Eating your own cooking: automatically linking wordnet synsets of two languages

Salil Joshi   Arindam Chatterjee   Arun Karthikeyan Karra
Pushpak Bhattacharyya
(1) IBM Research India, Bangalore, India
(2) Symantec Labs, Pune, India
(3) Oracle Labs, Hyderabad, India
(4) CSE Department, IIT Bombay, Mumbai, India
saljoshi@in.ibm.com, arindam.chatterjee23@gmail.com,
arun.karthikeyan.arun@gmail.com, pb@cse.iitb.ac.in

ABSTRACT

Linked wordnets are invaluable linked lexical resources. Wordnet linking involves matching a particular synset (concept) in one wordnet to a synset in another wordnet. We have developed an automatic wordnet linking system that is divided into a number of stages. Starting with a synset in the first language (also referred to as the source language), our algorithm generates a list of candidate synsets in the second language (also referred to as the target language). In consecutive stages, a heuristic is used to prune and rank this list. The winner synset is then chosen as the linkage for the source synset. The candidate synsets are generated using a bilingual dictionary (BiDict). Further, the earlier heuristics which we developed used BiDict to rank these candidate synsets. However, development of a BiDict is cumbersome and requires human labor. Furthermore, in several cases sparsity of the BiDict handicaps the ranking algorithm to a great extent. We have thus devised heuristics to eliminate the requirement of BiDict during the ranking process by using the already linked synsets. Once sufficient number of linked synsets are available, these heuristics outperform our heuristics which use a BiDict. These heuristics are based on observations made from linking techniques applied by lexicographers. Our wordnet linking system can be used for any pair of languages, given either a BiDict or sufficient number of already linked synsets. The interface of the system is easy to comprehend and use. In this paper, we present this interface along with the developed heuristics.

KEYWORDS: Wordnet Linking, Bilingual Dictionary, Resource reuse for linking.

1 Introduction

Wordnet Linking, as the name suggests, is the process of linking wordnet of one language to another. Efforts towards mapping synsets across wordnets have been going on for a while in various parts of the world. EuroWordNet (Vossen and Letteren, 1997) is one of the projects which is attempting to link wordnets across various European languages. Another effort towards wordnet linking can be found in the MultiWordNet (Pianta et al., 2002), aligning the Italian and the English language wordnets. Our linking process can be used for any pair of languages. Currently, we support linking for *Hindi to English*. Our wordnet linking system is automatic and involves deployment of heuristics to find the correct linkage. We initially developed BiDict (Bilingual Dictionary) based heuristics for ranking. The BiDict is a dictionary, mapping words from the source language to the target language. The usage of the such a dictionary has certain bottlenecks as follows:

1. **Human effort:** The development of the BiDict is cumbersome and requires considerable human efforts.
2. **Sparsity:** The BiDict is typically sparse in nature and hence in several cases the ranking results are error prone.
3. **Morphological issues:** The bilingual dictionary entries are categorized by their respective parts of speech. Hence in quite a few cases, the word forms differ in their morphology, which affects the performance of the system as the word form given in the synset and that in the dictionary do not match.

Due to the issues mentioned above, we had to find an alternative way of ranking the candidate synsets without using the BiDict. We decided to *eat our own cooking*. We devised a strategy to use the already linked synsets for this purpose. The heuristics based on this strategy perform better than the BiDict based heuristics. In our wordnet linking system, the user can select a source language (Hindi) synset which goes as input into the system, along with the desired heuristic (both with and without the usage of the BiDict), which is again opted by the user. The system outputs the top-5 candidate target language synsets, ranked by the chosen heuristic, which can be used for linking purposes.

The key features of the system are as follows:

1. **Minimized dependency on dictionary:** The system uses heuristics which use the information from already linked synsets, to generate candidate synsets of the target language. Since the bilingual dictionary has numerous bottlenecks in the process of pruning the candidate synsets, the tool provides an option of ranking the candidate synsets without such a knowledge source. This is beneficial for language pairs, where an efficient bilingual dictionary is not available that maps synsets, and a reasonable number of linked synsets are available.
2. **User friendliness:** Our system interface provides a nice visual experience. The design of the interface makes the operation of the interface completely clear to the end user.
3. **System independence:** The system is independent of the web browser and the operating system it is used on. Since the business logic is written in Java, the system can be easily ported on another machine.
4. **Ease of configuration:** The wordnets, BiDict and already linked synsets can be provided to the system through a simple change in the configuration. This is particularly useful in porting the system for wordnet linking for a different pair of languages.
5. **Easily interpretable output:** Our interface is designed in such a way that the user can easily understand the linking data being displayed. When the candidate synsets of the target language are exhibited, the user can click on each candidate, to view the details of the synsets

i.e., each portion of the candidate synset (*viz.* synset id, gloss, example, *etc.*), is displayed separately, making the output easily comprehensible.

This paper is organized as follows. Section 2 lays down the system architecture of our wordnet linking system, followed by section 3 describing the heuristics employed for ranking candidate synsets. In section 4 we describe the operation of the tool, followed by section 5 comparing the results of the heuristics against the baseline. We conclude the paper with section 6.

2 System Architecture

Figure 1 shows the basic architecture of the model adopted to achieve linking of a Hindi synset with an English synset. Given a Hindi synset, the gloss, examples and synonymous words are parsed depending on the parts-of-speech (POS) and a bag of words are obtained. This bag of Hindi words are then translated to English using a Hindi-English bilingual dictionary. Using these translated bag of words and the English WordNet, candidate English synsets are selected. Now on each of these candidate synsets, the heuristics are applied and the synset to be linked in the target language is generated.

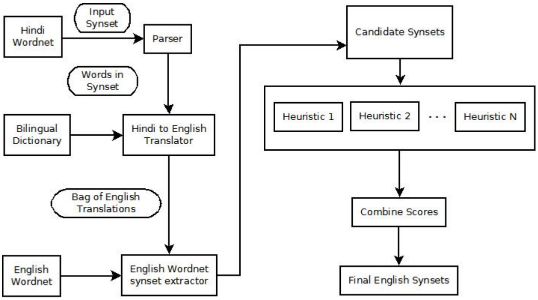


Figure 1: Hindi-English wordnet linking system architecture

3 Heuristics

As mentioned earlier, we have a set of heuristics for performing wordnet linking. Our initial heuristics made use of a bilingual dictionary, but due to the problems mentioned earlier, the results were error prone. We later developed heuristics which do not require any such dictionary for ranking purposes and yet provide comparable results in presence of already linked synsets. We present both types of heuristics in this section, each can be used under certain scenarios. Throughout this description, we follow the standard definitions for hypernym, hyponym, gloss, concept and synset (Fellbaum, 1998)

3.1 Heuristics based on Bilingual Dictionary

The heuristics that use the BiDict for ranking candidate synsets are as follows. These heuristics are particularly helpful in presence of a good quality BiDict:

1. **Monosemous Word Heuristic**- In this heuristic, the monosemous words in the source synset are only considered for obtaining the candidate synsets. First the translations of the monosemous words are obtained using the bilingual dictionary, then synsets containing these translations are chosen as candidate synsets.

2. **Single Translation Heuristic**- This heuristic is similar to the monosemous word heuristic. Here only those words which have single translations according to the bilingual dictionary are considered in obtaining the candidate synsets.
3. **Hyponymy/Hypernymy Word-bag Heuristics**- These heuristics rank candidates by finding the similarity of synset words of the hyponym/hypernym respectively, of the source synset with the candidate target synsets.
4. **Gloss/Synset/Concept Word-bag Heuristic**- These heuristics score the candidates based on the similarity between the words in the gloss or synset or concept respectively, on source and target sides.
5. **Synset/Concept Back Translation Heuristic**- These heuristics make use of synset or concept word translations respectively, from source to target language and vice versa. The candidates are ranked based on the combined score.

3.2 Heuristics based on already Linked Synsets

The primary aim of our work was to build a comprehensive, accurate and user friendly tool for wordnet linking. Since a low quality BiDict lowers the ranking performance of the system, we resorted to a strategy that avoids the usage of such a knowledge source for ranking, and makes use of the already linked synsets. The chief design strategy here was to find the closest synset from the existing set of linked synsets, to the synset in source language to be linked. The metric of *closeness* has been defined differently in different heuristics. These heuristics are particularly helpful in presence of sufficient number of already linked synsets.

3.2.1 Closest Common Synset Word-bag Heuristic

This heuristic uses the maximum intersection of the synset word-bags of the source synset to be linked and already linked synsets. Once this synset is found, its corresponding link on the target side is found. The candidate synsets are ranked based on the degree of intersection of synset words, with this synset in the target language. The heuristic score for each target language candidate synset is calculated follows:

$Score = |Target_{candidateSynsetWords} \cap Target_{closestSynsetWords}|$

3.2.2 Closest Common Concept Word-bag Heuristic

This heuristic uses the maximum intersection of the concept word-bags of the source synset to be linked and already linked synsets. It is similar to the Closest Common Synset Word-bag Heuristic, in operation. It uses the intersection of the concept word-bags instead of synset word-bags. Hence the heuristic score for each target language candidate synset is as follows:

$Score = |Target_{candidateConceptWords} \cap Target_{closestConceptWords}|$

3.2.3 Closest Hypernym Synset Heuristic

The total number of linked synsets from Hindi to English available is 24,124[1], which are manually inspected by our lexicographers. Among these 24,124 synsets, 18,441 (76.443%) are *nouns*. Hence a heuristic which performs well on nouns was desirable, as it would boost the accuracy of the entire system. More importantly, a strategy for finding similarity between the source language synset to be linked and the linked set of synsets was semantic relations available within the wordnet framework.

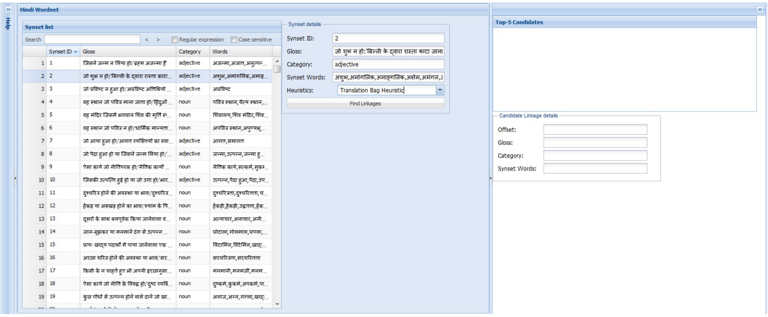
[1] as of June, 2012

Theoretically, the wordnet linking process should follow the hypernymy hierarchy of the source and target wordnets *i.e.,* if a synset *A* on source side is linked to synset *B* on target side, correspondingly synset *C* which is a hyponym of *A* in the source language wordnet should be linked to a hyponym of *B* on the target side. The heuristic score for each target language candidate synset can be calculated as follows:

$$Score = Distance_{hypernymy}(Target_{candidateSynset}, Target_{closestHypernym})$$

4 Interface Design

To support various web browsers on different operating systems, we have designed the web interface using standard open technologies. The interface runs using PHP5[2] on the server side, and for the GUI, we have used a javascript framework *viz.*, ExtJS v4.0[3] which provides a neat and aesthetic display to the user. Figure 2 shows the system interface. The main screen is divided into



	Synset ID	Gloss	Category	Words
1	1	जिसने जन्म न लिया हो:'ब्रह्म अजन्मा है'	adjective	अजन्मा,अजात,अनुत्पन्...
2	2	जो शुभ न हो:'बिल्ली के द्वारा रास्ता काटा...	adjective	अशुभ,अमांगलिक,अमाङ्...
3	3	जो प्रविष्ट न हुआ हो:'अप्रविष्ट अतिथियों ...	adjective	अप्रविष्ट
4	4	वह स्थान जो पवित्र माना जाता हो:'हिंदुओं ...	noun	पवित्र स्थान,चैत्य स्थान,...
5	5	वह मंदिर जिसमें भगवान शिव की मूर्ति स्...	noun	शिवालय,शिव मंदिर,शिव...
6	6	वह स्थान जो पवित्र न हो:'धार्मिक मान्यता...	noun	अपवित्र स्थान,अपुण्यभू...
7	7	जो आया हुआ हो:'आगत व्यक्तियों का स्वा...	adjective	आगत,समागत
8	8	जो पैदा हुआ हो या जिसने जन्म लिया हो:'...	adjective	जन्मा,उत्पन्न,जन्मा हु...
9	9	ऐसा कार्य जो नीतिपरक हो:'नैतिक कार्यों ...	noun	नैतिक कार्य,सत्कर्म,सुकर्...
10	10	जिसकी उत्पत्ति हुई हो या जो उगा हो:'भार...	adjective	उत्पन्न,पैदा हुआ,पैदा,उप...
11	11	दुश्चरित्र होने की अवस्था या भाव:'दुश्चरित्र...	noun	दुश्चरित्रता,दुश्चरितता,च...
12	12	हेकड़ या अक्खड़ होने का भाव:'श्याम के पि...	noun	हेकड़ी,हैकड़ी,उद्धतता,हेक...
13	13	दूसरों के साथ बलपूर्वक किया जानेवाला व...	noun	अत्याचार,अनाचार,अनी...
14	14	जान-बूझकर या मनमाने ढंग से उत्पन्न ...	noun	घोटाला,गोलमाल,घपला,...
15	15	प्रायः खाद्य पदार्थों में पाया जानेवाला एक ...	noun	विटामिन,विटैमिन,खाद्...
16	16	अच्छा चरित्र होने की अवस्था या भाव:'सच्...	noun	सच्चरित्रता,सच्चरितता
17	17	किसी के न चाहते हुए भी अपनी इच्छानुसा...	noun	मनमानी,मनमर्जी,मनम...
18	18	ऐसा कार्य जो नीति के विरुद्ध हो:'दुष्ट व्यक्ति...	noun	दुष्कर्म,कुकर्म,अपकर्म,पा...
19	19	कुछ पौधों से उत्पन्न होने वाले दाने जो खा...	noun	अनाज,अन्न,गल्ला,खाद्...

Synset ID: 2
Gloss: जो शुभ न हो:'बिल्ली के द्वारा रास्ता काटा जाना
Category: adjective
Synset Words: अशुभ,अमांगलिक,अमाङ्गलिक,अक्षेम,अमंगल,अ
Heuristics: Translation Bag Heuristic

Offset:
Gloss:
Category:
Synset Words:

Figure 2: Screen-shot showing the main interface of the system

neatly arranged panels, which can be collapsed, in case if the monitor resolution can not display the complete interface. The screen-shot shows that currently the *Help* panel on the left is in collapsed form. Each synset in the list displays the Synset ID, gloss, POS category and the constituent words fetched from Hindi Wordnet. Selecting a row in this list automatically populates the synset details in the form adjacent to the grid. The list of synsets can be searched for a particular synset.

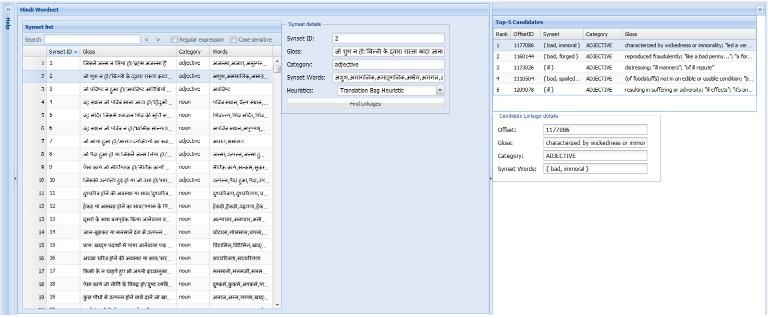
Once a source (Hindi) synset is chosen, the user simply needs to select a heuristic and submit the form. Based on the input synset and the heuristic, the system computes the candidate synsets, ranks them and returns top-5 candidates to the user. The outcome of this process is shown in figure 3.

5 Empirical evaluation

The system was tested on a total of 24,124 linked synsets manually inspected by our lexicographers. The results are shown in Table 1. We present the results for all the heuristics developed by us. Rows 1 to 9 summarize the performance of BiDict based heuristics, and rows 10 to 12 show the results obtained with heuristics which make use of already linked synsets. Row 13 compares the performance of all the heuristics against the random baseline, in which a random candidate synset is assigned as the linked synset. Clearly, the performance of heuristics is improved when they make use of already linked synsets for linking new synsets.

[2] http://php.net/downloads.php

[3] http://www.sencha.com/products/extjs/



	Synset ID	Gloss	Category	Words
1	1	जिसने जन्म न लिया हो:'ब्रह्म अजन्मा हैं'	adjective	अजन्मा,अजात,अनुत्पन्...
2	2	जो शुभ न हो:'बिल्ली के द्वारा रास्ता काटा...	adjective	अशुभ,अमांगलिक,अमाङ्...
3	3	जो प्रविष्ट न हुआ हो:'अप्रविष्ट अतिथियों ...	adjective	अप्रविष्ट
4	4	वह स्थान जो पवित्र माना जाता हो:'हिंदुओं ...	noun	पवित्र स्थान,चैत्य स्थान,...
5	5	वह मंदिर जिसमें भगवान शिव की मूर्ति स्...	noun	शिवालय,शिव मंदिर,शिव...
6	6	वह स्थान जो पवित्र न हो:'धार्मिक मान्यता...	noun	अपवित्र स्थान,अपुण्यभू...
7	7	जो आया हुआ हो:'आगत व्यक्तियों का स्वा...	adjective	आगत,समागत
8	8	जो पैदा हुआ हो या जिसने जन्म लिया हो:'...	adjective	जन्मा,उत्पन्न,जन्मा हु...
9	9	ऐसा कार्य जो नीतिपरक हो:'नैतिक कार्यों ...	noun	नैतिक कार्य,सत्कर्म,सुकर्...
10	10	जिसकी उत्पत्ति हुई हो या जो उगा हो:'भार...	adjective	उत्पन्न,पैदा हुआ,पैदा,उप...
11	11	दुश्चरित्र होने की अवस्था या भाव:'दुश्चरित्र...	noun	दुश्चरित्रता,दुश्चरितता,च...
12	12	हेकड़ या अक्खड़ होने का भाव:'श्याम के पि...	noun	हेकड़ी,हैकड़ी,उद्धतता,हेक...
13	13	दूसरों के साथ बलपूर्वक किया जानेवाला व...	noun	अत्याचार,अनाचार,अनी...
14	14	जान-बूझकर या मनमाने ढंग से उत्पन्न ...	noun	घोटाला,गोलमाल,घपला,...
15	15	प्रायः खाद्य पदार्थों में पाया जानेवाला एक ...	noun	विटामिन,विटैमिन,खाद्...
16	16	अच्छा चरित्र होने की अवस्था या भाव:'सच्...	noun	सच्चरित्रता,सच्चरितता
17	17	किसी के न चाहते हुए भी अपनी इच्छानुसा...	noun	मनमानी,मनमर्जी,मनम...
18	18	ऐसा कार्य जो नीति के विरुद्ध हो:'दुष्ट व्यक्ति...	noun	दुष्कर्म,कुकर्म,अपकर्म,पा...
19	19	कुछ पौधों से उत्पन्न होने वाले दाने जो खा...	noun	अनाज,अन्न,गल्ला,खाद्...

Synset details

Synset ID: 2

Gloss: जो शुभ न हो:'बिल्ली के द्वारा रास्ता काटा जाना

Category: adjective

Synset Words: अशुभ,अमांगलिक,अमाङ्गलिक,अक्षेम,अमंगल,3

Heuristics: Translation Bag Heuristic

Top-5 Candidates

Rank	OffsetID	Synset	Category	Gloss
1	1177086	{ bad, immoral }	ADJECTIVE	characterized by wickedness or immorality; "led a ver...
2	1160144	{ bad, forged }	ADJECTIVE	reproduced fraudulently; "like a bad penny..."; "a for...
3	1173026	{ ill }	ADJECTIVE	distressing; "ill manners"; "of ill repute"
4	1110504	{ bad, spoiled...	ADJECTIVE	(of foodstuffs) not in an edible or usable condition; "b...
5	1209078	{ ill }	ADJECTIVE	resulting in suffering or adversity; "ill effects"; "it's an...

Candidate Linkage details

Offset: 1177086

Gloss: characterized by wickedness or immor

Category: ADJECTIVE

Synset Words: { bad, immoral }

Figure 3: System interface showing the outcome of the linking process

Heuristic	Coverage	Accuracy
Monosemous Word	6644 (27.541 %)	4194 (63.131 %)
Single Translation	6100 (25.282 %)	3458 (56.692 %)
Gloss Word-bag	11298 (46.833 %)	5241 (46.393 %)
Hypernymy Word-bag	12127 (50.269 %)	5041 (41.570 %)
Hyponymy Word-bag	12127 (50.269 %)	5712 (47.108 %)
Synset Word-bag	12127 (50.269 %)	6068 (50.044 %)
Concept Word-bag	11298 (46.833 %)	5731 (50.737 %)
Synset Back Translation	12127 (50.269 %)	6133 (50.574 %)
Concept Back Translation	12127 (50.269 %)	6312 (52.052 %)
Closest Hypernym Synset	12127 (50.269 %)	9671 (79.758 %)
Closest Common Synset Word-bag	12127 (50.269 %)	9032 (74.482 %)
Closest Common Concept Word-bag	12127 (50.269 %)	6694 (55.203 %)
Random Baseline	12127 (54.071 %)	3024 (24.936 %)
Total number of synsets being mapped is 24,124		
*Average cardinality of candidate English synsets per Hindi synset is **25***		

Table 1: System Performance for different heuristics and random baseline

6 Conclusion

In this work, we presented a tool for wordnet linking which uses heuristic based approach. The necessity of a BiDict for ranking process was circumvented by designing new heuristics which made use of the already linked set of synsets from source to target languages. These heuristics perform at par with the heuristics based on the BiDict. Once quality linked data between two languages is available, tasks like WSD, Machine Translation, Cross-lingual Information Retrieval, *etc.*, can benefit from it. Our on-line system interface is simple yet user-friendly and allows the user to make use of several linking heuristics which we have developed. The system can be easily adapted for a new pair of languages by simply supplying the language wordnets along with either a bilingual dictionary or already linked synsets across the two languages.

In the future, we would like to support more languages. We would also like to provide the users a facility of adding new heuristics to our system for better comparison.

References

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Vossen, P. and Letteren, C. C. (1997). Eurowordnet: a multilingual database for information retrieval. In *In: Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.