, 事務 affairs, 國家 country, 運動隊伍 sport team, 地名 place name, 程度 degree, 開始 begin, 方式 style, 物體 objects</p> <p>T10-TW: 林書豪 Jeremy Lin, 比賽 contest, 尼克 Nicks, 球員 player, 報導 report, 記者 reporter, 今天 today, 表現 performance, 球隊 team, 今年 this year</p>
Politics (29024) AvLen: 327	<p>T5-SF: [組織 organization]->[官 officer]->[政治人物 politicians]->[表示 express]; [組織 organization]->[國家 country]->[政黨 Party]->[聚集 gather]->[政治人物 politicians]->[表示 express]; [組織 organization]->[國家 country]->[政黨 Party]->[政治人物 politicians]->[表示 express]; [組織 organization]->[國家 country]->[政黨 Party]->[聚集 gather]->[政治人物 politicians]; [組織 organization]->[實施 implementation]->[政治人物 politicians]->[表示 express]</p> <p>T10-SC: 國家 country, 人 human, 事務 affairs, 政治人物 politicians, 地名 Place name, 組織 organization, 機構 institution, 表示 express, 官 officer, 物體 objects</p> <p>T10-TW: 政府 government, 民進黨 DPP, 立委 Legislator, 總統 President, 國民黨 KMT, 記者 reporter, 中國 China, 報導 report, 民眾 people, 台北 Taipei</p>
Travel (22257) AvLen: 361	<p>T5-SF: [花草 flowers]->[旅遊 travel]->[康健 healthy]->[世界 world]->[花草 flowers]->[旅遊 travel]->[花草 flowers]->[經受 Withstand]->[世界 world]; [吸引 attract]->[旅客 travelers]->[享受 enjoy]; [歡迎 welcome]->[旅遊 travel]->[新聞 news]->[旅遊 travel]->[增多 increase]->[新聞 news]; [地名 place name]->[景點 attractions]->[道路 road]; [陸地 land]->[長度 length]->[景點 attractions]</p> <p>T10-SC: 人 human, 地名 place name, 國家 country, 事務 affairs, 景點 attractions, 地方 location, 方式 style, 物體 objects, 程度 degree, 道路 road</p> <p>T10-TW: 記者 reporter, 步道 trail, 遊客 tourist, 攝影 photography, 報導 report, 聯合報 United Daily News, 公園 park, 旅遊 travel, 提供 provided, 民眾 people</p>
Health (15845) AvLen: 355	<p>T5-SF: [導致 cause]->[疾病 disease]->[疾病 disease]->[疾病 disease]->[殺害 killing]->[傳染 infection]; [重要 important]->[氣 breath]->[口水 saliva]->[直接 direct]->[病人 patient]->[微生物 Microbiological]->[喉嚨 throat]->[排泄 excretion]->[傳染 infection]; [地名 place name]->[醫院 hospital]->[醫院 hospital]->[醫師 doctor]->[表示 express]; [幼兒 child]->[醫師 doctor]->[傳染 infection]->[微生物 Microbiological]; [疾病 disease]->[微生物 Microbiological]->[傳染 infection]->[禽 poultry]->[普通 common]</p> <p>T10-SC: 事務 affairs, 人 human, 國家 country, 疾病 disease, 物體 objects, 表示 express, 事件 event, 地名 place names, 醫師 doctor, 機構 institution</p> <p>T10-TW: 醫師 doctor, 治療 treatment, 可能 possible, 指出 noted, 發現 found, 報導 report, 醫院 hospital, 研究 research, 美國 USA, 患者 patients</p>
Education (15024) AvLen: 347	<p>T5-SF: [年級 grade]->[學生 student]; [學生 student]->[人 human]->[人 human]; [學生 student]->[指代 refers to]; [地名 place name]->[國小 elementary]->[老師 teacher]; [老師 teacher]->[人 human]</p> <p>T10-SC: 人 human, 事務 affairs, 國家 country, 地名 place name, 學生 student, 學校 school, 物體 objects, 機構 institution, 表示 express, 方式 style</p> <p>T10-TW: 大學 colleges, 學校 schools, 記者 reporter, 報導 report, 教育 education, 今年 this year, 國小 elementary, 老師 teacher, 教育部 Ministry of Education, 台北 Taipei</p>

Fig. 6 Statistics of our dataset. *T5-SF* top 5 semantic frames, *T10-SC* top 10 semantic classes, *T10-TW* top 10 topic words in LDA, *AvLen* average length of documents

names). Thus, under a fivefold cross-validation scheme, the increase in training data contributes to more representative frames for “Travel”. On the other hand, more training data does not help the frames in “Health” because of the existence of specific terms in this topic. Nevertheless, our method is robust in that it can achieve comparable performance regardless of the data split.

Moreover, we further conduct an experiment to evaluate performance with different number of semantic frames, as shown in Fig. 9. The SFTD can achieve the best performance when selecting the top 100 semantic frames with the highest sum of LLR values. We can see that the trends of F1-measure for topics “Health” and “Travel” are reversed when we select the top 100 frames, with an increase for “Health”

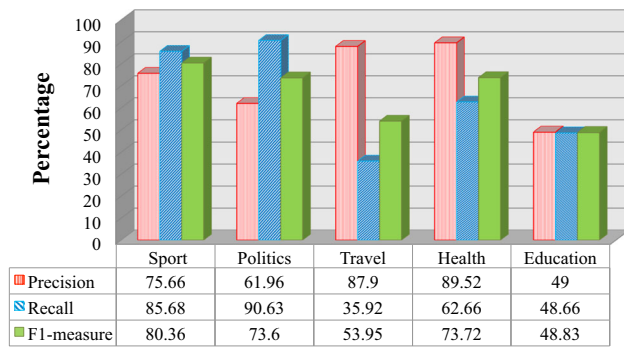


Fig. 7 Performance of our topic detection system on five topics

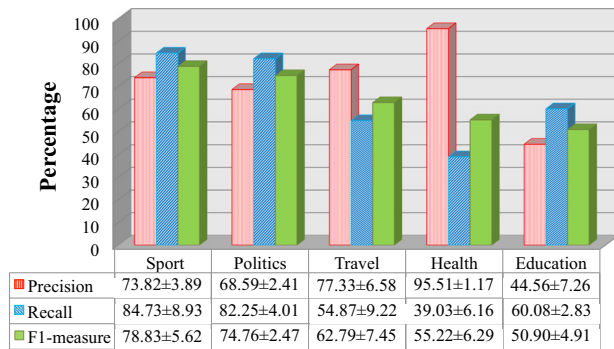


Fig. 8 Performance of SFTD via fivefold cross-validation

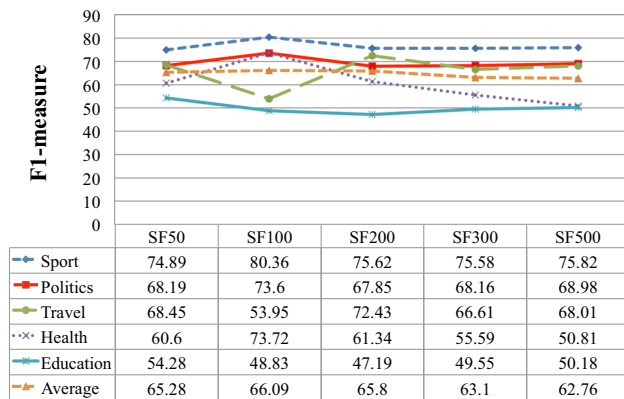


Fig. 9 Performance evaluation with different number of semantic frames

and decrease for “Travel”. But the overall performance fluctuation is not obvious. Thus, it shows that the number of frames does not necessarily affect the effectiveness of SFTD.

In addition, we compare our system to four other methods. Table 1 shows a comparison of different systems. As expected, the random baseline has the lowest performance across all methods, with PRF values of around 20 %. The Naïve Bayes classifier surpasses the random baseline by a considerable amount. VSM further outperforms the above-mentioned methods by around 20 %. As for the LDA and our SFTD, they have a comparable performance. In general, SFTD has higher precision and LDA has higher recall, which results in LDA having an overall higher F1-measure.

5 Discussion

The improvement from the random baseline to the Naïve Bayes classifier indicates that keyword information is crucial in detecting the topic of documents. The VSM benefits from weighing keywords in different topics by vectors to find unique words and leave out the less distinct ones in each topic, thus outperforms the Naïve Bayes classifier. However, since VSM considers similarity between two words as a cosine function with independent dimensions, it is difficult to represent relations among many words. On the other hand, when compared with LDA, our system has a higher precision and lower recall, which resulted in an overall lower F1-measure. It may be attributed to the use of Chinese word segmentation tool in LDA for constructing a word dictionary as background knowledge, in addition to a probabilistic graph with weighted edge representing between-word relations. In contrast, our system relies on topic-specific NE ontology for semantic class labeling and frame generation, which is constrained by the scope of the ontology. Moreover, some keyword information in the original document is discarded by the labeling process, which is retained in other keyword-based models. The missing crucial information can impair the coverage of our system. Despite the lower recall, our system is unique in that it can generate and accumulate knowledge during the process. We can capture crucial information beyond the

Table 1 Performance of five topic detection systems, presented with precision/recall/F1-measure

Topic	Random	Naïve Bayes	VSM	LDA	SFTD
Sport	25.90/19.59/22.19	76.61/59.65/67.07	94.76 /67.92/79.13	79.24/81.19/80.20	75.66/ 85.68 / 80.36
Politics	25.68/19.37/22.08	70.15/28.37/40.41	91.86 /48.69/63.65	73.15/89.11/ 80.35	61.96/ 90.63 /73.60
Travel	20.40/20.45/20.43	31.35/67.75/42.86	76.92/ 59.18 /66.89	80.82/57.67/ 67.31	87.90 /38.92/53.95
Health	13.72/19.16/15.99	53.56/35.26/42.52	57.49/78.92/66.31	72.10/ 90.78 / 80.37	89.52 /62.66/73.72
Education	12.97/19.50/15.58	19.73/49.76/28.25	29.04/70.08/41.07	46.26/ 77.78 / 58.01	49.00 /48.66/48.83
μ -Average	19.73/19.61/19.25	50.28/48.16/44.22	70.01/64.96/63.41	70.31/ 79.30 / 73.25	72.81 /65.31/66.09

Bold: the best across five systems

word-level for a topic over time, and generate frames that can capture the relations between them. Those generated semantic frames can describe the semantic relations within a document and assist in detecting the topic. We consider them as the foundation for a deeper understanding of topics that extends beyond the surface words.

Among the five topics, our system performs the best on the topic “Sports”. This is because there are plenty of specific nouns in the articles within this topic, such as “大聯盟 (MLB)” or “林書豪 (Jeremy Lin)”. In addition, unique sports terms like “先發 (Starter)” and “晉級 (Advance)” are also common. The integration of key terms and frames is the reason why semantic frames for the topic “Sports” are very stable and distinct, resulting in an overall higher F1-measure. As for “Politics”, we speculate that named entities of politicians and other organizations are common among articles about “Politics”. Thus, the frames in this topic are very extensive, which lead to a broader recall. Other methods that use only keyword information can achieve a higher precision. But, without long-distance information like those encoded by frames, the recall can be limited. Regarding the rest of the topics, although the SFTD can obtain the highest precision, the finite knowledge may be the cause of a restricted coverage. For example, the precision of the topic “Health” is 89.52 %, the highest among all five topics. We believe it is because specific terms are predominant in these topics, such as “Sarcoma (肉瘤)” in health-related articles or “日月潭 (Sun-moon Lake)” in travel-related ones. They are very competent in determining the topic of these documents. However, considering the fact that constructing the ontology of these fields requires extensive effort, we only include common entities in these topics. Consequently, the generated frames have limited length and scope. Nevertheless, under our framework, expanding and accumulating the knowledge base is easily done. Therefore, advancement of our system is foreseeable.

In sum, our approach can automatically generate frames that retain the benefit of knowledge-based approaches, such as high precision and knowledge accumulation, while retaining considerable amount of recall. It can continue to expand as more knowledge is incorporated into our resources. The SFTD is language independent in nature. Although we use Chinese for demonstration in this paper, it can be easily adopted for other languages such as English. For detecting topics of English documents, we only need to substitute language-specific knowledge sources for semantic class labeling. First, English NE ontology [e.g., Freebase (Bollacker et al. 2008)] can be adopted to label semantic class of named entities. The EHowNet has to be replaced by English lexical database (e.g., WordNet) to further label sense of all remaining words. We can then generate semantic frames using the aforementioned procedures. Therefore, our proposed SFTD is highly automated that integrates similar knowledge and reduces the total number of patterns through

pattern summarization. Hence, it has great potential in overcoming common disadvantages of other systems.

6 Concluding remarks

We propose a novel approach that utilizes knowledge sources and semantic frame generation for the topic detection task. It differs from popular machine learning methods as it can create a flexible and expansible topic-dependent knowledge base. Results showed that this approach can effectively detect the topic of articles, as well as assist the user in constructing background knowledge of each topic to better understand the essence of them. In the future, we will expand the ontology and include keyword information to improve the effect of semantic class labeling and frame generation. Moreover, we will reduce the human effort and rapidly broaden the coverage of the knowledge ontology through automatic construction. Furthermore, we will also modularize different mechanisms for the ease of use in other researches.

Acknowledgments This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 103-3111-Y-001-027.

References

- Alani H, Kim S, Millard DE, Weal MJ, Hall W, Lewis PH, Shadbolt NR (2003) Automatic ontology-based knowledge extraction from web documents. *Intell Syst IEEE* 18(1):14–21
- Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval: the concepts and technology behind search. Addison Wesley, New York
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proceedings of the 2008 ACM SIGMOD international conference on management of data*. ACM, pp 1247–1250
- Bun KK, Ishizuka M (2002) Topic extraction from news archive using tf*pdf algorithm. In: *international conference on web information systems engineering*. IEEE Computer Society, p 73
- CKIP (2009) An introduction to E-HowNet (E-HowNet technical report). Tech. rep, Academia Sinica
- Dong Z, Dong Q, Hao C (2010) HowNet and its computation of meaning. In: *Proceedings of the 23rd international conference on computational linguistics: demonstrations, association for computational linguistics*, pp 53–56
- García-Sánchez F, Martínez-Béjar R, Contreras L, Fernández-Breis JT, Castellanos-Nieves D (2006) An ontology-based intelligent system for recruitment. *Exp Syst Appl* 31(2):248–263
- Grineva M, Grinev M, Lizorkin D (2009) Extracting key terms from noisy and multitheme documents. In: *Proceedings of the 18th international conference on world wide web*. ACM, pp 661–670
- Hsu W, Chen Y, Wang Y (1998) A context sensitive model for concept understanding. In: *Proceeding of 3rd international conference on information theoretic approaches to logic, language, and computation*

- Lee CS, Jian ZW, Huang LK (2005) A fuzzy ontology and its application to news summarization. *IEEE Trans Syst Man Cybernet Part B Cybernet* 35(5):859–880
- Lee CS, Chang YC, Wang MH (2009) Ontological recommendation multi-agent for Tainan city travel. *Exp Syst Appl* 36(3):6740–6753
- Li S, Lv X, Wang T, Shi S (2010) The key technology of topic detection based on k-means. In: 2010 international conference on future information technology and management engineering (FITME), vol 2. IEEE, pp 387–390
- Lovász L (1993) Random walks on graphs: a survey. *Combinatorics, Paul erdos is eighty* 2(1):1–46
- Manning CD, Schütze H (1999) Foundations of statistical natural language processing, vol 999. MIT Press, Cambridge
- Nallapati R, Feng A, Peng F, Allan J (2004) Event threading within news topics. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, pp 446–453
- Scott S, Matwin S (1999) Feature engineering for text classification. *ICML (Citeseer)* 99:379–388
- Shih CW, Hsieh YL, Hsu WL (2014) Sense decomposition from e-hownet for word similarity measurement. In: The 3rd IEEE EM-RITE
- Tho QT, Hui SC, Fong ACM, Cao TH (2006) Automatic fuzzy ontology generation for semantic web. *IEEE Trans Knowl Data Eng* 18(6):842–856
- Wang MH, Lee CS, Hsieh KL, Hsu CY, Acampora G, Chang CC (2010) Ontology-based multi-agents for intelligent healthcare applications. *J Ambient Intell Humaniz Comput* 1(2):111–131
- Wu Y, Ding Y, Wang X, Xu J (2010) On-line hot topic recommendation using tolerance rough set based topic clustering. *J Comput* 5(4):549–556
- Zhang X, Wang T (2010) Topic tracking with dynamic topic model and topic-based weighting method. *J Softw* 5(5):482–489