AUTOMATIC MULTILINGUAL LEXICON GENERATION USING WIKIPEDIA AS A RESOURCE

Ahmad R. Shahid and Dimitar Kazakov
Department of Computer Science, The University of York, YO10 5DD, York, U.K.

Keywords: Multilingual Lexicons, Web Crawler, Wikipedia, Natural Language Processing, Web mining, Data mining.

Abstract: This paper proposes a method for creating a multilingual dictionary by taking the titles of Wikipedia pages in English and then finding the titles of the corresponding articles in other languages. The creation of such multilingual dictionaries has become possible as a result of exponential increase in the size of multilingual information on the web. Wikipedia is a prime example of such multilingual source of information on any conceivable topic in the world, which is edited by the readers. Here, a web crawler has been used to traverse Wikipedia following the links on a given page. The crawler takes out the title along with the titles of the corresponding pages in other targeted languages. The result is a set of words and phrases that are translations of each other. For efficiency, the URLs are organized using hash tables. A lexicon has been constructed which contains 7-tuples corresponding to 7 different languages, namely: English, German, French, Polish, Bulgarian, Greek and Chinese.

1 INTRODUCTION

The main goal of this project is to attract the attention to and demonstrate the feasibility of creating a multilingual dictionary using Wikipedia. Here, English, German, French and Polish were chosen for their wealth of information and another three languages to demonstrate that our program could also handle different writing systems and alphabets: Greek, Bulgarian and Chinese in this case. The technique can be applied to a number of other online resources where versions of the same article appear in different languages; one such example is the Southeast European Times news site (http://www.setimes.com/). In this case, and many others, the use of crawlers is unavoidable as an off-line version of the resource is not at hand.

2 LITERATURE REVIEW

There have been efforts in the past to build multilingual dictionaries with varying degrees of success, and the ones we know of are only extensions of bilingual dictionaries already available. Yet none of them had tried to use Wikipedia as the potential source of lexical information. Only very recently, there have been attempts to tap into multilingual dimension of Wikipedia, which has been used to identify named entities Richman *et al.* (2008) .

Lafourcade (1997) carried out two multilingual construction projects: French-English-Malay (FeM), and French-English-Thai (FeT). The FeM data was created by crossing of French-English and English-Malay lexical resources. the source language. For each word in French, one or several meanings in the target language were grouped together in so called blocks. Thus there were equivalent blocks for English, Malay and Thai. In the case that more than one translation existed, one entry was restricted to just one meaning, with extra entries for extra meanings. Sometimes only one entry was used even in the case of several alternative meanings, leaving it to the discretion of the lexicographers to decide which alternative meaning to use. Two kinds of dictionaries were targeted: the general dictionary with about 20,000 entries and a Computer Science domain specific dictionary with 5,000 entries.

Boitet *et al.* (2002) worked on the PAPILLON project. The project covered seven different languages: English, French, Japanese, Thai, Lao, Vietnamese, and Malay. They started with open source data, known as “raw dictionaries”. Some of them were monolingual: 4,000 French entries from

R. Shahid A. and Kazakov D. (2009).
AUTOMATIC MULTILINGUAL LEXICON GENERATION USING WIKIPEDIA AS A RESOURCE.
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 357-360
DOI: 10.5220/0001783003570360

UdM, and 10,000 Thai entries from Kasetsart University; some were bilingual: 70,000 Japanese-English entries, plus 10,000 Japanese-French entries in J. Breen's JDICT XML format, 8,000 Japanese-Thai entries in SAIKAM XML format, and 120,000 English-Japanese entries in KDD-KATE LISP format. Finally, there were 50,000 French-English-Malay entries in FeM XML format. The authors defined a *macrostructure*, with a set of monolingual dictionaries of word senses, called "lexies" linked together through a set of interlingual links, called "axies". In the next step the "raw dictionaries" were transformed into a "lexical soup", in the (Mangeot-Lerebours, 2001) intermediate DML format, which comprised of the XML schema and the namespace. The star-like structure thus created made it easier to add new languages.

Breen (2004) built a multilingual dictionary, JMdict, with Japanese as the pivot language and translations in several other languages. The project was an extension of an earlier Japanese-English dictionary project (EDICT: Electronic Dictionary) (Breen, 1995), that began in the early 1990s to create a Japanese-English dictionary, which grew to 50,000 entries by the late '90s. Yet its structure was found to be inadequate to represent the orthographical complexities of the language, as many Japanese words can be written with alternative *kanji* and *kana* and may have alternative pronounciations. The kanji came from the ancient Chinese and kana was a derivative of it. In modern use they are used to describe different parts of speech. For French translations they used two projects: 17,500 entries from *Dictionnaire français-japonais* (Desperrier, 2002) and 40,500 entries from French-Japanese Complementation Project at http://francais.sourceforge.jp/. For German translations they used WaDokuJT Project (Apel, 2002). XML (Extensible Markup Language) was used to format the file on account of the flexibility it provides. The JMdict XML structure contained an element type: <**entry**>, which in turn contained: the sequence number, the *kanji* word, the *kana* word, information, and translation information. The translation part consisted of one or more sense elements. The combining rules were used to weed out unnecessary entries. The rule stated in short: treat each entry as a triplet of kanji, kana and senses; if for any two or more entries, two or more members of the triplet are the same, combine them into one entry. Thus if the kanji and kana in different entries are included as alternative forms, and if they differ in sense, they are included as polysemous words. The *entry* also stored information regarding the meanings of the word in different languages. The JMDict file contained over 99,300 entries in both English and Japanese, while 83,500 keywords/phrases had German translations, 58,000 had French translations, 4,800 had Russian translations, and 530 had Dutch translations. A set of 4,500 Spanish translations was being prepared.

3 MULTILINGUAL LEXICON GENERATION

Since its humble beginnings in 2001, Wikipedia has emerged as a huge online resource attracting over 684 million visitors yearly by 2008. There are more than 75,000 active contributors working on more than 10,000,000 articles in more than 250 languages (Wikipedia, August 3, 2008). Each Wikipedia page has links to pages on the same topic in other languages.and combined in the form of 7-tuples, which are entries in the lexicon, detailing a word in English and its translations in the six other languages. The aim was to extract as many such 7-tuples as possible.

3.1 Web Crawler

A web crawler is a computer program that follows links on web pages to automatically collect data (hypertext) off the internet. We use it here to move from one Wikipedia article to another, collecting the above mentioned tuples of word/phrase translations in the process.

Our version of the web crawler takes the starting page as an input from the user. It visits the given page, and extracts all the links on that page and appends them to a list. Then it repeats the process for each link collected earlier, and visits them one by one, extracting the links and once again appending them to the list. Putting them at the end (e.g., making the list a queue) ensures that the search method adopted is Breadth First Search (BFS). In our context following BFS will explore a number of related concepts consecutively while Depth First Search (DFS) would drift-off any given topic. There may be technical aspects related to the use of memory by each approach but we will not discuss them here.

The BFS approach was used in the following experiments. With the BFS capability thus incorporated, other lists were defined that would keep track of all the web pages that have already been visited thus keeping the code from revisiting them and extracting repeatedly the same 7-tuples into the lexicon. Apart from ensuring that there was no redundancy within the lexicon, either purely numeric or had a *null* entry for any of the seven languages.

The program picks up a URL from the top of the queue, expands in terms of URLs by exploring new

English	German	French	Polish	Bulgarian	Greek	Chinese
Wikipedia	Wikipedia	Wikipédia	Wikipedia	Уикипедия	Βικιπαίδεια	维基百科
Encyclopedia	Enzyklopädie	Encyclopédie	Encyklopedia	Енциклопедия	Εγκυκλοπαίδεια	百科全书
English language	Englische Sprache	Anglais	Język angielski	Английски език	Αγγλική γλώσσα	英语
Venice	Venedig	Venise	Wenecja	Венеция	Βενετία	威尼斯
Film director	Regisseur	Réalisateur	Reżyser	Режисьор	Σκηνοθέτης	電影導演
Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Унифициран локатор на ресурси	Uniform Resource Locator	统一资源定位符
Web search engine	Suchmaschine	Moteur de recherche	Wyszukiwarka internetowa	Търсачка	Μηχανή αναζήτησης	搜索引擎
University	Hochschule	Université	Uniwersytet	Университет	Πανεπιστήμιο	大學
Monopoly	Monopol	Monopole	Monopol	Монопол	Μονοπώλιο	垄断
Computer	Computer	Ordinateur	Komputer	Компютър	Ηλεκτρονικός υπολογιστής	計算機
University of Oxford	University of Oxford	Université d'Oxford	Uniwersytet Oksfordzki	Оксфордски университет	Πανεπιστήμιο της Οξφόρδης	牛津大学
Population density	Bevölkerungsdichte	Densité de population	Gęstość zaludnienia	Гъстота на населението	Πυκνότητα πληθυσμού	人口密度
Presidential system	Präsidentielles Regierungssystem	Régime présidentiel	System prezydencki	Президентска република	Προεδρική Δημοκρατία	總統制
Dictatorship	Diktatur	Dictature	Dyktatura	Диктатура	Δικτατορία	专政
European Community	Europäische Gemeinschaft	Communauté européenne	Wspólnota Europejska	Европейска общност	Ευρωπαϊκή Κοινότητα	歐洲共同體
Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Беназир Бхуто	Μπεναζίρ Μπούτο	贝娜齐尔·布托
Thomas Edison	Thomas Alva Edison	Thomas Edison	Thomas Alva Edison	Томас Едисън	Τόμας Έντισον	托马斯·爱迪生
Art	Kunst	Art	Sztuka	Изкуство	Τέχνη	艺术
California	Kalifornien	Californie	Kalifornia	Калифорния	Καλιφόρνια	加利福尼亚州
Buddhism	Buddhismus	Bouddhisme	Buddyzm	Будизъм	Βουδισμός	佛教

Figure 1: A Snapshot of the Lexicon.

links on the given URL, and then extracts titles in the given languages. An essential first step was to separate the tracking of the URLs from the code itself. Thus a database in Access was created, comprising of 28 different tables. One table stored all the URLs to be visited, and the rest implemented a hash table, to store the already visited URLs, indexed by the first character of the page title. 26 different tables were created for the 26 different letters in the English alphabet. Another table was created, URLExtra, which would store all the URLs with Wikipedia page titles starting with anything but English letters, including numbers.

The program starts with a user-provided URL, then looks for potential URLs to be used for extraction of titles. These URLs are added to the queue. For each page that is visited and removed from the queue, the URLs of several pages (contained in that page) are typically added. Thus the number of URLs to be searched rises quickly and might yield a huge list. In order to avoid such a scenario and keep the size of the URL list within reasonable limits, an upper limit was set to the number of URLs at a time. Similarly a lower limit was set so that the program could be prescient and started looking for more URLs before it ran out of them. The lower and upper limits were set to 50 and 1,000 respectively. Thus barring an exhaustion of all potential URLs, the program would never run out of URLs to be searched for.

3.2 Results

Figure 1 shows a snap shot of the lexicon in a table. UTF-8 was used as the coding scheme, which makes possible writing characters in other non-English languages.

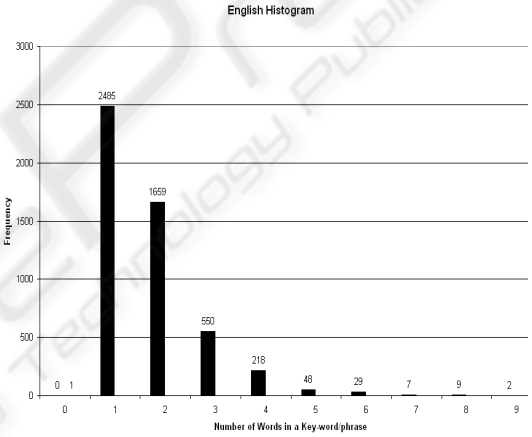

Figure 2: The English Histogram.

A total of 5 runs were carried out to get a total of 8,748 entries, out of which 5,006 were unique (57% of the total). In order to get that many entries the program visited 726,715 English language articles, not everyone of which was unique. Despite the checks on revisiting pages some of them were re-visited due to a bug in the code. Only a little more than 1% of the total had corresponding pages in all the other six languages. The crawler still had to visit more than a quarter of all Wikipedia articles in English.

Using the least-prolific language as a pivotal language would have saved us a lot of time in searching for lexical entries since most pages in the least-frequently occurring language would probably have corresponding pages in other languages, yet English was chosen as the pivotal language. The reason being that English is easy to play with, with its very familiar alphabet-set and thus building hashing tables was easily done. Also we were more interested in English

at the first instance and different languages were incorporated at a later stage.

Looking at Figure 2 one can see that unigrams make the bulk of entries (2,485 - almost 50% of the total), followed by bigrams (1,659 - 33% of the total).

In terms of semantics, the resulting lexicon is a mix of: toponyms, names of famous people, names of languages, and general concepts, such as “rock music” and “fire fighter”, among others.

4 USES OF A MULTILINGUAL DICTIONARY

It can help the lexicographers build traditional dictionaries, as a starting point for their work. Another important use is for Cross-Lingual Information Retrieval (CLIR). Prikola (1998), compared the performance of translated Finnish queries against English documents to the performance of original English queries against English documents, by using a general dictionary and a domain specific dictionary (a medical dictionary in this case). It was found that a cross-lingual IR system based on Machine Readable Dictionary (MRD) translation was able to achieve the performance level of monolingual IR.

The lexicon containing more than 5,000 entries is available at http://www-users.cs.york.ac.uk/∼ahmad/index.htm for free use under GNU Free Documentation License.

5 FUTURE WORK

A useful thing to do would be to create domain specific dictionaries based on the categories defined within Wikipedia, according to which each article belongs to one or more categories. A domain could be defined using a set of categories and only those articles could be used for building the lexicons belonging to that particular domain.

ACKNOWLEDGEMENTS

This research was carried in the period Nov 2007 - July 2008, and was partly sponsored by the Higher Education Commission in Pakistan.

REFERENCES

Apel, U. (2002). WaDokuJT - A Japanese-German Dictionary Database. In *Papillon 2002 Seminar*, Tokyo.

Boitet, C., Mangeot-Lerebours, M., and Serasset, G. (2002). The PAPILLON Project: Cooperatively Building a Multilingual Lexical Data-base to Derive Open Source Dictionaries & Lexicons. In *Proceedings of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop*, Taipei.

Breen, J. (1995). Building an Electronic Japanese-English Dictionary. In *Japanese Studies Association of Australia Conference*.

Breen, J. (2004). JMdict: a Japanese-Multilingual Dictionary. In *Coling 2004 Workshop on Multilingual Linguistic Resources*, pages 71–78, Geneva.

Desperrier, J.-M. (2002). Analysis of the Results of a Collaborative Project for the Creation of a Japanese-French Dictionary. In *Papillon 2002 Seminar*, Tokyo.

Lafourcade, M. (1997). Multilingual Dictionary Construction and Services Case Study with the Fe* Projects. In *Proc. PACLING'97*, pages 173–181.

Mangeot-Lerebours, M. (2001). *Environnements Centraliss et Distribus pour Lexicographes et Lexicologues en Contexte Multilingue.* PhD thesis, Universite Joseph Fourier.

Pirkola, A. (1998). The Effects of Query Structure and Dictionary Setups in Dictonary-Based Cross-Language Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63, Melbourne.

Richman, A. and Schone, P. (2008). Mining Wiki Resrouces for Multilingual Named Entity Recognition. In *Proceedings of ACL-08: HLT*, pages 1–9, Columbus, Ohio, USA.