Towards Automatic Knowledge Acquisition from Text Based on Ontology-centric Knowledge Representation and Acquisition

Yu-Sheng Lai
Industrial Technology Research Institute,
Tainan, Taiwan, R.O.C.
laiys@itri.org.tw

Ren-Jr Wang
Industrial Technology Research Institute,
Hsinchu, Taiwan, R.O.C.
rjwang@itri.org.tw

ABSTRACT
With the development of the Semantic Web and ontology technologies, many ontologies have been built or will be built before long. Based on the ontologies, we attempt to investigate the technology of automatic knowledge acquisition from text. This paper presents an ontology-centric framework for knowledge representation and acquisition, called *iOkra*. By combining NLP technologies with replaceable ontologies, the framework is able to acquire different domain knowledge from natural language input. The acquired knowledge is represented in the form of instances and statements associated with the ontologies.

Categories and Subject Descriptors
I.2.4 Knowledge Representation Formalisms and Methods

I.2.7 Natural Language Processing

Keywords
Natural language processing, knowledge representation, knowledge acquisition, ontology

INTRODUCTION
Knowledge acquisition traditionally requires various specialists in logic, linguistic, philosophy, etc. Although many facilities have been developed for enabling these specialists to collaborate, large number of handcrafting task is still unavoidable. Automatic knowledge acquisition from text seems to be a pleasant aspect because of wealthy textual documents and data. However, no fully satisfactory approaches to automatic knowledge acquisition from text have been proposed.

In [3], Berners-Lee and the co-authors claimed that "the Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." It indicates that the data on the Semantic Web is not only human-readable but also machine-understandable.

Technically, researchers use ontologies to describe the semantics of websites. The W3C defines the Semantic Web as "the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined" and has been developing it in collaboration with many researchers and organizations. A document that specifies usage scenarios, goals and requirements for a web ontology language (OWL) has been proposed [6]. OntoWeb Network involving most European Union (EU) members has been integrating academic and industrial resources to promote interdisciplinary work and strengthen the European influence on Semantic Web standardization efforts such as those based on RDF and XML.

The Semantic Web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding [8]. With the development of the Semantic Web and ontology technologies, many ontologies have been built or will be built before long. This paper proposes an ontology-centric framework (see Fig. 1) that integrates natural language processing (NLP) technologies and the ontologies to automatically acquire knowledge from natural language input.

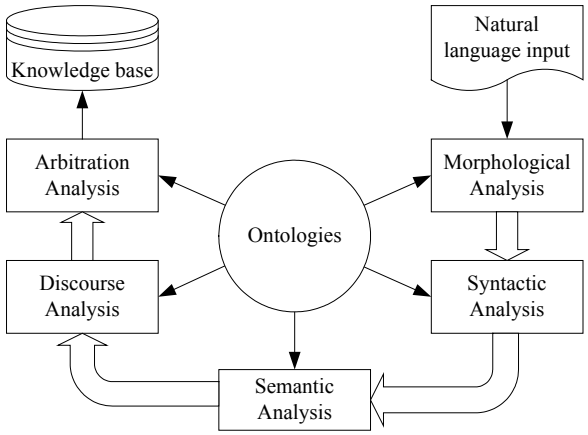

Figure 1. The proposed framework for knowledge acquisition.

K-CAP'03, October 23-26, 2003, Sanibel Island, FL, USA.

THE FRAMEWORK

As illustrated in Fig. 1, the framework called *iOkra* is expected to automatically acquire knowledge from natural language input, to represent the knowledge in the form of instances and statements associate with the ontologies, and to store the acquired knowledge into knowledge base.

The central ontologies comprise two kinds of ontologies: linguistic ontologies and domain ontologies. The main characteristic of linguistic ontologies is that they are bound to the semantics of grammatical units, such as words, nominal groups, etc. [5]. The domain ontologies provide varied ontological information, which might be domain-specific, task-oriented, or use-desirable.

In the framework, the natural language input is processed through several modules including morphological, syntactic, semantic, and discourse analyses and arbitration module.

- ✧ **The morphological analysis** splits the input text into words and connects to the ontologies for each word. The connections provide syntactic and semantic information for the following analyses.
- ✧ **The syntactic analysis** performs a semantic case frame parsing. The information-based case grammar [4] is adopted to suggest parts of the thematic roles, such as agent, patient, theme, goal, etc., in each sentence.
- ✧ **The semantic analysis** finds the remaining roles out and identifies the statements, cf. RDF statements, namely the concept for each word and the relations between the word concepts, according to the ontologies.
- ✧ **The discourse analysis** addresses the contextual issues, such as ellipsis and anaphora resolutions, which is currently an initial and on-going task and will be not presented in the following of this paper.
- ✧ **The arbitration module** quantifies all possible statements to reconcile conflicts, produces final result statements, and stores the results into a knowledge base, which is in a form of statements associated with the ontologies.

ONTOLOGY-BASED KNOWLEDGE REPRESENTATION

What is knowledge representation (KR)? Allen considered that "knowledge representation means different things to different researchers [2]." For some, it concerns the structure of the language used to express the knowledge. For others, it concerns the content of sentences. Herein we are interesting in the meaning representation of sentences.

Stevens et al. presented an ontology-based knowledge representation system for bioinformatics since they believed that "the combination of an ontology with associated instances is what is known as a knowledge base [9]," in which the instances indicate the things represented by concepts. Similar to the notion, we represent knowledge in an ontology-based representation system.

The Ontologies

An ontology basically consists of a set of concepts that represent classes of objects, and a set of binary relations defined on concepts. A special transitive relation *subClassOf* represents a subsumption relationship between concepts. The subsumption relations structure a taxonomy for the ontology. In addition to the taxonomy, an ontology typically contains a set of axioms explicitly or implicitly. The axioms enhance the ontology for reasoning.

Maedche and Staab proposed an ontology-learning framework [8] for the Semantic Web. In their case, they formally defined an ontology as an 8-tuple $<L, C, H_C, R, H_R, F, G, A>$, in which the first primitive L denotes a set of strings that describe lexical entries for concepts and relations, the middle 6 primitives structure the taxonomy of the ontology, and the last primitive A is a set of axioms that describe additional constraints on the ontology. The axioms make implicit facts more explicit. Based on the same definition, two ontologies: a linguistic ontology and a domain ontology, are currently in iOkra.

Linguistic Ontology

Following the DAML+OIL specification, Lai et al. constructed a Chinese lexical ontology call CLO [7]. To improve the ability in Traditional Chinese language processing, we define an amended version that has altered by a wide margin. Major amendments are as follows:

1. The approach to real world applications such as information extraction and knowledge acquisition, we make an adjustment in taxonomy. "人 (person)," "事 (affair)," "時 (time)," "地 (place)," "物 (thing)" are five basic entities in documents (Chen et al., 1998). Therefore we define the five entities plus two additional concepts "屬性 (attribute)" and "數量 (quantity)" as the upmost concepts.
2. To increase the compatibility with other ontology editors, such as OilEdit, the concept Lexicon in CLO is eliminated from the amendment. Some of the lexical entries are changed into instances. Others are moved to new, more proper position.
3. To enhance the expression power in linguistics, some thematic roles, such as theme, goal, range, etc., are interpreted as relations between concepts and added to the ontology.

Domain Ontology

For different domains, one term could be interpreted as many different meanings. For example, "大陸 (mainland)" means a country - China in a hard news article, but also means a corporation name - CEC in a stock news article. It

means different ontologies are required for different domains, even for different tasks.

Addressing the problem of knowledge representation and acquisition from the news articles of Taiwan stock market, we create an ontology that aims at the terminology of Taiwan stock market, such as industrial categories, corporation names, product names, people names, proper nouns, etc. Most of them are collected from the WWW and are organized into the domain ontology automatically. A small number are reorganized or modified manually.

Instance and Statement

The *iOkra* represents ontology-based knowledge consisting of two components: instance and statement. An instance is a specific description of a concept. For example, "台積電 (TSMC)" is an instance of concept "公司 (corporation)." A statement specifies a relationship between instances. For example, the concept "公司 (corporation)" has a "董事長 (board chairman)" relation to the concept "自然人 (natural person)," and "張忠謀 (Morris C.M. Chang)" is the board chairman of the corporation "台積電 (TSMC)." Fig. 2 is a conceptual graph describing the relationships.

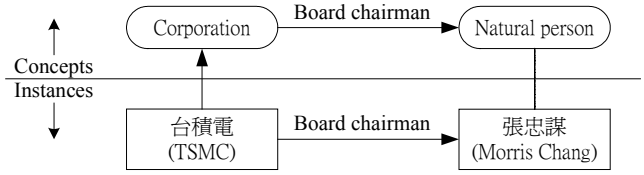

Figure 2. A conceptual graph describing instances and statements

ONTOLOGY-CENTRIC KNOWLEDGE ACQUISITION FROM NATURAL LANGUAGE INPUT

In the following, we will describe the three NLP modules: morphological, syntactic, and semantic analysis, and their cooperation with ontologies.

Morphological Analysis

A word segmentation algorithm is used for morphological analysis. It splits a sentence into a sequence of words. The words are possibly the words in the general ontology, the proper nouns in the domain ontologies, or compound words from a grammatically word-formation process. For example, the sentence "聯電1月29日至2月27日處分聯發科股票150張 (UMC sold 150 kilo-shares of MediaTek stocks during 1/29 to 2/27.) " can be split into words: "聯電 (UMC)," "1月29日 (1/29)," "至 (to)," "2月27日 (2/27)," "處分 (sold)," "聯發科 (MediaTek)," "股票 (stock)," "150," and "張 (kilo-shares)."

The corporation names "聯電 (UMC)" and "聯發科 (MediaTek)" come from the domain ontology, the dates, "1月29日 (1/29)" and "2月27日 (2/27)" and the numeral determinatives (ND), "150" and "9678萬 (96.78 million)," from a word-formation process.

Syntactic Analysis

A shallow syntactic analysis is performed in this module due to the lack of full Chinese grammar. The analysis is divided into two phases. In the first phase, a phrase-formation process is performed. A parser based on the CYK algorithm [1] is used to concatenate words into phrases. For example, the three words "1月29日 (1/29)," "至 (to)," "2月27日 (2/27)" in Table 1 can be combined to form the phrase "1月29日至2月27日 (1/29 to 2/27)."

In the second phase, we use the Information-based Case Grammar (ICG) to recognize some of the thematic roles of each of the words in a sentence. The thematic roles are defined in the general ontology and are represented as relations. For example, a basic pattern in the ICG AGENT[{NP, PP[由]}]<VC2<GOAL[NP] denotes that a verbal head with the syntactic category "VC2" has two thematic roles: agent and goal. The agent could be an NP (noun phrase) or a PP[由] (preposition phrase led by "由 (by or through)"), and should occur on the left-hand side of the head. The goal could be an NP (noun phrase) and should occur on the right-hand side of the head.

A head-driven approach is performed to recognize the thematic roles using the basic patterns. We design an automaton, called ICG-machine, to perform the recognition process. It is somewhat different from the Mealy machine. An enhanced scanning algorithm enables the ICG-machine to scan an input and output all acceptance paths. Besides, it is able to scan a fragmental input and output partially matched paths while no fully matched paths exist. For the basic patterns of each of the syntactic categories, we create an ICG-machine to perform the recognition process. For example, there are five basic patterns for the syntactic category "VC2." The basic patterns can be used to create an ICG-machine as illustrated in Fig. 3.

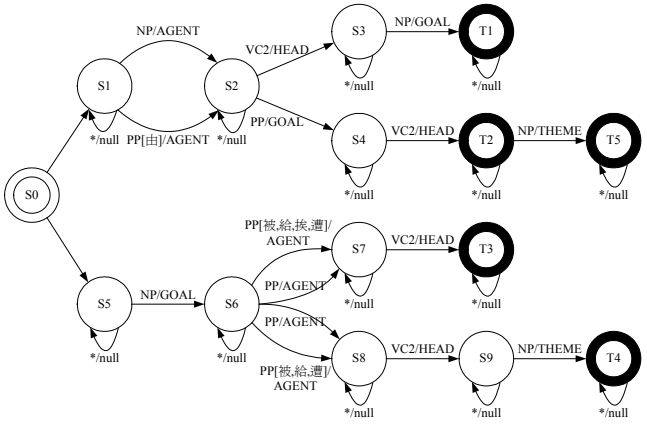

Figure 3. The ICG-machine created from the basic patterns of the syntactic category “VC2.”

Table 1. The words and phrases in the sentence "聯電1月29日至2月27日處分聯發科股票150張(UMC sold 150 kilo-shares of MediaTek stocks during 1/29 to 2/27.)" associated with their possible syntactic and semantic categories

Word	***Syntactic category***	***Phrase***	***Syntactic category***	***Semantic category***
聯電(UMC)	Nba	聯電(UMC)	NP (Nba)	corporation
1月29日(1/29)	Ndabd	1月29日至2月27日(1/29 to 2/27)	NP (Ndabf)	duration
至 (to)	P61			
2月27日(2/27)	Ndabd			
處分(sold)	VC2, Nac	處分(sold)	VC2, NP (Nac)	sell
聯發科(MediaTek)	Nba	聯發科(MediaTek)	NP (Nba)	corporation
股票(stock)	Nab	股票(stock)	NP (Nab)	stock
150	ND	150張(150 kilo-shares)	DM	quantity
張(kilo-shares)	Nbc, Nfa, VC2			

For example, the machine scans an input: "*NP1 NP2 VC2 NP3 NP4 DM*," then four matched paths shown in Table. 2 are outputted.

Table 2. The matched paths by inputting the syntactic sequence "NP_1 NP_2 *VC2* NP_3 NP_4 *DM*" into the ICG-machine shown in Fig. 3

Path	NP_1	NP_2	VC_2	NP_3	NP_4	*DM*	Terminal
1	agent	-	head	goal	-	-	T1
2	agent	-	head	-	goal	-	T1
3	-	agent	head	goal	-	-	T1
4	-	agent	head	-	goal	-	T1

Semantic Analysis

In case the word "處分(sold)" with the syntactic category "VC2" is a head of the sentence, the agent is probably "聯電 (UMC)" or "1月29日至2月27日 (1/29 to 2/27)" and the goal is probably "聯發科 (MediaTek)" or "股票 (stock)" since all of them are noun phrases. (See Table 2) In other words, there are two candidates respectively to the agent and the goal. However, the agent and the goal are unique to the head in this case. Here are syntactic ambiguities.

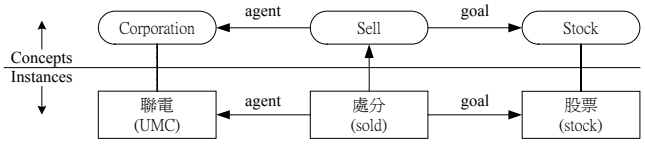

Figure 4. A conceptual graph describing the relationships among the three words "聯電(UMC)," "處分(sold)," and "股票(stock)."

The ontologies are used to resolve the ambiguities. In the general ontology, the concept "sell" has an "agent" relation to the concept "corporation." The phrase "1月29日至2月27日(1/29 to 2/27)" cannot be an instance of the concept "corporation." Therefore the word "聯電(UMC)" is the agent. The same as the reason, the goal is "股票 (stock)." Some of syntactic ambiguities can be resolved due to specific constraints of the ontology. The domain ontology affords the same functionality too. After recognition of the thematic roles, the sentence can be interpreted as a conceptual graph shown in Fig.4.

Presently, three unknown roles: "1月29日至2月27日 (1/29 to 2/27)," "聯發科 (MediaTek)," and "150張(150 kilo-shares)" have not been identified yet. A common characteristic of languages – "local dependency" exists in text everywhere. Using the characteristic, we find the nearest relations between unrecognized and recognized words. Thus, three additional relationships can be found. The head "處分(sold)" has a "time" relation to the duration "1月29日至2月27日 (1/29 to 2/27)." The word "股票 (stock)" has a "corporation-of-issue" relation to the word "聯發科 (MediaTek)" and a "quantity" relation to the phrase "150 張 (150 kilo-shares)." Fig. 5 shows the full relationships among the members of the sentence.

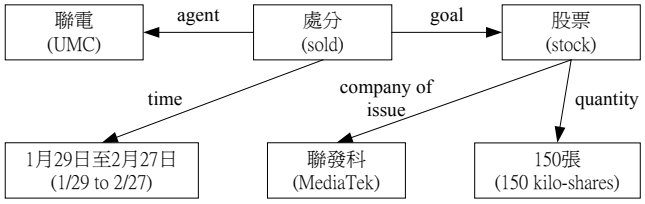

Figure 5. The relations within the sentence "聯電1月29日至2月27日處分聯發科股票150張(UMC sold 150 kilo-shares of MediaTek stocks during 1/29 to 2/27.)."

EXPERIMENTS AND DISCUSSION

To initially evaluate the performance of iOkra in automatic knowledge acquisition, we conduct an experiment on a collection of 501 news titles randomly selected from Yahoo!股市 (tw.stock.yahoo.com). Each of the titles may consist of one or more clauses and is manually annotated as

a set of instances and statements. The evaluation metrics used in this experiment includes: recall rate, precision rate, and F-measure. The experiment is conducted on the titles, statements, and concepts, in which a correct title means all the statements in the title must be fully recognized. The experimental result is shown in Table 3.

Table 3. Experimental result for overall test data

	Title	*Statement*	*Concept*
Recall rate	65.86%	78.21%	86.80%
Precision rate	66.00%	83.73%	91.85%
F-measure	65.93%	80.88%	89.25%

The test data contains some titles that cannot be split into words correctly by the automatic word segmentation process. Therefore we conduct an additional experiment on the titles that can be split correctly. There are totally 391 tiles in this set. The experimental result is shown in Table 4.

Table 4. Experimental result for the titles that can be split correctly by the automatic word segmentation process

	Title	*Statement*	*Concept*
Recall rate	70.84%	81.46%	89.35%
Precision rate	71.02%	86.45%	94.60%
F-measure	70.93%	83.88%	91.90%

By an analysis on errors, we summarize the errors in two aspects: NLP technologies and ontology engineering. In NLP, there are three major problems as follows:

1. **Ellipsis and anaphora problem**. Many titles consist of several clauses. Some of the clauses share a common word.
2. **Unknown word problem**. Many new created words, translated names, loanwords, etc. occur in the title.
3. **Word segmentation problem**. As shown in Tables 3 and 4, many errors result from the word segmentation.

For iOkra, several derived research topics on the ontology field are described as follows:

1. **Consistency between different ontologies**. In a multi-ontology-supported system, how to maintain the consistency between different ontologies is a well-known important issue.
2. **Integration between ontology and knowledge base**. In an ontology-based knowledge system, one, either ontology or knowledge base, is changed, another should do something to correspond to the change.
3. **Cross-domain knowledge**. Text knowledge may be cross two or more domains. How to acquire and represent such knowledge is still a problem.

CONCLUSION AND FUTURE WORK

This paper presents an ontology-centric knowledge representation and acquisition framework, called iOkra. Combining NLP technologies with replaceable ontologies, the framework is able to automatically acquire knowledge from natural language input. Based on iOkra, a prototypical document annotation system is constructed. By using different domain ontologies, the system is able to automatically annotate text documents of different domains.

A preliminary experimental result shows the system performance at title level achieves 65.93% in F-measure, 80.88% at statement level, and 89.25% at concept level. Without considering the errors from word segmentation, the performance is as follows: 70.93% at title level, 83.88% at statement level, and 91.90% at concept level. In the future, we will work on the research topics mentioned above.

ACKNOWLEDGEMENTS

This paper is a partial result of Project A321XS1A10 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C. The authors would like to thank the CKIP Group of Sinica, R.O.C. for providing the ICG.

REFERENCES

[1] Aho, A.V. and Ullman, J.D. *The Theory of Parsing, Translation, and Compiling*, Prentice Hall, Englewood Cliffs, N.J., 1972

[2] Allen, J. *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, 1994.

[3] Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web, *Scientific American*, 2001.

[4] Chen, K.J. and Huang, C.R. Information-based Case Grammar, *In Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, University of Helsinki, Finland, 2 (1990), 54-59.

[5] Gomez-Perez, A., Fernandez-Lopez, M. and Corcho, O. Technical Roadmap D.1.1.2, OntoWeb, 2002.

[6] Heflin, J. *Web Ontology Language (OWL) Use Cases and Requirements-working draft 3*, W3C, 2003.

[7] Lai, Y.S. Wang, R.J. and Hsu, W.K. A DAML+OIL-Compliant Chinese Lexical Ontology, *In Proceedings of the 19th International Conference on Computational Linguistics*, 2 (2002), 1238-1242.

[8] Maedche, A. and Staab, S. Ontology Learning for the Semantic Web, *IEEE intelligent Systems*, 16, 2 (2001), 72-79.

[9] Stevens, R., Goble, C.A. and Bechhofer, S. Ontology-based Knowledge Representation for Bioinformatics, *Briefings in Bioinformatics*, 1, 4 (2000), 398-416.