Chapter 6

A Proposed Hybrid Approach for WSD

This chapter presents a Hybrid approach for WSD which combines information from multiple knowledge sources namely WordNet and Corpus. Collocational features are extracted from the texts in the Corpus and semantic relation are obtained from WordNet. The Corpora used in this chapter for experimentation are Bengali and Nepali corpora for Tourism , Health and Mixed Domain. A Mixed Domain Corpus consists of texts belonging to a wide variety of subjects- short stories, abridged novels, newspaper articles, sports topics etc.

6.1 Introduction

There are basically four important parameters which facilitate WSD to a large extent when applied on a corpus. These are:-

- Domain/Corpus Specific Sense Distribution
- Dominant Concepts within a domain
- Conceptual Distance
- Semantic Graph Distance

6.1.1 Domain/Corpus Specific Sense Distribution

Domain specific frequent senses of a wordcan be gleaned from sense tagged corpora. Domain/Corpus specific sense may vary from WordNet first sense(baseline).

For e.g. let us take the wordসুবিধা. The WordNet first sense forসুবিধা is সুবিধা_3350(convenience) which has the gloss “freedom from difficulty, hardship or

effort" whereas the most frequent or the most appropriate sense for সুবিধা in the Tourism domain is সুবিধা_28213(facility) which has the gloss " a service that an organization or a piece of equipment offers you". So we see that the WordNet first sense may not be the first sense in a specific domain.

6.1.2 Dominant concepts within a domain

A synset node in Wordnet Hypernymy is said to be dominant if the sub tree of synsets below it frequently occurs in the domain corpora. The dominant concepts for the Tourism and Health domains are listed in Table 6.1

Table 6.1 Dominant concepts for the Tourism and Health domains

Tourism	Health
{place,country,city,area}	{doctor,nurse}
{Mode of transport}	{patient}
{Flora-Fauna}	{disease}
{Fine arts}	{treatment}
{facility}	

6.1.3 Conceptual Distance

Conceptual Distance between two synsets (S1 , S2) is given by

$$\frac{\text{Length of the path between S1 and S2 in the WordNet hierarchy}}{\text{Height of the lowest common ancestor of S1 and S2 inthe WordNet hierarchy}} \quad \ldots\ldots(6.1)$$

Conceptual Distance increases with path length between two synsets and it is inversely proportional to the height of the common ancestor .

Exposition of Contextual Distance with an example:- "এরপর থেকে ১৯৯৬ খ্রীষ্টাব্দে স্বাধীনতা অর্জনের আগে অবধি এই রাজয়াংশ ব্রিটিশ গায়ানা হিসাবে পরিচিত ছিল। ১৮৩৪ খ্রীষ্টাব্দের **দাস** প্রথার অবসানের পর ব্রিটিশ **জমিদার**রা ভারতবর্ষ থেকে মজুর আমদানী করা শুরু করে " (from Gyan Nidhi corpus of TDIL[6]). Here **জমিদার** is a monosemous word having synset id 4365 is the context word and **দাস** which is a polysemous word is the target word i.e. the correct sense of **দাস** in the given context needs to be disambiguated. The

word**জমিদার**_4365 can be considered as the seed word for disambiguation. Let us draw the Hypernymy-Hyponymy graph for both the target and the context words.

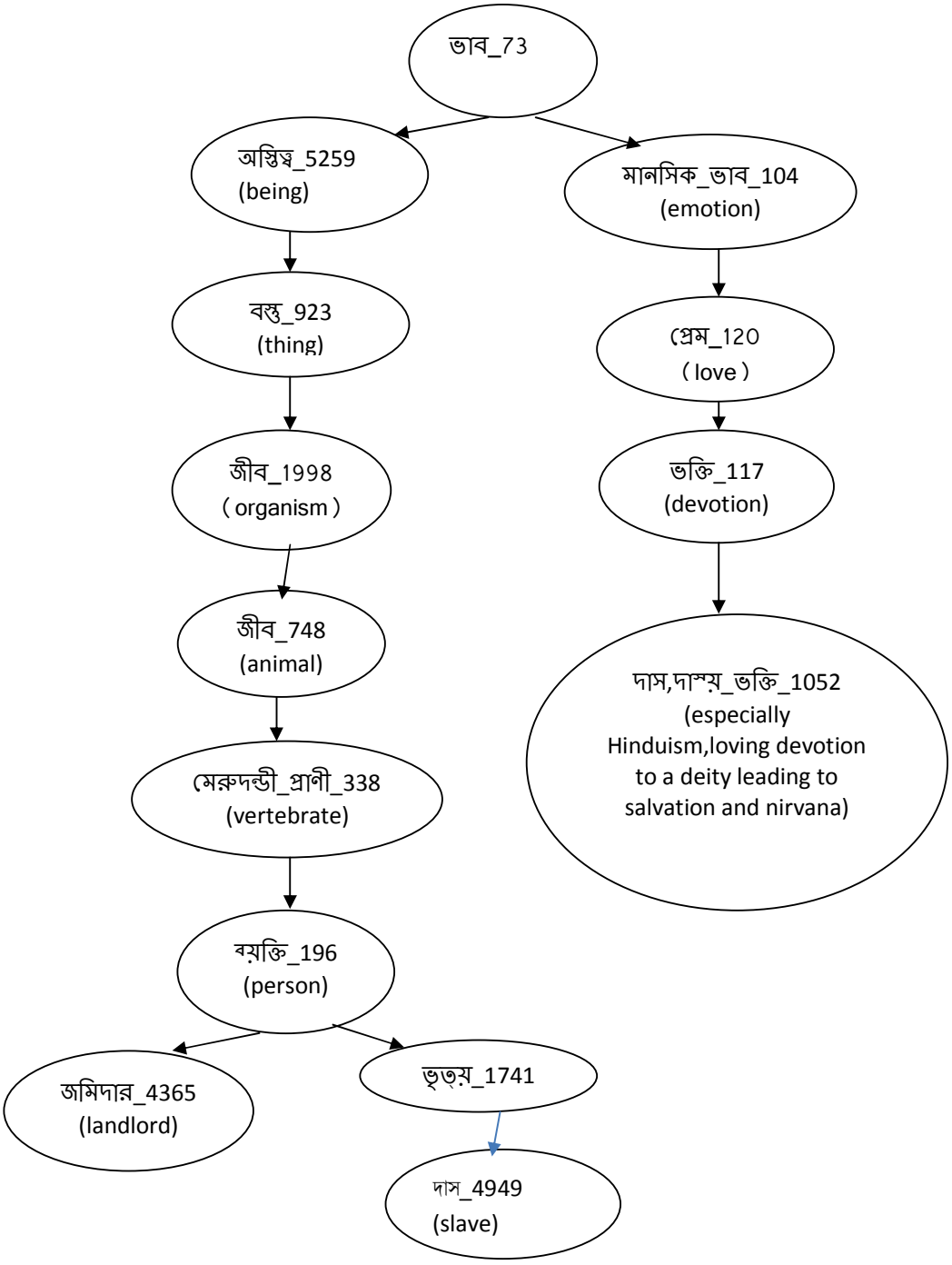

Figure 6.1 Hypernymy-hyponymy graph for contextual distance example

It can be easily seen from Fig 6.1 that the conceptual distance between জমিদার_4365 (landlord) and দাস_4949 (slave) given by (1) is 3/6 = 0.5 whereas the conceptual distance between জমিদার_4365 (landlord) and দাস_1052 given by (1) is 11/0 = ∞ (the common ancestor of জমিদার_4365 and দাস_1052 is ভাব_73 which is the root of the graph and height of the root is 0).

Therefore we can say that the conceptual distance between জমিদার_4365 (landlord) and দাস_4949 (slave) is less than the conceptual distance between জমিদার_4365 (landlord) and দাস_1052. Hence the synset দাস_4949 (slave) should be given a higher rank compared to the synset দাস_1052.

6.1.4 Semantic Graph Distance

It is defined as the shortest path length between any two synset nodes in the WordNet graph. An edge on the shortest path can be any semantic relation in the WordNet (as opposed to conceptual distance where the edges are only hyponymy-hypernymy relations.

Exposition of Semantic Graph Distance with an example :- Let us consider the sentence "কোহিমাতে মোকো চুঙ্গ সার্কিট হাউসে পর্যটকদের থাকার ভালো সুবিধা আছে" (from Gyan Nidhi Corpus of TDIL[6]).

English Gloss of the above example:- Kohima Moko Chung Circuit House tourists stay good facilities available (In the Moko Chung Circuit House of Kohima , good facilities are available for lodging of tourists)

WordNet captures the semantic relation (MODIFIES_NOUN) between the synset {ভালো_32 (having desirable properties)} and {ক্রিয়া_150 (denoting an action or a state) } as well as the semantic relation (HYPONYMY) between the synsets {সুবিধা_28213 (service)} and {ক্রিয়া_150}. This helps us in inferring the relationship between synsets ভালো_32 and সুবিধা_28213. Figure 6.2 shows an example of Semantic Graph Distance.

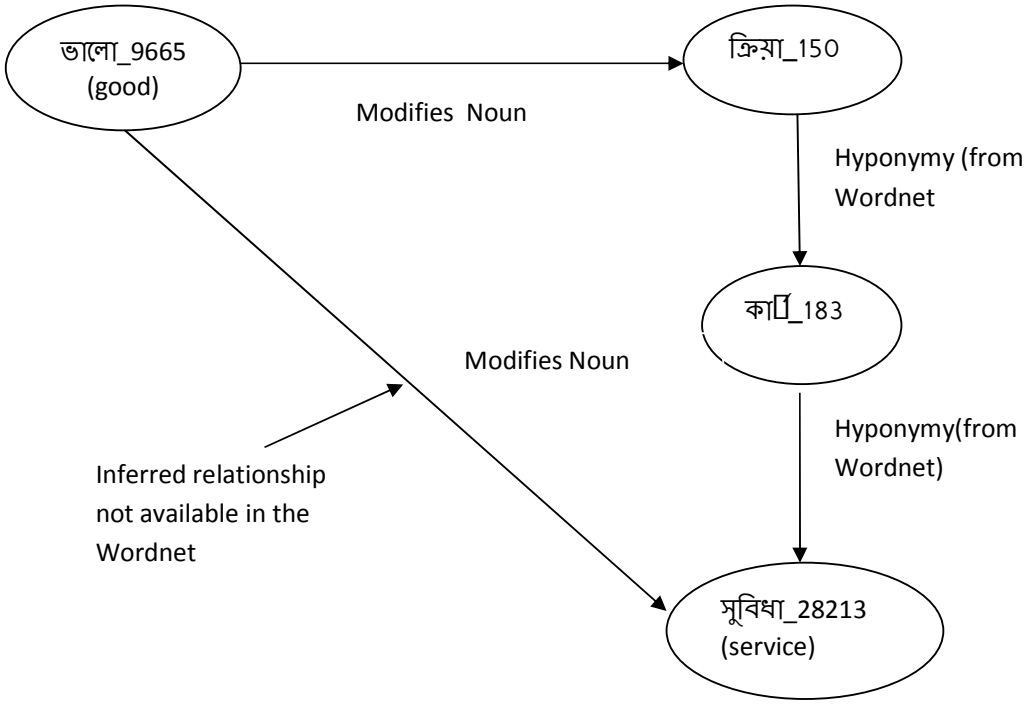

Figure6.2 Example of Semantic Graph Distance

6.2 Sense information from Overlap Approach

As discussed earlier nearby words or context words provide strong and consistent clues as to the sense of a target word and there is a strong tendency for words to exhibit only one sense in a given collocation. This sense may be obtained through the Overlap Approach[14]. The target word disambiguated by a context word can be considered to be in a knowledge_concurrence relationship

For e.g. let us take the following sentence from a parallel Hindi-Bengali Health corpus[6]

ওষুধ পড়ার লক্ষণ মূত্র ত্যাগে অসুবিধে (The symptom of medicine administration is having difficulty in urinating)

दवाई पड़ने होनेवाले लक्षण मूत्र त्याग में परेशानी

If we consider লক্ষণ as the target word , we may form a collocation vector around লক্ষণ which would contain the entries [ওষুধ , পড়ার , মূত্রত্যাগে , অসুবিধে]

If we look up the entry for the synset ওষুধ_3735 in the Bengali WordNet , we find that the entry contains the following Gloss and Example sentence:-

রোগীকে সুস্থ করা অথবা রোগের চিকিৎসা বা রোগ প্রতিরোধকরার জন্য বিধিপূর্বকতৈরী করাও যুধের মিশ্রণ (Drugs prepared in a statutory manner for ameliorating the condition of a patient or for treatment or prevention of a disease)

নিয়মিত ওষুধ সেবন করলেই রোগ সারে (Regular intake of medicine can cure a disease)

The synset লক্ষণ having synset_id 21794 has the following Gloss:-

শারীরিক অবস্হা বাক্রিয়ায় হওয়া সেই পরিবর্তন যাকোনো রোগী ভোগ করে এবং যা কোনো না কোনো রোগের পরিচয় দেয় (any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease)

So from gloss and example sentences above we observe that there is one root word which is common between them and that is রোগ(disease) . So লক্ষণ in the sense রোগের_লক্ষণ should get a higher score.

The idea that a context word or a word in the collocation of a target word may disambiguate a target word has been incorporated in the proposed algorithm Hy_WSD where we have formed collocation of the words around the target word. If there is a sense overlap between words in the collocation and the target word which we call knowledge_concurrence, we give knowledge_concurrence parameter a high value(close to 1) and use it as one of the parameters to calculate the weight between the neurons of the words in the collocation and the neuron of the target word. The weight is eventually used in a scoring function to score the competing senses of a target polysemous word . The knowledge_concurrence parameter may be obtained from the Overlap_App algorithm which is described next.

6.2.1 Implementationof Overlap Approach (Overlap_App)

The Overlapbased approach (Overlap_App) to determine knowledge_concurrence consists of the following steps :-

Preprocessing phase*:*In the pre-processing phase a tokenizer parses the Bengali and Nepali sentence in the corpus into words based on the space between words.

Context Selection: The Overlap_App uses the words of the sentence within a window of size +/- 5 surrounding the target word as context leaving aside function words like conjunctions, articles, prepositions etc . Let this collection be B.

Finding senses of the target word: The Overlap_App finds all the possible senses of target word with the help of the Bengali WordNet ,Nepali Wordnet as the case may be and forms a collection of words from:

- Glosses of the synsets of the collection in B and Gloss of the target word .
- Example sentences of the synsets of the collection in B and the example sentences of the target word
- Glosses of Hypernyms, hyponyms, meronyms, troponym of the collection in B and Gloss of Hypernyms, hyponyms, meronyms, troponym of the target word (upto 3 levels)
- Example Sentences of Hypernyms, hyponyms, meronyms, troponym of the collection in B and example sentences of Hypernyms ,hyponyms, meronyms, troponym of the target word (upto 3 levels)

Let this collection be called as C_i where i is the i^{th} sense.

Determining the Winner Sense:- Find the maximum number of overlapping words in C_i. The value of i for which collectionC_i has the maximum number of overlapping words is the winner sense.The Fig 6.3 shows a snapshot of the implementation of the Overlap_App approach [89].

6.3 A Proposed Hybrid Algorithm for WS (Hy_WSD)

In proposing an algorithm for WSD we have used the energy function of an asynchronous Hopfield network [81] because it lends itself quite gainfully for adaptation to our purpose.

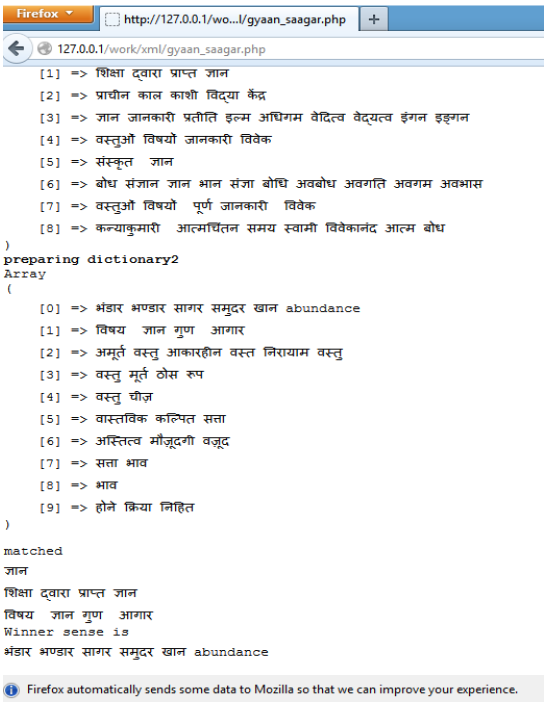

Figure 6.3: A snapshot showing implementation of the Overlap_App

The energy function is E= $-\sum \theta_i S_i + \sum_{i=1}^{N} \sum_{j>1}^{N} W_{ij} * S_i * S_j$ ..... (6.2)

where θ_i is the threshold for neuron i , S_i and S_j are the activations for neurons i and j , W_{ij} is the weight of the connection between neurons i and j , N is the total number of neurons in the network.

The asynchronous Hopfield network is a fully connected bidirectional network of bi-polar neurons (0/1 or 1/-1). At any point of time a randomly chosen neuron examines its input , compares it with the threshold value and attains the value 0/1 or 1/-1 depending on whether the input is greater , less or equal to the threshold value. The collection of 0/1 or 1/-1 states forms the states of the network and each state is associated with a scalar value called the energy E of the network which is given by equation 6.2 above. Energy is a distinctive feature of the Hopfield network providing for convergence , stability etc.

We notice from the energy equation is that there is a clear separation between self activation and the interaction among the neurons. Here's how we adapt equation 6.2 for our purpose. We use a scoring function SCF to be used in place for E to score the neurons. Then we find out that value of i which maximizes the score of the sense of the target word given by the synset Syn_i . The scoring function (SCF) thus becomes

$SCF = (\theta_i S_i + \sum_{j \in J} W_{ij} * S_i * S_j$)..... ...(6.3)

where J is the set of disambiguated words , i is the set of the senses of the target synset Syn_i,θ_i is Belongingness to DominantConcept(S_i), S_i is the corpus frequency of Synset i(neuron i) which is given by P(Syn_i|word) and W_{ij} is the weight between two neurons which is a function of onceptual distance, Semantic Graph Distance and also the Knowledge_ concurrence between two synsets obtained from Overlap_App.

Mathematically W_{ij} is given by

W_{ij}=$1/ConceptualDistance(Syn_i, Syn_j)$*$1/SemanticGraphDistance(Syn_i$,S yn$_j$) * Knowledge_Concurrence(Syn_i, Syn_j)..... (6.4)

Algorithm:-

1. For each POS tag of the content words in the corpus, a training set is constructed.
2. Each training example and test example is represented as a feature vector containing POS of the word and its neighbouring words, local collocations,semantic relations based WSD parameters and knowledge_concurrence.
3. The monosemous words in the sentenceserve as the seed set for disambiguation.
4. Next disambiguate the remaining words in the sentence in ascending order of their polysemy.
5. for each word(synset) in 2 exhibiting polysemy do
   for i = 1 to N // N is the number of multiple senses of the word to be disambiguated
   for j = 1 to J // J is the number of disambiguated words in the sentence
   $SCF_i = (\theta_i S_i + \sum_{j \in J} W_{ij} * S_i * S_j$)
6. Find the value of i which maximizes SCF_i.
7. Select the synset Syn_i as the winner sense.

The Figure 6.4 shows a snapshot of implementation of the proposed algorithm Hy_WSD

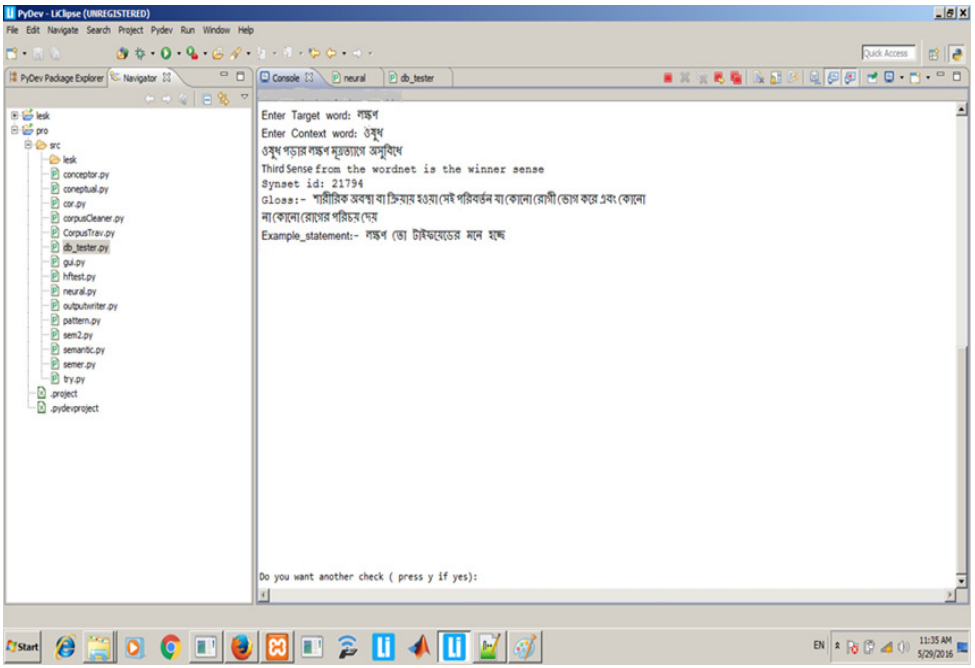

Figure 6.4 Snapshot of implementation of the proposed algorithm

6.4 Experimental Set Up

Experiments were performed on datasets obtained from TDIL[6] and also some part of the text obtained from [6] was manually translated by two Nepali lexicographers well versed with the Nepali language.The datasets belonged to three different domain namely Tourism , Health and Mixed. The languages of the corpus were Bengali and Nepali.

Table 6.2 Total number of Polysemous and monosemous words in 2 domains of Bengali corpus

Category	Tourism(Polysemous Words)	Health(Polysemous Words)	Tourism(Monosemous Words)	Health(Monosemous Words
Noun	37,314	10,655	12958	4992
Verb	11,048	6468	1093	301
Adjective	13,702	4825	5779	1750
Adverb	4279	1211	2485	1278
All	66343	23159	22,315	8321

Table 6.3 Total number of Polysemous and monosemous words in 2 domains of Nepali corpus

Category	Tourism (Polysemous Words)	Health (Polysemous Words)	Tourism (Monosemous Words)	Health (Monosemous Words)
Noun	29757	10473	12115	4135
Verb	9032	4929	808	283
Adjective	11159	4107	5452	1423
Adverb	3721	1438	2321	1017
All	53669	20947	20696	6858

6.5 Experimental Results and Discussion

We use the Hybrid algorithm Hy_WSD proposed above in the following three settings[43]:-

Source setting :- We train the algorithm on one domain (say Mixed) and test it on another domain (say Health) .

Target setting :- We train and test the algorithm on the target domain only. This will give the best performance or the zenith performance.

Customized setting :-We insert some amount of target data into the training set and observe how gradual increase of the size of insertion of target data affects the overall performance . The goal here is to reach the zenith performance.

We report the results in the settings mentioned above using the following four frameworks for source and target data:-

- Mixed to Health where Mixed is used as the source domain and Health is used as the target domain.
- Mixed to Tourism where Mixed is used as the source domain and Tourism is used as the target domain.
- Tourism to Health where Tourism is used as the source domain and Health is used as the target domain.
- Health to Tourism where Health is used as the source domain and Tourism is used as the target domain.

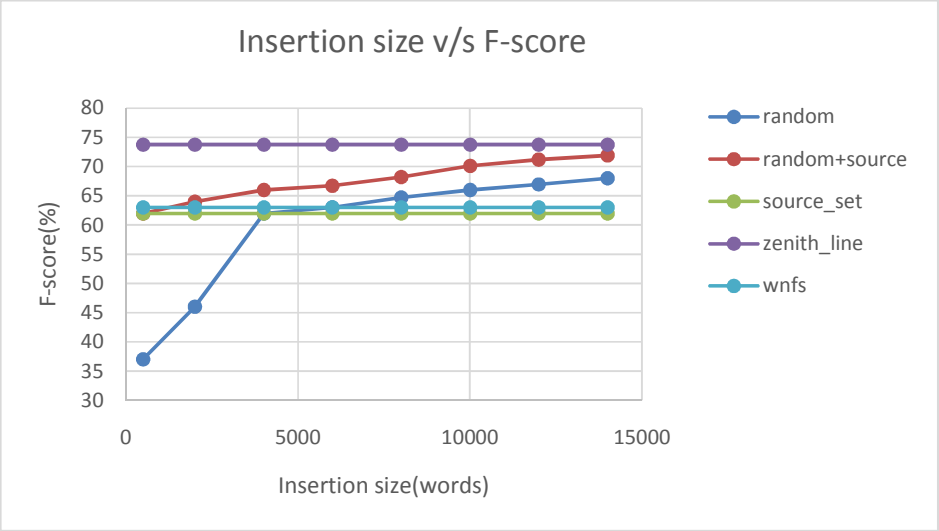

Figure 6.5 Supervised training from Mixed to Tourism(Bengali)

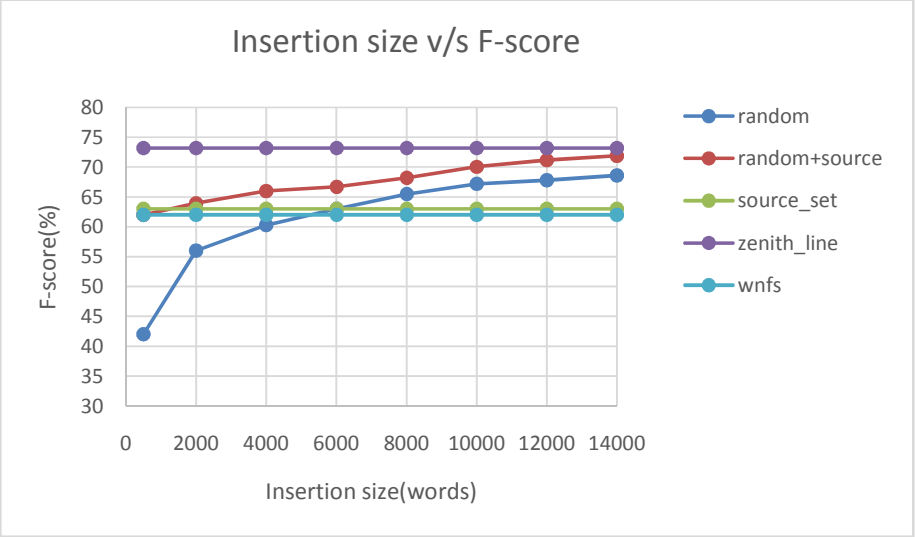

Figure 6.6: Supervised Training from Mixed to Health(Bengali)

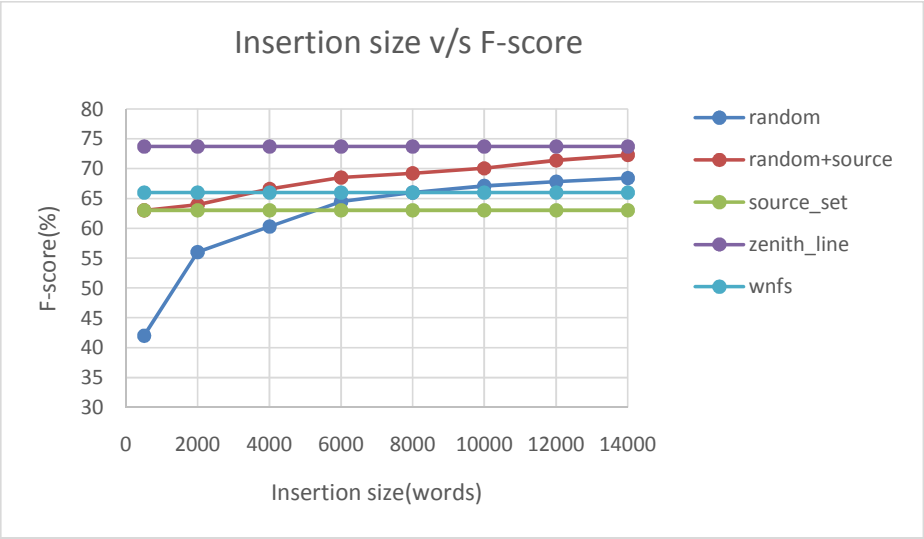

Figure 6.7: Supervised Training from Tourism to Health (Bengali)

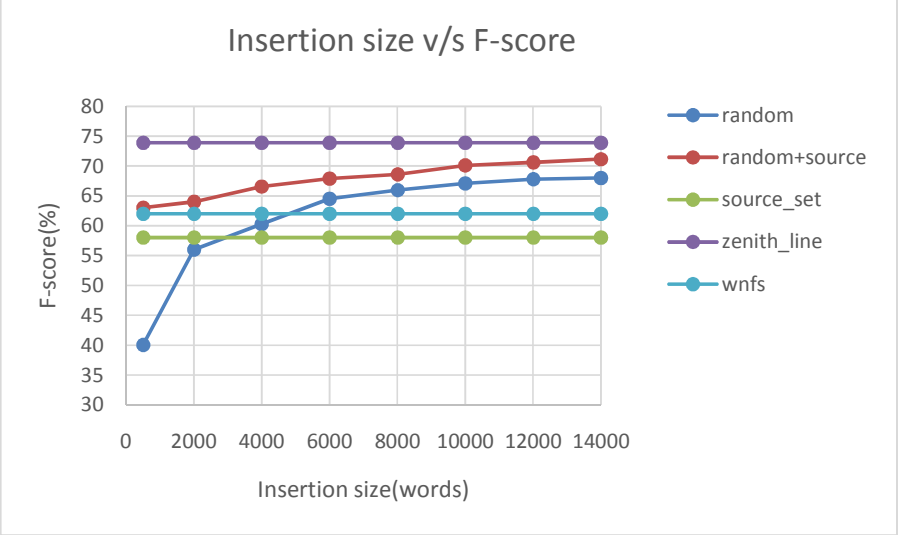

Figure 6.8 : Supervised training from Health to Tourism (Bengali)

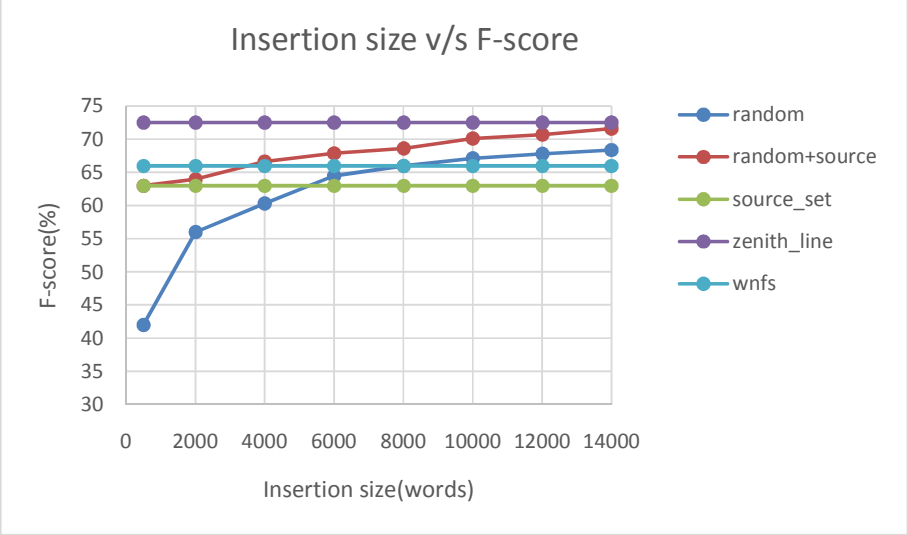

Figure 6.9: Supervised training from Tourism to Health (Nepali)

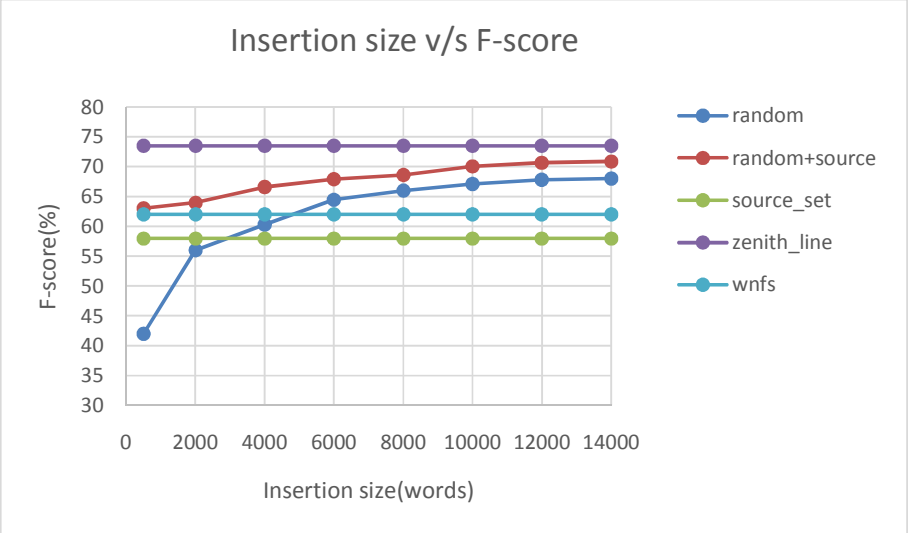

Figure 6.10: Supervised training from Health to Tourism (Nepali)

In each of the figures 6.5-6.10 , there are five plots which signify the following:-

Random:- It means that a random amount of tagged data (x) was taken from the target domain for trainingand x was varied from 1000-14000 words. No sense tagged data was taken from the source domain.

Random+source:- It means that a random x amount of tagged data was taken from the target domain along with the entire data from source domain during training. Again x was varied from 1000-14000 words. For e.g. if training is from Tourism to Health then random + source is random + tourism.

Source_set:- It reports the F-score when training was done on source data and testing was done on target data without mixing any data from the target domain.

Zenith_line :- Partition was made of the target data. Training was done on one partition and testing on the other. It reports the average F-score when training is done on one partition of the target data and testing is done on the other partition. Similarly the partition which was considered for training is used for testing and vice versa.

Wnfs:- This represents the F-score when the first sense from the WordNet is selected. It is a standard baseline.

Table 6.4 Comparing the performance of WordNet first sense(wnfs) baseline , PPR with proposed algorithm(Hy_WSD) on Tourism and Health domain of Bengali corpus

Domain	Algorithm	P%	R%	F%
Tourism	WNFS	64	64	64
	PPR	53.1	53.1	53.1
	Hy_WSD	74	72	73
Health	WNFS	64	64	64
	PPR	51.1	51.1	51.1
	Hy_WSD	73	69.10	71

Table 6.5 : Comparing the performance of WordNet first sense(wnfs) baseline , PPR with proposed algorithm(Hy_WSD) on Tourism and Health domain of Nepali corpus

Domain	Algorithm	P%	R%	F%
Tourism	wnfs	63	63	63
	PPR	53.1	53.1	53.1
	Hy_WSD	71	67	69
Health	wnfs	63	63	63
	PPR	51.1	51.1	51.1
	Hy_WSD	70	66.11	68

From the Table6.4 and Table6.5 it can be seen that in case of both Bengali and Nepali corpus for Tourism and Health Domain , the proposed hybrid algorithm Hy_WSD results in an improved performance over the standard baseline WordNet first sense (Wnfs) and also over the standard knowledge based algorithm known as Personalized Page Rank algorithm (PPR).

6.5 Observations from experiments

Following are the observations from experiments conducted:-

- Infusion of some amount of data from target domain with source domain data generates better performance. As the amount of insertion size increases, the performance approximated to the zenith line performance.
- The performance of Random+source is better than Only_Random indicating that source domain data helps in enhanced performance.
- In the customized setting when training is done on data from a specific source domain (for e.g. Tourism) mixed with some amount of data from another specific domain (for e.g., Health), then the performance noticed is almost the same as when training is done on data from Mixed Domain combined with some amount of data from another specific domain. This is encouraging because if we have data from one domain (for e.g. Dom_A) then a WSD module can be built for another domain (for e.g. Dom_B) by infusion of some amountof data from $Dom_B$with entire data from Dom_A. This means that we would not require data from Mixed domain(by definition contains varied set of words)

which is a costly affair to obtain. Datasets from two specific domains would suffice to give us the same performance level as when Mixed Domain was available.

- The size of the corpus also matters . If we look at the Random+Source plot in each of the figures 6.5-6.10 we find that it plateaus after insertion of 12,000-14,000 words(the X-axis in the figures 6.5-6.10). The size of the tourism and health corpora were bigger in both Bengali and Nepali languages. So the ratio of insertion (words) to the total size of the corpus was less in case of Tourism than in Health. But if the size of the corpus in Tourism and Nepali had been equivalent, then the ratios would also have approximated each other.

6.5 Chapter Summary

This chapter discusses various WSD parameters and also introduces a new parameter knowledge_concurrence to calculate the weight between a sense in a collocation formed around the target word and the target word itself. A neural network is used where the neurons embed the synsets or the senses . The Hybrid algorithm for WSD proposed in this chapter Hy_WSD results in enhanced performance over the standard baseline and the knowledge based algorithm PPR . But of course this comparison is not absolute in the sense that the standard algorithm PPR and standard baseline (Wnfs) were tested on different data sets than the proposed algorithm. We have noticed that Supervised training from specific domain to another specific domain gives almost the same level of performance as supervised training from mixed domain to a specific domain. It is also observed from experiments that some amount of infusion of target data which are sense tagged with source data which are also sense tagged might result in overall improvement of performance.