An Automatic Bilingual Corpora Generator

Siti Nordianah Hai Hom [1], Azniah Ismail [2]
[1]Department of Computing, Faculty of Art, Computing and Creative Industry, Universiti Pendidikan Sultan Idris, dianah_83@yahoo.com
[2] Department of Computing, Faculty of Art, Computing and Creative Industry, Universiti Pendidikan Sultan Idris, azniah@fskik.upsi.edu.my

Abstract

Bilingual corpora that contains similar documents of two different languages are examples of essential resources for Natural Language Processing (NLP) tasks including Cross-Lingual Information Retrieval (CLIR) and machine translation. Nevertheless, these resources could also be useful for many processes in learning languages. We introduce an automatic bilingual corpora generator that builds corpus resources from the web. This generator involves the use of the *in-domain terms* (IDT), in which the terms can be thought of as the most important contextually relevant words. The method used is simple yet practical, and makes acquiring resources from web sources more than just collecting texts and pasting them all together. However, as an on-going project, the system has not been fully implemented and evaluated. In this paper, the researchers emphasizes more on the prototype of the system in terms of appearance and display. For example, the generator shall be built on a web-based system that gives different options to users on how they would like to observe the acquired texts.

Keywords Bilingual Corpora, in-domain-term (IDT)

INTRODUCTION

Building linguistic resources, automatically has become a common research topic in the natural language processing field due to the unavailability or less availability of many required resources. These tasks would normally require learning from other linguistic resources; hence, the task itself may introduce the chicken-egg-problem specifically in acquiring linguistic resources.

The process of automatic building is risky as its effectiveness is highly dependent on the quality of the linguistic resources that are being used. Al-Onaizan et al. (1999) and Fung and Cheung (2004) emphasize that good quality outcomes may only be achieved if good quality linguistic resources with sufficient amount are present. Otherwise, the outcomes of the automated process may not be good. However, although good quality linguistic resources may be available, getting sufficient amount of resources for the automatic building process may be another area of serious concern. In this paper, we would like to propose a method that would consider both issues; the quality and the amount of acquired resources. Moreover, we would also like to bring forward the interface design of the prototype of the system.

BACKGROUND

Corpora are the main resources required to learn translation pairs. Somers (2001) has noted "fully annotated aligned multilingual parallel corpora are becoming increasingly available through various coordinated international efforts". However, Somers was also concerned about the number of different languages featured, which, according to him, "...is still rather small". Insufficient text collections in terms of their amounts or the domain coverage would probably threaten any extraction attempt. Hence, acquiring corpora is an issue that requires serious attention at that time. In the current situation, the scenario has not much changed.

PROBLEM STATEMENT

Research related to bilingual corpora is not eagerly pursued thus far; nonetheless, there is a pressing need for this study. We are concerned that there is not much corpora builder available and whether the existing system is really able to accommodate different users' needs.

RESEARCH OBJECTIVE

These research objectives were formulated to guide the study as follows:

- to develop a web-based prototype system for easy access and display.

RESEARCH METHODOLOGY

> "Prototype is an easily modifiable and extensible working model of a proposed system, not necessarily representative of a complete system, which provides users of the application with the physical representation of key parts of the system before implementation"
>
> (Futrell, 2002: page)

According to Futrell (2002), prototype model is suitable for system development model for this system. Nevertheless, the flow in prototype model needs modification from the original source for the system necessary. (See Figure 1 for modified prototype model.) In this paper, researcher focus on the design phase of this system to give different options to users how they would like to observe the acquired texts.

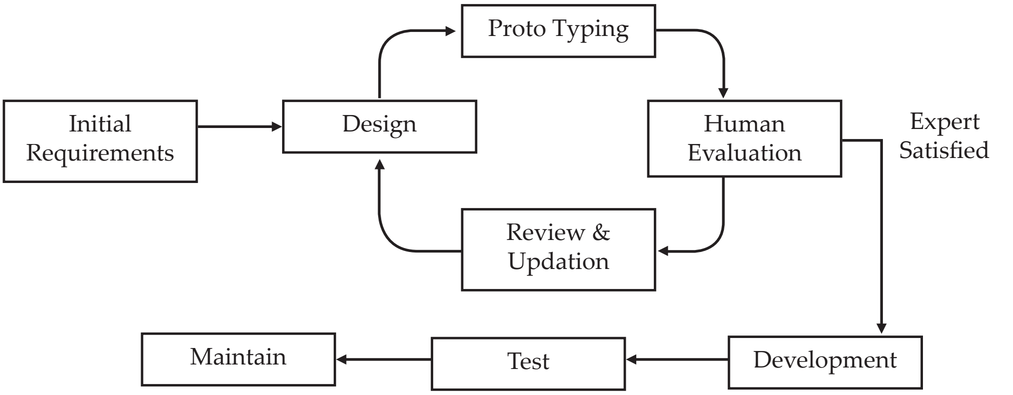


Figure 1 Prototype Model
Source: Prototype model

The development of the system is optimally designed based on the prototype approach. There are four main sections developed in the system:

A. Home
B. About
C. Corpora
D. Contact

In general, the main objectives for the design is to be easy for the user to understand and to control the entirety of the application; the color scheme chosen for the system is appropriate; relevant information and specific functions are displayed for intended users; and multimedia elements is appropriately used.

RESULT

In this section, we present the fundamental design of each section available in the prototype. Figure 2 shows the transition diagram that represents the four main sections in the system. The most important part of the design is of the corpora page which tends to accommodate different user requests for texts (i.e., content of corpora) viewing. (See *Part C Corpora Page* in this paper for details.)

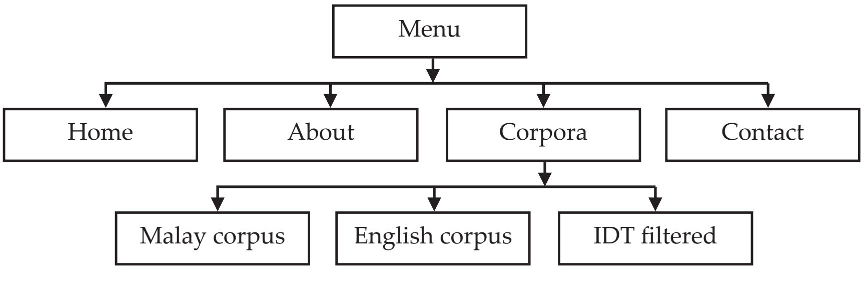


Figure 2 Transition Diagram

However, we maintain a similar layout for each section to allow a simple and uniform web page style throughout the system. Figure 3 shows the standard layout proposed for each section in the system.

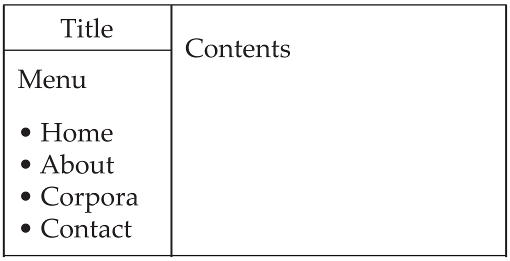


Figure 3 Layout Diagram

To build the prototype of the system, we used the PHP language and Adobe Dreamweaver CS3. We also make use of free services from Cooltext.com to create title text and simple animations including logo for the prototype system. The followings are examples of pages that we built for this system:

A. Home Page

Figure 4 shows the main (home) page for the system. Apart from the main menu, this page also contains a welcoming note for users. This page plays the role as an introductory page of the system.

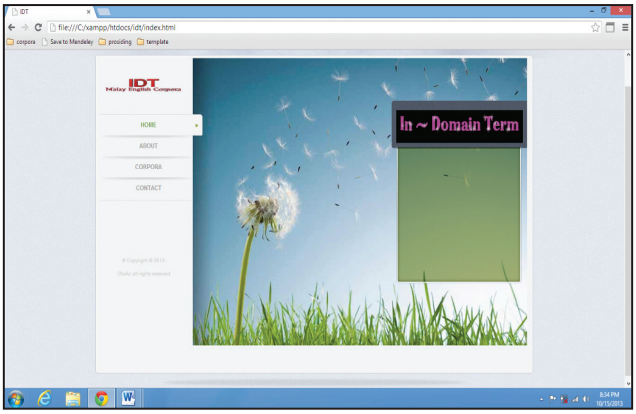


Figure 4 Home page

B. About Page

Figure 5 shows the about page, in which, a brief explanation of this IDT special project and some guidelines about how to browse and use the system are provided.

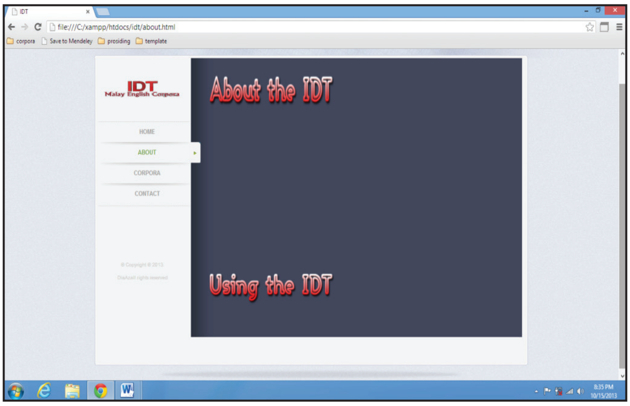


Figure 5 About page

C. Corpora Page

Figure 6 shows the most important page in the system, i.e., the corpora page. This page contains three subsections: 1. malay corpora, 2. english corpora and 3. bilingual corpora using IDT. By using the Malay corpora sub menu, a user can find sentences or context words for any Malay words that occur in the malay texts (if any) such as *sekolah*. Likewise, the English corpora sub menu allows a user to find sentences or context words for any English words that occur in the English texts such as *school*. To display these words and texts, the system provides a few options for the user to choose from, including in form of collocate texts, wordlist text, word cloud and *chinese whisper* clustering.

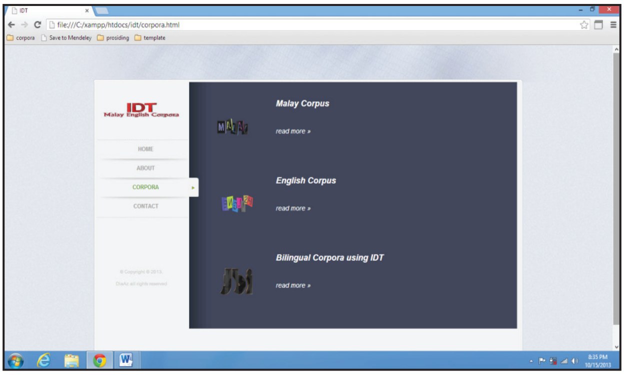
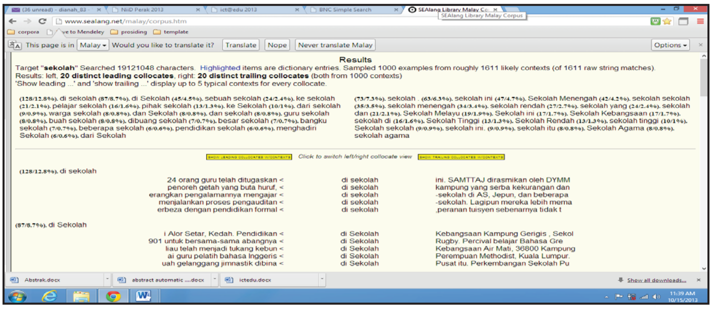


Figure 6 Corpora page

The type of displays selected by users is usually depending on their needs. Some users might be looking for context words of their requested word, whereas the others might want to have access to all the sentences that contain the requested word. We attempt to design the system to fulfill this part of user request. Yet, the number of visual display options we want to include in the real system is not finalized.

Figure 7 shows a display example of the collocate texts for the word *sekolah*. Collocate text is a sequence of words or terms that co-occur more often than would be expected by chance. Generally, collocation is useful and important in any language learning, and this type of display can be considered as a basic display for most text viewer.


(36 unread) - dianah_83 -
NiiD Perak 2013
ict@edu 2013
BNC Simple Search
SEAlang Library Malay Co
SEAlang Library Malay Corpus
www.sealang.net/malay/corpus.htm
corpora
ve to Mendeley
prosiding
template
This page is in Malay Would you like to translate it? Translate Nope Never translate Malay
Options
Abstrak.docx
abstract automatic ....docx
ictedu.docx
Show all downloads...
11:39 AM
10/15/2013

Results

Target "**sekolah**" Searched 19121048 characters. Highlighted items are dictionary entries. Sampled 1000 examples from roughly 1611 likely contexts (of 1611 raw string matches).
Results: left, **20 distinct leading collocates**, right: **20 distinct trailing collocates** (both from 1000 contexts)
'Show leading ...' and 'show trailing ...' display up to 5 typical contexts for every collocate.

(128/12.8%). di sekolah (87/8.7%). di Sekolah (45/4.5%). sebuah sekolah (24/2.4%). ke sekolah (21/2.1%). pelajar sekolah (16/1.6%). pihak sekolah (13/1.3%). ke Sekolah (10/1%). dari sekolah (9/0.9%). warga sekolah (8/0.8%). dan Sekolah (8/0.8%). dan sekolah (8/0.8%). guru sekolah (8/0.8%). buah sekolah (8/0.8%). dibuang sekolah (7/0.7%). besar sekolah (7/0.7%). bangku sekolah (7/0.7%). beberapa sekolah (6/0.6%). pendidikan sekolah (6/0.6%). menghadiri Sekolah (6/0.6%). dari Sekolah

(73/7.3%). sekolah . (63/6.3%). sekolah ini (47/4.7%). Sekolah Menengah (42/4.2%). sekolah sekolah (35/3.5%). sekolah menengah (34/3.4%). sekolah rendah (27/2.7%). sekolah yang (24/2.4%). sekolah dan (21/2.1%). Sekolah Melayu (19/1.9%). Sekolah ini (17/1.7%). Sekolah Kebangsaan (17/1.7%). sekolah di (16/1.6%). Sekolah Tinggi (13/1.3%). Sekolah Rendah (13/1.3%). sekolah tinggi (10/1%). Sekolah sekolah (9/0.9%). sekolah ini. (9/0.9%). sekolah itu (8/0.8%). Sekolah Agama (8/0.8%). sekolah agama

SHOW LEADING COLLOCATES W/CONTEXTS *Click to switch left/right collocate view* SHOW TRAILING COLLOCATES W/CONTEXTS

(128/12.8%). di sekolah

24 orang guru telah ditugaskan <	di sekolah	ini. SAMTTAJ dirasmikan oleh DYMM
penoreh getah yang buta huruf, <	di sekolah	kampung yang serba kekurangan dan
erangkan pengalamannya mengajar <	di sekolah	-sekolah di AS, Jepun, dan beberapa
menjalankan proses pengauditan <	di sekolah	-sekolah. Lagipun mereka lebih mema
erbeza dengan pendidikan formal <	di sekolah	,peranan tuisyen sebenarnya tidak t

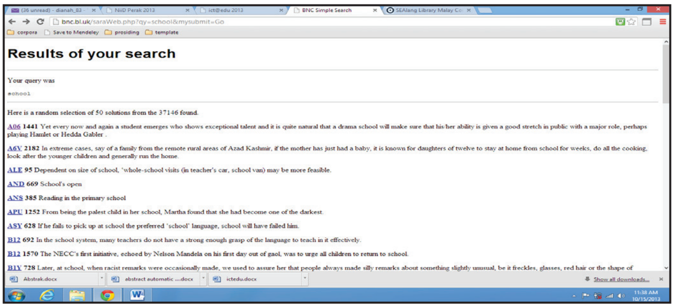
(87/8.7%). di Sekolah

i Alor Setar, Kedah. Pendidikan <	di Sekolah	Kebangsaan Kampung Gerigis , Sekol
901 untuk bersama-sama abangnya <	di Sekolah	Rugby. Percival belajar Bahasa Gre
liau telah menjadi tukang kebun <	di Sekolah	Kebangsaan Air Mati, 36800 Kampung
ai guru pelatih bahasa Inggeris <	di Sekolah	Perempuan Methodist, Kuala Lumpur.
uah gelanggang jimnastik dibina <	di Sekolah	Pusat itu. Perkembangan Sekolah Pu

Figure 7 Collocates text display

Source: SEAlang Library Malay

Figure 8 shows the example of the wordlist text display for word *school*. This page displays all search results for the request word in the form of sentences and a user can find the full article that contains the sentences when the link provided is pressed.



Results of your search

Your query was

`school`

Here is a random selection of 50 solutions from the 37146 found.

A06 **1441** Yet every now and again a student emerges who shows exceptional talent and it is quite natural that a drama school will make sure that his/her ability is given a good stretch in public with a major role, perhaps playing Hamlet or Hedda Gabler .

A6V **2182** In extreme cases, say of a family from the remote rural areas of Azad Kashmir, if the mother has just had a baby, it is known for daughters of twelve to stay at home from school for weeks, do all the cooking, look after the younger children and generally run the home.

ALE **95** Dependent on size of school, 'whole-school visits (in teacher's car, school van) may be more feasible.

AND **669** School's open

ANS **385** Reading in the primary school

APU **1252** From being the palest child in her school, Martha found that she had become one of the darkest.

ASY **628** If he fails to pick up at school the preferred 'school' language, school will have failed him.

B12 **692** In the school system, many teachers do not have a strong enough grasp of the language to teach in it effectively.

B12 **1570** The NECC's first initiative, echoed by Nelson Mandela on his first day out of gaol, was to urge all children to return to school.

B1Y **728** Later, at school, when racist remarks were occasionally made, we used to assure her that people always made silly remarks about something slightly unusual, be it freckles, glasses, red hair or the shape of

Figure 8 Wordlist text display

Source: British National Corpus

Another different text display that we include in this prototype is shown in Figure 9. A visual presentation of keywords drawn from a long text, visually differentiated based on their positions and frequencies of use in that text relating to the frequency of the words appearing in one or more documents. One can click on any word appearing in the cloud to obtain detailed information about its relativity. The larger the word, the more frequent the term. This type of display could assist users in identifying important words that occur in the context of word that they requested earlier as the system should be able to omit less useful words such as function words like *yang* and *untuk*.

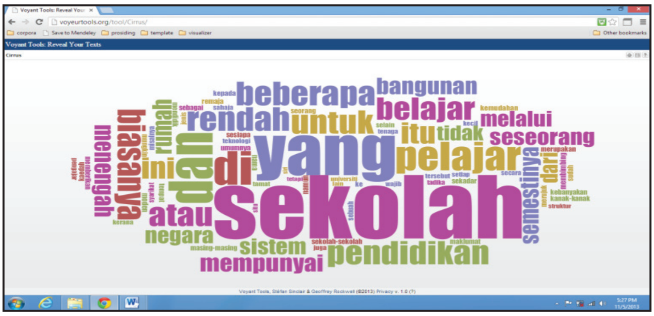
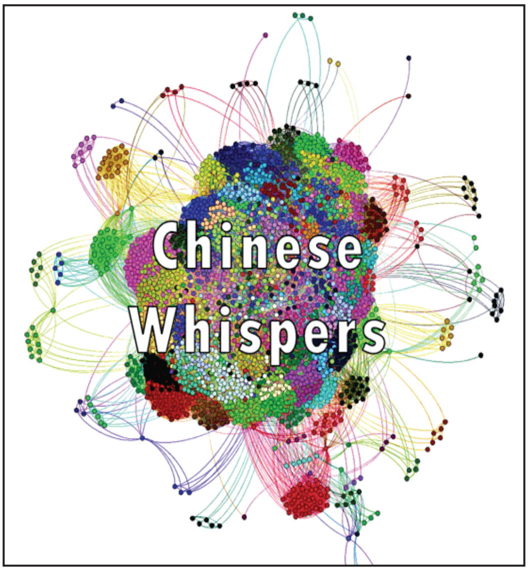


Figure 9 Word Cloud
Source: Cirrus Word Cloud

Figure 10 shows a more complex display based on an unsupervised algorithm which operates on the word co-occurrence network. By weighting the edges with the number of co-occurrences between words and then performing k-nearest neighbors clustering in the resulting graph until stabilization is achieved some related clusters might be identified Biemann (2005). According Biemann (2006), in an evaluation of relatively long documents with clean text, Chinese Whispers performed very well at identifying major European languages. The same technique can be applied to identify clusters of context words or, perhaps, of similar sentences



.

Figure 10 Chinese Whisper Clustering
Source: Cirrus Word Cloud

D. Contact Page

Last but not least, the contact page is the place where a user can contact the administrator for any related purposes, such as if any help is needed, or need further explanation or to give any feedback about the system or the project. (*See Figure 11 for an example of the page.)*

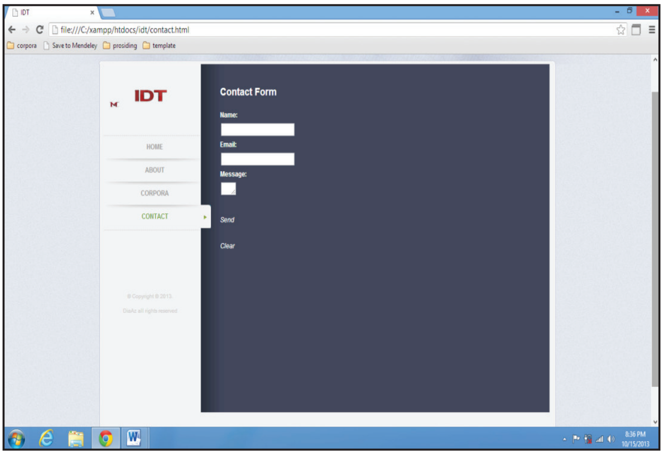


Figure 11 Contact page

CONTRIBUTION

The success of developing this automatic bilingual corpora generator would provide contributions not only to the research community, but also, generally, to anyone involves in learning languages.

CONCLUSION

We have developed a web-based prototype system for a bilingual corpora generator. This research is an on-going project. Currently, our focus is on the display features, which the system offers to users, using different options from basic to complex type of displays. Having these features provides the users the flexibility of viewing the acquired texts according to their preferences.

REFERENCES

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N., and Yarowsky, D. (1999). *Statistical machine translation. Technical Report Center for Language and Speech Processing,* Place: John Hopkins University.

Azniah {Formatting Citation} Ismail (2012). Minimally Supervised Techniques for Bilingual Lexicon Extraction, Ph.D Thesis. York University.

Biemann, C. (2006). Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Languange Processing Problems. In Proceeding of the

Human Languange Technology- North American Chapter of the Association for Computational Linguistics (HLTNAACL).

Biemann, C., Teresniak, S. (2005). Disentangling from Babylonian Confusion – Unsupervised Languange Identification. In Proceedings of Conference on Intelligent Text Processing and Coutational Linguistics (CICLing).Place: Publisher

Chinese Whispers Clustering. Retrieved from https://marketplace.gephi.org/plugin/chinese-whispers-clustering/

Cirrus Word Cloud. Retrieved from http://voyeurtools.org/tool/Cirrus/

Cooltext Graphics Generator. Retrieved from http://cooltext.com/?gclid=CMTyoYPemLoCFWsF4godSUgA9g

Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In Proceedings of the 20th International Conference on Computational Linguistics (COLING): Place: Date

Futrell, R. T., Shafer, D.T. & Shafer, L. (2002), *Quality Software Project Management*. Place: Prentice Hall.

Lou, B. (2009). British National Corpus. Retrieved from http://www.natcorp.ox.ac.uk/

Prototype model. Retrieved from http://csebrules.blogspot.com/2011/01/assignment-2-task-2-prototyping-model.html

Prototype model. Retrieved from http://istqbexamcertification.com/what-is-prototype-model-advantages-disadvantages-and-when-to-use-it/

SDLC – Incremental Model (2009), Quality Testing. Retrieved from http://www.qualitytesting.info/profiles/blogs/sdlc-incremental-model

SEAlang Library Malay. Retrieved from http://www.sealang.net/malay/corpus.htm

Somers, H. (2001). Bilingual parallel corpora and language engineering. In Anglo-Indian Workshop Language Engineering for South-Asian Languages (LESAL). Place:Date.