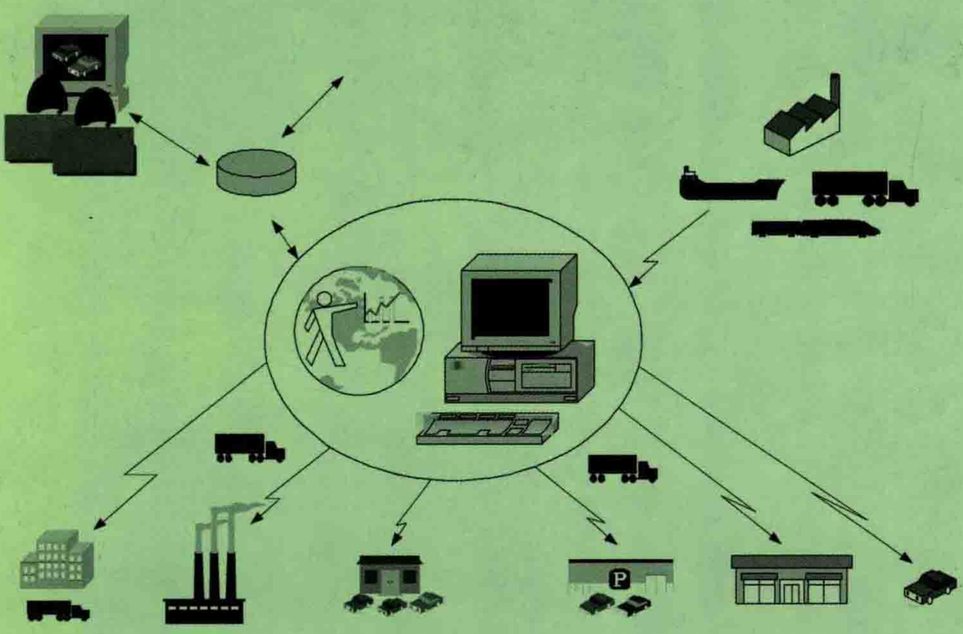
The Internet Challenge: Technology and Applications

Edited by
Günter Hommel and Sheng Huanye



Kluwer Academic Publishers

The Internet Challenge: Technology and Applications

Proceedings of the 5th International Workshop
held at the TU Berlin, Germany, October 8th-9th, 2002

Edited by

GÜNTER HOMMEL
Technische Universität Berlin,
Berlin, Germany

and

SHENG HUANYE
Shanghai Jiao Tong University,
Shanghai, China

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-0903-8

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands.

The Internet Challenge: Technology and Applications

Workshop Co-Chairs

Günter Hommel, Technische Universität Berlin
Sheng Huanye, Shanghai Jiao Tong University

Program Committee

Martin Buss
Kurt Geihs
Sergei Gorlatch
Hans-Ulrich Heiß
Günter Hommel (Chair)
Sheng Huanye
Adam Wolisz
Fritz Wysotzki

Organizing Committee

Wolfgang Brandenburg
Michael Knoke

Workshop Secretary

Gudrun Pourshirazi
Silvia Rabe
TU Berlin
Institut für Technische Informatik
und Mikroelektronik
Einsteinufer 17
10587 Berlin
Germany

Preface

The International Workshop on "The Internet Challenge: Technology and Applications" is the fifth in a successful series of workshops that were established by Shanghai Jiao Tong University and Technische Universität Berlin. The goal of those workshops is to bring together researchers from both universities in order to exchange research results achieved in common projects of the two partner universities or to present interesting new work that might lead to new cooperation.

The series of workshops started in 1990 with the "International Workshop on Artificial Intelligence" and was continued with the "International Workshop on Advanced Software Technology" in 1994. Both workshops have been hosted by Shanghai Jiao Tong University. In 1998 the third workshop took place in Berlin. This "International Workshop on Communication Based Systems" was essentially based on results from the Graduiertenkolleg on Communication Based systems that was funded by the German Research Society (DFG) from 1991 to 2000. The fourth "International Workshop on Robotics and its Applications" was held in Shanghai in 2000 supported by VDI/VDE-GMA and GI.

The subject of this year's workshop has been chosen because both universities have recognized the fact that internet technology is one of the major driving forces for our economies now and in the future. Not only the enabling technology but also challenging applications based on internet technology are covered in this workshop. The workshop covers the scope from information extraction, data analysis, e-learning and e-trading over the areas of robotics, telepresence, communication techniques and ends with metacomputing, electronic commerce, quality of service aspects and image retrieval.

At TU Berlin the German Research Society (DFG) has been funding a new Graduiertenkolleg on "Stochastic Modeling and Analysis of Complex Systems in Engineering" since 2000. Results from this Graduiertenkolleg but also from other projects funded by different institutions are presented in this workshop. The workshop is supported by the special interest group of GI and ITG "Communication and Distributed Systems". Financial support by DAAD for the bilateral exchange of scientists between our universities is gratefully appreciated. We also gratefully recognize the continuous support of both universities that enabled the permanent exchange of ideas between researchers of our two universities.

Berlin, June 2002
Günter Hommel, Sheng Huanye

Contents

without any language barriers from the Internet is the challenge for many scientists and experts on different research areas. Because of the complexity of natural languages, accurate information retrieval and robust information extraction still remain tantalizingly out of reach.

Information extraction is the process of identifying relevant information where the criteria for relevance are predefined by the user in the form of a template. Below is a passage of investment news in Chinese:

本报讯，世界最大的芯片制造巨头英特尔宣布增加在华投资。该公司宣布向位于上海浦东外高桥的生产制造企业新增投资 3.02 亿美元，使其在上海的封装/测试厂投资总额达到 5 亿美元。这次新追加的投资将引进技术和设备，用于验证、测试和封装最新的支持英特尔奔腾 4 处理机平台的英特尔 845 芯片组[1]。

The filled template corresponding to the above news is:
Company name:英特尔
Company to be invested: 生产制造企业
Its place: 上海浦东
Newly invested money: 3.02 亿美元
Amount of invested money: 5 亿美元
Currency: 美元
Content of investment: 用于验证、测试和封装最新的支持英特尔奔腾 4 处理机平台的英特尔 845 芯片组

Even if Information Extraction seems to be a relatively mature technology, it still suffers from a number of unsolved problems that limited the application only on the predefined domain. Many effects have been focused on such issues [1][2][3].

In this paper, we introduce an investment information extraction system, which is oriented to the multilingual information extraction. German, English and Chinese have applied in this experimental system. In the following, the system architecture is first described, then, some features of the system are introduced, such as predefined templates and dynamic acquired templates, language-independent templates and language-dependent patterns. Finally some evaluations and conclusions will be given.

[1] Translation: News report, The world biggest chip maker INTEL company has announced that it will invest another 302 million US$ to the Production and Manufactory enterprise situated in Shanghai Pu Dong economic zone. The amount of investment in Shanghai INTERL package and test factory has reached 500 million US$. The newly appended money will be used in buying new devices and technologies for evaluation and test the new 845 chips supporting for Pentium 4 process platform.

2. SYSTEM ARCHITECTURE

The system (shown in figure 1) consists of three parts: user query processing, extraction based on templates and patterns and dynamical acquisition:

- User query processing provides two possibilities for the users. One is template-based keyword. The other is natural language question, such as who invested in some company? Which company has been invested by INTEL company? There are question templates defined in the system. If the user inputs other questions, the system picks up the most similar question to process. All the queries can be made in one of the three languages: Chinese, English and German.
- Extraction module receives a user query and searches the tagged corpus for relevant contents according to predefined templates and patterns. The templates are language independent, it is defined by the event in reality. However patterns describe the extraction rules related to each slot, they are different from one natural language to another.
- Dynamic acquisition extracts the templates and the patterns from the tagged corpus to complement the predefined templates made by human being. It makes the extraction system easy to adapt to other domains.

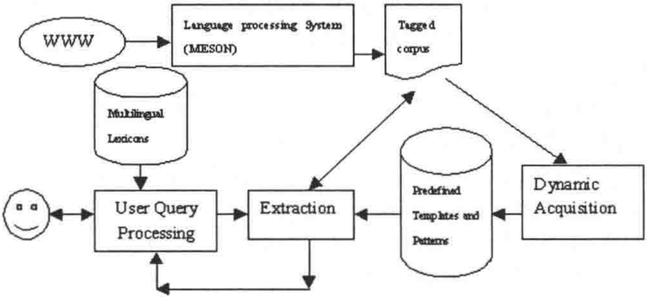
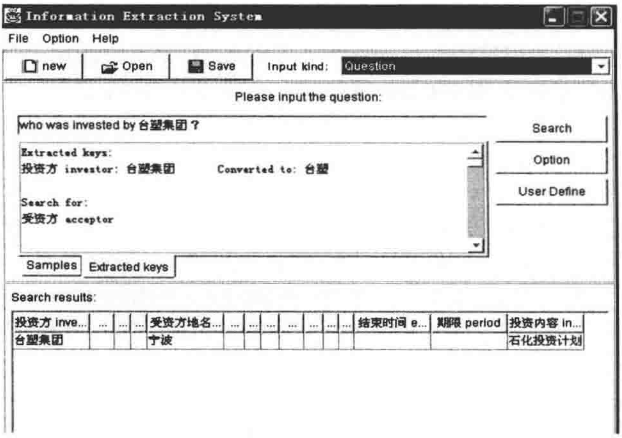


Figure-1 Architecture of the system

The corpus is collected from financial news on the Internet, it is tagged by a shallow parsing system called MESON[2], with English and German shallow parsing ability. For Chinese, It integrates a modern Chinese Automatic Segmentation and Part-of-Speech (POS) Tagging System[3], some Chinese grammar rules[4] for company name, person name, money, dates and other name entities (NE) are added in the MESON. The interface of the system is in the figure 2. It is realized by Java and XML.



Input kind: Question

Please input the question:

who was invested by 台塑集团 ?

Extracted keys:
投资方 investor: 台塑集团 Converted to: 台塑

Search for:
受资方 acceptor

Search results:

投资方 inve...	...	...	...	受资方地名...	...	...	...	...	...	...	...	结束时间 e...	期限 period	投资内容 in...
台塑集团				宁波										石化投资计划

Figure-2 Question-Interface of the system

Comparing with other information extraction systems, the experimental system has some features in the following:

1. Integrating some components already developed instead of starting from scratch. The system consists of components of different functions, either borrow from others or develop by ourselves.
2. Combining static information with dynamic information, such as predefined templates and dynamic acquired templates, template-based keywords and question query in user query processing.
3. Focusing on multilingual texts to realize a multilingual information extraction based on language independent templates and language dependent patterns. Therefore, uniform processing can be achieved for different languages. In the following, some features are detailed.

[2] MESON was developed by Markus Beck in German Research Center for Artificial Intelligence (DFKI), Saarbruecken, Germany.

[3] Modern Chinese Automatic Segmentation and POS Tagging System was developed by Shan Xi University in CHINA.

[4] Written by Edith Klee and Tannja Scheffler from Saarbruecken University.

3. PREDEFINED TEMPLATES AND DYNAMIC ACQUIRED TEMPLATES

In the system, we first defined two templates about the investment event by hand. One template is about the normal activity of investment such as the example in Introduction, the other is about the investment on stock market, which concerns shares, the unit price and so on. Through dynamic acquired process, we can get the templates about merge, acquisition and so on, as long as texts describing such events appear in the corpus. The process is described in the following:

1. The user inputs an event name, all examples about this event in the corpus will be identified.
2. Show the first example with identified actors of the event, time of the event, location of the event and so on, i.e. define slots of the template, and also the type of slots according to the POS or NE automatically.
3. User can make some corrections on those identified slots. The correction is made on all of the examples related to this event automatically.
4. Show the next example, and do step 2 and step 3.
5. Repeat until the final template is correct and complete.

In fact, more examples will give the new template a wide coverage, it needs of cause more human help. On the other hand, the template can be generated automatically without user interaction, however, the precision is unsatisfied. We should balance between the two factors. For example, there is an event: acquisition. The news report is as following:

Isaac 公司将以换股形式收购软体公司, 收购总价格为 8 亿 1000 万美元[5]

After tagging, the passage became:

[FIRM-NP Isaac 公司] ("将" ("将" NIL . :D)) ("以" ("以" NIL . :P)) ("换" ("换" NIL . :V)) [NP 股 形式] ("收购" ("收购" NIL . :V))[FIRM-NP HNC 软体公司] ("，" ("，" NIL . :W)) [NP 收购 总 价格] ("为" ("为" NIL . :P)) [MONEY 8 亿 1000 万美元] ("。" ("。" NIL . :W))

The system identifies the actor of the acquisition is Isaac 公司(Isaac company), the type of the actor is FIRM-NP, the company to be acquired is HNC 软体公司 (HNC soft company) and its type is also FIRM-NP. The money of acquisition is: 8 亿 1000 万美元 (810 million). From the above example, the system generates three slots for this event, after analyzing the second example, the system may find another 2 slots, such as the place of the

[5] Translations: Isaac company will acquire the HNC company by stock holding. The amount of the acquisition is 810 million US$.

acquisition, and the date. The slots will be more and more complete as the system analyzing more examples. During the whole process, a user can make correction before the template is generated. Finally the system generates the new template of acquisition event with XML format in the following:

```
<template event="acquisition" id="t1001">
<slot id="s1001001" name="actor of the acquisition" type="FIRM-NP">
<slot id="s1001002" name="acquired company " type="FIRM-NP">
<slot id="s1001003" name="amount of money" type="MONEY">
<slot id="s1001004" name="acquisition date" type="DATE">
<slot id="s1001005" name="place of the acquisition" type="PLACE">
<slot id="s1001006" name="the percentage of acquisition " type="M">
<slot id="s1001007" name="country of the acquired company"
type="COUNTRY">
```

4. LANGUAGE-INDEPENDENT TEMPLATES AND LANGUAGE-DEPENDENT PATTERNS

In order to realize a multilingual information extraction system, templates combing patterns is our solution to solve the multilingual problems. The structure of the templates and patterns is in the following:

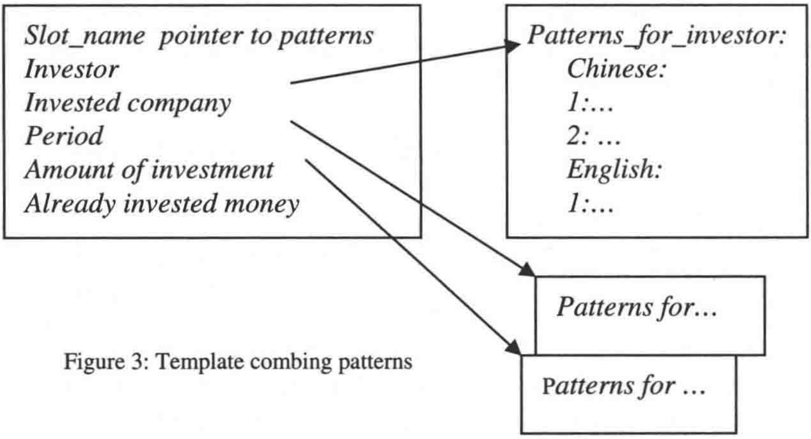


Figure 3: Template combing patterns

Templates and patterns are written in XML in order to have unified resources and platform-independent realization. For example, some patterns of Chinese and English for the actor of investment is listed in the following:

```
<template id="t001" event="investment">
    <slot id="s001001" name="investor" type="FIRM-NP">
      <pattern language="Chinese">
        <stuff type="FIRM-NP"/>
        <substitutable>投资</substitutable>
        <phrase type="MONEY"/>
      </pattern>
      <pattern language="English">
        <stuff type="FIRM-NP"/>
        <select>
          <li>are</li>
          <li>is</li>
        </select>
        <substitutable>interested in</substitutable>
        <substitutable>investing in</substitutable>
        <select>
          <li>
            <phrase type="PLACE"/>
          </li>
          <li>
            <phrase type="FIRM-NP"/>
          </li>
        </select>
      </pattern>
```

In the above, *stuff type* means the type of the slot, *substitutable* means the word can be substituted by its synonym or other morphological forms, *select* means one of them can be selected, *phrase type* means the type of a phrase followed, *removable* means the word may be removed, *notext* means there is no text between two words. With those parameters, we can describe all kinds of patterns in multilingual form.

5. SYSTEM EVALUATION AND ANALYSIS

The training corpus consists of 80 news reports collected from the Internet. There are about 400 sentences in the corpus. The test corpus consists of 50 news reports. The test result for extraction is shown in the table 1. Let *all* be the total number of extracted slot content, *Act* the number of correct and *false* number of wrong answer, *miss* the number of omitted one. We define the precision is Act/all, recall is Act/ (all+miss).

Table -1. The result of extraction

	Precision	Recall	P & R
Investor name	83%	76%	80%
Invested name	64%	58%	61%
Location of the investor	69%	69%	69%
Location of the invested	88%	88%	88%
Stock buyer	100%	90%	95%
Stock seller	100%	75%	88%
Amount of money	90%	77%	83%
date	100%	89%	94%

The performance is OK according to the state-of-the-art and to the time spent on development. However, we analyze the remaining errors as following:

- Some errors are caused due to the absence of reference resolution and discourse analysis. At the moment, in our system, there is no co-reference resolution and discourse analysis, the system cannot recognize the real entities of pronouns, such as, it, this and so on. Some key information is across the sentence boundary. Therefore, some slots has been omitted or misunderstood.
- Some errors are caused by the diversities of natural languages, the complexity of the order. The more patterns, the higher the precision.

For the template dynamic generation, according to the test, we find that without human corrections, the precision is 85.27%, the recall is 78.01%, if a user makes corrections on 5 examples, the precision is 88.55%, the recall 82.27%, if the user makes corrections on 10 examples, the precision and recall remain the same, if the user corrects more examples, for example, 15 or 20 examples, the precision is 89.31%, and the recall is 83.57%. Therefore, a user makes corrections on 5 examples, both the precision and the recall will rise 3% or so. Not too much human involvement is needed in the dynamic generation.

CONCLUSION

In this paper, we describe the multilingual investment information extraction system based on templates and patterns. Currently the user queries can be in Chinese or English or German. But the tagged corpus is only in