8.14 Automatic Construction of Telugu Thesaurus from Available Lexical Resources

M. Santhosh Kumar and Kavi Narayana Murthy, University of Hyderabad

Abstract—This paper is about automatic construction of a Telugu Thesaurus from a variety of available lexical resources. A thesaurus links semantically related words and helps in the selection of most appropriate words for given contexts. A thesaurus contains synonyms (words which have basically the same meaning) and as such, an important tool for many applications in NLP too. Yet many of the major languages of India have no thesauri till date. Recent work has focused on automatic or semi-automatic construction of thesauri from annotated corpora and other available lexical resources. Annotated corpora and other lexical resources are limited in Indian languages, and hence, many of the techniques used for other languages of the world are not applicable at present to Indian languages. However bilingual dictionaries are available. It has been shown that a thesaurus can be constructed automatically and efficiently from bilingual dictionaries with little human effort [11]. Here we present the algorithm used for this and show its application for the automatic construction of a Telugu thesaurus from available lexical resources. We include examples from the Telugu thesaurus so constructed.

Index Terms—Thesaurus, Dictionary, Indexing

I. 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 thesaurus when we have an idea, some concept or 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 some how 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 location 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 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.

M. Santosh Kumar and Kavi Narayana Murthy are with the Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh (e-mail: santosh_ai@yahoo.co.in, knmuh@yahoo.com)

II. AUTOMATIC CONSTRUCTION OF THESAURUS

The biggest challenge in constructing a thesaurus, therefore, is in identifying the 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 can not 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 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 words) text corpora for most major languages of India but hardly any parallel corpora or annotated corpora. There are of course no word nets 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.

III. DICTIONARY VERSUS THESAURUS

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 [P6,P7] 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 [P9]. 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 Telugu thesaurus constructed automatically from bilingual English-Telugu dictionaries. It may be noted that there is hardly any large scale lexical resource available today for Telugu, although Telugu is a major language spoken by many people in India. 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.

IV. METHODOLOGY

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 dictionary entry with head word W
  For each category i = C1, C2, ... Cn
    For each meaning j = M1, M2, ... Mp
      For each synonym k = S1, S2, ... Sq
        push W into index(i,j,k)

#Create the thesaurus index:

For each word W
For all HW = index(i,j,W)
synset(W,i,j) = synset(W,i,j) 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.

V. THESAURUS FOR TELUGU

Telugu is the second most spoken language in India, one of the twenty-two official languages of the Republic of India and one of the official languages of the state of Andhra Pradesh. Telugu has a vast and rich literature dating back to many centuries. Yet there is no widely available electronic thesaurus till date. In this work, a thesaurus for Telugu was generated automatically starting from two English-Telugu dictionaries. One was of these was developed by C.P.Brown and the other was developed here as a part of a machine aided translation project earlier. These dictionaries give more or less substitutable equivalents rather than elaborate descriptions or precise definitions and are therefore suitable for our purpose here.

More than 30,000 root words were extracted from the above two bilingual dictionaries. Inverted indexing as described above resulted in the thesaurus in less than a second. Below we show excerpts from the thesaurus. For each word in Telugu, corresponding synonyms are listed based on their category and also its sense in English.

Synset	Category	Sense
telivi	N	Judiciousness
buddhi_sukSmata naipuNi	N	Sagacity
sami:kSanamu ja:gratta	N	Discretion
Kaus'alamu pravi:Nata ne:rpu	N	Skill
jN~a:namu buddhi telivi	N	wisdom

Table 1: Synonyms of the word "vive: kamu" (wisdom).

Here are some statistics. Total number of Telugu words is 30361. Average number of synonyms per word is 1.39. This could be higher if the dictionaries gave more number of equivalents. Maximum number of synonyms for a word in the thesaurus is 28 (for the word 'peMcu', see table-2). Maximum categories for a word are 5 (for the word 'le:du' - Adjective, Noun, Adverb, Determiner and Interjection, see table-3). The synset with maximum number of synonyms for a word in particular category is 9 (see table-4). The total number of synsets found in the thesaurus is 27558.

Table 2: synset of the word "peMcu", is the biggest synset among all the synset in the developed thesaurus

Synset	Category	Sense
sa:gu_ce:yu	VT	cultivate
perugu edugu	VT	Grow
ekkuva_ceyyi vRddhi_ceyyi vRddhi_poMdiMcu	V	Increase
vRddhi_ceyyi pogaDu ki:rtiMcu vRdhdhi_ceyyi peMci_cu:piMcu	V	magnify
po:SiMcu	V	nourish
po:SiMcu pro:tsahiMcu	V	foster
po:SiMcu	V	cherish
hecca_veyyi guNiMcu abhivRdhdi_avu	V	multiply
sa:gadi:yu	V	elongate
merugu_parucu	V	enhances
paikettu pogaDu	V	exalt
po:SiMcu vRddhi_ceyyi pado:nnati_kalpiMcu protsahiMcu	V	promote
merugu_parucu	V	enhance
balaMga:_toyyi	V	boost
po:SiMcu ne:rpiMcu	V	nurture
hecciMcu vRddhi_ce:yu nilabeTTu le:vanettu	V	raise

Table 3: Synonyms of the word "le:du (no)" which is accounting for 5 categories

Synset	Category	Sense
ka:du	Det	No
ka:du ka:du	Adv	no Not
ka:du paiga le:kuMDaTaM	N	No Nay Haven't
ka:du	I	nope
e:di_ka:du	Adj	no

Table 4: A synset which contains maximum number of synonyms for a category

Synset	Category	Sense
muTTaDi avaro:dhaM aDDaMki digbadhaM a:TaMkaM muTTaDiMcu digbaMdhaM_ceyyi digbaMdhaM_ce:yu cuTTu_muTTaDi	N	blockade

VI. CONCLUSION

It has been shown earlier that a thesaurus can be automatically and efficiently constructed from a good dictionary with little human effort. In this paper we show the application of this idea for the automatic construction of a Telugu thesaurus. The necessary code was developed in Perl under Linux. The program takes less than a second to construct the thesaurus. The method holds promise since it is relatively easy to develop electronic dictionaries and other lexical resources 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 Telugu thesaurus constructed automatically from existing English-Telugu dictionaries. To the best of our knowledge, there was no thesaurus for the Telugu so far. 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 Telugu is a serious issue for large scale quantitative evaluation of the current work. Lexical resources for Telugu are slowly getting developed and systematic, large scale quantitative evaluations may be possible in future.

REFERENCES

[1] W. 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.

[11] Kavi Narayana Murthy, "On Automatic Construction of a Thesaurus" Proceedings of ICSLT-O-COCOSDA 2004 International Conference, Vol-1, pp 191-194, November 2004, New Delhi

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.