5th Workshop on Multilingual Lexical Databases, 30 Aug - 1 Sept 2004, Grenoble, France.

Building A Specialised Multilingual Dictionary from General Monolingual Dictionaries

Choy-Kim CHUAH

Computer-Aided Translation Unit (UTMK),
School of Computer Sciences,
Universiti Sains Malaysia, MALAYSIA
kimc@cs.usm.my

Abstract. There have been efforts to link entries to produce an "enriched" bilingual, or even multilingual dictionary. To do this, we mainly make use of the definition of the dictionary. In this paper, we report how we can easily prepare the first draft of a specialised dictionary of plants/animals from two monolingual dictionaries. The dictionaries in question, however, have necessarily to be rich in taxonomic information, and the entries, names of plants/animals. If the dictionaries used are in different languages, we not only obtain a specialised dictionary, but one that is multilingual. Using simple manipulations, and the simple fact that the scientific names of plants and animals are unique, i.e. there is only one scientific name to a plant or animal, we can link entries, and ultimately, dictionaries. Unlike word entries of other types which do not have a scientific name, we were able to dispense with the use of a dictionary's linguistic content when linking.

1 Introduction

In the past, information used to be written mainly on cards and stored in drawers. With the years, many of these repositories have reached a size where simple retrieval is no longer possible. With "265 drawers containing 290,996 index cards" and "68 million biological specimens" in her possession, the Natural History Museum in London has found it necessary to computerise (see Downton, et al, 2003), or lose this wealth of information which had been painstakingly archived over the decades.

Today, specialised databases are being constructed by individuals and groups world-wide for various reasons. One such knowledge database being constructed serves to keep an inventory of crocodiles and turtles in Borneo (Das & Ismail, 2002). For others, the database may be augmented with various types of information. *LucID*, for example, is not just a database of taxonomic information, but is a "multimedia expert system" which functions as a diagnostic tool to help non-taxonomists identify biological specimens correctly (Norton, Patterson & Schneider, 2000).

Efforts to build specialised databases has even transcend borders. One such project is that which seeks to draw up an inventory of medicinal plants in the Asia-Pacific region (Batugal, 2003). This four-year project (2002-2005) which involves 14 countries is sponsored by the Korean government. Other than developing a catalogue of

medicinal plants with images, identification characteristics, etc., the project also hopes to encourage collaborations among nations to conserve and promote the use of such medicinal plants.

Because of the effort and time involved, projects to construct databases should not only be collaborations such as that just mentioned, but should also look into "recycling" pertinent information from whatever resources available.

Available in the market are general monolingual dictionaries which have seen many man-hours of hard work. These dictionaries which are targeted at different groups of users contain information of different degrees of technicality. While some contain only general information intended for the layman, other dictionaries which have seen many revisions may now be "semi-technical" in nature. For names of compounds, the definitions may include chemical formulae, while that of names of plants and animals may contain taxonomic information.

benzene ...: a colorless volatile flammable toxic liquid aromatic hydrocarbon C_6H_6 used in organic synthesis, as a solvent, and as a motor fuel – ...

cockroach *n* ...: any of an order (Blattaria) of chiefly nocturnal insects including some that are domestic pests

(examples taken from *Webster's New Colegiate Dictionary*, 1977)

Compiling a specialised dictionary not only requires much time and effort, but the dictionary compilers themselves have to be domain specialists. Now, if we can relieve these specialist-compilers of this tedious task, the time saved can be better spent on research.

Given this scenario, we wondered if we can "naively" use two general monolingual dictionaries rich in taxonomic information to get out the first draft of a specialised multilingual dictionary of plants and animals.

In this paper, we report the linking of entries from two general dictionaries via their taxonomic content, to get out the first draft of a list of names of insects in English and Malay. We intend this draft to be the "seed" database of information about insects for managing urban pests.

2 General Monolingual Dictionaries

At the Computer-Aided Translation Unit in Universiti Sains Malaysia, we have access to data from two digital dictionaries, viz. the *Kamus Dewan* (Dewan Bahasa dan Pustaka, 1994), a Malay dictionary, and a copy of a dictionary which had been processed and disambiguated for senses by Guo (1989). While this dictionary is said to be the *Longman Dictionary of Contemporary English* (LDOCE), we however found the data to be that of *WordNet* (WN). Hence, we will refer to this content as having come from WN.

2.1 Kamus Dewan (KD)

The *Kamus Dewan* (Dewan Bahasa dan Pustaka, 1994) is one of the most comprehensive Malay dictionary ever compiled. For each entry, not only are its various derivative forms given, but also idiomatic expressions which contain the entry word. The origin of the word is also provided.

During the compilation of the *Kamus Dewan*, many dictionaries and monographs on flowers, trees, and fishes were consulted. After cross-referencing, this information was incorporated into the entries. Consequently, the dictionary becomes quite enriched with taxonomic information on plants and animals. We note that this dictionary was manually compiled by a group of lexicographers.

Now, we want to use as much of this information to build our first seed database of names of insects.

2.1.1 Synonyms and variant spellings

In an examination of a digital copy of the *Kamus Dewan*, we found many entries, especially those of names of plants and animals, to have quite a few synonyms and variant spellings.

Consider the entries for cockroach which were extracted. We found seven entries to be about cockroaches.

cecunguk 1. (Sunda) lipas; …
coro Jw lipas
kacoak Jk lipas
kecoak Jk lipas; **kecuak**.
kecuak Jk lipas; → **kecoak**
kepuyuk Mn lipas, kacoa
lipas sj serangga berwarna perang gelap, berbadan leper dan bujur, mempunyai sesungut yg panjang dan bahagian mulut menggigit, kepuyuk, *Periplaneta* spp.; ~ *kudung* sj lipas, *Periplaneta orientalis*; …

On the basis of their similarity in spelling, we would say that **kacoak** and **kecuak** are variant spellings of **kecoak**. And, because **cecunguk**, **coro**, **kecoak** and **kepuyuk** share the same single-word definition of **lipas** 'cockroach', we say that they are its synonyms. We note however that while we assume these latter four words to be synonymous with **lipas**, we are unable to say for sure if they are indeed its synonyms, or its hyponyms, or they refer to similar insects of a different kind.

Until and unless we have their scientific names, we cannot be sure if the layman has correctly attached the correct scientific name, or even definition, to the insect. A case in point, consider the definition of **labah** taken from the dictionary.

labah; labah-labah, lelabah sj serangga berkaki lapan yg membuat sarang dgn menyirat benang yg keluar drpd badannya (dan sarang ini dapat memerangkap mangsanya); jenis-jenisnya: ~ *beruk*, ~ *lotong*, dll; …

Here, **labah-labah** 'a spider' has been incorrectly defined as "an insect with eight legs ...". This definition, in fact, contradicts with the definition of **serangga** 'insect', viz. "a small animal with ... and six legs ..."

serangga sj binatang kecil yg badannya bertekung tiga dan kakinya enam serta beruas-ruas (spt nyamuk, lalat, belalang, dll); ....

Scientific names of plants and animals are unique. There is only one scientific name to a plant or animal. Hence, if two entries do not share the same scientific names, it is impossible that they refer to the same entity. While it is only the entomologist who with the help of the layman can determine if a given insect is indeed that of a given scientific name, we can help the entomologist set up a knowledge database by getting out the first seed database for him to verify. In this way, enthusiasts in dictionary compilation can contribute to other domains, and help lexicographers - general or specialised - improve on the quality of the dictionaries produced.

We attribute the high presence of synonyms in the *Kamus Dewan* to dialectical differences, and/or the result of assimilation and borrowing from other languages spoken in the region. This poses a little bit of a problem as we have first to determine if they are true synonyms or not. In any case, our interest is to get out a seed database, and not correct the content. The *Kamus Dewan* gives no less than twenty abbreviations for the dialects and languages used in the encoding.

2.1.2 Impoverished definitions

While we accept that the *Kamus Dewan* is not intended to be a specialised dictionary, we note that many of the definitions are rather "impoverished". For example, many entries have just *sj serangga* meaning 'a type of insect' as definition. Consider some examples given here.

kelulut III sj serangga, *Laccifer lacca*.
sesorok sj serangga, *Gymnogryllus elegans*.
api II; api-api sj serangga, kunang-kunang.
cengkerik sj serangga, jengkerik, keridik.
sambah; sambah-sambah Br sj serangga, gegancung, mentadak-mentadu.
walang II Jw sj belalang; ~ sangit sj serangga, cenangau, pianggang.

If not for the scientific names given, or the additional synonym(s), a non-domain specialist would, for sure, erroneously conclude **kelulut III** and **sesorok** to refer to the same entity, when they in fact are different. *Kelulut* refers to scale insects, while *sesorok* refers to crickets. If not for the scientific name given, the definition alone would not be enough to differentiate one entry from the other.

Note that by convention, scientific names, in whatever language, have necessarily to be written in italics, e.g. *Laccifer lacca*, *Gymnogryllus elegans*, or underscored if italics are not used, e.g. Laccifer lacca, Gymnogryllus elegans.

2.2 WordNet (WN)

From the processed copy a dictionary, obtained from Guo (1989), we found the data to be quite rich in taxonomic content.

Using the "raw" data that we obtained from Guo (1989), we found 36 entries with the word "cockroach".

```
ldoce(american_cockroach, 1, n, ... large, reddish, brown,
  'free-flying', cockroach, originally, from, southern,
  unite, state, but, now, widely, distribute, ...
...
ldoce(periplaneta, 1, n, ... cosmopolitan, genus, of,
  large, cockroach, ...
ldoce(periplaneta_americana, 1, n, ... large, reddish,
  brown, 'free-flying', cockroach, originally, from,
  southern, unite, state, but, now, widely, distribute,
  ...
```

After processing and merging some of the information that could be merged, we obtained:

```
american_cockroach 1n ... large reddish brown 'free-
  flying' cockroach originally from southern unite state
  but now widely distribute ...
periplaneta 1n ... cosmopolitan genus of large cockroach ...
periplaneta_americana 1n ... large reddish brown 'free-
  flying' cockroach originally from southern unite state
  but now widely distribute ...
```

and determined `american_cockroach` and `periplaneta_americana` to co-refer.

From the taxonomic information provided in some of the entries, we were able to put together a bit about their taxonomy by just organising the content into three columns as shown in Table 1.

And given that *lipas* in Malay is an entity of the genus *Periplaneta*, we can link the English definition of *Periplaneta* to that of *lipas*.

Of course, should the scientific name not be provided, we would not be able to discover so much. Still, we were able to quickly put together a small database within the span of about a few weeks.

Now, using another dictionary, this time a bilingual dictionary, such as the *Kamus Inggeris-Melayu Dewan* which is not rich in taxonomic information, we were able to link more of the entries in all the three dictionaries to constitute our basic specialised multilingual database of entomological names.

Table 1. Merged entry *lipas* 'cockroach'

Order/Family/ Genus	Common name	Definition
Order: Dictyoptera		in some classification replace by the order here suborder blattodea cockroach and manteodea mantids in former classification often subsume under a much broader order orthoptera
Suborder: Blattaria, Suborder: Blattodea		cockroach in some classification consider an order
Family: Blattidae		domestic cockroach
Genus: *Periplaneta*		cosmopolitan genus of large cockroach
Periplaneta americana	American cockroach	large reddish brown free-flying cockroach originally from southern unite state but now widely distribute
Periplaneta australasiae	Australian cockroach	widely distribute in warm country
Blattella germanica		small 'light-brown' cockroach bring to unite state from europe a common household pest…
Blattella		small cockroach …
Blatta orientalis	Asiatic cockroach, blackbeetle	dark brown cockroach originally from orient now nearly cosmopolitan in distribution,…
Blatta		type genus of the blattidae cockroach infest building worldwide …
Blaberus		giant cockroach …
	cockroach, roach	any of numerous chiefly nocturnal insect some are domestic pest
	giant cockroach	large tropical american cockroach
	dictyopterous_insect	cockroach and mantids
	dictyopteran	of or relating to or belong to the order dictyoptera

Table 2. Merged entry *lipas* 'cockroach'

Entry	Malay definition	English definition
lipas	sj serangga … perang gelap,… kepuyuk, *Periplaneta* spp.; ~ *kudung* sj lipas, *Periplaneta orientalis*; …	cosmopolitan genus of large cockroach …

3 Conclusion

In this naïve way, we were able obtain about 50 bilingual entries of entomological names for the first version of our database. We give some entries here.

Table 3. Some Linked entries

Scientific name	Definition from WordNet	Definition from Kamus Dewan
Aedes aegypti	yellow-fever_mosquito; mosquito that transmit yellow fever and dengue	nyamuk; nyamuk aedes (rimau); sj nyamuk
Anopheles spp.,	malaria mosquitoes distinguish by the adult head-downward stance and absence of breathe tube in the larvae	tiruk III; nyamuk tiruk sj nyamuk yg membawa kuman penyakit malaria
Camponotus gigas	carpenter ant	temenggung III; semut temenggung; sj semut hitam yg besar
Glossina spp	tsetse fly, tsetse, tzetze fly, tzetze; blood-sucking african fly transmit sleep sickness etc.	lalat I; lalat tsetse; sj lalat
Musca domestica	housefly; common fly that frequent human habitation and spread many disease	lalat I; lalat rumah; sj lalat yg biasa terdapat di rumah

We note that because the names of the fauna and flora in the *Kamus Dewan* are predominantly from this region of the world, and that in WordNet are from very different regions, we were not able to link as many entries than if the dictionaries contain words of plants and animals from the same region.

Some possible applications of such a database of names of insects is important especially with the movement of people across the globe, pests and diseases are bound to be transmitted as well. With a database of names of insect pests, urban entomologists, like those working on traditional medical and medicinal plants may be assured that they are talking about the same thing.

4 Acknowledgements

We thank Dewan Bahasa dan Pustaka and also WordNet for using some examples that were taken from their dictionaries.

References

1. Batugal, Pons. 2000. *Inventory and Documentation of Medicinal Plants in 14 Asia Pacific Countries*. (www.herbamalaysia.net/portal/articles/articles_database2/Dr%20Pons%20Batugal.ppt)
2. Das, Indraneil & Ismail, Ghazally. 2002. *Crocodiles and Turtles of Borneo*. (http://www.arbec.com.my/crocodilesturtles/introduction.php.) Sarawak, Malaysia: Institute of Biodiversity and Environmental Conservation, Universiti Malaysia Sarawak.
3. Dewan Bahasa dan Pustaka 1992. *KamusInggeris-Melayu Dewan:An English-Malay Dictionary*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
4. Dewan Bahasa dan Pustaka 1994. *Kamus Dewan*. Third edition. Kuala Lumpur: Dewan Bahasa dan Pustaka.
5. Downton, A. C., Lucas, S. M., Patoulas, G., Beccaloni, G. W., Scoble, M. J. & Robinson, G. S. 2003. "Computerising Natural History Card Archives." In: *Proceedings ICDAR 2003 7th International Conference on Document Analysis and Recognition*. August 3-6, 2003, Edinburgh, Scotland. Volume 1: 354-358.
6. Guo, Chengming. 1989. *Deriving a Natural Set of Semantic Primitives from Longman Dictionary of Contemporary English*. PhD. dissertation. New Mexico State University.
7. Norton, Geoff A., Patterson, David J. & Schneider, Margaret. 2000. "LucID: A Multimedia Educational Tool for Identification and Diagnostics". *CAL-laborate* Vol. 4, June 2000 (http://science.uniserve.edu.au/pubs/callab/vol4/norton.html.) New South Wales, Australia: The University of Sydney.
8. Woolf, Henry Bosley (editor-in-chief). 1977 *Webster's New Colegiate Dictionary*. Springfield, Massachusetts: Merriam-Webster.