Expert Systems with Applications 37 (2010) 341–347

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling

Sheng-Hao Hung [a], Chia-Hung Lin [b], Jen-Shin Hong [a,*]

[a] Department of Computer Science and Information Engineering, National ChiNan University, Taiwan
[b] Department of Management Sciences, R.O.C. Military Academy, Taiwan

ARTICLE INFO

Keywords:
Web mining
Commonsense knowledge retrieval
Event-based knowledge
Semantic role labeling
Is capable of

ABSTRACT

A sophisticated commonsense knowledgebase is essential for many intelligent system applications. This paper presents a methodology for automatically retrieving event-based commonsense knowledge from the web. The approach is based on matching the text in web search results to designed lexico-syntactic patterns. We apply a semantic role labeling technique to parse the extracted sentences so as to identify the essential knowledge associated with the event(s) described in each sentence. Particularly, we propose a semantic role substitution strategy to prune knowledge items that have a high probability of erroneously parsed semantic roles. The experimental results in a case study for retrieving the knowledge is "capable of" shows that the accuracy of the retrieved commonsense knowledge is around 98%.

1. Introduction

Commonsense knowledge refers to beliefs or propositions that appear to be obvious to most people, without dependence on any specific esoteric knowledge. Examples include "a cat is an animal," "dog is a canine," "a dog is capable of eating," "fire is hot," etc. By further incorporating certain inference algorithms, commonsense knowledge can be widely applied in real life intelligent systems, such as agents, expert systems, dialog understanding systems, etc. For decades, there has been a thirst in the artificial intelligence research community for a large-scale commonsense knowledgebase.

To build a large-scale commonsense knowledgebase is an extremely complicated task and usually requires investing enormous amounts of human effort. In the literature, a number of projects have been conducted to construct such a large-scale commonsense knowledgebase, including WordNet (Miller, 1995), Cyc project (Lenat, 1995), ConceptNet (Liu & Singh, 2004), etc. WordNet is one of the most popular and widely used lexical resources today. It is a collection of English words organized into discrete senses that refer to particular interpretations of a word in a particular context. The collected words are interlinked by a small set of semantic relations, including hypernyms (IsA), hyponyms (reverse of IsA), meronym (part–whole) relationships, etc. Although WordNet appears to be an easy to use lexico resource with high precision entries, the limited number of semantic relations provided makes it difficult to apply to many real life applications that require richer semantic resources. The scope of such a desired semantic resource certainly needs to extend beyond primitive lexical knowledge to encompass more general commonsense knowledge.

The Cyc knowledgebase is specifically designed for artificial intelligence applications, which require fine-grained and sophisticated inferences. The Cyc project, begun in 1984, aims to construct a large-scale commonsense knowledgebase for general artificial intelligence applications. Cyc formalizes commonsense knowledge into a logical framework expressed by the "CycL," which attempts to produce unambiguous logical formulations for human knowledge (Lenat & Guha, 1991). The concepts and semantic relations in the Cyc project are created by human experts in a rigorous way with the goal of providing a solid foundation of sophisticated inferences for application systems. However, due to the complexity of human commonsense, much work is still needed before this knowledgebase is finished. So far, the unavailability of its full content to the public makes it a prohibitively difficult option for real life applications.

ConceptNet is the knowledgebase created in the OMCS project (Liu, Lieberman, & Selker, 2002). The ConceptNet knowledgebase is built from the contributions of thousands of volunteers across the web. It includes a wide range of commonsense concepts interlinked by about twenty semantic relations. The knowledge structure allows a variety of primitive inferences, such as temporal, spatial, affective, etc. Roughly speaking, the relations in ConceptNet can be classified into two categories, including "entity-based knowledge" and "event-based knowledge." Entity-based knowledge, for example "a cat is an animal," refers to certain attributes of an individual entity, or between two entities. The IsA, propertyOf, partOf, MadeOf, and OftenNear in ConceptNet belong to this category. On the other hand, event-based knowledge, for example, "a dog chases a cat," refers to knowledge associated with a

* Corresponding author. Tel.: +886 49 2915225; fax: +886 49 2915226.
E-mail address: jshong@ncnu.edu.tw (J.-S. Hong).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.060

particular event (i.e., action). The UsedFor, CanDo, CapableOf belong to this category. By aggregating two primitive events, more complicated event-based knowledge, such as FirstSubeventOf, LastSubeventOf, EffectOf in ConceptNet, can be achieved. The abundance of the semantic relations supported in ConceptNet makes it of practical value in many applications on intelligent systems (Liu, Lieberman, & Selker, 2003, 2002; Liu, 2002; Liu & Singh, 2002; Lieberman & Liu, 2002).

Overall, both Cyc and ConceptNet projects require significant amounts of human effort to develop the knowledgebase. As compared to the immense scope of general human commonsense knowledge, the coverage of such a manually authored knowledgebase is severely restricted. How to efficiently accumulate knowledge using a more automatic mechanism is worthwhile to explore. To this end, over the past few years, automatic knowledge retrieval using "text mining" approaches has attracted much attention in researches involving question answering, information extraction, text summarization, text understanding, and many other areas. The most commonly used technique for text mining-based knowledge retrieval is the "lexico-syntactic pattern matching" technique. We refer to such an approach as a "pattern-based" approach for knowledge retrieval. The basic idea of the pattern-based technique is to search for sentence segments in a corpus that match a given lexico-syntactic pattern representing a desired semantic relation. For example, a pattern "NP_1 such as NP_2, NP_3, NP_4 ... etc." often indicates that the entity NP_2, NP_3, and NP_4, is a member of the class NP_1. Based on this technique, various projects have been conducted to extract different kinds of knowledge, including hypernym, hyponym, part–whole relations, named entities, and many others (Berland & Charniak, 2002; Etzioni et al., 2005; Girju, Badulescu, & Moldovan, 2006; Hearst, 1992).

Our research aims to develop automatic techniques that can be applied to accumulate commonsense knowledge automatically to enrich ConceptNet. Such automatic procedures provide the only credible prospect for acquiring world knowledge on the scale needed to support real life applications. The focus of this study is to develop practical web-scale text mining techniques for retrieving event-related commonsense knowledge. In principle, given the broad coverage of the web document repositories, pattern-based commonsense knowledge retrieval could be conducted rather effectively. However, for retrieving event-based knowledge, the data redundancy in the massive collections of available web pages means that a particular event is likely to be phrased in many different ways in differing contexts and documents. How to extract different kinds of event-based commonsense knowledge out of free texts with high precision is a significant challenge. We propose an approach integrating the lexico-syntactic pattern matching technique and the semantic role labeling techniques (SRL hereafter, (Gildea & Jurafsky, 2002)) to retrieve and identify essential elements associated with an event. A case study on the knowledge of "is capable of" will be conducted to justify the applicability of the proposed approach.

This paper is organized as follows: In Section 2, the conventional text mining techniques for knowledge retrieval will be discussed. Section 3 describes the overall framework of the event-based commonsense knowledge retrieval process. A case study on the retrieval of the knowledge of "is capable of" will be described. Experiments for evaluating the proposed methodology will be presented in Section 4. Section 5 concludes the study.

2. Knowledge retrieval using text mining approaches

Over the past few years, there have been a significant number of studies attempting to automatically retrieve knowledge using text mining approaches. The purpose is to automatically find the relationships between concepts (a lexical item or phrase) so that the process of building semantic resources can be fully or partially automated. Many of the studies retrieve knowledge from certain machine-readable dictionaries (MRD) (e.g., Ahlswede & Evens, 1988; Alshawi, 1987; Byrd et al., 1987; Jensen & Binot, 1987; Klavans, Chodorow, & Wacholder, 1990; Markowitz, Ahlswede, & Evens, 1986; Nakamura & Nagao, 1988). The commonsense knowledgebase MindNet (Richardson, Dolan, & Vanderwende, 1998) was constructed by such a dictionary-based text-mining approach. These studies retrieved semantic relations between pairs of concepts by matching the sentences in dictionaries to a set of handcrafted lexico-syntactic patterns. The semantic relations that have been investigated in these studies include synonym, hypernym, hyponym, location, material, part_of, etc. In general, dictionary definitions are written mostly based on certain variations on a manageably small number of basic syntactic forms. This property of definitions makes the pattern-based approach particularly applicable. Overall, projects that center on extracting knowledge from MRDs have shown much success but are inherently limited, since the amount of entries within a dictionary is restricted.

In order to increase the scope of the coverage of the commonsense knowledge, many studies turned to the use of more large-scale free-text resources, especially the web. Most of these approaches are similar in spirit to the MRD-based knowledge mining approaches. Hearst (Hearst, 1992) described a method for the automatic acquisition of the hyponym lexical relation from unrestricted text using a small set of hand-crafted lexico-syntactic patterns. A number of studies have been conducted to retrieve "part–whole" (meronym) relations, for example (Berland & Charniak, 2002; Girju et al., 2006). Etzioni et al. (2004, 2005) used the pattern-based technique to retrieve named entities. As compared with the structured presentation of dictionary definitions, free text is extremely varied in syntactic form. The target knowledge may often be phrased in many different ways in different documents. Therefore, the difficulty for web-scale text mining lies in finding lexico-syntactic patterns that reliably indicate the semantic relation of interest while keeping the noise manageable.

In addition, there are a number of problems with the pattern-based approach to knowledge extraction from the web, which often results in the extraction of semantic relations that are incorrect or worthless. An obvious problem is that caused by erroneously written text on the web. For example, sentences or phrases such as "The dog can fly" or "Dolphin's wing" appear on the web. These texts are often syntactically correct but unfortunately are not compliant with general commonsense. The automatic detection and correction of this type of error usually requires human-level natural language understanding or more sophisticated world knowledge, which is beyond the capabilities of current natural language processing technologies. Another problem associated with the pattern-based approaches is that knowledge in a single sentence many often require certain "contextual information" to identify the concepts that the sentence is referring to. For example, in a sentence such as "*A tool such as a compiler, assembler, or debugger is essential for a student to learn basic computer skills*," it is difficult for the computer to know that the *compiler, assembler, debugger* in this context is referring to a type of programming language tool.

Furthermore, one other commonly found type of noise in pattern-based knowledge mining approaches is caused by a situation where a particular syntactic pattern may actually refer to a number of different semantics. For example, for retrieving a meromym using a pattern matching approach, there are a number of traps due to the many ambiguities in certain syntactic patterns in a sentence. Examples of these ambiguities include genitive constructions (Mary's hand vs. Mary's novel), noun compounds (spoon handle vs. linen bag), verbs (Kate has a sister vs. Kate has green eyes), etc. (Girju et al., 2006). As another example, patterns indicating a hyponym are often used for other purposes. For instance, the pattern "x including y_1,y_2,..., and y_n" which often indicates a hyponym, can also be used instead to indicate "group" membership relations (e.g., "*A family including parents and children need...*). A number of techniques based on the semantic relationships between the extracted terms have been proposed to filter out such noise. For example, Girju et al. presented a learning method to discriminate whether or not a pattern contains a part–whole relation (Girju et al., 2006). Cederberg and Widdows applied a technique for filtering out many incorrect and spurious extracted relations using a variant of latent semantic analysis (Cederberg & Widdows, 2003). Etzioni et al automatically evaluated the plausibility of the candidate knowledge it extracts using computed *pointwise mutual information* (PMI) statistics (Etzioni et al., 2004, 2005). The above techniques use semantic relationships between the extracted terms to filter out the possible noise for a syntactic pattern. Their experimental results show improvement in the accuracy.

For all the success of the previous knowledge mining studies, the direct application of previously developed pattern-based approaches to automatically extract "event-based" knowledge is nevertheless more difficult. A conceptual level description of an "event" often consists of a set of event-related conceptual characteristics in a scene, such as the subject, action, object, place, time. In an English sentence, a verb (predicate) indicates an event and expresses the relationship between entities. The verb's syntactic arguments generally are associated with the participants of the event. A "semantic role" is the relationship that a syntactic argument has with the verb. Since an event can be referenced using varieties of lexical items with different syntactic realizations, it is often difficult to find a clear-cut segment in a sentence to completely represent the associated knowledge of the event. As such, it is difficult to directly apply a conventional pattern-based approach alone to extract event-based knowledge.

Many studies within the information extraction community have been devoted to the extraction of event-based knowledge. Detecting events automatically and obtaining a semantic representation for them would be equivalent to creating a "Who did what to whom when and where" interpretation of the text. The event extraction task was one of the main tasks in the Automatic Content Extraction (ACE) (ACE, 2003) program and the Message Understanding Conference (MUC) that preceded it (MUC, 2008). Various studies aimed to retrieve the relationships between participants, times, and locations in events, and label the extracted events to reflect those relationships. Aone and Ramos-Santacruz (2000) presented a scalable event extraction system that relies on text matching between sentences and a bag of pre-defined lexico-syntactic patterns that specify the syntactic and semantic restrictions on the verb's arguments for the target event. The lexico-syntactic patterns were obtained manually from examples collected by knowledge experts. The provision of manually created rules for different events can be a tedious task since usually too many patterns among the events are observed. In practice, due to the versatile syntactic realizations of the events, it is rather difficult to come up with a thorough list of syntactic patterns that give a satisfactory recall rate.

Various studies developed machine learning or probability based techniques to automatically generate the syntactic patterns for a number of specific events (e.g., Filatova & Hatzivassiloglou, 2003; Xu, Uszkoreit, & Li, 2006). Most of these approaches to event extraction focused on mentions textually anchored by verbs. However, many mentions of events surface as noun phrases. For example, events can be expressed with a single noun phrase such as "war" or "strike." (Creswell, Beal, Chen, Cornell, & Nilsson, 2006) developed a supervised method for detecting nominal event mentions that increase the recall of event extraction. These studies aimed to develop techniques that can detect specific events that might be distributed among different sentences, even in different articles. A high recall rate of event extraction is preferred in these studies. Although these techniques show reasonable performances, the main difficulty in applying them in commonsense knowledge retrieval applications is the lack of labeled sentences or documents for events with a widely varying scope. In addition, these techniques usually require the use of sophisticated natural language processing techniques (such as syntactic parsing, named entity recognition, conference resolution techniques (Kehler, Appelt, Taylor, & Simma, 2004; Ng & Cardie, 2002; Ponzetto & Strube, 2006), etc.), as well as machine learning packages. Due to the lack of public-available and properly integrated tools, they might not be readily applicable in the application of event-based commonsense extraction within the community of intelligent system developers.

Aiming to retrieve event-based commonsense knowledge from the web, this study integrates a pattern-based knowledge retrieval approach with a semantic role labeling technique to identify the semantic roles of the event(s) in the retrieved sentence. The proposed methodology for sentence-level atomic event-based knowledge can serve as a foundation to facilitate the discovery of more complicated event-based knowledge. Since at this stage there exists no semantic role labeling engine that can achieve very high precision semantic role identification, we specifically develop a "semantic role plausibility verification technique" to prune knowledge items that potentially have a high probability of erroneously parsed semantic roles. The following section presents the methodology we propose.

3. Methodology

To efficiently and accurately extract event-based knowledge from web documents, we propose a framework based on lexico-syntactic pattern matching and semantic role labeling techniques. The core processes of the framework include raw sentence extraction, semantic role identification, semantic role plausibility verification, and knowledge distillation. In the first stage, raw sentences are crawled through web search engines using a variety of lexico-syntactic patterns. The raw sentences are then parsed using a SRL engine to extract the semantic roles associated with the event(s) in each sentence. The semantic roles of each event form a "knowledge item" that possibly contains the desired knowledge. To raise the accuracy rate of the semantic role labeling, each knowledge item is further processed using a "semantic role substitution" technique to verify the plausibility of its semantic roles. A knowledge item that fails to pass the process of semantic role plausibility verification is discarded. In the final stage, various heuristic knowledge-filtering rules are applied to remove knowledge items that require contextual information or refer to specific events other than commonsense knowledge. In the following, we elaborate the underlying principles and detail operations for the above core processes. An evaluation will be given in Section 4 to verify the applicability of the proposed approach. We use an illustrative example for extracting knowledge of "*is capable of*" to detail the proposed knowledge mining process. Fig. 1 shows the overall picture of the proposed framework.

3.1. Stage 1: raw sentence extraction based on lexico-syntactic pattern matching

In the first stage, the purpose is to collect a large number of raw sentences possibly containing the target knowledge through web search engines. Web queries are formulated based on a set of lexico-syntactic patterns. A query response from a web search engine often includes a list of snippets containing the specified lexico-syn-

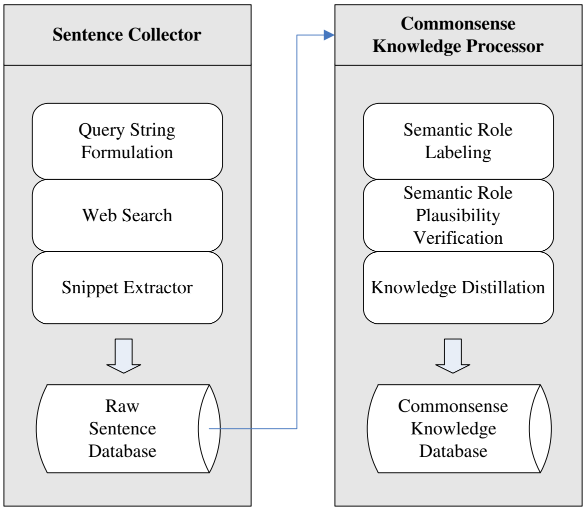

Fig. 1. Overall framework of the knowledge extraction process.

tactic patterns. From the snippets, sentences that appear to be syntactically complete are extracted for subsequence processing, while those without correct punctuation to mark the endings are discarded.

The design of the lexico-syntactic patterns is one of the most critical processes for the success of a typical knowledge mining process. In principle, automatic discovery of a semantic relation must start with a thorough investigation of the lexical terms and syntactic forms used to reliably express that target knowledge. For the retrieval of commonsense knowledge of "*is capable of*," obvious lexico-syntactic patterns that potentially identify the capability of the subject at least includes

(1) subject + "*is capable of*" + *verb* (present participle) (e.g., "A soldier is capable of fighting the enemy")
(2) subject + "*is able to*" + *verb* (e.g., "The soldier is able to fire the gun")
(3) subject + *modal verbs* (can, will, should) + *verb* – (e.g., "A soldier can fight the enemy.")

Note that patterns 1 and 2 unambiguously identify the commonsense knowledge of "*is capable of*." However, it appears that not many sentences with these patterns appear on the web. Therefore, this study also applied patterns 3 to crawl for more raw sentences. In pattern 3, expressions with modal verbs such as "can" and "should" are ambiguous between at least three kinds of modality: deontic (used to affect a situation, by giving permission), epistemic (express the speaker's opinion about a statement) and dynamic (describing a factual situation about the subject of the sentence). In principle, only the dynamic modality unambiguously identifies the capability of the subject of the sentence. Nevertheless, based on our observation of the web search results, it appears that most of the crawled sentences belong to the dynamic modality. Therefore, we have not addressed any mechanism to filter out those with denotic or epistemic modalities.

3.2. Stage 2: semantic role identification

In an English sentence, a semantic role is the relationship that a syntactic argument has with the verb. One of the most commonly used schemes for specifying semantic roles is the PropBank annotation scheme (Palmer, Kingsbury, & Gildea, 2005). In the PropBank annotations, the arguments of a verb are labeled sequentially from ARG0 to ARG5, where ARG0 is usually the subject of a transitive verb; ARG1, its direct object, etc. A variety of adjunctive arguments, such as ARGM-LOC, for locative information, and ARGM-TMP, for temporal information, are also tagged. For each crawled sentence, the semantic roles of it are kept in a database as a knowledge item. For a sentence with multiple verbs, the associated semantic roles for different verb are regarded as distinct knowledge items. As an illustrative example, the semantic roles for the sentence "The dog barked at a cat in the park last night" based on the PropBank style markup is given as:

$[_{\text{ARG0}}\text{The dog}]\ [_{\text{Verb}}\text{bark}]\ [_{\text{ARG1}}\text{a cat}]\ [_{\text{ARGM-LOC}}\text{in the park}]\ [_{\text{ARGM-TMP}}\text{last night}]$

Semantic role labeling techniques automatically identify the semantic roles of a sentence. In principle, automatically tagging the semantic roles with high precision is difficult since an event can often be referenced using a variety of lexical items with different syntactic realizations. In the literature, there are a number of studies proposing different methodologies for this purpose, for example (Gildea & Jurafsky, 2002; Koomen, Punyakanok, Roth, & Yih, 2005; Pradhan, Ward, Hacioglu, Martin, & Jurafsky, 2004). However, none of these techniques could possibly achieve a significantly high accuracy for identifying all the semantic roles in domain-dependent free-text sentences. The performances usually get worse when the sentences to be parsed have rather distinct linguistic writing styles compared to the sentence in the annotated training corpus. This could cause a problem if we intend to use an SRL tool to parse the sentences extracted from the web, where various writing styles exist. The following stage provides a noise reduction strategy to prune knowledge with wrong semantic roles.

3.3. Stage3: semantic role plausibility verification using "semantic role substitutions"

Since the current state-of-the-art SRL techniques cannot achieve perfect semantic role identification, to ensure a high precision rate for the common sense knowledge retrieved, we propose a technique to prune knowledge items that contain semantic roles which are very likely to be erroneously parsed. For web-scale text mining, since we typically accumulate quite a large number of raw sentences, the proposed noise reducing strategy naturally emphasizes the precision rate rather than the recall rate (which actually is much difficult to measure in such a context).

Most existing SRL tools are based on machine learning techniques with which various low level syntactic features extracted from annotated corpus are used as the training set. These tools feed the training data into a typical machine learning package for predicting the semantic roles of possible unseen sentences. There are a number of issues that may affect the accuracy of the prediction, at least including the features selected, the quality and quantity of the training sets, and the performance of the classifier applied. In practice, the applied manually-labeled sentences in the training set may not cover a sufficiently wide spectrum of the lexical terms and syntactic patterns in typical English writing. In many instances, the parsed semantic roles of a sentence may be different from those of a corresponding fictitious sentence that is composed by substituting a specific semantic role in the given sentence. In such a case, it is likely that the original semantic roles parsed are questionable. In general, by repeating this semantic role substitution process for different semantic roles of a sentence, we would be able to roughly estimate the plausibility of the semantic role parsing results of the given sentence. Knowledge items that fail to survive the plausibility verification process should be pruned so as to prevent a probably erroneous instance in the knowledgebase.

Based on the above idea, we propose a strategy to prune knowledge items with a high probability of questionable semantic roles.

The objective is to raise the accuracy to as high as above 90% so as to keep the final knowledge item reliable. Since our application of retrieving event-based knowledge mainly focuses on subject (ARG0), object (ARG1), time (ARGM-TMP), and location (ARGM-LOC), the pruning process mainly concerns the verification of these four arguments in the parse results.

3.4. Plausibility verification strategy

Based on our investigation on the results of the knowledge items, most of the erroneously parsed ARG0 or ARG1 are "contaminated" with irrelevant words that are supposed to be part of the ARGM-TMP or ARGM-LOC. For example, the parsed ARG1 using the applied SRL tool for a sentence "George Bush could play piano several years ago." might be erroneously given as "piano several years ago." Therefore, a main concern in the semantic role substitution strategy is to compose fictitious substitution sentences based on verbs that have a high probability to correctly separate an ARGM-TMP or ARGM-LOC that was erroneously coupled in either one of the semantic roles originally obtained. To this end, from the records in our database containing more than 5,60,000 knowledge items, we observed that the verbs "locate" and "find" gave the highest number of instances of ARGM-LOC, while "see" and "get" gave the highest number of instances of ARGM-TMP. Therefore, we chose these four verbs as the basis for composing the fictitious substitution sentences. Table 1 lists the substitution patterns we designed to be applied for ARG0, ARG1, ARGM-TMP, and ARGM-LOC, respectively. These patterns are applied to automatically compose fictitious substitution sentences by substituting a specific semantic role in the original sentences. For example, to estimate the plausibility of ARG1, we compose two fictitious substitution sentences, including *"George Bush can get* + ARG1" and "*Bill Clinton can find* + ARG1." We observed that for sentences in such two syntactic forms, the applied SRL tool typically correctly identify ARG0 as "George Bush" or "Bill Clinton". If the originally parsed ARG1 is questionable, we often get a different ARG1 result for the fictitious substituted sentences. Presumably, if the parsed ARG1s of the two fictitious substitution sentences are both identical to that of the original sentence, it is most likely that the parsed ARG1 for the original sentence is correct. In the evaluation section, we will report the experiments verifying the improvement of accuracy using the proposed semantic role substitution strategy.

The following gives an illustrative example to demonstrate the detail procedures of the proposed plausibility verification strategy for semantic roles. Consider a case for parsing the semantic roles of a sentence "The dog chased a ball in the Hyde Park last night." Assuming a knowledge item containing the parsed semantic roles for the sentence using a given SRL tool are as the following:

ARGO : "the dog"; ARG1 : a"ball";
ARGM-LOC : "in the Hyde"; ARGM-TMP : "Park last night".

Step 1: To estimate the plausibility of the ARG0 "the dog", the following two fictitious substitution sentences are composed, including "the dog can see a ball" and "the dog can locate a ball". The parsed ARG0s of these two sentences are both identical to "the dog". In such a case, proceed to Step 2.

Step 2: For the case of the ARG1 "a ball", the following two fictitious substitution sentences are composed, including "George Bush can get a ball" and "Bill Clinton can find a ball." The parsed ARG1s of these two sentences are both identical to "a ball". In such a case, proceed to Step 3.

Step 3: For the cases of ARGM-LOC "in the Hyde", the following two fictitious substitution sentences are composed, including "Gorge Bush can find a ball in the Hyde" and "Bill Clinton can locate a ball in the Hyde." The parsed ARGM-LOCs of these two sentences are both identical to "in the Hyde". In such a case, proceed to Step 4.

Step 4: Finally, to estimate the plausibility of the originally parsed ARGM-TMP "Park last night", the following two fictitious sentences are composed, including "Gorge Bush can see a ball Park last night." and "Bill Clinton can get a ball Park last night." The parsed ARGM-TMPs for these two fictitious sentences are both given as "last night" which unfortunately differs from that of the original sentence. In such a case, it is very likely that the originally-parsed semantic roles are questionable. Therefore, this knowledge item is discarded.

3.5. Stage4: knowledge distilling

After pruning the knowledge items that have a high probability of erroneously parsed semantic roles, the remaining ones will go through a distilling process to remove knowledge items that might not be regarded as commonsense knowledge. After looking at the knowledge items retrieved, a number of straightforward filtering rules are applied to eliminate the following noises to ensure the correctness of the knowledge item:

1. Knowledge requiring certain contextual information to identify the exact semantic (e.g., ARG1 contains a pronoun such as "it, "that," "she," "him," "those," etc.
2. Sentences that appear to be syntactically correct but are describing certain particular instances of an event in certain uncommon settings. For example, the information implied in the following knowledge item, [ARG0 A person] [Target tracks] [ARG1 many bits of information about historical incidents as well as getting a feel for a time and place], appears to be semantically correct. Nevertheless, it should not be regarded as general commonsense knowledge. In practice, we found that when the ARG1 of a knowledge item contains a long list of words, the event knowledge usually refers to certain specific events rather than commonsense knowledge. This type of knowledge items are discarded in the distilling process.

4. Evaluations

This section presents the evaluations results of the plausibility verification strategy and the commonsense knowledge retrieval

Table 1
Substitution patterns applied in the process of semantic role plausibility verification.

	Substitution pattern 1	Substitution pattern 2
ARG0	ARG0 + *can see a ball*	ARG0 + *can locate a ball*
ARG1	*Gorge Bush can get* + ARG1.	*Bill Clinton can find* + ARG1
ARGM-LOC	*Gorge Bush can find a ball* + ARGM-LOC	*Bill Clinton can locate a ball* + ARGM-LOC
ARGM-TMP	*Gorge Bush can see a ball* + ARGM-TMP	*Bill Clinton can get a ball* + ARGM-TMP

approach. A sentence extraction system was implemented using C#. The experiment aimed to retrieve the "*is capable of*" commonsense knowledge for a number of subjects, including farmer, dog, fish, person, adult, and bird. Based on the syntactic patterns described in the previous section and about a thousand verbs listed in PropBank Propbank (2008), we issued a significant number of queries to various web search engines. The raw sentences were trimmed from the snippets of the web search results for the queries. To identify the semantic roles for each of the raw sentences crawled, we applied a publicly available automatic semantic labeling engine – ASSERT (Pradhan et al., 2004). To speed up the computations of the semantic role-labeling using ASSERT, a cluster of distributed workstations were set up to parse the crawled sentences in parallel. The crawled raw sentences with their associated parsed semantic roles were managed in an SQL database server.

To investigate the performance of the proposed plausibility verification technique for semantic role parsing, in the first experiment, we conducted an experiment to compare the accuracy rates of ARG0, ARG1, ARGM-TMP, and ARGM-LOC before and after the verification process. Within the event-based knowledge items parsed from the crawled sentences, we generated 4 different data sets, each containing 500 randomly chosen knowledge items with ARG0, ARG1, ARGM-TMP, ARGM-LOC, respectively. One human adjudicator was asked to judge whether the parsed ARG0, ARG1, ARGM-TMP, and ARGM-LOC in both the original and pruned knowledge items were correct or not. Table 2 gives the evaluation results. The original accuracy rate for the ARG0 results appears to be rather high (98%). It appears that the SRL tools applied can easily identify the subject of the raw sentences crawled by using the designed syntactic patterns. After the pruning process, the accuracy rate of ARG0 was slightly raised from 98% to nearly 100%. On the other hand, for the ARG1, the applied SRL tool achieved a low accuracy rate of only 63%. The pruning technique successfully raises the accuracy rate to about 94%. As for the ARGM-TMP and ARGM-LOC, the original accuracy rate is about 89% and 86%, respectively. The proposed pruning technique raises the accuracy rate to almost 98%.

The second evaluation experiment aimed to investigate the accuracy rates for the distilled commonsense knowledge. For each of the 6 selected subjects, we randomly selected 1000 knowledge items in the database. Each knowledge item first went through the process of semantic role plausibility verification. The knowledge items surviving this semantic role plausibility verification then went through the knowledge distilling process. Table 3 presents the results. The survival rates for the semantic role plausibility verification range from 55% (552/1000) for "bird" to 66% (666/1000) for "person." Such survival rates roughly conform to the accuracy rate for the ARG1 as shown in Table 2.

After the knowledge distilling process, the remaining knowledge items were investigated to evaluate the accuracy rate for the retrieved knowledge. One human adjudicator was asked to judge whether the knowledge item was acceptable as commonsense knowledge. On average, the results indicate that the accuracy rate of the knowledge retrieved was above 95%. Such an accuracy rate is satisfactory and comparable to that achieved by the volunteer-based knowledgebase construction in ConceptNet.

Table 4 lists a number of examples of the reasonable commonsense knowledge retrieved. Table 5 lists a number of examples of the unreasonable commonsense knowledge retrieved. It appears that the major problems with the unreasonable knowledge retrieved at least include the following:

Table 4
Examples of reasonable commonsense knowledge retrieved.

ARG0	Verb	ARG1
Dog	Eat	Foods
Dog	Bite	Someone
Fish	Enter	The pool
Police	Check	The taxi's licenses
Police	Capture	A criminal
Person	Hack	The system
Person	Enjoy	Hiking and wildlife watching
Soldiers	Stop	The violence
Soldier	Obey	The army

Table 5
Examples of unreasonable commonsense knowledge retrieved.

item	ARG0	Verb	ARG1
1	Fish	Detect	Exposure to high
2	Dog	Extract	More value
3	Person	Lack	The last thing
4	Person	Remember	The forms and qualities of Kr
5	Police	Contact	ITP
6	Fish	Translate	Any spoken knowledge
7	Fish	Use	A playstation
8	Dog	Use	A search engine
9	Dog	Link	A new word
10	Dog	Talk	Homepages on the internet

Table 2
Evaluation results of the plausibility verification technique for semantic role parsing.

	Original semantic roles			Semantic roles after pruning		
	Total items	Correct items	Accuracy rate (%)	Total items	Correct items	Accuracy rate (%)
ARG0	500	491	98	444	443	100
ARG1	500	316	63	303	285	94
ARGM-TMP	500	443	89	295	286	97
ARGM-LOC	500	430	86	351	344	98

Table 3
Evaluation results for the commonsense knowledge retrieval approach.

Subject	Farmer	Dog	Fish	Person	Adult	Bird
Original knowledge item	1000	1000	1000	1000	1000	1000
Knowledge item after plausibility verification	607	618	562	666	574	552
Knowledge item after distilling	472	477	379	500	377	410
Accurate commonsense knowledge item	460	467	372	487	370	401
Accuracy rate (%)	97	98	98	97	98	98

- Syntactically incomplete, e.g., item 1.
- Referencing to contextual information, e.g., item 2,3.
- Dependent on domain-specific terminologies, e.g., item 4, 5.
- Incompliant to general world knowledge, e.g., item 6,7,8,9,10.

5. Conclusions

In this study, we proposed an approach by applying semantic role labeling and lexical-syntactic pattern matching techniques for the automatic retrieval of event-based commonsense knowledge from the web. In particular, we proposed a strategy for semantic role plausibility verification, based on a semantic role substitution strategy, which significantly pruned knowledge items with a high probability of erroneously parsed semantic roles. The evaluation results showed that the proposed approach could automatically accumulate commonsense knowledge efficiently with very high accuracy rates that are close to 98%. The achieved accuracy rates are comparable to that of ConceptNet, which is completely human annotated. With the proposed methodology for retrieving the sentence-level event-based commonsense knowledge, future work involves further developing techniques for the automatic acquisition of more complicated event-based knowledge, such as FirstSubeventOf, LastSubeventOf, EffectOf, EventForGoalEvent, EffecOfIsState, DesiresEvent EventRequiresObject, and DesiresNotEvent, from free-text.

References

ACE (2003). Automatic content extraction. National Institute of Standards and Technology. Available at: http://www.nist.gov/speech/tests/ace, last visited at: January 17th 2008.

Ahlswede, T., & Evens M. (1988). Parsing vs. text processing in the analysis of dictionary definitions. In *Proceedings of the 26th annual meeting of the association for computational linguistics* (pp. 217–224).

Alshawi, H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. *American Journal of Computational Linguistics, 13*(3), 195–202.

Aone, C., & Ramos-Santacruz, M. (2000). REES: A large-scale relation and event extraction system. In *Proceedings of the sixth conference on applied natural language processing* (pp. 76–83). Seattle, Washington.

Berland, M., & Charniak, E. (2002). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 57–64).

Byrd, J. R., Calzolari, N., Chodorov, S. M., Klavans, L. J., Neff, S. M., & Rizk, O. (1987). Tools and methods for computational linguistics. *Computational Linguistics, 13*(3/4), 219–240.

Cederberg, S., & Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Conference on natural language learning (CoNLL2003)* (pp. 111–118).

Creswell, C., Beal, J. M., Chen, J., Cornell, L. T., Nilsson, L., & Srihari K. R. (2006). Automatically extracting nominal mentions of events with a bootstrapped probabilistic classifier. In *Proceedings of the COLING/ACL 2006 main conference poster sessions* (pp. 168–175).

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., et al. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence, 165*(1), 91–134.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., et al. (2004). Web-scale information extraction in KnowItAll. In *Proceedings of the 13th international world-wide web conference* (pp. 100–110). New York.

Filatova, E., & Hatzivassiloglou, V. (2003). Domain-independent detection, extraction, and labeling of atomic events. In *Proceedings of the fourth international conference on recent advances in natural language processing (RANLP-2003)*. Borovets, Bulgaria.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245–288.

Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part–whole relations. *Computational Linguistics, 32*(1), 83–135.

Hearst, A. M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on computational linguistics* (pp. 539–545).

Jensen, K., & Binot, J. (1987). Disambiguating prepositional phrase attachments by using online dictionary definitions. *Computational Linguistics, 13*(3/4), 251–260.

Kehler, A., Appelt, D., Taylor, L., & Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of 2004 North American chapter of the association for computational linguistics annual meeting* (pp. 289–296).

Klavans, J., Chodorow, M., & Wacholder, N. (1990). From dictionary to knowledge base via taxonomy. In *Proceedings of the sixth conference UW centre for the new OED* (pp. 110–132).

Koomen, P., Punyakanok, V., Roth, D., & Yih, W. (2005). Generalized inference with multiple semantic role labeling systems. In *Proceedings of the ninth conference on computational natural language learning (CoNLL2005)* (pp. 181–184).

Lenat, B. D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33–38.

Lenat, B. D., & Guha, V. R. (1991). The evolution of CycL, the Cyc representation language. *ACM SIGART Bulletin, 2*(3), 84–87.

Lieberman, H., & Liu, H. (2002). Adaptive linking between text and photos using common sense reasoning. In *Conference on adaptive hypermedia and adaptive web-based systems* (pp. 2–11).

Liu, H. (2002). Semantic understanding and commonsense reasoning in an adaptive photo agent. Master of Engineering Thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Liu, H., & Singh, P. (2002). MAKEBELIEVE: Using commonsense knowledge to generate stories. In *Proceedings of the 18th national conference on artificial intelligence, AAAI 2002* (pp. 957–958). Edmonton, Alberta, Canada.

Liu, H., Lieberman, H., & Selker, T. (2002). GOOSE: A goal-oriented search engine with commonsense. In *Proceedings of the 2002 international conference on adaptive hypermedia and adaptive web based systems*. Malaga, Spain.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the seventh international conference on intelligent user interfaces (IUI 2003)* (pp. 125–132). Miami, Florida.

Liu, H., & Singh, P. (2004). ConceptNet – A practical commonsense reasoning toolkit. *BT Technology Journal, 22*(4), 211–226.

Markowitz, J., Ahlswede, T., & Evens, M. (1986). Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting of the association for computational linguistics* (pp. 112–119). New York.

Miller, A. G. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

MUC (2008). In *Message understanding conference*. National Institute of Standards and Technology. Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, last visited at: Jan. 17th 2008.

Nakamura, J., & Nagao, M. (1988). Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12th international conference on computational linguistics* (pp. 459–464). Budapest, Hungry.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to conference resolution. In *Proceedings of 40th annual meeting on association for computational linguistics (ACL-02)* (pp. 104–111).

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–106.

Ponzetto, P. S., & Strube, M. (2006). Semantic role labeling for conference resolution. In *Companion volume of the proceedings of the 11th meeting of the European chapter of the association for computational linguistics* (pp. 143–146).

Pradhan, S., Ward, W., Hacioglu, K., Martin, H. J. & Jurafsky, D. (2004). Shallow semantic parsing using support vector machines. In *Proceedings of the human language technology conference/North American chapter of the association for computational linguistics* (pp. 233–240).

Propbank (2008). Available at: http://www.cs.rochester.edu/~gildea/PropBank/Sort/, last visited at: Jan. 17th 2008.

Richardson, D. S., Dolan, B. W., & Vanderwende, L. (1998). MindNet: Acquiring and structuring semantic information from text. In *Proceedings of the 17 international conference on computational linguistics* (pp. 1098–1102).

Xu, F., Uszkoreit, H., & Li, H. (2006). Automatic event and relation detection with seeds of varying complexity. In *Proceedings of the AAAI workshop event extraction and synthesis* (pp. 12–17). Boston.