Bilingual Lexicon Extraction From Internet

Fang Li
Dept. Of Computer Science,
Shanghai Jiao Tong University, Shanghai, China 200030
lifang@cs.tu-berlin.de

Wilhelm Weisweber
Dept. Of Computer Science
Technische Universität Berlin, Berlin, Germany
ww@cs.tu-berlin.de

Abstract

This paper introduces an experimental system which can extract translations of words and phrases from the Internet through alignment on parallel WWW pages. The automatic extraction takes place online, is language independent and incrementally formed after a post-editing step by a human being. Actually the experimental system can extract words and phrases between pairs of the languages English, German and Chinese. It is a simple and effective way and is quite different from the traditional extraction techniques.

Introduction

The Internet provides a vast source of information. Compared with traditional dictionaries, databases or repositories, Web information is dynamic, semi-structured and can be represented in many forms, shared over multiple sites and platforms. Extraction lexicons from the Internet can reflect the changes on lexicons in many areas such as education, technologies, economics, socials or politics (Fung & Lo Yuen Yee 1998). It seems that the Internet can be regarded as one of the largest, up-to-date multilingual corpora for natural language processing.

On the Internet, any individual or institution can create WWW sites with an unrestricted number of documents and links. Because the Internet is an international medium shared by many people in the world, many documents are presented in their mother tongues and one or more translations in other languages. These pages have the same content and most of them have similar document styles and structures (Resnik 1998, Li, Sheng & Weisweber 1999). WWW pages which are translations of each other are called parallel. In general a link to a parallel page appears on the main (entry) page of each site. Bilingual pairs of words and phrases can be found on these parallel pages through alignment.

Many traditional lexicon extraction techniques include tokenization, part-of-speech tagging, disambiguation, calculating frequencies, some use shallow parsing and so on. Such methods are more or less language specific. For different languages they need different methods for tokenization and parsing which vary with the usage by human beings and it is difficult to have a broad coverage. In this paper a method which does not rely on the specific syntax of a language is presented. We rely on the markup tag for segmentation, use an alignment algorithm for aligning the segments of parallel pages and finally extract word and phrase translations in the alignment result. In the following a markup language on the Internet is first discussed and then the whole system will be explained. After that the method of extraction is given in detail.

1 Markup Languages

Markup languages describe the format of information on the Internet. The most popular one is HTML which is used for publishing hypertext on the World Wide Web. There is a famous search engine named "Yahoo". It has many parallel pages with different natural languages. We take some source code from its English and German WWW pages as an example.

English version:

```
<html><head><title>Yahoo!</title><base
href=http://www.yahoo.com/></head><body>
<center><form
action=http://search.yahoo.com/bin/search>
<ahref="/homet/?http://auctions.yahoo.com">
<b>Yahoo! Auctions</b></a><br><small>
<ahref="/homet/?http://list.auctions.yahoo.co
m/27813category.html">Pokemon</a>
......
<input type=submit value=Search>
<a href=r/so>advanced search</a>
</td></tr></table><table border=0
cellspacing=0 cellpadding=4 width=600><tr>
<td nowrap align=center><small>
<a href=r/sh>Shopping</a>.....
<a href=r/os><b>Auctions</b></a> -
<a href=r/yp>Yellow Pages</a> -
<a href=r/ps>People Search</a> -
<a href=r/mp>Maps</a> -
<a href=r/ta>Travel</a> -
<a href=r/cf>Classifieds</a> -
<a href=r/pr>Personals</a> -
.....
```

German version:

```
<html><head><title>Yahoo!
Deutschland</title><base
href=http://de.yahoo.com/r/></head>
<body><center><formaction=http://de.search.ya
hoo.com/search/de>
<ahref="/home/kleinanzeigen/?http://de.classi
fieds.yahoo.com/de/">Y! Kleinanzeigen</a><br>
<ahref="/home/klein_autos/?http://de.classifi
eds.yahoo.com/de/car/">Autos</a>

<INPUT type=submit value="Suche
starten"> <a href=ba>Erweiterte Suche
</a></td></tr></table>
<table border=0 cellpadding=4 cellspacing=0
width=630><tr><td align=center>
<b><ahref="/home/shopping_promo/?http://de.sh
opping.yahoo.com/">Yahoo! Shopping</a>
....
<a href=ac><b>Auktionen</b></a> -
<a href=cp>Jobs</a>,
<a href=cl>Autos</a>,
<a href=cq>Immobilien</a> -
<a href=tx>Reisen</a> -
<a href=ja>Branchenbuch</a> -
.....
```

The formatting information is included in <...>. The text information occurs between <...> and </...>. From this some conclusions can be drawn:

- In HTML documents and other markup documents, some words and phrases are already marked up by a tag such as "Yellow pages" by a link <a href = r/yp>, "Yahoo" by a title <title> on the above. Many different tags can be used.

- The formatting of a document can give some hints on the content of the document, while content itself is usually expressed by words or phrases.

- The English and German versions above have a similar style which means that the formatting information is similar. The contents are translations of each other. If both pages are aligned, some translations of words or phrases can be found, such as:

  advanced search ------- erweiterte Suche
  Auctions ------ Auktionen
  Yellow Pages ----- Branchenbuch
  Travel ------- Reisen
  .....

Based on the above observations, lexical information including phrases can be extracted from the Internet which hosts a huge amount of documents and other kinds of information.

2. System Design

The structure of our experimental system is described in fig.1 at the end of this article. It is realised in JAVA running on a Unix Solaris platform. The input of the system are URL addresses of parallel pages. The output are bilingual lexicons. The system consists of three modules: analysis, alignment and extraction.

2.1 Analysis Module

In this module, a markup document is analysed. Some useless tags are filtered out for simplicity, some special characters like for example "&...;" are transfered, text segments between tags <...> and </...> and their tags are extracted. The tag is very important because it gives some hints on the text segment. The text segments are extracted one by one from the markup document online and finally are saved in a temporal file. The same work will subsequently be done on the parallel page. The result of this module is bitext generated from the Internet.

2.2 Alignment Module

Alignment is based on two criterions. One is the similarity of the tag. The other is the lexical information of the text segment. In many cases parallel WWW pages have the same style and the same content because they publish the same information in different languages. Suppose there are two parallel pages like for example in German and English, one is called source, the other will be target. These parallel pages can be found by different search engines.

The similarity of the tags is estimated by equation (1)

$$\mathrm{Tag\_Sim}(S_i, T_j) = \frac{2 * \mathrm{LCS}(S_i, T_j)}{\mathrm{length}(S_i) + \mathrm{length}(T_j)} \quad (1)$$

S_i is the tag of the i-th source text unit and T_j is the tag of the j-th target text unit. $LCS(S_i,T_j)$ is the longest common subsequence of S_i and T_j. Equation (1) represents the similarity between two tags. If the two tags are completely identical, the value will be 1. If two tags are completely different, the value will be 0. For example for the tags S_i= <H1> and T_j= <H2>, equation (1) yields 0.5.

The similarity of the source and target text unit according to the content is estimated by word pairs through looking up a dictionary. More word pairs are found, more likely the both text units are translation each other. The following equation is used to describe the similarity according to the lexical information:

$$\text{Word\_pairs}(S_i, T_j) = \frac{2 * \text{NOWP}(S_i, T_j)}{\text{NOW}(S_i) + \text{NOW}(T_j)} \quad (2)$$

In this equation S_i is the i-th source text segment and T_j is the j-th target text segment. NOW(X) is the number of words in X text unit. NOWP(S_i, T_j) is the number of translation pairs occurring in S_i and T_j. A dictionary is used in order to find translation pairs in these two segments. In the beginning the dictionary is empty. The value of equation (2) will be 0 in the first run. After some results have been found and put into the dictionary the value of equation (2) may still be 0 when no translation pair is found in the dictionary, but it is greater than 0, otherwise.

The similarity of source and target text unit is calculated according to the tag and lexical information in the following:

$$\text{Sim}(S_i, T_j) = \text{Tag\_Sim}(S_i, T_j) + \text{Word\_pairs}(S_i, T_j).$$

Then dynamic programming (Gale & Church 1993) is applied on the alignment of the whole bilingual text in order to get the most similarity alignment.

$$\text{Alignment}(S_i, T_j) = \max\Big(\sum_{i,j} \text{Sim}(S_i, T_j)\Big)$$

The result of this module is a set of aligned text segments.

3 Bilingual Lexicon Extraction

3.1 Method Description

Among the alignment results, there are six cases: one-to-zero, one-to-one, one-to-two, two-to-one, two-to-two, zero-to-one. Only one-to-one are possible to be translation pairs. Among one-to-one alignment results, some are word to word, phrase to phrase, sentence to sentence or even paragraph to paragraph. Word and phrase translation pairs are automatically extracted by checking the number of words in the translation pairs. For example, an English phrase should not contain more than four words because in alphabetic languages we adopt phrases to have a limited number of words. For non-alphabetic languages such as Chinese phrases are estimated by restricting the length of a text segment. No word segmentation is applied for non-alphabetic language.

Long text segments such as long phrases, sentences, or paragraphs will be automatically segmented using stop words. For example:

Dekan und Fachbereichsverwaltung
→ Dean and Administration of the Department

segmented into:

(Dekan, Fachbereichsverwaltung)
→ (Dean, Administration, Department)

During the post-editing, people can choose which word or phrase correspond to which word or phrase in the both groups.

After post-editing the translation pairs, which have been automatically extracted, they are put into the dictionary as lexical information used for alignment in the next run.

3.2 Result Analysis

We have investigated about 80 pairs of Web pages which were found by many search engines on the Internet. The precision of alignment is 86.11%, the precision of lexicon acquisition is 70.68% because some information is not suitable for expanding the dictionary, e.g. dates, email addresses, fax numbers, abbreviations and some other segments which are not phrases. In these translation lexicons 55.38% of the entries are word translations and 44.62% are phrases translations.

Compared with a commercial German-English Dictionary (Langenscheidts Taschenwörterbuch English containing more than 80000 entries) and with an electronic German-English Dictionary (about 3.6 MB with more than 114000 entries) the translation lexicons extracted from the Internet, we find that only 37.68% of the entries can be found in the dictionary, 34.08% can be found in the online dictionary; the other 62.32% of the entries can not be found in the commercial dictionary and 65.92% can not be found in the electronic dictionary.

Analysing the translation lexicons extracted from the Internet, we made some observations: There are

- many proper names for people, companies, place and goods, such as *Albert Einstein*, *SAP*, *Bodensee*, *Neuschwanstein* and so on.
- many complex words (composites), for example, *Barockkirchen*, *Basisinstallationen*, *Benutzerberatung* and so on. Using composites is a specific feature of the German language .
- some newly created words especially in the area of information and communication technology, such as, webmaster, webteam, infocenter, kernel-hacking and so on.
- many context translations such as *Ausländer* translating into *international students* in the university, *Fahrpläne* translating into *bus and train timetable* , *nach oben* translating into *back to top* and so on.

Table 1 and table 2 at the end of this article show some excerpts from the result of the system: German-English and English-Chinese.

Conclusion

Lexical information has always been playing an important role in natural language processing, machine translation and Internet applications such as information retrieval, data mining and so on. The extraction of bilingual words and phrases from the Internet will become more and more important as the Internet is widely used and people can easily access the Internet. Such online extracted translation pairs really reflect the usage of words, their morphological forms, their collocations and different senses.

Languages are very rich such that people can use them very flexible. They can use different words to express the same meaning and on the other hand, many single words have more than one meaning. So, a word and its meaning is a real many-to-many relationship. This causes some difficulties for natural language processing systems. Although some traditional dictionaries have been existing for a long time, there are so many differences in context translations that dictionaries can not include all of them (Sue J. Ker & Jason S. Chang 1997). For intelligent human beings it is no problem, but e.g. for machine translation, this is a big problem. Extraction translation lexicons from the Internet can make some effects to solve the problem.

The precision of the extraction of translation lexicons depends on the precision of alignment in our system. While dictionaries (the results of our experimental system) can improve the precision of alignment, they also increase the overhead of processing time.

However parallel WWW pages are only a part of all documents in the Internet. Many WWW pages use dynamic JAVA or frames. The system cannot align these pages. More and more companies use different styles, different content pages with different languages, instead of parallel pages. This gives us a new challenge in the future.

Acknowledgements

We would like to thank Prof. Sheng Huanye and Prof. Yao Tianfang for their support on the work since beginning and suggestive comments on the paper. Many thanks to Dr. Robert Tolksdorf and Dr. Manfred Stede for their help during the realisation of the system. At last we are grateful to the unknown reviewers for their valuable comments and suggestions.

References

Fung, Pascale and Lo Yuen Yee (1998), ”An IR Approach for translating New Words from Nonparallel, Comparable Texts”, pp. 414-420 in Coling-ACL 98

Fang Li, Huanye Sheng and Wilhelm Weisweber (1999), “Extracting and aligning Bilingual Text from Internet Resources”, in Proceedings of the 5th Natural Language Processing Pacific Rim Symposium 1999, Nov.5-7, Beijing, China.

Resnik, Philip (1998), “A preliminary investigation into Mining Web for Bilingual Text”, in AMTA-98, October 1998.

Sue J. Ker, Jason S. Chang, “A class-based Approach to word ” pp 313-343, Computational Linguistics volume 23, Number 2, 1997.

William A. Gale, Kenneth W. Church (1993), “A program for Aligning Sentences in Bilingual Corpora”, pp 75-102, Computational Linguistics vol. 19 March 1993.

Wu, Dekai (1995), “Large-Scale Automatic Extraction of an English-Chinese translation Lexicon”, Machine Translation, 9: 3-4, pp. 285-313, 1995.

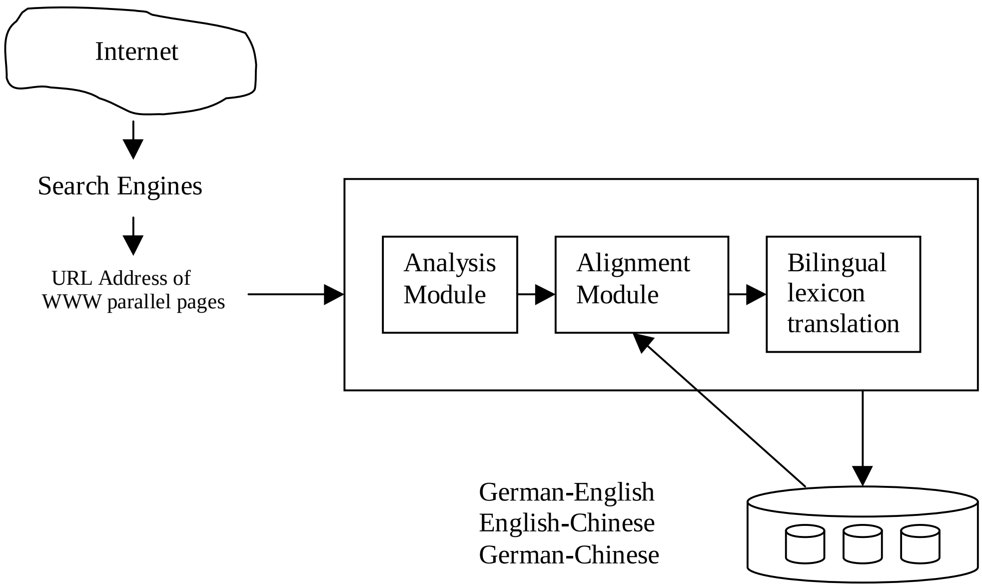

Fig. 1 Architecture of the system

German	English
abgeschlossene Projekte	completed projects
Aktuelles	current event
Aktuelles	latest
Aktuelles	news
Aktuelles	what's new
Albert Einstein	Albert Einstein
Arbeitssicherheit	work safety
Ausländer	international students
Auslandsreferat	international office
bekannt	famous
bekannt	known
Bewerbungsunterlagen	application forms
Bibliothekskataloge	library catalogues
Bibliothek	library
Bodensee	Lake Constance
deutsche Gerichte	German courts
OS2	OS2
shell-scripting	shell scripting
SAP	SAP
Studienberatung	study programmes counseling
Studienfächer	subjects
Studiengänge	study programmes
Studierende	students
Studium	studying

Studium	courses
webmaster	webmaster
webteam	webteam

Table 1 German-English translation Lexicons from Internet

English	Chinese
chat	专题沙龙
chat	聊天室
copyright reserved	版权所有
coverages	新闻报道
directory	网站分类
ebox	网邮
ebuddy	网伴
ebuilder	网建
email	电邮
enews	新闻夹
exactmatch	精确查寻
fashion model	时装模特
fashion	时尚
gallery home	图库主页
health	健康
health	医药
help	问题求助
help	查询
highlights	热门景点
highschools	中学
homepages	个人网页
home	家庭
home	主页
hot categories	热门专题
hot movie	热门影视
hot nations	热门国家
hot stars	热门人物

Table 2 English-Chinese(GB2312 encoding) translation Lexicons from Internet