On Automatic Construction of a Thesaurus

Kavi Narayana Murthy
Department of Computer and Information Sciences,
University of Hyderabad, Hyderabad, INDIA
email: knmuh@yahoo.com

Abstract

A thesaurus links semantically related words and aids in the selection of most appropriate words for given contexts. As such, it is a very valuable tool. Yet many of the major languages of India have no thesauri till date. Constructing a thesaurus is a difficult and time consuming task. Recent work has focused on automatic or semi-automatic construction of thesauri from annotated corpora and other available lexical resources. Corpora and other lexical resources available in Indian languages are very limited and hence many of these techniques are not applicable at present. However, bilingual dictionaries exist, or are being developed with applications such as automatic translation in mind. In this paper we show that a thesaurus can be constructed automatically and efficiently from a bilingual dictionary with little human labor. We show examples from a Kannada thesaurus constructed automatically from a bilingual dictionary.

Keywords*: Thesaurus, Dictionary, Indexing*

1 Introduction

In very general terms, a thesaurus has been defined as a treasury or a storehouse; hence, a repository, especially of knowledge; often applied to a comprehensive work, like a dictionary or encyclopedia. More specifically, a thesaurus is a book containing a classified list of synonyms, organized to help you find the word you want but cannot think of.

We go to a thesaurus when we have an idea, some concept or a meaning in our mind but we are unable to get just the right word that fits our need. We have some word on hand but we somehow feel that there should be a better word, a word that says more precisely what we wish to say, a word that is best for the current context. A thesaurus usually contains an index from where we can start. We look up the index for the tentative word we have with us, a word that approximates what we wish to say but not quite exactly. The index tells us which locations in the thesaurus we need to look up. We go to those locations and hopefully we will get the word that we are looking for. At times, we get more ideas and we may want to continue searching from the words we just got and we may go on several rounds in this fashion. Given this broad idea, it is not necessary that a thesaurus be constructed strictly in terms of synonyms. Any word that is semantically related in some way to the given word can be linked. In fact by going beyond the strict notion of synonym, we may be able to produce a more general and more useful resource. In fact WordNet is just such an extension.

The biggest challenge in constructing a thesaurus, therefore, is in identifying words that are semantically related to one another. Manual construction of thesauri is a tedious and time consuming task. Manually constructed thesauri also tend to suffer from problems of bias, inconsistency and limited coverage. In addition, thesaurus developers cannot keep up with constantly evolving language usage and cannot afford to build new thesauri for many new sub-domains that NLP techniques are being applied to. There is a clear need for automatic construction of thesauri.

Recent work has focused on automatic or semi-automatic construction of thesauri from parallel corpora, annotated corpora, and other available lexical resources. See for example [4, 2, 1, 3, 10, 5]. However, these techniques are not applicable to Indian languages at present since corpora and other lexical resources available in electronic form are extremely limited, although there is some recent interest in developing such resources. There are small (about 3 Million word) plain text corpora for most major languages of India but hardly any parallel corpora or annotated corpora. There are of course no wordnets etc. as yet. There are no significant computational grammars or syntactic parsers for any of these languages. Electronic dictionaries are, however, available in many languages.

Here we show that a bilingual dictionary is one good source that can be tapped. Dictionaries are more readily available in Indian languages compared to other forms of electronic resources. A bilingual dictionary, especially of the kind developed with applications such as automatic translation, tends to list target language equivalents for each source language word. In doing so, these dictionaries actually group together related words. It should therefore be possible to extract this hidden structure and build a thesaurus. This is the main idea in this paper.

The content of a thesaurus is very similar to that of a dictionary. A dictionary is typically organized in, say, alphabetical order so that you can quickly locate the word of interest and then you can get the correct spelling, pronunciation, meanings, usage, etymology and other such pieces of information associated with the word in question [6, 7]. A thesaurus, on the other hand, could be organized in terms of an ontology - a hierarchy of concepts, and the words are structured into groups that convey a specific meaning. The difference between a dictionary and a thesaurus, therefore, is more of structure and organization rather than that of content. Both the dictionary and the thesaurus contain words of a given language and their meanings.

Given this, it makes a lot of sense to consider a dictionary and a thesaurus as simply two different views of the same data, rather than as two entirely different entities. It appears to be a good idea to store the words only once and provide two different indexing mechanisms, one to use the words as a dictionary, and another to use the same words as a thesaurus [9]. Some kind of a thesaurus can thus be automatically and very efficiently constructed from a dictionary and such a thesaurus can be practically very useful. In this paper we show that a thesaurus can be constructed automatically and efficiently from a bilingual dictionary with little human labor. We show examples from a Kannada thesaurus constructed automatically from a bilingual English-Kannada dictionary also developed by the author. It may be noted that there is hardly any large scale lexical resource available today for Kannada although Kannada is a major language spoken by more than 50 Million people. An automatically constructed thesaurus may not be as good as one that is carefully handcrafted by lexicographers. But it can serve an immediate need. Also, a thesaurus so generated can be viewed as a raw material for further research and development.

2 Automatic Construction of Thesauri

To construct a thesaurus automatically, the data needed include the words of the language, the grammatical categories and other relevant features, and the meanings. Different words may have same spellings and a word may have many meanings (homonymy and polysemy). It is important to keep these things in mind while developing a thesaurus. Perhaps the best single source of all these required pieces of information is the dictionary itself. We now give the skeleton of an algorithm to show the basic idea:

```
#ALGORITHM:

#INPUT: A DICTIONARY
#OUTPUT: A THESAURUS

#First Create a Reverse Index:

For each dict. entry with head word W
   For each category i = $C_1$, $C_2$, ... $C_n$
     For each meaning j = $M_1$, $M_2$, ... $M_p$
       For each synonym k = $S_1$, $S_2$, ... $S_q$
         index(i,j,k) = W

# Create the thesaurus index:

For each word W
   For all HW = index(i,j,W)
     synset(i,j,W) = synset(i,j,W) Union (i,j,X) for all index(i,j,X) = HW
```

Note that the algorithm keeps the synsets separately for each category and each meaning and thus users should be able to locate the word they are looking for without mixing up different grammatical categories or different senses of a given word.

The algorithm has been implemented efficiently using suitable data structures and hashing techniques. It takes only a few minutes to generate the complete thesaurus on a desktop personal computer.

3 A Thesaurus for Kannada

To the best of our knowledge, Kannada, a language spoken by more than 50 million people and with vast and rich literature dating back to many centuries, has no thesaurus till date. A thesaurus for Kannada was generated automatically as described above starting from an English-Kannada dictionary. This dictionary was developed by the author [8] for the purpose of machine translation from English to Kannada. As such, the dictionary gave more or less substitutable equivalents rather than elaborate descriptions or precise definitions. Further, the dictionary was designed to contain a large number of synonyms since the choice of translated words is best left to the human post editor looking at the output of the machine and he must be given adequate choices to select from. These form the ideal conditions for automatic construction of thesauri by our technique.

We give below a few examples from the thesaurus:

huDugi :

Synset	Category	Sense
baalike	n	GIRL
kanye	n	GIRL

mane :

Synset	Category	Sense
kaTTaDa	n	BUILDING
sadana	n	HOUSE
gRha	n	HOUSE
kaTTaDa	n	HOUSE
nivaasa	n	RESIDENCE
vaasasthaana	n	RESIDENCE
vaasa	n	HABITATION
iruvu	n	HABITATION
biiDu	n	HABITATION
vasati	n	HABITATION

Clearly, the words we get from the thesaurus are not exactly synonyms in the strictest sense of the word. But the whole idea of a thesaurus is to provide a tool to the user to explore the semantic space of words by offering terms that are related in some way to the given word. Users are often not looking for exact synonyms, they are in fact looking for terms that may mean a somewhat different thing, the term that fits the particular usage on hand. This is why a thesaurus of this kind has been found be very useful in multilingual contexts and for translation. We give more examples below:

cikka:

Synset	Category	Sense
saNNa	a	LITTLE
svalpa	a	LITTLE
koMca	a	LITTLE
tusa	a	LITTLE
kiriya	a	LITTLE
giDDa	a	SHORT
kuLLa	a	SHORT
mooTu	a	SHORT
saNNa	a	SMALL
puTTa	a	SMALL
kiriya	a	SMALL
kSudra	a	SMALL
saNNa	a	TINY
puTaaNi	a	TINY
kaDime	a	LESS
eLeya	a	YOUNG
hareyada	a	YOUNG
yauvanaavastheya	a	YOUNG
yuvakanaada	a	YOUNG
anubhavavillada	a	YOUNG

The word ’cikka’ can be used in the sense of YOUNG and it is conceivable that in the context of someone being still too young, a connotation of lack of experience is involved. Thus ’anubhavavillada’ - ’not experienced’ is surely not a synonym but something that is semantically related to the given word.

nooDu:

Synset	Category	Sense
paris’iilisu	v	LOOK
diTTisu	v	LOOK
kaaNu	v	LOOK
tooru	v	LOOK

Observe how the transitive and intransitive senses are both included. Appropriate grammatical features from the dictionary can be used to show such variations.

eeLu:

Synset	Category	Sense
sapta	n	SEVEN
huTTu	v	RISE
udayisu	v	RISE
heccu	v	RISE
eddeeLu	v	RISE
udbhavisu	v	RISE

The different senses are indicated here by English words. It is possible to view these as abstract definitions of word meanings rather than as words of any particular language. We have chosen English words here since they are readily available from the bilingual dictionary we started with.

4 Conclusions

We have shown that a thesaurus can be automatically and efficiently constructed from a good dictionary with little human effort. The method holds promise since it is relatively easy to develop electronic dictionaries and other lexical resources are not yet available for many Indian languages. The quality of the thesaurus depends on the quality of the dictionary we start from. It is also possible to use this tool to verify the quality of a dictionary and hence correct, enhance, enrich and otherwise improve the dictionary itself. The automatically constructed thesaurus can also be taken as a starting point for developing a better thesaurus.

We have shown extracts from the Kannada thesaurus constructed automatically from our English-Kannada dictionary. To the best of our knowledge, there was no thesaurus for the Kannada so far. The thesaurus has been used for post-editing of machine translated output [8] and also as an independent tool by researchers and students. The thesaurus has been found to be useful.

Only informal and limited manual evaluations have been carried out so far but the results are very encouraging. Lack of other thesauri, word-nets, sense tagged corpora, parallel corpora etc. for Kannada is a serious issue for large scale quantitative evaluation of the current work. Lexical resources for Kannada are slowly getting developed and systematic, large scale quantitative evaluations will be possible soon.

The algorithm has since been cast as a general purpose tool for thesaurus constructing. Given a good bi-lingual dictionary we can get a first-cut thesaurus out automatically.

References

[1] Z. Chen, S. Liu, L. WenYin, G. Pu, and W.-Y. Ma. Building a web thesaurus from web link structure. Technical Report MSR-TR-2003-10, Microsoft Research, 2003.

[2] J. Curran. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing Philadelphia*, pages 222 – 229, PA, USA, 2002.

[3] H. Dejean, E. Gaussier, and F. Sadat. Bilingual terminology extraction: an approach on a multilingual thesaurus applicable to comparable corpora. In *Proceedings of COLING - 2002*, 2002.

[4] E. A. Fox, J. T. Nutter, T. Ahlswede, M. Evens, and J. Markowitz. Building a large thesaurus for information retrieval. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 101–108, Austin, TX, 1988. ACL.

[5] J. Jannink and G. Wiederhold. Thesaurus entry extraction from an on-line dictionary, 1999.

[6] Narayana Murthy Kavi. Electronic dictionaries and computational tools. *Linguistics Today*, 1(1):34–50, 1997.

[7] Narayana Murthy Kavi. An indexing technique for efficient retrieval from large dictionaries. *Proceedings of National Conference on Information Technology NCIT-97, 21-23 December 1997, Bhubaneswar*, 1997.

[8] Narayana Murthy Kavi. Mat: A machine assisted translation system. *Proceedings of the Fifth Natural Language Pacific Rim Symposium, NLPRS-99, 5-7 November, Beijing, China*, 1999.

[9] Sivasankara Reddy A, Narayana Murthy Kavi and Vasudev Varma. Object oriented multipurpose lexicon. *International Journal of Communication*, 6(1 and 2):69–84, 1996.

[10] T. Takenobu, I. Makoto, and T. Hozumi. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI-95*, 1995.