

Linking Monolingual Resource with Bilingual Resource

Akshar Bharati
Rajeev Sangal

Vibhav Agarwal

Sandipan Dandapat
Dipti Misra Sharma

International Institute of Information Technology Hyderabad
{vibhav,sandipan}@students.iiit.net, {sangal,dipti}@iiit.net

1. Introduction

This paper outlines the algorithm to link lexical resources available freely on the Net and Shabdaanjali, a free bilingual dictionary developed in-house at the Language Technology Research Centre, IIIT Hyderabad (LTRC). The algorithm also uses an application to combine the information available in Shabdaanjali and WordNet. WordNet gives different synset for a particular word and Shabdanjali has the senses of the word. The structuring of the hierarchy in which the senses appear in the dictionaries are different from the hierarchy of the synset appear in the WordNet. The challenge was to map the correct sense from Shabdaanjali to the closest synset from the synset list of WordNet.

2. The Problem

We used three sources of lexical information: WordNet, a semantic net of words arranged conceptually and programmed to offer output in terms of synonyms, antonyms, hyponyms, hypernyms, meronyms, etc.; a monolingual dictionary consisting of over a hundred thousand entries, an English-Hindi dictionary developed at LTRC. Each of the resources had their own strength and weaknesses, but the biggest problem stemmed from the fact that often, the number of senses allotted to words in each resource was different. For example, if Shabdaanjali listed only two senses of the noun Plant, WordNet listed four senses. How were we map the appropriate senses in each of this resource automatically to built more informative lexical resources for the future application?

3. Methodology

To link Shabdaanjali and WordNet, we have used three approaches:

- (a) Dictionary Based Approach, measure the similarity between various Shabdaanjali senses of a given word with the WordNet synset of that word by using dictionaries.

- (b) Corpus Based Approach, uses example sentences related to each sense given in Shabdaanjali to measure the similarity between sense and WordNet synset.
- (c) Combined Approach, is a combination of previous two approaches, Dictionary-based approach and Corpus Based approach.

3.1 Dictionary Based Approach

The first approach to link Shabdaanjali and WordNet was based on sense similarity. It measures the similarity between various Shabdaanjali senses of a given word with the Wordnet synsets of that word by using dictionaries.

We used two dictionaries:

Shabdaanjali - An English-Hindi dictionary that consists of more than 27,000 entries.

Brahad Hindi Shabdakosha - A big Hindi-Hindi dictionary that consists of more than 1,00,000 Hindi entries.

The second dictionary Brahad Hindi Shabdakosha was used for enhancement of results.

For a given Shabdanjali sense sh_1 , the similarity is calculated between sh_1 and a WordNet Synset by looking the senses of words of Synset in Shabdaanjali. If sense of the words of Synset and sh_1 is not matched then we go to Brahad Hindi Shabdakosha for further matching and a score is given. According to the score, mapping of sh_1 with a Synset was decided.

The words occurring in a Synset provide an essential clue towards the unique meaning conveyed by that synset. So, cluster of words of a synset was made to determine similarity between a sense of a given word with that cluster.

For example, "plant" has two senses in Shabdaanjali and four synsets in WordNet. The senses of "plant" in Shabdanjali are: "vanaspati" and "sanyantra"

And Synsets of "plant" in WordNet are

Synset 1: plant, works, industrial plant, building complex, complex

Synset 2: plant, flora, plant life, life form, organism, being, living thing

```

For a word 'w', do the following:
Store Shabdaanjali senses of 'w' in a list SHW
Make clusters of synsets of 'w' and store them in CLUSTER
For each entry i of SHW, do the following

    Set Scorei,j to '0'
    For each cluster j of 'w', do the following

        Match Shabdanjali senses of each word
        of CLUSTERj with SHWi and if senses
        are matched then
            Scorei,j = 7 + (1/4)
        If some part of the senses is matched,
        give some penalty in the score
            Scorei,j = 7 + (1/4) * penalty
        If Scorei,j is less than threshold value then set
        Scorei,j = 0
        If Scorei,j = 0, Go to Hindi-Hindi dictionary for
        Shabdaanjali sense of each word of CLUSTERj,
        and do the following

            If senses are matched then
                Scorei,j = 4 + (1/4)
            If some part of the senses is matched,
            give some penalty in the score
                Scorei,j = 4 + (1/4) * penalty

        If Scorei,j is less than threshold value then set
        Scorei,j = 0

    Return Scorei,j

```

Synset 3: plant, contrivance, stratagem, dodge

Synset 4: plant, actor, histrion, player, thespian, role player

It is clear that cluster #1 is conveying sense of *vanaspati(botanical)* and cluster #2 is conveying sense of *sanyantra(industrial plant)*. Cluster #3 and cluster#4 are not supposed to match with any of two sense of plant given by Shabdaanjali.

The matching algorithm is given in above box.

For example, when the algorithm was tested on *plant*, we got following scores between two senses of *plant* and each Synset cluster.

```

Score(vanaspati, Synset#1) = 0
Score(sanyantra, Synset #1) =0
Score(vanaspati, Synset#2) = 7.25
Score(sanyantra, Synset #2) = 0
Score(vanaspati, Synset#3) = 0
Score(sanyantra, Synset #3) = 0
Score(vanaspati, Synset#4) = 0
Score(sanyantra, Synset #4) = 0

```

As the scores shows vanaspati is matching with Synset #2, to whom it should have matched but sanyantra is not matching with any Synsets. To

evaluate the algorithm, we proposed an evaluation criterion

3.2 Corpus Based Approach

In the previous approach, we had used the senses of words given in Shabdaanjali. The second approach for linking Shabdaanjali and WordNet uses example sentences related to each sense given in Shabdaanjali to measure the similarity between sense and WordNet Synset. Using the WordNet, the words in the hypernymy and hyponymy synsets are grouped together to form semantic cluster for each sense of the noun and verb. The verbs co-occurring in the example sentence of a sense is used to obtain the most likely synset for that sense of the noun word. Same thing was done to get most likely synsets for the senses of verb words by using co-occurring nouns. The algorithm uses the corpus based statistical technique for selecting semantic cluster relevant to each sense of the noun and verb. We used British National Corpus (BNC).

3.2.1 Method and Implementation

The intuition of verb-noun relation is that the verb, which is correlated to a noun, also correlates with the hyponymy and hypernymy of that noun. In a given sentence verb can be used to find out the sense of the noun present in the sentence. Same thing is followed for noun-verb relation.

For example, consider the two senses of 'plant', as '*buildings for carrying on industrial labor*' or '*a living organism lacking the power of locomotion*'. Only the sense of 'buildings for carrying on industrial labor' is available in the context of '*they built a large plant to manufacture automobiles*' because the verb 'built' helps in identifying the meaning of its object plant.

The words co-occurring in the sentence provide essential clue towards the intended meaning of the polysemous words. The verb-noun co-occurring pair determines the meaning of the polysemous nouns and the noun-verb co-occurring pair determines the meaning of the polysemous verbs.

As we have discussed earlier, different synsets of a word conveys different senses of that word. It was the basis of our first approach i.e. dictionary based approach and again it is the basis of this approach.

The method proposed here is basically based on two observations

- (a) Different senses of the words appear in different semantic clusters.

- (b) The relation between the verb and nouns; verbs which co-occur with a noun also co-occur with the words in the hyponymy and hypernymy synsets of that word.

For example, consider the hyponymy synsets of plant. It is likely to find instances of (build, factory) and (build, manufacturing plant) in British National Corpus and less probable to get instances of (build, acrogen) and (build, aquatic) in British National Corpus. Thus, by combining the lexical knowledge acquired from the IS-A taxonomy and word-word association, a usable estimate of conditional probability can be obtained.

This system takes example sentence of a word from Shabdaanjali as input. Preprocessing was done to find out co-occurring verbs and nouns in the example sentence using Brill Part of Speech Tagger and English Morphological Analyzer. The set of co-occurring verb-noun and noun-verb pairs were obtained from the syntactically tagged British national corpus (BNC). Again preprocessing was done using English Morphological Analyzer. After obtaining the co-occurring verb-noun and noun-verb pairs, probability of each sense with each semantic cluster was calculated.

Conditional probability was calculated as

$$P(\text{noun}|\text{verb}) = \frac{\text{freq}(\text{noun, verb})}{\text{freq}(\text{verb})}$$

The probability of a semantic cluster with the co-occurring verb can be determined as:

$$P(\text{SemCluster}_i | \text{verb}) = \frac{1}{Z} \sum_{w \text{ in all Semcluster}_i} (P(w | \text{verb}))$$

Where 'Z' is the normalizing factor taken over all words in all semantic clusters.

$$Z = \sum_{w \text{ in all SemCluster}} P(w | \text{verb})$$

We have calculated probability for nouns. Probability for verb can be calculated by same method. So, whole algorithm is summarized as: Preprocess the British National Corpus by using English Morphological Analyzer to obtain verb-noun co-occurring pairs and also noun-verb co-occurring pairs.

For a given word 'w', do the following

- Obtain example sentence related to each sense of 'w' from Shabdaanjali

- Run Brill Part of Speech Tagger and English Morphological Analyzer to get co-occurring nouns and verbs.
- Construct Semantic clusters $SemCluster_i$ for each sense i of 'w'
- For each Semantic cluster obtain the probability with each co-occurring verbs and nouns.
- If this probability is less than defined threshold value, ignore it.
- Select the sense of 'w' for which probability is greater than threshold value

3.3 Combined Approach

This approach is a combination of previous two approaches, Dictionary-based approach and Corpus-based approach. In the Dictionary-based approach, we had used the senses of the Shabdaanjali and in the Corpus-based approach; we had used the example sentences related to each sense. In Combined approach, we have used both the senses and the example sentences.

The combined approach is actually the combination of the results of both algorithms. As we had seen, the results of Dictionary-based approach were much better than the results of Corpus-based approach. So, the results of Dictionary-based approach were used as the base on which results of Corpus-based approach were added.

The algorithm of combined approach is:

```

For a given word 'w', do the following
Store Shabdaanjali senses of 'w' in a list SHW
For each entry 'i' of SHW, do the following
  For each Synset 'j' of 'w', do the following
    Set Scorei,j to 0
    Run Dictionary-based algorithm
    If Scorei,j is less than 7 then do the
    following
      Run Corpus-Based algorithm and
      store result of that algorithm in
      CORP_SCOREi,j
      If CORP_SCOREi,j is less than
      threshold value then set Scorei,j
      and CORP_SCOREi,j to 0 else
      set Scorei,j = 7 +
      CORP_SCOREi,j
return Scorei,j

```

4. Experiments and Results

The algorithms have been tested separately for nouns and verbs. Also, separate threshold value was defined for nouns and verbs. The algorithms were tested on 65 randomly chosen nouns and 45 randomly chosen verbs.

In Dictionary Based Approach average precision and recall for noun are 85% and 73% respectively and verb it is 80% and 71% respectively, which shows the effectiveness of this

approach in aligning Shabdaanjali with WordNet. Also, this approach is of very less complexity compared to the other approaches.

The evaluation criterion for Corpus Based Approach is same as we discussed in Dictionary-based approach. This algorithm was also separately tested for nouns and verbs. It was tested on same Data Set on which Dictionary-based algorithm was tested. Also separate threshold values were defined for nouns and verbs. These threshold values had no relation with the threshold values defined in Dictionary-based Approach.

The average precision obtained was 31% and average recall obtained was 95%. After introducing a threshold value of 0.1, the average precision and recall became 49% and 57%.

The average precision obtained was 26% and average recall obtained was 96%. After introducing a threshold value of 0.3, average precision became 39% and average recall became 29%.

These results were not so good but it is said, "Research calculates success in failure". How much this was true in our task, was proved later when we combined both approach i.e. Dictionary-based approach and Corpus-based approach.

Combined approach, using both the strategies, improved the results to 90.75% precision and 64.26% recall for nouns and 89.23% precision and 63.3% recall for verbs.

As seen from the results this combined approach is optimal and suggested for application in word sense disambiguation and machine translation.

The evaluation results are given in following table.

Approach	Lexical Category	Precession (%)	Recall (%)
Dictionary Based	Noun	85.71	72.89
	Verb	80.35	71.42
Corpus Based	Noun	49	57
	Verb	30	29
Combined	Noun	90.75	72.89
	Verb	89.23	63.3

Table1: Results of evaluation

5. Further Development

The above algorithms are also implemented on Roget's thesaurus instead of WordNet. Roget's thesaurus also gives different synset of a particular word with a particular category. Clusters are created

from the synsets. Above algorithm are executed over the cluster defined from Roget's thesaurus. As the algorithm create a matrix of weight assigned to each synset for each sense in Shabdaanjali. Now to assign a synset for a sense of Shabdaanjali, it goes through the assignment problem algorithm and assign and assign the best synset for a sense in Shabdaanjali. If all the weights are zero for a sense of Shabdaanjali then no synset is assigned to the sense.

The challenge was to map the correct word from the synset list to the closest synonym of the ambiguous word in the target language. Algorithm gives a sense and the best synset, which fits to the sense.

6. Conclusion

In this paper, we presented a working algorithm that uses both bilingual and monolingual recourses and attempts to map the various senses of a word across languages. Using standard techniques of efficiency testing, we calculated the precession and recall of the algorithm at work and found that the average efficiency stood at 77.15% for noun and 76.27% for verbs.

References

- [1] Pianta, E., Bentivogli, L., and Girardi C. MultiWordnet: Developing an Aligned Multilingual database. *1st International Global WordNet Conference*, Mysore, India, 2002.
- [2] Carpuat M., Ngai G., Fung P., Church W. K. Creating A Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet, *1st International Global WordNet Conference*, Mysore, India, 2002.
- [3] Lafourcade M. and Prince V. Relative Synonymy and Conceptual Vector, *6th NLP Pacific Rim Symposium*, 2001, Japan. 2001
- [4] Narayan D. K. and Bhattacharyya P. Using verb-Noun Association for Word Sense Disambiguation , *Int Conf Knowledge Based Computer Systems*, Mumbai, India, 2002.
- [5] Dekang Lin - Automatic Retrieval and Clustering of Similar Words, *COLING-ACL98*, Montreal, Canada, 1998.
- [6] *Shabdanjali : A Free English-Hindi Bilingual Dictionary*, LTRC, IIIT Hyderabad (2001), <http://www.iiit.net/ltrc/Dictionaries/Shabdanjali/dict-README.html>