Semantic Frame-based Statistical Approach for Topic Detection

Yung-Chun Chang[1,2] **Yu-Lun Hsieh**[2,3] **Cen-Chieh Chen**[2,3]
Chad Liu[2] **Chun-Hung Lu**[4] **Wen-Lian Hsu**[2]
[1]Department of Information Management, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Department of Computer Science, National Chengchi University, Taiwan
[4]Innovative Digitech-Enabled Applications & Services Institute, III, Taiwan
{changyc,morphe,can,hsu}@iis.sinica.edu.tw, [4]enricoghlu@iii.org.tw

Abstract

We propose a statistical frame-based approach (FBA) for natural language processing, and demonstrate its advantage over traditional machine learning methods by using topic detection as a case study. FBA perceives and identifies semantic knowledge in a more general manner by collecting important linguistic patterns within documents through a unique flexible matching scheme that allows word insertion, deletion and substitution (IDS) to capture linguistic structures within the text. In addition, FBA can also overcome major issues of the rule-based approach by reducing human effort through its highly automated pattern generation and summarization. Using Yahoo! Chinese news corpus containing about 140,000 news articles, we provide a comprehensive performance evaluation that demonstrates the effectiveness of FBA in detecting the topic of a document by exploiting the semantic association and the context within the text. Moreover, it outperforms common topic models like Naïve Bayes, Vector Space Model, and LDA-SVM.

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 of automatically detecting topics and tracking related documents from document streams such as online 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, which is an active research area in information retrieval (IR).

Linguistic information provides useful features to many natural language processing (NLP) tasks, including topic detection (Nallapati, 2003). Such information is usually represented as rules or templates. The main advantages of the rule-based approach are its high precision as well as the capability of knowledge accumulation. When confronting a new domain, they can be adapted by adding rules that exploit the missing knowledge. However, only a limited number of cases can be captured by a single rule, and increasing the number of rules could create undesired conflicts. Thus, the inflexibility of rule-based systems has put their competence for NLP tasks in doubt.

On the other hand, there are several machine learning-based approaches. For instance, Nallapati et al. (2004) attempted to find characteristics of topics by clustering keywords using statistical similarity. The clusters are then connected chronologically to form a time-line of the topic. Furthermore, many previous methods treated topic detection as a supervised classification problem (Blei et al., 2003; Zhang and Wang, 2010). These approaches can achieve substantial performance without much human involvement. However, to manifest topic as-

Copyright 2014 by Yung-Chun Chang, Yu-Lun Hsieh, Cen-Chieh Chen,
Chad Liu, Chun-Hung Lu, and Wen-Lian Hsu
28th Pacific Asia Conference on Language, Information and Computation pages 75–84

sociated features, one often needs to annotate the features in documents, which is rarely done in most machine learning models (Scott and Matwin, 1999). Those models have encountered bottlenecks due to knowledge shortage, data sparseness problem, and inability to make generalizations. Once the domain is changed, the models need to be re-trained 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.

In light of this, we propose a flexible frame-based approach (FBA), and use topic detection as a case study to demonstrate its advantages. FBA is a highly automated process that integrates similar knowledge and reduces the total number of patterns through pattern summarization. Furthermore, a matching mechanism allowing insertion, deletion, and substitution (IDS) of words and phrases is employed together with a statistical scoring mechanism. To create linguistic patterns with higher level of generality, we adopt the dominating set algorithm to reduce 350,000 patterns to a total of 500. Dominating set has been used extensively in network routing researches, e.g., Das and Bharghavan (1997), Du et al. (2013), and adopted in NLP related tasks such as text summarization (Shen and Li, 2010).

In the training phase, we consider keywords, context, and semantic associations to automatically generate frames. Thus, the obtained frames can be acknowledged as the essential knowledge for each topic that is comprehensible for humans. Results demonstrated that our method is more effective than the following approaches: the word vector model-based method (Li et al., 2010) and the latent Dirichlet allocation (LDA) method (Blei et al., 2003), a Bayesian networks-based topic model widely used to identify topics.

The structure of this paper is as follows. We discuss some of the previous work that apply statistical NLP methods to the topic detection problem in Section 2. Section 3 describes in detail the architecture and components of our system. Section 4 presents the performance comparison of various systems, and . Finally, we conclude our work in Section 5.

2 Related Work

Much work have been done on topic detection, or, a more general task like automatic text categorization. Most of them are concerned with the assignment of texts into a set of given categories, and rely on some measures of the importance of 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 VSM 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 also being exponentially proportional to the ratio of documents that contain the term in the channel itself.

Several researches have adopted machine learning-based approaches. Some formulate this task as a supervised classification problem (Blei et al., 2003; Zhang and Wang, 2010), in which a topic detection model is used to assign (i.e. classify) a topic to a document using a manually tagged training corpus. Nallapati et al. (2004) attempted to uncover characteristics of topics by clustering keywords using a statistical similarity measure into groups, each of which represents a topic. Wu et al. (2010) uses 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. The advantage of these methods is that they require little human involvement to acquire sizable outcome. However, they are faced with problems like data sparseness, knowledge accumulation, and the incapability to make generalizations. As we observed in the experiments, less than 1% of the keywords and semantic tags dominate the majority of the content. Thus, generalization of the surface words into a more abstract level, like the one in our approach,

can substantially decrease the sparseness. Moreover, the models of such approaches need to be re-trained or re-tuned to obtain satisfactory results when applying to a different domain. Such problem can be easily tackled in our approach by including more knowledge in the knowledge base. Besides, a more comprehensive linguistic knowledge can also be encoded and utilized in the proposed system. The hierarchical nature of our semantic features is necessary for a deeper understanding of the natural language.

One of the resources that is related to the organization of human knowledge is ontology. It is the conceptualization of a domain into a human understandable and machine-readable format consisting of entities, attributes, relationships, and axioms (Tho et al., 2006). It can also be used repeatedly, making it a very powerful method for representing domain knowledge. Ontology related applications have been involved 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. 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 Tainan City travel and further integrated fuzzy inference with ant colony optimization to recommend a personalized travel route that effectively meets the tourist's requirements to enjoy Tainan City. Some document detection methods made use of ontology and utilized the structured information in Wikipedia to enhance their performance (Grineva et al., 2009). Other ontologies like the WordNet may be included in the proposed system to further extend the scope of its knowledge.

Our method differs from existing approaches in a number of aspects. First, the FBA 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, 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.

3 System Architecture

We define the topic detection task as the following. Let $W = \{w_1, w_2, \cdots, w_m\}$ be a set of words, $D = \{d_1, d_2, \cdots, d_k\}$ be a set of documents, and $T = \{t_1, t_2, \cdots, t_n\}$ be a set of topics . Each document d is a set of words such that $d \subseteq W$. Our goal is to decide the most appropriate topic t_i for a document d_j, although one or multiple topics can be associated with each document. Our system mainly consists of three components, Semantic Class Labeling (SCL), Semantic Frame Generation (SFG), and Semantic Frame Matching (SFM), as shown in Figure 1. The SCL first uses prior knowledge of each topic to mark the semantic classes of words in the corpus. Then the SFG generates frames for each topic. These frames are stored in the topic-dependent knowledge base to provide domain-specific knowledge for our topic detection. During detection, an article is first labeled by the SCL as well. Then, the SFM applies an alignment-based algorithm which utilizes our knowledge base to calculate the similarity between each topic and the article to determine the main topic of this article. Details of these components will be explained in the following sections.

3.1 Semantic Class Labeling, SCL

First of all, the documents undergo the semantic class labeling process. Most Chinese topic detection researches rely on the error-prone word segmentation process. By contrast, our system labels words with their semantic classes, enabling us to extract representative semantic features. We adopt a novel labeling approach that utilizes various knowledge sources like dictionaries and Wikipedia. Since keywords within a topic are often considered as important information, we used the log likelihood ratio (LLR) (Manning and Schütze, 1999), an effective feature selection method, to learn a set of topic-specific keywords. Given a training dataset, LLR employs Equation (1) to calculate the likelihood of the assumption that the occurrence of a word w in topic T is not random. In (1), T denotes the set of documents of the topic in the training dataset;

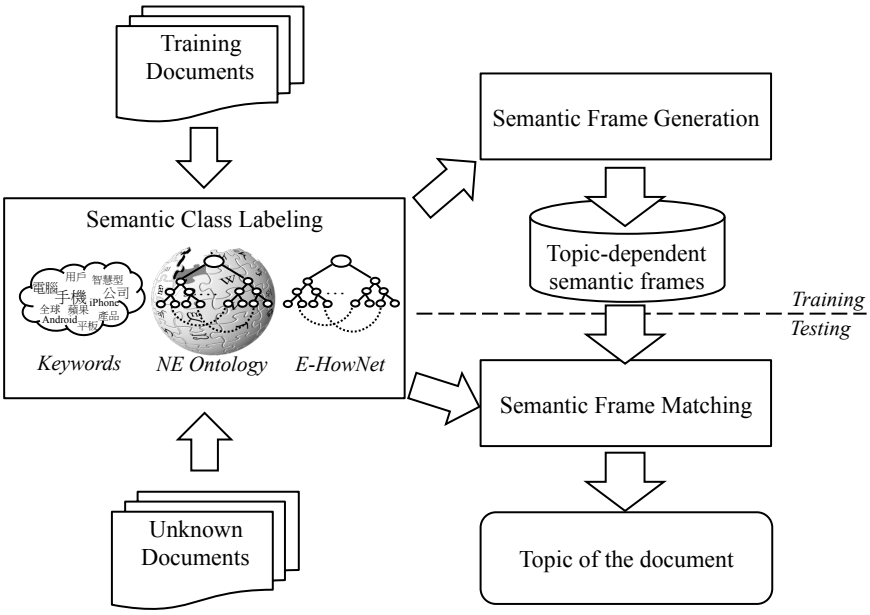

Figure 1: Architecture of our semantic frame-based topic detection system

$$-2log\left(\frac{p(w)^{N(w\wedge T)}(1-p(w))^{N(T)-N(w\wedge T)}p(w)^{N(w\wedge\neg T)}(1-p(w))^{N(\neg T)-N(w\wedge\neg T)}}{p(w|T)^{N(w\wedge T)}(1-p(w|T))^{N(T)-N(w\wedge T)}p(w|\neg T)^{N(w\wedge\neg T)}(1-p(w|\neg T))^{N(\neg T)-N(w\wedge\neg T)}}\right) \quad (1)$$

$N(T)$ and $N(\neg T)$ are the numbers of on-topic and off-topic documents, respectively; and $N(w \wedge T)$ is the number of document on-topic having w. The probabilities $p(w)$, $p(w|T)$, and $p(w| \wedge T)$ are estimated using maximum likelihood estimation. A word with a large LLR value is closely associated with the topic. We rank the words in the training dataset based on their LLR values and select the top 1,000 to compile a topic keyword list.

Recognizing named entities from text can facilitate document comprehension and improve the performance of identifying topics (Bashaddadh and Mohd, 2011). Therefore, we construct the Named Entity Ontology semi-automatically by using Wikipedia for semantic class labeling. Wikipedia category tags are used to label NEs recognized by the Stanford NER tools. We select the category tag to which the most *topic paths* are associated, and use them to represent the main semantic label of NEs in documents. Topic paths can be considered as the traversal from general categories to more specific ones. Thus, more topic paths may indicate that this category is more general. For example, Wikipedia has a page titled "勒布朗-詹姆斯(LeBron James)", and within this page, there are a number of category tags such as "邁阿密熱火隊球員(Miami Heat players)" and "美國籃球運動員(American basketball players)". For these two category tags, there are five and nine topic paths, respectively. Suppose "美國籃球運動員(American basketball players)" is the category with the most topic paths, our system will label "勒布朗-詹姆斯(LeBron James)" with the tag "[美國籃球運動員(American basketball players)]". In this way, we can transform plain NEs to a more general class, and increase the coverage of each label. In addition, we further integrated E-HowNet (Chen et al., 2005) to capture even richer semantic context. It is an extension of the HowNet (Dong et al., 2010) with the purpose of creating a structured representation of knowledge and semantics. It connects approximately 90 thousand words of the CKIP Chinese Lexical Knowledge Base and HowNet, and included extra frequent words that are specific to Traditional Chinese. It also contains a different formulation of each word to bet-

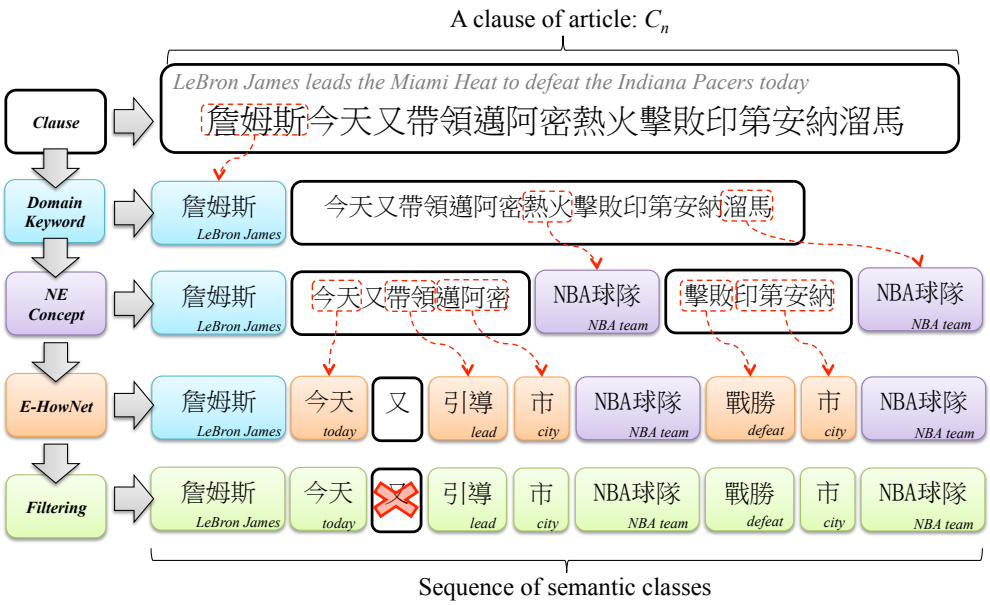

Figure 2: Semantic class labeling process

ter fit its semantic representation, as well as distinct definition of function and content words. 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 E-HowNet, let's take "手術 (Operation)" for example. It is defined as the following:

Simple Definition:
{affairs|事務: CoEvent = {開刀|HaveOperation}}
Expanded Definition:
{affairs|事務: CoEvent = {split|破開: purpose = {doctor|醫治}}}

We can see that the definitions in E-HowNet enable us to combine or dissect the meaning of words by using its semantic components. Therefore, we use it to label the remaining texts with their sense labels after all the NEs have been tagged.

To illustrate the process of SCL, consider the sentence C_n = "詹姆斯今天又帶領邁阿密熱火擊敗印第安納溜馬 (LeBron James leads the Miami Heat to defeat the Indiana Pacers again today)", as shown in Figure 2. First, "詹姆斯 (LeBron James)" is found in the keyword dictionary and tagged. Then, NEs like "熱火 (Heat)", "溜馬 (Pacers)" are found in NE ontology and tagged as "[NBA球隊 (NBA teams)]". Finally, other terms like "邁阿密 (Miami)", "今天 (today)", and "擊敗 (defeat)" are labeled with their corresponding E-HowNet senses. Evidently, the SCL can not only prevent errors caused by Chinese word segmentation, but also group the synonyms together. This enables us to generate distinctive and prominent semantic classes for a topic in the next stage.

3.2 Semantic Frame Generation, SFG

Semantic frame generation aims to automatically generate representative frames from sequences of semantic class labels and keywords. We observed that the rank-frequency distribution of semantic classes followed Zipf's law (Manning and Schütze, 1999), which was also the case for normalized frequency of semantic frames. Thus, we only used the most frequent 1,000 semantic frames ($\approx 0.5\%$) to dominate the tail of distribution. These frames can be regarded as the fundamental knowledge for a certain topic, and can be understood by computers as well as humans. Knowledge of such quality cannot be easily achieved in ordinary machine-learning models. To illustrate, consider the topic "Technology" and one of the automatically-acquired frames "[利用 (use)]-[iPhone (Tech-keyword)]-[看

(look)]-[網路術語 (Internet terminology)]". We can think of various semantically similar sentences that were covered by this frame, e.g., "使用 iPhone 來瀏覽部落格 (use iPhone to browse weblog)" or "善用 iPhone 察看電子郵件 (utilize iPhone to check email)".

The dominating set algorithm is adopted for SFG, and it has been proven that finding the dominating set on a graph is NP-hard (Garey and Johnson, 1979). Thus, several approximations have been proposed (Guha and Khuller, 1998; Kuhn and Wattenhofer, 2005; Shen and Li, 2010, i.a.). We also implemented an approximation based on the greedy algorithm. First of all, we construct a directed graph $G = \{V, E\}$, in which vertices V contains all semantic frames $\{SF_1, \cdots, SF_m\}$ in each topic, and edges E represent the dominating relations between frames. If a frame SF_x dominates SF_y, there is an edge $SF_x \rightarrow SF_y$. There are three criteria for constructing the dominating relations. First, only high frequency frames were selected for the dominators. Secondly, in general, longer frames dominate shorter frames, except for those mentioned in the following rule. Lastly, shorter frames would only be dominated if their head and tail semantic classes are identical to those of longer frames. The intermediate semantic classes could be skipped, as they can be identified as insertions and given scores based on their statistical distribution in this topic during the matching process. An illustration of a dominating frame and some dominated frames are shown in Table 1. Using dominating set to find frequent patterns on semantic graphs can help us capture the most prominent and representative frames within a topic. Afterwards, the dominating frames undergo a selection process that is similar to our keyword extraction method mentioned above. We use the LLR to discriminate semantic classes between topics. Given training data comprised of different topics, the LLR calculates the likelihood that the occurrence of a semantic class in the topic is not random. Those with a larger LLR value are considered as closely associated with the topic. Lastly, we rank the frames based on a sum of semantic classes LLR values and retain the top 100 from approx. 350,000 frames. By doing so, we can reduce the number of frames to 0.2% while keeping the most prominent and distinctive ones. Moreover, such reduction of the frames allows the execution of more sophisticated text classification algorithms, which leads to improved results. Existing algorithms cannot be executed on the original semantic class graph because the excessive execution times required makes them impractical (Baeza-Yates and Ribeiro-Neto, 2011). Therefore, selecting semantic frames closely associated with the topic would improve the performance of topic detection.

Dominating Frame:						
[player]	[team]	[person]	[player]	[news]		[speed]
Dominated Frames:						
-	[team]	-	[player]	-	[average]	[speed]
[player]	-	-	[player]	-	[attack]	[speed]
[player]	[equip]	[speed]	[player]	-	-	-
[player]	[team]	-	-	-	[attack]	[speed]
[player]	[team]	-	-	-	[attack]	[speed]
			⋮			
-	[team]	[person]	[player]	[news]	-	-
[player]	[team]	-	-	-	[average]	[speed]

Table 1: Illustration of a dominating frame and some dominated frames in the topic "Sports" generated by SFG.

3.3 Semantic Frame Matching, SFM

During matching, an unknown article is first labeled by SCL and a alignment-like algorithm (Needleman and Wunsch, 1970) is applied to determine the similarity between the article and the frames derived by SFG. It enables a single frame to match multiple semantically similar expressions. The SFM compares all sequences of semantic classes in an article to all the frames in each topic, and calculates the sum of scores for each topic. Unlike normal templates that involve mostly rigid left-right relation, we consider them as scoring criteria during frame alignment. The topic t_i with the highest sum of scores defined in (2) is considered as the winner.

$$Topic = \arg\max_{t \in Topic} Score(Document, t_i), \quad (2)$$

where

$$Score(Document, t_i) = \sum_{sf_i \in SF_{topic}, sl_j \in SL_{document}} \Delta(sf_i, sl_j) + LLR(k, t_i), \quad (3)$$

in which

$$\Delta(sf, sl) = \sum_i \sum_j \Delta(sf \cdot sc_i, sl \cdot sc_j), \quad (4)$$

where sc_i and sc_j represent the i^{th} semantic class of sf and j^{th} semantic class of sl, respectively. We use a keyword score computed from the LLR mentioned in Section 3.1, denoted as $LLR(k, t_i)$ in (3). As for scoring of the matched and unmatched components in frames, the details are as follows. If $sf \cdot sc_i$ and $sl \cdot sc_j$ are identical, we add a matched score obtained from the frequency of the semantic class in a topic times a normalizing factor $\lambda = 100$, as in (5).

$$Matched(sc) = \lambda \frac{f_{sc}}{\sum_{i=1}^{m} f_{sc_i}} \quad (5)$$

Otherwise, the score of insertions and deletions are added. An insertion, defined as (6), can be accounted for by the inversed entropy of this class, representing the uniqueness or generality of this class among topics. And a deletion, defined as (7), is computed from the log frequency of this class in this topic. It denotes the importance of a class in a topic. The detailed algorithm is described in Algorithm 1.

$$Insertion(sc) = -\frac{1}{\sum_{i=1}^{m} P(t_i) log_2\big(P(t_i)\big)} \quad (6)$$

$$Deletion(sc) = -log \frac{f_{sc}}{\sum_{i=1}^{m} f_{sc_i}} \quad (7)$$

Algorithm 1 Semantic Frame Matching

Input: A semantic frame $F = \{S_1, ..., S_m\}$, S: semantic class; A sequence of semantic class from a clause $C = \{s_1, ..., s_n\}$

Output: Matching score σ between F and C

```
pos ← 0;
for i = 1 to m do
    pos ← current matched position in C;
    if found s_j = S_i in C after pos then
        σ ← σ+ MatchedScore(S_i);
        isMatched ← true;
    end if
end for
if isMatched = false then
    σ ← σ-(insertion or deletion) score of S_i;
end if
```

4 Performance Evaluation

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 Yahoo! Chinese news website between the year 2010 and 2014. It contains a total of 140,000 documents with six different topics, and the number of documents of each topic is included in the parentheses, i.e., "Sports" (28,920), "Politics" (29,024), "Travel" (22,257), "Technology" (27,032), and "Education" (15,024). For each topic, 10,000 documents are selected as the training data, while the rest are used for testing. The evaluation metrics used are the precision, recall, and F_1-measure. A random baseline and three widely-used methods are also implemented and evaluated for comparison. The first is the Naïve Bayes classifier (Manning and Schütze, 1999), which is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features (denoted as Naïve Bayes). Another is a vector space model-based method (Salton et al., 1975) that is an algebraic model for representing text documents as vectors of identifiers (denoted as VSM). The last is a probabilistic graphical model which uses the LDA model as document representation to train an SVM to classify the documents as either topic relevant or irrelevant (Blei et al., 2003) (denoted as LDA-SVM). Details of these implementations are as follows. The dictionary required by Naïve Bayes, VSM and LDA-SVM is constructed by removing stop words according to a Chinese stop word list provided by Zou et al. (2006), and retaining tokens that make up 90% of the accumulated frequency. In other words, the dictionary can cover up to 90% of the tokens in the corpus. As for unseen events, we use Laplace smoothing in Naïve Bayes

and VSM, which is a common add-one smoothing method. And an LDA toolkit is used to perform the detection of LDA-SVM.

4.2 Results

A comparison of the five topic detection methods is displayed in Table 2. Our FBA system achieved the best performance on the topic "Politics", with the precision, recall, and F_1-measure scores of 78.37%, 92.12%, and 84.69%, respectively. Nevertheless, performances with high precision and low recall were found in the topics "Travel" and "Technology", as the FBA system obtained precisions over 90% with recalls only around 40%. On the contrary, the FBA system showed lower precisions of 57% and 72% and higher recalls of 95% and 93% for the topics "Sports" and "Health", respectively. Overall, the FBA system achieved an average precision of 78.17%, average recall of 69.39% and an average F_1-measure of 69.14%.

To further investigate the competence of our system, four other methods were also evaluated for comparison. As expected, the random baseline has the lowest performance among all methods with average P/R/F values around 17%. The Naïve Bayes classifier significantly outperforms the random baseline. Nevertheless, in the topics "Travel", "Technology", and "Education", this method obtained a relatively lower recall compared with others. On the other hand, VSM surpasses the overall performance of Naïve Bayes by about 20%. It is worth noting that VSM shares some of the low recall topics of the Naïve Bayes method, while acquiring the highest precision scores in three out of the six topics. For the topic "Technology", it has the best P/R/F scores of 93%, 50%, and 65%, respectively. As for the LDA-SVM, the difference is not as obvious. It achieved an improvement over the VSM's average F_1-measure by 4%. It also obtained the highest recalls among all systems in two of the six topics: "Travel" and "Education". Finally, the FBA outperforms LDA-SVM in the overall F_1-measure by 2%. In general, FBA has a higher precision while LDA-SVM has a higher recall, and FBA achieved the highest overall F_1-measure of all methods compared.

4.3 Discussion

To begin with, we provide an analysis of the difference in the average performance among different methods. The improvement in performance from the random baseline to the Naïve Bayes classifier indicates that keyword information is indispensable. The VSM benefits from weighing keywords in different topics by vectors in order to discover unique words and leave out less distinctive ones in each topic, thereby outperforming the Naïve Bayes classifier. However, since VSM considers similarity between two words as a cosine function with independent dimensions, it is difficult to represent the relations among many words.

On the other hand, when compared with the LDA-SVM method, our system has a higher precision and lower recall, resulting in a subtle increase of overall F_1-measure over the LDA-SVM. It may be attributed to the use of Chinese word segmentation tool in LDA-SVM for constructing a word dictionary as background knowledge, in addition to a probabilistic graph with weighted edge representing between-word relations. By contrast, our system relies on a NE database for semantic class labeling and frame generation, which is constrained by the scope of the data. Moreover, some keyword infor-

Topic	Random	Naïve Bayes	VSM	LDA-SVM	FBA
Sport	24.45/16.62/19.79	57.09/55.81/56.45	**94.76**/67.92/79.13	94.40/85.85/**89.92**	57.15/**95.06**/71.38
Politics	24.85/16.94/20.15	47.67/78.50/59.31	**91.86**/48.69/63.65	80.34/82.94/81.62	78.37/**92.12**/**84.69**
Travel	15.95/17.00/16.46	30.86/15.88/20.97	76.92/59.18/66.89	80.58/**62.11**/**70.16**	**91.06**/43.87/59.21
Technology	21.96/16.82/19.05	73.32/27.52/40.02	**92.87**/**50.39**/**65.33**	70.56/47.38/56.69	92.68/40.47/56.34
Health	10.28/16.26/12.59	38.43/69.65/49.53	57.49/78.92/66.31	44.41/70.56/54.51	**71.56**/**93.00**/**80.88**
Education	10.15/16.07/12.44	46.88/46.50/46.69	29.04/70.08/41.07	37.18/**82.06**/51.17	**78.19**/51.82/**62.33**
μ-Average	17.94/18.29/16.75	49.04/48.98/45.50	73.82/62.53/63.73	67.91/**71.82**/67.35	**78.17**/69.39/**69.14**

Table 2: Precision/Recall/F_1-measure(%) and micro-average of different topic detection systems. The highest numbers among all systems are in bold.

mation in the original document is discarded by the labeling process, which is retained in other keyword-based models. Potentially crucial information may be abandoned in this manner and impair the coverage of our system. Despite the slightly lower recall, our system is unique in the ability to generate and accumulate knowledge during the process. This enables us to capture essential information beyond the word-level for a topic, and generate frames that can capture the relations between them. The generated frames can describe the semantic relations within a document and assist in detecting the topic. We consider them as the foundation for a more profound understanding of topics that extends beyond the surface words.

Of the six topics, our system performed best on the topic "Politics" due to the abundant specific nouns in the articles of this topic, such as "民主黨 (Democratic Party)" or "歐巴馬 (Obama)". In addition, unique political terms like "參議員 (Senator)" and "內閣 (Cabinet)" are also common. The integration of key terms and frames contributes to the stability and uniqueness of the semantic frames of this topic, resulting in a higher overall F_1-measure. As for the topics "Sports" and "Health", we speculate that the NEs of athletes or disease names and other organizations are common among these articles. Thus, the frames in these topics are very extensive, leading to a broader coverage and higher recall. Other methods simply relying on keyword information can achieve a higher precision. Nonetheless, without long-distance information such as those encoded by frames, the recall can be limited. Regarding other topics, although the FBA can obtain the highest precision, insufficient knowledge may be the major cause of a restricted coverage. For example, the precision of the topic "Technology" is 92.68%, the highest among all topics. We believe this is due to the fact that specific technological terms, such as "iPhone" or "微軟 (Microsoft)", are predominant in these topics. Terms of such are very competent in determining the topic of these documents. However, considering the fact that novel terms are emerging frequently, we will have to integrate new knowledge into our system. Fortunately, under our framework, expanding and accumulating the knowledge base is easily done. Therefore, the advancement of our system is foreseeable.

Interestingly, it can be observed that the topics "Travel" and "Technology" generally have lower recall, regardless of the system used. This may be due to the fact that context information in these topics is hard to be captured by the current systems. Using only the word it self or word-related features is not enough. Even for a semantically-based system like the LDA-SVM or FBA, such information is still not fully encoded. Further research on the integration of richer and wider semantic context may be fruitful.

In sum, our approach can automatically generate frames that retain the benefit of knowledge-based approaches, including high precision and knowledge accumulation, while retaining considerable amount of recall. It can be continuously upgraded as more knowledge is incorporated. Hence, it has great potential in overcoming common disadvantages of other systems.

5 Concluding Remarks

This research proposes the FBA, a flexible and automatic approach to the topic detection task based on knowledge sources and automatic frame generation. It differs from popular machine learning methods as it can create an adaptable and extensible topic-dependent knowledge base, while preserving the accuracy of rule-based models. Results showed that FBA can effectively detect the topic of articles, as well as assist the user in constructing background knowledge of each topic in order to better understand the essence of them. In the future, we plan to expand this approach to include more topics, and even apply it to other applications in NLP. Also, further studies can be done on combining statistical models into different components in FBA.

Acknowledgment

This study is conducted under the NSC 102-3114-Y-307-026 "A Research on Social Influence and Decision Support Analytics" of the Institute for Information Industry which is subsidized by the National Science Council.

References

Harith Alani, Sanghee Kim, David E Millard, Mark J Weal, Wendy Hall, Paul H Lewis, and Nigel R Shadbolt. 2003. Automatic ontology-based knowledge

extraction from web documents. *Intelligent Systems, IEEE*, 18(1):14–21.

R. Baeza-Yates and B. Ribeiro-Neto. 2011. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

Omar Mabrook A Bashaddadh and Masnizah Mohd. 2011. Topic detection and tracking interface with named entities approach. In *Semantic Technology and Information Retrieval (STAIR), International Conference on*, pages 215–219. IEEE.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Khoo Khyou Bun and Mitsuru Ishizuka. 2002. Topic extraction from news archive using tf*pdf algorithm. In *Web Information Systems Engineering, International Conference on*, page 73. IEEE Computer Society.

Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, and Yi-Jun Chen. 2005. Extended-HowNet: A representational framework for concepts. In *Proc. 2nd IJCNLP*.

Bevan Das and Vaduvur Bharghavan. 1997. Routing in ad-hoc networks using minimum connected dominating sets. In *Proc. ICC'97, Towards the Knowledge Millennium.*, volume 1, pages 376–380. IEEE.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. Hownet and its computation of meaning. In *Proc. the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 53–56. ACL.

Hongjie Du, Ling Ding, Weili Wu, Donghyun Kim, PanosM. Pardalos, and James Willson. 2013. Connected dominating set in wireless networks. In Panos M. Pardalos, Ding-Zhu Du, and Ronald L. Graham, editors, *Handbook of Combinatorial Optimization*, pages 783–833. Springer New York.

Francisco García-Sánchez, Rodrigo Martínez-Béjar, Leonardo Contreras, Jesualdo Tomás Fernández-Breis, and Dagoberto Castellanos-Nieves. 2006. An ontology-based intelligent system for recruitment. *Expert Systems with Applications*, 31(2):248–263.

Michael R Garey and David S Johnson. 1979. *Computers and intractability: A Guide to the Theory of NP-Completeness*. Freeman San Francisco.

Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. 2009. Extracting key terms from noisy and multitheme documents. In *Proc. the 18th International Conference on World Wide Web*, pages 661–670. ACM.

Sudipto Guha and Samir Khuller. 1998. Approximation algorithms for connected dominating sets. *Algorithmica*, 20(4):374–387.

Fabian Kuhn and Roger Wattenhofer. 2005. Constant-time distributed dominating set approximation. *Distributed Computing*, 17(4):303–310.

Chang-Shing Lee, Young-Chung Chang, and Mei-Hui Wang. 2009. Ontological recommendation multi-agent for tainan city travel. *Expert Systems with Applications*, 36(3):6740–6753.

Shengdong Li, Xueqiang Lv, Tao Wang, and Shuicai Shi. 2010. The key technology of topic detection based on K-means. In *Future Information Technology and Management Engineering (FITME), International Conference on*, volume 2, pages 387–390. IEEE.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*, volume 999. MIT Press.

Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. 2004. Event threading within news topics. In *Proc. the 13th CIKM*, pages 446–453. ACM.

Ramesh Nallapati. 2003. Semantic language models for topic detection and tracking. In *Proc. HLT-NAACL Student Research Workshop*, pages 1–6.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sam Scott and Stan Matwin. 1999. Feature engineering for text classification. In *Proc. ICML-99, 16th International Conference on Machine Learning*, pages 379–388. Morgan Kaufmann Publishers.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proc. the 23rd International Conference on Computational Linguistics*, COLING '10, pages 984–992, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quan Thanh Tho, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. 2006. Automatic fuzzy ontology generation for semantic web. *Knowledge and Data Engineering, IEEE Trans. on*, 18(6):842–856.

Yonghui Wu, Yuxin Ding, Xiaolong Wang, and Jun Xu. 2010. On-line hot topic recommendation using tolerance rough set based topic clustering. *Journal of Computers*, 5(4).

Xiaoyan Zhang and Ting Wang. 2010. Topic tracking with dynamic topic model and topic-based weighting method. *Journal of Software*, 5(5):482–489.

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, and Lu Sheng Wang. 2006. Automatic construction of chinese stop word list. In *Proc. the 5th WSEAS International Conference on Applied Computer Science*, pages 1010–1015.