Structured POI data Extraction from Internet News

Hua-Ping Zhang
School of Computer Sciences
Beijing Institute of Technology
Beijing, P.R.C 100081
Email: kevinzhang@bit.edu.cn

Qian Mo
Beijing Technology and Business
University, Beijing, P.R.C 100048

He-Yan Huang
School of Computer Sciences
Beijing Institute of Technology
Beijing, P.R.C 100081

Abstract: **POI (Point of Interest) data is key resources for GPS application. Manual POI collection is expensive and time consuming. This paper presents a novel approach that automatically extracts structured POI data from Internet news articles. The procedure includes erasing noisy news document with POI linguistic features, making lexical analysis on the remaining texts using ICTCLAS2010, identifying time expression and the full name of POI location and organization, extracting the relationship between entities, and getting structured data given a POI event based on extraction modeling. The POI extraction model is computed with the term frequency and word distance, without any syntax analysis, scenario template or relationship induction. Consistency and validity check were employed to optimize result. Open testing with experiment conducted on 1,000 news articles, the precision is 97.30% and recall is 75.48%. The approach has been applied in industrial POI collection. POI oriented event extraction is effective.**

Keywords：information extraction; extraction model;relation extraction;POI ICTCLAS2010

I. INTRODUCTION

A point of interest, or POI, is a specific point location that someone may find useful or interesting. A name or description for the GPS POI is usually included. And other information such as the related products or services, even a telephone number may also be attached. Information about a specific cafe or parking lot at a given street is commonly used POI information. While the end user tries to locate his destination, such data cannot have any little mistake, and any change on POI must be updated as soon as possible. GIS data suppliers have to keep hundreds of vehicles running and recording any change at each location from morning till night. Without any hint or schedule, such aimless circling is expensive and time-consuming.

With the development of Internet, related change tends to be instantly announced on the news articles or BBS pages. We present a simple illustration with the news[1] from the official website from the Central Government of China. Given in figure 1, some important attributions related to the event should be extracted shown in the TABLE I. POI Entity is defined as the subject of POI event, such as a hotel, a road, or a supermarket. Moreover, the output location should be specific for geographic navigation. For example, news about “a movie star participated the open ceremony of a shop” would be discarded if there was no description about the shop address. Event type is defined as the POI category. It is represented with feature words, such as 通车（*or traffic open*）, 道路封闭（*or traffic close*） and 限行（*or traffic restraint*）. Timestamp is not the publish date, but the date of event occurrence. Such structured data in the table is more flexible for further utilization than unstructured news. In common sense, the structured data could be regarded as GPS POI data.

> 12 月 6 日，车辆驶上聚（源）青（城山）路。当日，四川省都江堰市聚源镇至青城山道路通车。该路是都江堰启动灾后重建首个新建道路项目，道路全长 10.75 公里，全线设计时速 60 公里/小时，采用一级公路标准，双向 6 车道，并按 8 级抗震设防。新华社记者 刘海 摄
>
> *Or English: Dec.6, the vehicles riding on Ju Qing Road. On the day, In the Dujiangyan City, Sichuan Province, the road from Juyuan Town to Qing Cheng Mountain has been opened to traffic. The road, which has a total length of 10.75 kilometers, is the first rebuilt road project. It is warranted that the maximum designed speed of 60 km/h, and use one-class road standard with two-way six lanes and with 8 seismic intensity protection. Pictures taken by Xinhua News Agency reporter Hai Liu.*

Figure 1. News related to POI data change

TABLE I. POI EVENT ATTRIBUTIONS

POI Entity	聚青路 or *Ju Qing Road*
POI Location	四川省都江堰市聚源镇至青城山 or *from Juyuan Town to Qing Cheng Mountain*
Event Type	通车 *or traffic open*
Timestamp	12 月 6 日 *or Dec.6*

How to automatically generate POI data from news articles? There are three aspects of this problem have to be addressed. Firstly, Chinese named entities in POI data are hard to recognize due to the word segmentation problem. Furthermore, recognition of the specific POI location or organization name is more complicated than the general location name. From the sentence “位于城关区皋兰路 38 号的兰州市第二家‘竞彩’加盟店正式开门纳客。” (or *The second lottery franchise store was formally opened at No. 38 Gaolan Road,*

[1] URL：http://www.gov.cn/jrzg/2009-12/06/content_1481290.htm

Chengguan District), the output POI entity should be “兰州市第二家‘竞彩’加盟店”(or *The second lottery franchise store*) and the POI location should be “兰州市城关区皋兰路38号”, which was generated from the four separate tokens “兰州市”(Lanzhou City), “城关区”(Chengguan District), “皋兰路”(Gaolan Road) and “38 号”(No. 38). Secondly, POI extraction has to tackle some tough natural language processing (NLP) tasks, such as disambiguation, temporal inference and co-reference resolution. For instance, “成都银行重庆分行本月底在渝开业”(or *Chongqing Branch of Bank of Chengdu was opened in Yu at the end of this month*). Here, “成都”(or *Chengdu*) and “重庆”(or *Chongqing*) are not referred as a city name, but a bank name. Here, both “重庆”(or *Chongqing*) and “渝”(or *Yu*, abbreviation of Chongqing) are co-referred as POI location. The precise POI timestamp should be referred from “本月底”(or *at the end of this month*) and the publish time. Last one but not least, there should be taken more account on relationship between different entities. After the lexical analysis and named entity recognition, there are ambiguous entities, locations and time expressions. It needs further analysis to filter irrelevant candidates and extract the precise description on a given POI event.

This paper put forward a novel approach that aims to extract structured POI data from open news articles automatically. It has solved such problems mentioned above. The performance is competitive by precision and recall. Afterwards, the proposed solution has been practically applied in a GIS data supplier. The remainder of the paper is organized as follows. Section 2 gives a brief survey on related works of information extraction, focusing on the classification and contrast. Section 3 presents our approach of extracting POI data from open news. And Section 4 reports on our experimental results. The paper concludes with an illustration of practical system.

II. Related work

The extraction of data, structure and relation from open noisy, unstructured web sources is a challenging task, which has engaged a veritable community of researchers from NLP, information retrieval and database [1] [2] [12]. IE as a subject and standards of evaluation and success up to MUC-5 were surveyed in [16], and broadly one can say that the field grew very rapidly when ARPA, the US defense agency, funded competing research groups to pursue IE, based initially on scenarios like the MUC-4 terrorism events[17][18].

In the past few years, previous approaches to the problem of information extraction were categorized into three types by the structure of web pages: structured, unstructured and semi-structured.

Structured Approach. Structured webpage is defined as web pages with predefined and strict format. The page tends to be dynamically generated from underlying database. Such information can easily be correctly extracted using the frame model. Usually quite simple matching techniques are efficient for extracting provided the page template is known. Structured approach focused on DOM tree with html tags and template learning based on sample pages. Structured approach usually made use of hand-made wrappers using general-purpose programming languages [3] [4]. Besides web page structure, the structure of web sites also plays key roles in extraction [11]. Reference [15] puts forward augmenting automatic information extraction with visual perceptions.

Unstructured Approach. Unstructured webpage is made up of free text with natural language, such as news articles, technical reports. IE systems for unstructured web pages has generally used natural language techniques and the extraction rules are typically based on patterns involving syntactic relations between words, part-of-speech, and phrases or even named entities [5]. The rules or patterns often hand-made or learned automatically from training examples tagged with the right label by experts. Natural language understanding on unrestricted domain is far from the practice. However, information extraction for a special purpose can work if we can well define the priori knowledge. Domain dependent knowledge or rules can be trained with machine learning algorithm, such as probabilistic model, conditional random field model [2] [6] [7][10]. Semantic or ontology often used in free text extraction [13].

Semi-structured Approach. Semi-structured webpage, such as product introduction page or academic paper, is an intermediate between structured record of format data and unstructured text. For instance, a paper has structured meta-data such as title, author, affiliation and contacts, while the full content is free texts. Semi-structured approach often utilizes heuristic-based wrappers on structured data and natural language techniques on texts. Structured result could help disambiguation on free texts. Reference [8] described an approach of automatic information extraction from semi-structured web pages by pattern discovery.

More recently, Reference [9] proposed domain-independent information extraction from web tables. However, unstructured extraction usually depends on the given domain and application [6].

The tools for information extraction include HTML-aware tools, NLP-based tools, wrapper induction tools, modeling-based tools and ontology-based tools.

This paper is the first report focusing on extracting POI data update from unstructured news articles. It involves location, organization and time expression identification, temporal inference, relationship extraction and event extraction. In addition, the practical system was divided into two individual components: extractor and knowledge base. Therefore, the extractor could be applied in similar domains with the appropriate knowledge base.

III. POI data extraction and integration

A. Architeture

The POI data extraction and integration is divided into four main stages. It includes: text preprocessing, recognition of full entity name, POI extraction modeling and result optimization. The architecture is illustrated in Figure 2.

In the architecture, there are two individual parts: the extractor and knowledge base. The extractor is designed for general purpose. In this work, the knowledge is POI related. However, knowledge representation is domain independent. Therefore, the approach could be extended to information extraction on similar domain.

B. *Text Preprocessing*

This procedure includes erasing noisy news document with POI linguistic features, and then making lexical analysis on the remaining texts, identifying time expression, location and organization entities using shareware ICTCLAS2010[2], which is one of the most popular Chinese lexical analyzer.

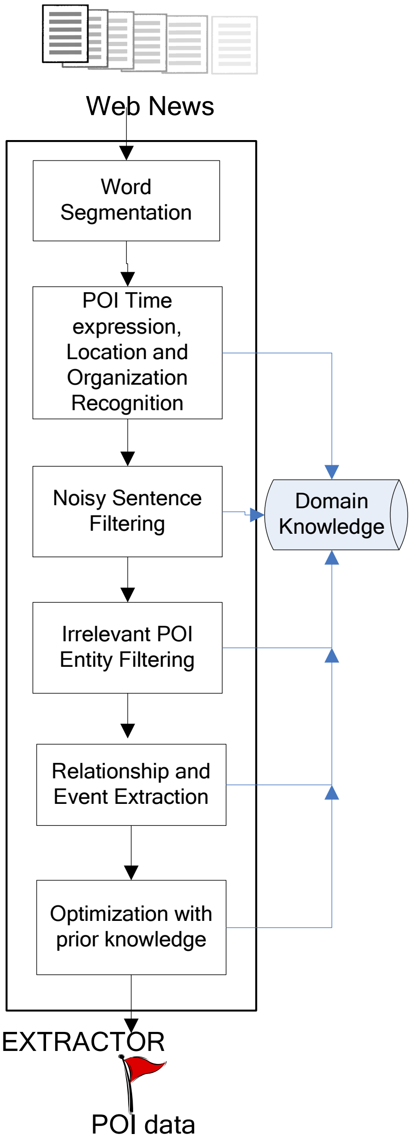

Web News
Word Segmentation
POI Time expression, Location and Organization Recognition
Noisy Sentence Filtering
Irrelevant POI Entity Filtering
Relationship and Event Extraction
Optimization with prior knowledge
Domain Knowledge
EXTRACTOR
POI data

Figure 2. Architecture of POI Extraction

After lexical analysis, all the sentences and words are segmented, and time expressions are marked with '/t', location entities are marked with '\ns' or '\nsi'.

Based on ICTCLAS2010, a location lexicon with 640,000 entries was imported as the user-defined lexicon of ICTCLAS2010. It can improve the precision and recall of base location name recognition.

C. *Recognition of full entity name*

Traditional Chinese lexical analyzer usually produces tokens with small granularity. For instance, the time expression "12 月 6 日"(or Dec. 6) was not tokenized into "12 月 6 日/t", but "12 月/t 6 日/t". Meanwhile, a whole POI address "四川省都江堰市聚源镇" will be segmented into "四川省/ns 都江堰市/ns 聚源镇/nsi". The problem with organization name is much more severe.

```
12 月/t 6 日/t ，/wd 车辆/n 驶上/v 聚/v （/wkz 源/ng ）/wky 青/a （/wkz 城山/ns ）/wky 路/n 。/wj 当日/t ，/wd 四川省/ns 都江堰市/ns 聚源镇/nsi 至/p 青城/ns 山/n 道路/n 通车/vi 。/wj 该/rz 路/n 是/vshi 都江堰/ns 启动/v 灾/n 后/f 重建/v 首/m 个/q 新建/v 道路/n 项目/n ，/wd 道路/n 全长/n 10.75/m 公里/q ，/wd 全线/n 设计/vn 时速/n 60/m 公里/小时/n ，/wd 采用/v 一级/b 公路/n 标准/n ，/wd 双向/b 6/m 车道/n ，/wd 并/cc 按/p 8/m 级/q 抗震/vn 设防/vn 。/wj 新华社/nt 记者/n  刘海/nr  摄/vg
```

Figure 3. Lexical Result of the News Sample

Hence, based on lexical analysis result from ICTCLAS2010, time expression was recognized using regular grammar and POI location and organization entities was combined with sequential tokens using heuristic knowledge.

In this stage, we use a rule-based method to find the maximal Noun phrase of POI entities. By scanning the word and its part of speech, we get the full entity name of POI road, as shown in Fig. 4.

```
四川省/ns 都江堰市/ns 聚源镇/nsi 至/p 青城/ns 山/n 道路/n
```

Figure 4. Sample of POI full road name

However, rule-based method usually occurs some mistakes, therefore an evaluation function was used to select the optimal phrase.

$$W_e = \log(LEN_e) \cdot \log(TF_e) \qquad (1)$$

In formula (1), W_e is the weight of the POI candidate entity name *e*. LEN_e is the count of word in *e*. TF_e is the frequency of the entity name *e*. Finally, the entity of the maximal weight is selected.

After lexical analysis and recognition of POI location, organization and time expressions, the POI item candidates

[2] It can be downloaded from http://hi.baidu.com/drkevinzhang

were selected. At the same time, any sentence without entities, time expressions or POI feature words are discarded as useless information.

D. POI extraction modeling

POI extraction model is designed to compute the coherence measure that a POI attribution *a* to a given POI event feature word *f*.

$$Measure(f,\ a) = \log(1+1/(Distance(f,\ a)+\alpha))+\log(\beta+TF_f)+\log(\beta+TF_a),\ \text{where } \alpha \text{ and } \beta \text{ are smoothing factor.} \quad (2)$$

In formula (2):

- *Measure*(*f*, *a*) is the measurement that POI attribution *a* is coherent to the given event featured word *f*.
- *a* is POI attribution, such as POI location, organization and time expression, recognized from the lexical analysis.
- And *f* is the feature word of given POI event. The sample is listed in Table II. Event feature words are a small part of domain ontology. The features are simply generated both from POI experts and entropy-based feature selection algorithm on given samples.
- *Distance*(*f*, *a*) is the count of words between *a* and *f*.
- TF_f and TF_a are the frequencies of the feature word *f* and POI attribution *a* respectively.

In the extraction model, only term frequency and the distance from POI location, organization or time expression to the given feature word are introduced,. There is no further natural language technique, such as partial parsing, syntax analysis, relationship induction, or semantic interpreter. Except the event feature words, no relation or template knowledge was used. Therefore, the extraction modeling is domain independent.

It can be proved that the POI event feature word is the center for extraction. However, one article could have several feature words. Different POI events can be measured with the extraction models. The most possible POI with the maximum score would be chosen. Similarly, the model can find the most possible attribution among all candidates for a given POI event.

TABLE II. POI EVENT FEAT URE SAMPLE

Category	Sub-Category	Feature sample
Building	Opening	开业(Open)|开张(Open)|开放(open)|营业(Open)|运营(on business) |成立(built)|落户(accomplished)|落成(built)|挂牌(on business)|投入使用(in use)|揭牌(in use)|建成(built)|开建(start to build)|开工(start to build)|新建(newly built)|动工(newly built)|完工(built)|竣工(built)
	Update	搬迁(moved)|迁移(moved)|合并(merged)|更名(rename)|迁址(moved)|改建(modified)|扩建(scaled)|搬家(moved)
	Close	关闭(closed)|倒闭(shut down)|停业(stopped)|歇业(stopped)|拆除(removed)|爆破(exploded)|关门(close)
Road	Update	通车(traffic open)|新增(traffic open)|建成(built)|完工(built)|竣工(built)|拓宽(widen)|开通(open)|开工(start to build)|开建(start to build)|启动(start to build)|奠基(start to build)|拆除(remove)|改造(modify)|更名(rename)|变更(modify)|更改(modify)|命名(named)|交工|新增|建成|施工(under construction)
	Limitation	限行(restraint)|封闭(close)|限速(speed limit)|单行(one way limit)|禁行(no entry)|禁止通行(no entry)|调整(change)

E. Result Optimization

Given a POI event feature, all the candidate location, organization and time expression is generated and sorted by its coherence measure value. Some optimization strategy was employed in this stage. Firstly, temporal inference was performed according to publish date and time expression. Then, the outdated POI event was filtered by judging as useless. Secondly, consistency and validity check would be used to filter illegal POI. Such rules are collected with the POI suppliers. For instance, POI would be removed if its location is out of range. In our work, any data outside the mainland of China will be discarded.

Finally, the structured POI data would be extracted from a given free news article.

IV. EXPERIMENT AND RESULTS

On 1,000 open news articles, several experiments were conducted to validate the effectiveness of the proposed approach.

A. Data Sets

A web crawler system is designed to collect Internet News from the following search engines: Google, Baidu and Sogou. Many GPS POI keywords are used to help find the Appropriate News.

We asked 3 human evaluators to label ground truth data and 2000 news are used for training data collection and 1000 news for testing data collection.

B. *Evaluation Measure*

In our method, evaluation is relatively simple. We could use traditional evaluation method in Information Retrieval.

We use precision (P) and recall(R) to measure the performance:

$$P = |C \cap R| / |R| \quad (3)$$

$$R = |C \cap R| / |C| \quad (4)$$

where R is the set of results returned by our system, and C is the set of manually tagged correct results.

C. *Experimental Design*

As it is well known, the result of information extraction is hard to compare except for experiments with the same tasks on a given test set. What's more, other available systems such as GATE[19], RAPIE[20] and SRV[21], could not be adapted for POI extraction tasks.

One key requirement for making IE a usable technology is developing the ability to produce IE systems rapidly without using the full resources of an NLP research laboratory. The most recent MUC had introduced a task, "co-reference evaluation", with the goal of stimulating more fundamental NLP research. Therefore, the BASELINE experiment is designed only with lexical analysis and co-occurrence in one sentence.

D. *Result and Analysis*

On the basis of BASELINE, each experiment was conducted with an additional NLP module based on the above one. The results are given as follows.

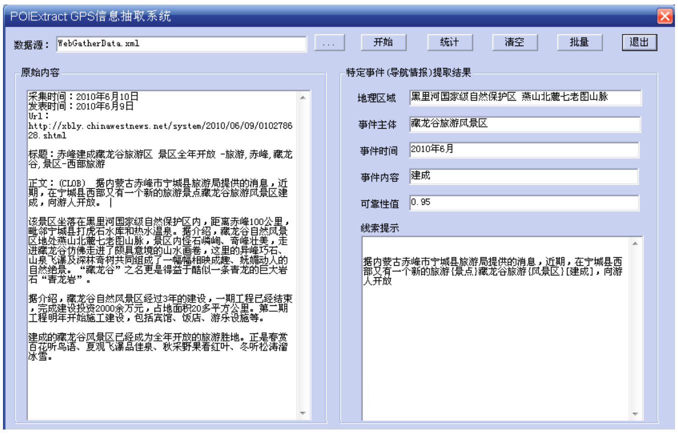
TABLE III. EXPERIMENTAL RESULTS

Different NLP modules	P	R
BASELINE: lexical analysis and co-occurrence in one sentence	97.44	24.51
+noisy sentence removing	98.31	37.42
+irrelevant entity filtering	95.59	41.94
+consistency between event and entities	96.05	47.10
+filtering if no time or location expression	94.87	47.74
+extended dictionary in POI field	94.44	54.84
+time optimization	92.71	57.42
+location optimization	93.44	73.55
+short news removal	97.30	75.48

Compared with BASELINE, the proposed POI extraction method achieved better performance in terms of both precision and recall. From the TABLE III, it indicated that each NLP module solved different problems and improved the performance in POI extraction. And location optimization is the most effective.

V. CONCLUSIONS

This paper presents a novel approach that automatically extracts structured GPS POI data from news articles. Open testing with experiment conducted on 1,000 news articles, the precision is 97.30% and recall is 75.48%. The method within POI oriented event extraction is effective. The approach has been applied in a practical system named POIExtractor, illustrated in the following figure.



数据源：WebGatherData.xml

采集时间：2010年6月10日
发表时间：2010年6月9日
Url：
http://xbly.chinawestnews.net/system/2010/06/09/010278628.shtml

标题：赤峰建成藏龙谷旅游区 景区全年开放 -旅游,赤峰,藏龙谷,景区-西部旅游

正文：(CLOB) 据内蒙古赤峰市宁城县旅游局提供的消息，近期，在宁城县西部又有一个新的旅游景点藏龙谷旅游风景区建成，向游人开放。

该景区坐落在黑里河国家级自然保护区内，距离赤峰100公里，毗邻宁城县打虎石水库和热水温泉。据介绍，藏龙谷自然风景区地处燕山北麓七老图山脉，景区内怪石嶙峋、奇峰壮美，走进藏龙谷仿佛走进了颇具意境的山水画卷，这里的异峰巧石、山泉飞瀑及深林奇树共同組成了一幅幅相映成趣、妩媚动人的自然绝景。"藏龙谷"之名更是得益于酷似一条青龙的巨大岩石"青龙岩"。

据介绍，藏龙谷自然风景区经过3年的建设，一期工程已经结束，完成建设投资2000余万元，占地面积20多平方公里。第二期工程明年开始施工建设，包括宾馆、饭店、游乐设施等。

建成的藏龙谷风景区已经成为全年开放的旅游胜地。正是春赏百花听鸟语、夏观飞瀑品佳泉、秋采野果看红叶、冬听松涛溜冰雪。

地理区域：黑里河国家级自然保护区 燕山北麓七老图山脉
事件主体：藏龙谷旅游风景区
事件时间：2010年6月
事件内容：建成
可靠性值：0.95

据内蒙古赤峰市宁城县旅游局提供的消息，近期，在宁城县西部又有一个新的旅游{景点}藏龙谷旅游{风景区}[建成]，向游人开放

Figure 5. POIExtractor Illustration

This work has successfully solved the problem of POI extraction from open Internet news. The future work focuses on extending the extractor to other domains such as restaurant introduction and product review.

ACKNOWLEDGMENT

The author would like to thank Prof. Qun Liu and Dr. Yang Liu who provided useful suggestions and discussions.

REFERENCES

[1] Eikvil L., Information Extraction from World Wide Web - A Survey - .Technical Report 945, Norvegian Computing Center, 1999

[2] Sarawagi S., Automation in information extraction and integration, Tutorial of VLDB,2002

[3] A. Tengli, Y. Yang, and N. L. Ma. Learning table extraction from examples. In Proc. 20th COLING, pp. 987{993. COLING, Aug. 2004.

[4] Y. A. Tijerino, D. W. Embley, D. W. Lonsdale, Y. Ding, and G. Nagy. Towards ontology generation from tables. World Wide Web, 8(3):261{285, 2005.

[5] D. Downey,M. Broadhead, and O. Etzioni. Locating Complex Named Entities in Web Text. In Proc. of IJCAI, 2007.

[6] S. Sekine. On-demand information extraction.In Procs. of COLING, 2006.

[7] Y. Shinyama and S. Sekine.Preemptive information extraction using unrestricted relation discovery. In Proc. of the HLT-NAACL, 2006.

[8] Chang C-H., Hsu C.-N. and Lui, S.-C. Automatic information extraction from semi-structured web pages by pattern discovery. Decision Support Systems Journal, 35(1): 129-147 ,2003

[9] W Gatterbauer, P Bohunsky, M Herzog, B, "Towards domain-independent information extraction from web tables" Proceedings of the 16th international conference on World Wide Web (WWW2007),2007.

[10] A. Culotta, A. McCallum, and J. Betz. Integrating probabilistic extraction models and relational data mining to discover relations and patterns in text. In Proceedings of HLT-NAACL, New York, NY, 2006.

[11] K. Lerman, L. Getoor, S. Minton, and C. A. Knoblock. Using the structure of web sites for automatic segmentation of tables. In Proc. SIGMOD, pp. 119{130. ACM, June 2004.

[12] B. Liu and K. C.-C. Chang. Editorial: special issue on web content mining. SIGKDD Explorations, 6(2):1{4, 2004.

[13] B. Parsia and P. F. Patel-Schneider. Meaning and the Semantic Web. In Proc. IRW at 15th WWW, May 2006.

[14] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible web document analysis for delivery to narrow-bandwidth devices. In Proc. 6th ICDAR, pp. 1074{1078. IEEE, Sept. 2001.

[15] K. Simon and G. Lausen. ViPER: augmenting automatic information extraction with visual perceptions. In Proc. 14th CIKM, pp. 381{388. ACM, Nov. 2005.

[16] J. Cowie, & W. Lehnert, Information Extraction, in (Y. Wilks, ed.) Special NLP Issue of the Comm. ACM. 1996

[17] DARPA. Proceedings of the Third Message Understanding Conference (MUC-3), San Diego, California.Morgan Kaufmann. , 1991

[18] DARPA. Proceedings of the Fourth Message Understanding Conference (MUC-4), McLean, Virginia Morgan Kaufmann. , 1992.

[19] H. Cunningham, R. Gaizauskas & Y. Wilks, GATE: a general architecture for text extraction, University of Sheffield, Computer Science Dept. Technical memorandum, 1995

[20] M. E. Califf, R.J. Mooney, Relational Learning of Pattern-Match Rules for Information Extraction. Proceedings of the ACL Workshop on Natural Language Learning, Spain, July 1997.

[21] D. Freitag, Multistrategy learning for Information Extraction.. Proceedings of the 15th International Conference on Machine Learning(ICML-98),Madison,Wisconsin, July 1998.