The Design and Implementation of Chinese Semantic Search Engine Based on FAQ Corpus and Ontology Construction from Information Extraction

Wen-Chih Chen, Lu-Ping Chang and Shi-Jim Yen
Advanced e-Commerce Technology Lab., Institute for Information Industry, ROC
National DongHwa University, Taiwan
E-mail : {wjchen, clp}@iii.org.tw

Abstract

In the paper, we propose FAQ corpus and Ontology construction to implement Chinese semantic search engine. These frequently asked questions would be split into Subject term, Attribute term, Intention term and interrogative term. The Ontology construction is from information extraction and How Net. Information extraction consists of human concept extraction, event relationship extraction, time concept extraction, location concept extraction and entity concept extraction.

Keyword: *FAQ Corpus, Ontology, information extraction, Semantic Search Engine*

1. INTRODUCTION

1.1 Search engine

Up to now, there are two revolution of search engine. The first generation search engine is based on Keyword search. The Yahoo and AltaVista are two famous first generation search engines. This type of search engines accepts the Keywords or Keywords with Boolean operation composed. And providing results by searching the huge databases of web sites. The advantages of this kind search engine are easy using and high recall rate. The disadvantage is huge results. Many portals provide search mechanism and the different between these search mechanisms is ranking technology.

The Excite portal's search engine has the ability of semantic understanding. But It's not suitable for commercialized cause of the exception search results of technology bottlenecks. Sun, IBM and Microsoft concern on researches on this area but they don't release the relative services or productions. The third generation search engine should more intelligence and can understand the connotation of queries. And can retrieve the useful and exact results which contained in users' mind. In this paper we propose a construct third generation search engine technology. The search based on Natural Language Process and Ontology representation is intelligence and has knowledge process ability.

1.2 Ontology

In order for a Search engine to make statements and ask queries about a subject domain, it must use a conceptualization of that domain. A domain conceptualization names and describes the concepts that may exist in that domain and the relationships among those concepts. It therefore provides a vocabulary for representing and communicating knowledge about the domain.. The research of Ontology construction use the Natural Language Process(NPL), Machine learning, data mining and information retrieval to archive the goal。 And The methodology of Ontology construction could be divided into following types:

(1).Ontology construction using Wordnet.[1][2][3][4]

(2).Using clustering for Ontology construction [5][6]:

(3).Ontology construction from information extraction & retrieval. [7][8][9]

(4).Ontology construction from ground instances.[10][11][12]

1.3 HowNet

HowNet is an on-line common-sense knowledge base. Like WordNet, HowNet is a kind of ontology. HowNet handles inter-conceptual relations and inter-attribute relations of concepts as connoting in Chinese lexicons and their English equivalents. The design of HowNet is based on that all physical and non-physical matters undergo a continual process of motions and changes in a specific space and time. The motions and changes are usually reflected by a change in state that in turn, is manifested by a change in the value of some attributes.

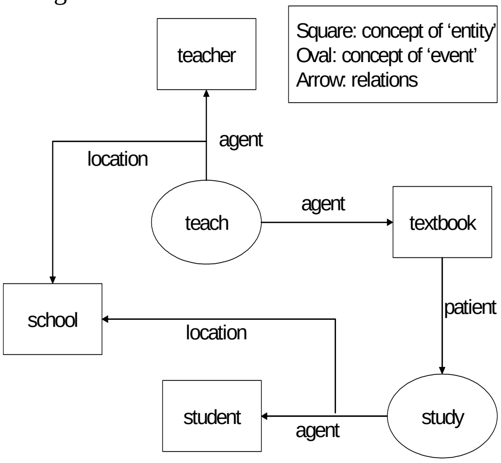

Figure 1. a simple example of relations between concepts in How Net

The explicated relations of HowNet include hypernymy-hyponymy, synonymy, antonymy, metonymy, part-whole, attribute-host, material-product, converse, dynamic role and concept co-occurrence, etc. The basic unit of meaning in HowNet is called sememe that cannot be further decomposed. The coverage of the set of sememes was tested against polysyllabic concepts to identify additional sememes. Eventually, a total of 1,503 sememes were found and organized hierarchically. The top-most level of classification in HowNet thus includes: entity|實體, event|事件, attribute|屬性 and attribute value|屬性值. The Knowledge Dictionary is created by referring to the most common dictionaries. A common-sense Knowledge Dictionary constituting a knowledge system describes general concepts and map out the relations among concepts. The latest version (HowNet 2000) covers over 110,000 concepts in the Dictionary.

2. THE PROPOSED METHOD/ ARCHITECTURE

2.1 System Architecture

In the following section, the design and implementation of FAQ corpus could be described in detail. As well as how could the semantic text query be matched in FAQ corpus. If semantic text query could not be matched with question in FAQ corpus, the intention would be extracted from the query. If semantic text query is near keyword, the kernel of semantic search is just like general keyword search. Or the query would be tried to match ontology and the search results would be more retrenched and precise. The design of architecture of Semantic Search Engine is as figure 2.

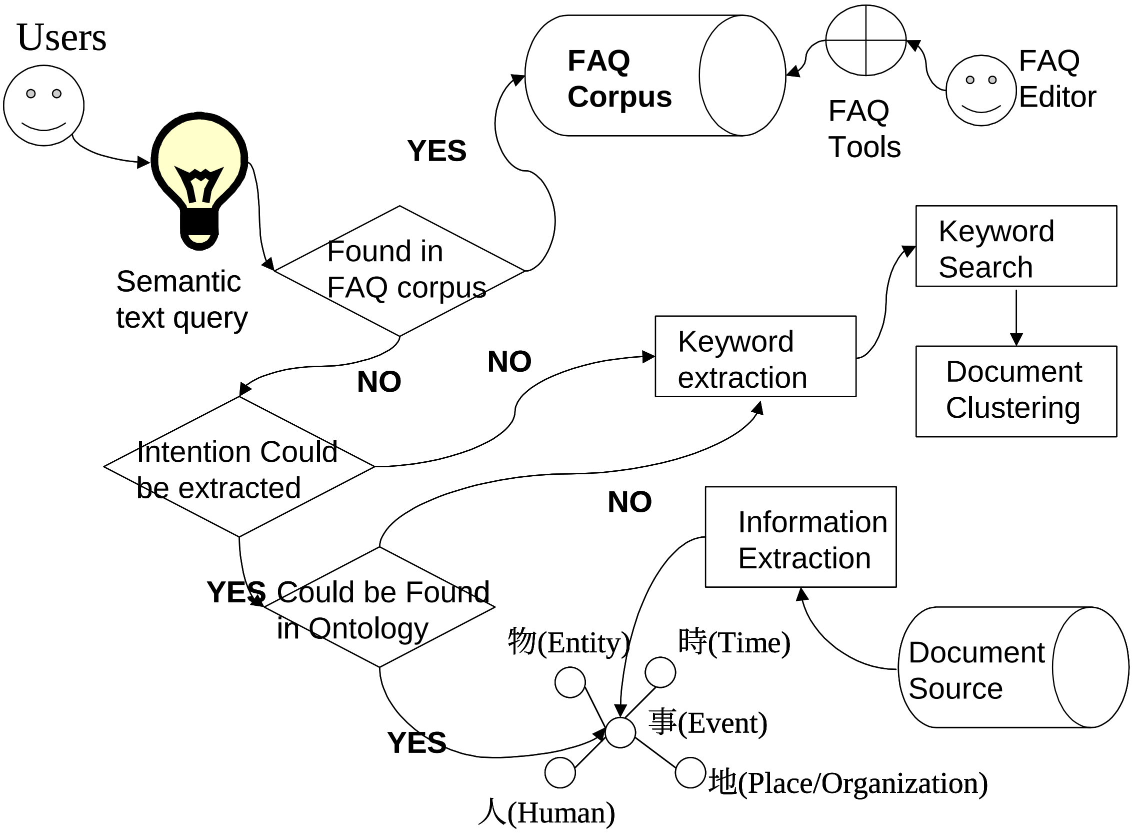

Figure 2. Architecture of Semantic Search Engine

2.2 FAQ corpus

In the design stage of semantic search engine, we collect about one thousand and two hundred frequently asked questions by questionnaire. These questions include one thousand frequently asked questions for employees and two hundred frequently asked questions for organization outsider. These frequently asked questions would be split into 主詞(Subject) term, 屬性詞(Attribute) term, 意圖詞(Intention) term and 疑問詞(interrogative) term. The knowledge of semantic search would be retrieved by Subject, Attribute, Intention and interrogative terms.

2.3 Information Extraction

2.3.1 人(Human) Concept extraction

Chinese Human names have a one-character surname (or rarely, two characters) that comes at the start of the name. The following shows three different types:

(1) Single character ”陳”,”林”,”李”'.

(2) Two characters “歐陽”,” 諸葛”,” 尉遲”

(3) Two surnames together “蔣宋”,”陳吳”,”李曾”

Most given names are two characters and some rare ones are single characters. Some of the two characters given names can be regarded as compounded words. Unfortunately neither single word in the given names nor compounded words can serve other functions in Chinese. Complicating combinations increasing the difficulty of name identification. There is not a limited set of given names but surnames come from a limited set of possibilities.

Theoretically, every Chinese person name has a one or two characters surname that comes at the start of the name and has one or two characters given name. Every Chinese character can be considered as names rather than a fixed set. Thus the length of Chinese person names ranges from 2 to 6 characters.

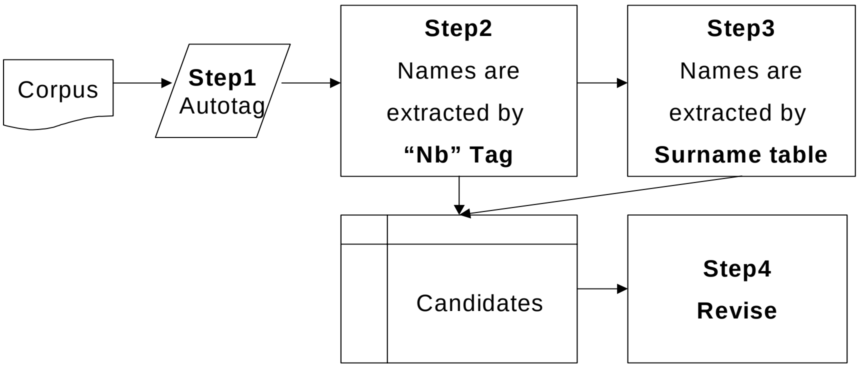

Figure 3 Chinese Person names extraction

Table I Relationship between concepts

Relationship	**Mapping to Event/事件 class in HowNet**	**Example**
Place/Organization Relationship	The classes contain “location” or “direction” in HowNet or have “VCL” or “P” tag by AutoTag .	situated|處于{existent,location}, LeadTo|通往{existent,direction}, 前往(VCL)
Time Relationship	The classes contain “time” or“duration” in HowNet or have “P” tag by AutoTag.	begin|起始{experiencer,~time}
Interactive Relationship	The classes contain in “act|行動” subclass in HowNet or have “VB”, “VC”, “VI”, “VJ” tag by AutoTag.	catch|捉住{agent,patient}, add|增加{agent,patient,quantity}, 舉行(VC), 買到(VC)

2.3.2 事(Event) Relationship extraction

The 事(Event) definition of this paper is the relationship between Human, Time, Place/Organization, entity. This definition is similar but not so complex to the Event/事件 class in HowNet. Hence, the Event/事件 class and corpus are analyzed to get the three different type of relationships in our event. The relationships are showed in Table I.

(1) Place/Organization Relationship: The Place/Organization Relationship is the relationship between Human and Place/Organization or Entity and Place/Organization. For example the corpus “陳水扁出現在總統府”. The pattern “陳水扁” is human and pattern “總統府” is Place/Organization. The pattern “出現” defined in HowNet is “appear|出現 {existent,~location}”. Therefore, the pattern “出現” is relationship(event) between Human(陳水扁) and Place/Organization(總統府).

(2) Time Relationship: The Time Relationship is the relationship between Human and Time or Entity and Time. For example the corpus “李登輝將在日本停留兩天”. The pattern “李登輝” is human and pattern “兩天” is Time. The pattern “停留” defined in HowNet is “stay|停留{agent,location,TimeIni,TimeFin,duration}”. Therefore, the pattern “停留” is relationship(event) between Human(李登輝) and Time (兩天).

(3) Interactive Relationship: The Interactive Relationship is the relationship between Human and Human, Entity and Entity or Human and Entity. For example the corpus “林正義卻攜帶機密文件叛逃”. The pattern “林正義” is human and pattern “機密文件” is Entity. The pattern “攜帶” defined in HowNet is “bring|攜帶 {agent,patient}”. Therefore, the pattern “攜帶” is relationship(event) between Human(林正義) and Entity (機密文件).

2.3.3 時(Time) Concept extraction

After decomposing corpus we can find some clues for Chinese time extraction.

(1)We using CKIP Autotag to help us segmenting a corpus into phrases and it is also introduced to provide part-of-speech information.

(2)According to analysis Autotag results, we can find most Chinese times consist with “Nd” tag.

(3)We collected 8 Keywords that imply time including “年(Year)”, ”月(Month), ”日(day), “天(day)”, ”時(hour)”, “點(clock)”, “分(minute)” and “秒(second)”.

(4)According to analysis Autotag results, we can find that Chinese times consist with seven kinds of tags. These tags are called legal tag ("Neu", "Nf", "Nes", "VCL", "Di", "FW", "D").

(5)The Chinese times can divided into three types:

(a) absolute time: “民國九十一年五月二十八日”.

(b) relative time: “昨天(yesterday)”, “當日(today)”… etc. The relative time should be transferred to absolute time. There are two types of relative time and each has different translation rule. The two different types of relative time are showed in following table II:

(c) duration time: “九十一年四月至五月”. The duration time is concatenated two times by pattern with “(P)” tag.

The figure 4 is our process of extracting Chinese 時(time) concept.

Table II Relative time Tag

Translation rule	**Relative time**
Based on previous time	“當日”, “隔日”, “當天”, “隔天”, “當晚”
Based on time of corpus	“昨天”, “前天”, “上旬”, “下旬”, “上月”, “下月”, “上個月”, “下個月”, “上一個月”, “下一個月”, “前一個月”, “今年”, “前年”, “去年”, “上個年度”, “下個年度”

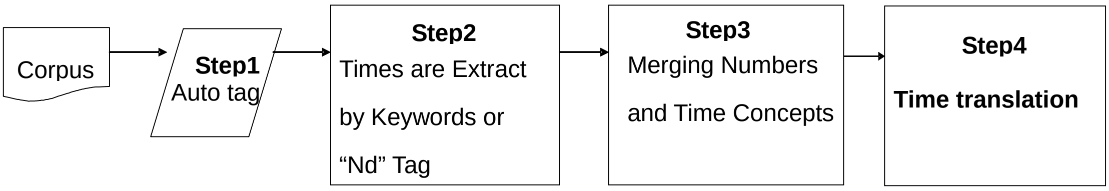

Figure 4 Process of Time Extraction

2.3.4 地(Place/Organization) Concept extraction

The 地(Place/Organization) concept can include a wide variety of types, and consequently can be one of the hardest types of entities to identify. We will first define a 地(Place/Organization) at the beginning of this section. 地(Place/Organization) identified by the extraction process include countries, states, provinces, cities, towns, directions, organizations, islands, and named geographic features (mountains, valleys, etc.) Next section, we will describe our method for extracting 地(Place/Organization) concept in detail.

The Entity Extraction Process

After decomposing corpus we can find some clues for Chinese 地(Place/Organization) concept extraction.

(1)We use CKIP Autotag to help us segment a corpus into phrases and it is also introduced to provide part-of-speech information.

(2)According to analysis Autotag results, we can find most Chinese locations consist with “Nc” tag . In Chinese grammar, “Nc” tag usually combines with other tags to describe the location more detail. (E.g. 蘇澳(Nc)東北方(Ncd) ; 台北(Nc)國際(Nc)金融(Na)大樓(Na)二十三樓(Nc)). After analyzing the grammar, we found that there are only 6 kind of tags in four groups of tags we usually use them in most corpus. The four groups are showed below:

Group	Tags	Example
A	Ncd	南部(Ncd)山區(Nc)
B	Nc	澎湖(Nc)
C	Nb , VC, FW	台北市(Nc)立(VC)陽明(Nb)醫院(Nc)3(FW)樓(Nc)
D	Na	金融(Na)大樓(Na)

(3) We collected 26 Keywords that imply location including”樓(floor)”;”街(street)”;”區(district)”;”學院(college)”;….

The figure 5. is our process of extracting named locations.

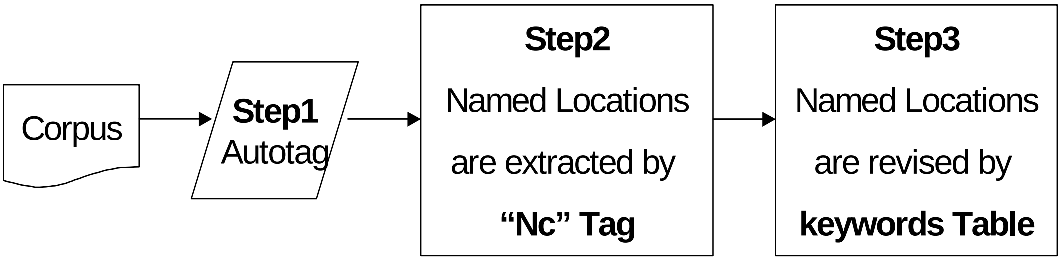

Figure 5. Process of Named Locations Extraction

Table III. The five Heuristic rules for Entities extraction

Rule Number	Rule contain	Example
Rule1	Entity is a Nouns “Na”	資料(Na)
Rule2	There is a ”數詞定詞 Neu” or “數量定詞 Neqa” in front of entity to describe the amount of the entity. 物前面通常會加一”數詞定詞 Neu” or “數量定詞 Neqa” 加以描述此物的數量	三(Neu) 座(Nf) 單打(Na) 冠軍(Na) 錦標(Na)
Rule3	物前面通常會加 ” transitive verb 及物動詞 VB,VC,VI,VJ” 以描述使用此物	台北(Nc) 累積(VC) 雨量(Na) 為(VG) 六(Neu) 公厘(Nf)
Rule4	There is a “的 DE” in front of entity to describe the hypernym-hyponym relations of entity. 物前面通常會加 “的 DE” 以描述此物屬於的從屬關係	大雨(Na) 特(D) 報(VC) 的(DE) 標準(Na)
Rule5	There is a “adverb (D)” in back of entity. 物的後面會加副詞 (D)	家長(Na) 也(D) 領(VC) 著(Di) 小(VH) 朋友們(Na)

2.3.5 物(Entity) Concept extraction

We will first define a Chinese Entity (CE) at the beginning of this section. Entities are every things exclusive Human, Place/Organization, time and event. The following shows two different types of entity:

(1) Thing: ”physical”, ”mental” and “internet”.

(2) component: “part” and ”fittings”.

The two types of entity are defined in HowNet “Entity|實體” class. In this paper we adopt the definition of “Entity|實體” class in HowNet to fit our definition. Just some classes of the “Entity|實體” class are used. Below we will describe our method for extracting entities in more detail.

The Entity Extraction Process

After decomposing corpus we can find some clues for Chinese Entity extraction.

(1) We using CKIP Autotag to help us segmenting a corpus into phrases and it is also introduced to provide part-of-speech information.

(2)According to analysis Autotag results, we can find Chinese entities consist with five kinds of Heuristic rules. These Heuristic rules are showed in Table III.

3. EVALUATIONS

We collect five thousand pieces of Chinese news from www.chinanews.com from March to May 2002. First, five people are responsible for 人(human) concept extraction, 事 (event) relationship extraction, 時 (time) concept extraction, 地(location) concept extraction and

物(entity) concept extraction manually. These results of manual extraction are thought as standard answers. According to heuristic rules of information extraction given above section, the precision/recall rate is as table IV.

Table IV the precision/recall rate of information extraction

Information Extraction	Precision Rate	Recall Rate
人 (human) concept	0.92	0.96
事 (event) relationship	0.85	0.88
時(time) concept	0.98	0.96
地 (location) concept	0.83	0.80
物 (entity) concept	0.82	0.84

4. SYSTEM APPLICATIONS

Advanced e-Commerce Technology laboratory (ACT) is one department of the Institute for Information Industry. Its fundamental R & D features technology of personalization, knowledge exploration technology, workflow management and enterprises workflow integration skills, intelligent agent technology, corporate knowledge portal-site technology, knowledge management technology, n-tier application structure technology, and web services technology.

According to FAQ corpus and ontology construction given above, we applied them in the ACT e-Service site. The kernel Semantic Search Engine of ACT e-Service is named Knowmation Instant Semantic Search (KISS). The KISS has a corpus which contains about one thousand frequently asked questions for employees and two hundred frequently asked questions for ACT outsiders’. KISS parses incoming questions, matches the queries created from the parse trees against its knowledge base and presents the appropriate information segments to the user.

The KISS system could be referred to as another text based call center. It accepts natural language semantic text query and it usually outputs satisfactory answers. In our experimental case, every employee of ACT submits about four semantic text queries from. The satisfactory degree is divided into 5 ranges inclusive most agree(5 points), little agree(4 points), no comments(3 points), little disagree(2 points), most disagree (1 points). The KISS receives satisfactory degree by online questionnaire. The average scores of two hundred queries are about 4.26. And 71 percent semantic text queries could be matched in FAQ corpus successfully.

In real scenario of KISS, we notice that most users often omit an interrogative of query sentence. For example, users often submit a query like “創新前瞻的簡報範本”(brief sample of innovative and Prospective Technologies Project) instead of a complete interrogative sentence “哪裡可以找到創新前瞻計劃的簡報範本”(Could you tell me which location to find brief sample of innovative and Prospective Technologies Project). In other words, users often submit a query pattern such as subject followed by preposition followed by attribute instead of a complete interrogative sentence.

5 CONCLUSION

When it comes to search engine, most users think google search engine. Nowadays, google gets second in hit rates among world search engines. According to experimental results, if you submit a keyword search in google search engine, you will get what you want among ten search results by google page rank algorithms. Somebody argues that keyword search still works well and semantic search often needs extra words such as verb and interrogative. There is no doubt with the usefulness of google. But keyword is always a keyword. Human communication does not merely rely on keywords. For example, if you would like to know the answers to question ”國外出差要如何結報” (How could I close foreign business travel fees), you only ask “國外出差”(foreign business travel) or 出差(business travel), you will get another question ”國外出差的什麼”(what subject for foreign business travel) from other. Therefore, keyword search is just the first step to filter mass information. People have to spend many efforts perusing many documents and getting what they really want.

In this paper, we propose a near human knowledge representation by automatically extracting 人 (person) 事 (event) 時 (time) 地 (location)物(entity) concept and relationship. Or you could think 人(person)事(event)時(time)地(location)物(entity) concept and relationship as new or condensed hownet methodology. In the KISS system, we propose a semantic search engine architecture, FAQ corpus analysis, and ontology construction. When you submit semantic text query, the KISS outputs similar questions from FAQ corpus or extracts knowledge from ontology.

In the next stage, the intelligent spider will be implemented for searching Chinese pages through internet. Our ultimate aim is to set up the first Chinese “askjeeves” web site in Taiwan.

ACKNOWLEDGEMENT

This research was supported by the III Innovative and Prospective Technologies Project of Institute for Information Industry and sponsored by MOEA , ROC

REFERENCES

1. A. Wagner: “Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis”. 14th European Conference on Artificial Intelligence
2. E. Agirre, O. Ansa, E. Hovy, D. Martinez: “Enriching very large ontologies using the WWW”. 14th European Conference on Artificial Intelligence
3. R.H.P. Engels, B.A. Bremdal, R. Jones: “CORPORUM: a workbench for the Semantic Web”. 12th European Conference on Machine Learning (ECML'01)/ 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)
4. M. Kurematsu, N. Nakaya, T. Yamaguchi : “Acquiring Conceptual Relationships from a MRD and Text Corpus”. 12th European Conference on Machine Learning (ECML'01)/ 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)
5. P. Clerkin, P. Cunningham, C. Hayes : “Ontology Discovery for the Semantic Web Using Hierarchical Clustering”. 12th European Conference on Machine Learning (ECML'01)/ 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)
6. G.Bisson, C. Nedellec and D. Canamero: “Designing Clustering Methods for Ontology Building - The Mo'K Workbench”. 14th European Conference on Artificial Intelligence
7. D. Faure and T. Poibeau: “First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX”. 14th European Conference on Artificial Intelligence
8. A. Todirascu, F. de Beuvron, D. Galea, F. Rousselot: “Using Description Logics for Ontology Extraction”. 14th European Conference on Artificial Intelligence
9. G. de Chaelandar and B. Grau: “SVETLAN' - A System to Classify Words in Context”. 14th European Conference on Artificial Intelligence
10. H. Suryanto and P. Compton: “Learning Classification taxonomies from a classification knowledge based system”. 14th European Conference on Artificial Intelligence
11. A.B Williams and C. Tsatsoulis: “An Instance-based Approach for Identifying Candidate Ontology Relations within a Multi-Agent System”. 14th European Conference on Artificial Intelligence
12. M. Kavalec, V. Svatek, P. Strossa: “Web Directories as Training Data for Automated Metadata Extraction”. 12th European Conference on Machine Learning (ECML'01)/ 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)
13. Kenneth Ward and Patrick Hank**s**., “Word Association Norms, Mutual Information, and Lexicography.” *Computational Linguistics* 16(1), 1990, pp. 22-29.
14. J. S. Chang, et al., “Large-Corpus-Based Methods for Chinese Personal Name Recognition.” *Journal of Chinese Information Processin*g 6.3 1992, pp. 7-15.
15. J. S. Chang, S. D. Chen, S. J. Ker, Y. Chen, and J. Liu, “A multiple-corpus approach to recognition of proper names in Chinese texts.” *Computer Processing of Chinese and Oriental Languages*, 8(1), 1994, pp. 75-85.
16. J. S. Chang, “Automatic lexicon acquisition and precision-recall maximization for untagged text corpora.” *Ph.D. Thesis, Dept. of Electrical Engineering, National Tsing-Hua University,*Taiwan, 1997.
17. C. R. Huang, “Introduction to CKIP Balanced Corpus.” *Proceedings of ROCLING VIII*, 1995, pp. 81-99.
18. H. H. Chen , J. C. Lee, “The Identification of Organization Names in Chinese Texts.” *Communication of Chinese and Oriental Languages Information Processing Society*, **4**(2), Singapore, 1994,pp. 131-142.
19. J. C. Lee, Y. S. Lee and H. H. Chen, “Identification of Personal Names in Chinese Texts,” *Proceedings of ROCLING VII*, 1994, pp. 203-222.
20. H. H. Chen and Y. Y. Wu, “Aligning Parallel Chinese-English Texts Using Multiple Clues.” *Proceedings of 2nd Pacific Association for Computational Linguistic Conference*, 1995, pp. 39-48.
21. H. H. Chen and J. C. Lee, “Identification and Classification of Proper Nouns in Chinese Texts,” *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.