Value for Money: Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD

Mitesh M. Khapra **Saurabh Sohoney** **Anup Kulkarni** **Pushpak Bhattacharyya**
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay
{miteshk,saurabhsohoney,anup,pb}@cse.iitb.ac.in

Abstract

Sense annotation and lexicon building are costly affairs demanding prudent investment of resources. Recent work on multilingual WSD has shown that it is possible to leverage the annotation work done for WSD of one language (S_L) for another (T_L), by projecting Wordnet and sense marked corpus parameters of S_L to T_L. However, this work does not take into account the cost of manually cross-linking the words within aligned synsets. Further, it does not answer the question of *"Can better accuracy be achieved if a user is willing to pay additional money?"* We propose a measure for ***cost-benefit analysis*** which measures the *"value for money"* earned in terms of accuracy by investing in annotation effort and lexicon building. Two key ideas explored in this paper are (i) the use of ***probabilistic cross-linking model*** to reduce manual cross-linking effort and (ii) the use of ***selective sampling*** to inject a few training examples for hard-to-disambiguate words from the target language to boost the accuracy.

1 Introduction

Word Sense Disambiguation (WSD) is one of the most widely investigated problems of Natural Language Processing (NLP). Previous works have shown that supervised approaches to Word Sense Disambiguation which rely on sense annotated corpora (Ng and Lee, 1996; Lee et al., 2004) outperform unsupervised (Veronis, 2004) and knowledge based approaches (Mihalcea, 2005). However, creation of sense marked corpora has always remained a costly proposition, especially for some of the resource deprived languages.

To circumvent this problem, Khapra et al. (2009) proposed a WSD method that can be applied to a language even when no sense tagged corpus for that language is available. This is achieved by ***projecting Wordnet and corpus parameters*** from another language to the language in question. The approach is centered on a novel synset based multilingual dictionary (Mohanty et al., 2008) where the synsets of different languages are aligned and thereafter the words within the synsets are manually cross-linked. For example, the word W_{L_1} belonging to synset S of language L_1 will be manually cross-linked to the word W_{L_2} of the corresponding synset in language L_2 to indicate that W_{L_2} is the best substitute for W_{L_1} according to an experienced bilingual speaker's intuition.

We extend their work by addressing the following question on the economics of annotation, lexicon building and performance:

- *Is there an optimal point of balance between the annotation effort and the lexicon building (i.e. manual cross-linking) effort at which one can be assured of best value for money in terms of accuracy?*

To address the above question we first propose a probabilistic cross linking model to eliminate the effort of manually cross linking words within the source and target language synsets and calibrate the resultant trade-off in accuracy. Next, we show that by injecting examples for most frequent hard-to-disambiguate words from the target domain one can achieve higher accuracies at optimal

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 555–563,
Beijing, August 2010

cost of annotation. Finally, we propose a measure for ***cost-benefit analysis*** which identifies the optimal point of balance between these three related entities, viz., cross-linking, sense annotation and accuracy of disambiguation.

The remainder of this paper is organized as follows. In section 2 we present related work. In section 3 we describe the Synset based multilingual dictionary which enables parameter projection. In section 4 we discuss the work of Khapra et al. (2009) on parameter projection for multilingual WSD. Section 5 is on the economics of multilingual WSD. In section 6 we propose a probabilistic model for representing the cross-linkage of words within synsets. In section 7 we present a strategy for injecting hard-to-disambiguate cases from the target language using selective sampling. In section 8 we introduce a measure for **cost-benefit analysis** for calculating the value for money in terms of accuracy, annotation effort and lexicon building effort. In section 9 we describe the experimental setup. In section 10 we present the results followed by discussion in section 11. Section 12 concludes the paper.

2 Related Work

Knowledge based approaches to WSD such as Lesk's algorithm (Lesk, 1986), Walker's algorithm (Walker and Amsler, 1986), Conceptual Density (Agirre and Rigau, 1996) and PageRank (Mihalcea, 2005) are less demanding in terms of resources but fail to deliver good results. Supervised approaches like SVM (Lee et al., 2004) and k-NN (Ng and Lee, 1996), on the other hand, give better accuracies, but the requirement of large annotated corpora renders them unsuitable for resource scarce languages.

Recent work by Khapra et al. (2009) has shown that it is possible to project the parameters learnt from the annotation work of one language to another language provided aligned Wordnets for two languages are available. However, their work does not address the question of further improving the accuracy of WSD by using a small amount of training data from the target language. Some similar work has been done in the area of domain adaptation where Chan et al. (2007) showed that adding just 30% of the target data to the source data achieved the same performance as that obtained by taking the entire source and target data. Similarly, Agirre and de Lacalle (2009) reported a 22% error reduction when source and target data were combined for training a classifier, compared to the case when only the target data was used for training the classifier. However, such combining of training statistics has not been tried in cases where the source data is in one language and the target data is in another language.

To the best of our knowledge, no previous work has attempted to perform resource conscious **all-words multilingual Word Sense Disambiguation** by finding a trade-off between the cost (in terms of annotation effort and lexicon creation effort) and the quality in terms of F-score.

3 Synset based multilingual dictionary

A novel and effective method of storage and use of dictionary in a multilingual setting was proposed by Mohanty et al. (2008). For the purpose of current discussion, we will refer to this multilingual dictionary framework as *MultiDict*. One important departure in this framework from the traditional dictionary is that **synsets are linked, and after that the words inside the synsets are linked**. The basic mapping is thus between synsets and thereafter between the words.

Concepts	L1 (English)	L2 (Hindi)	L3 (Marathi)
04321: a youthful male person	{malechild, boy}	{लड़का *(ladkaa)*, बालक *(baalak)*, बच्चा *(bachchaa)*}	{मुलगा *(mulgaa)*, पोरगा *(por-gaa)*, पोर *(por)*}

Table 1: Multilingual Dictionary Framework

Table 1 shows the structure of MultiDict, with one example row standing for the concept of *boy*. The first column is the pivot describing a concept with a unique ID. The subsequent columns show the words expressing the concept in respective languages (in the example table, *English, Hindi and Marathi*). After the synsets are linked, cross linkages are set up manually from the words of a synset to the words of a linked synset of the pivot language. For example, for the Marathi word मुलगा *(mulgaa)*, "a youthful male person", the

correct lexical substitute from the corresponding Hindi synset is *लड़का (ladkaa)*. The average number of such links per synset per language pair is approximately 3.

4 Parameter Projection

Khapra et al. (2009) proposed that the various parameters essential for domain-specific Word Sense Disambiguation can be broadly classified into two categories:

Wordnet-dependent parameters:

- belongingness-to-dominant-concept
- conceptual distance
- semantic distance

Corpus-dependent parameters:

- sense distributions
- corpus co-occurrence

They proposed a scoring function (Equation (1)) which combines these parameters to identify the correct sense of a word in a context:

$$S^* = \arg\max_i(\theta_i V_i + \sum_{j \in J} W_{ij} * V_i * V_j) \qquad (1)$$

where,

$$\begin{aligned} i &\in \textit{Candidate Synsets} \\ J &= \textit{Set of disambiguated words} \\ \theta_i &= BelongingnessToDominantConcept(S_i) \\ V_i &= P(S_i|word) \\ W_{ij} &= CorpusCooccurrence(S_i, S_j) \\ &\quad * 1/WNConceptualDistance(S_i, S_j) \\ &\quad * 1/WNSemanticGraphDistance(S_i, S_j) \end{aligned}$$

The first component $\theta_i V_i$ of Equation (1) captures influence of the corpus specific sense of a word in a domain. The other component $W_{ij} * V_i * V_j$ captures the influence of interaction of the candidate sense with the senses of context words weighted by factors of co-occurrence, conceptual distance and semantic distance.

Wordnet-dependent parameters depend on the structure of the Wordnet whereas the ***Corpus-dependent parameters*** depend on various statistics learnt from a sense marked corpora. Both the tasks of (a) constructing a Wordnet from scratch and (b) collecting sense marked corpora for multiple languages are tedious and expensive. Khapra et al. (2009) observed that by ***projecting relations*** from the Wordnet of a language and by ***projecting corpus statistics*** from the sense marked corpora of the language to those of the target language, *the effort required in constructing semantic graphs for multiple Wordnets and collecting sense marked corpora for multiple languages can be avoided or reduced.* At the heart of their work lies the *MultiDict* described in previous section which facilitates parameter projection in the following manner:

1. By linking with the synsets of a pivot resource rich language (Hindi, in our case), the cost of building Wordnets of other languages is partly reduced (semantic relations are inherited). The Wordnet parameters of Hindi Wordnet now become projectable to other languages.

2. For calculating corpus specific sense distributions, $P(Sense\ S_i|Word\ W)$, we need the counts, $\#(S_i, W)$. By using cross linked words in the synsets, these counts become projectable to the target language (Marathi, in our case) as they can be approximated by the counts of the cross linked Hindi words calculated from the Hindi sense marked corpus as follows:

$$P(S_i|W) = \frac{\#(S_i, marathi\_word)}{\sum_j \#(S_j, marathi\_word)}$$

$$P(S_i|W) \approx \frac{\#(S_i, cross\_linked\_hindi\_word)}{\sum_j \#(S_j, cross\_linked\_hindi\_word)}$$

The rationale behind the above approximation is the observation that within a domain sense distributions remain the same across languages.

5 The Economics of Multilingual WSD

The problem of multilingual WSD using parameter projection can be viewed as an economic system consisting of three factors. The first factor is the cost of manually cross-linking the words in a synsets of the target language to the words in the corresponding synset in the pivot language. The second factor is the cost of sense annotated data from the target language. The third factor is the accuracy of WSD The first two factors in some

sense relate to the cost of purchasing a commodity and the third factor relates to the commodity itself.

The work of Khapra et al. (2009) as described above does not attempt to reach an optimal cost-benefit point in this economic system. They place their bets on manual cross-linking only and settle for the accuracy achieved thereof. Specifically, they do not explore the inclusion of small amount of annotated data from the target language to boost the accuracy (as mentioned earlier, supervised systems which use annotated data from the target language are known to perform better). Further, it is conceivable that with respect to accuracy-cost trade-off, there obtains a case for *balancing* one cost against the other, *viz.*, the cost of cross-linking and the cost of annotation. In some cases bilingual lexicographers (needed for manual cross-linking) may be more expensive compared to monolingual annotators. There it makes sense to place fewer bets on manual cross-linking and more on collecting annotated corpora. On the other hand if manual cross-linking is cheap then a very small amount of annotated corpora can be used in conjunction with full manual cross-linking to boost the accuracy. Based on the above discussion, if k_a is the cost of sense annotating one word, k_c is the cost of manually cross-linking a word and A is the accuracy desired then the problem of multilingual WSD can be cast as an optimization problem:

$$\begin{aligned} minimize \quad & w_a * k_a + w_c * k_c \\ s.t. \quad & \\ & Accuracy \geq A \end{aligned}$$

where, w_c and w_a are the number of words to be manually cross linked and annotated respectively. Ours is thus a 3-factor economic model (cross-linking, annotation and accuracy) as opposed to the 2-factor model (cross-linking, accuracy) proposed by Khapra et al. (2009).

6 Optimal cross-linking

As mentioned earlier, in some cases where bilingual lexicographers are expensive we might be interested in reducing the effort of manual cross-linking. For such situations, we propose that only a small number of words, comprising of the most frequently appearing ones should be manually cross linked and the rest of the words should be cross-linked using a probabilistic model. The rationale here is simple: invest money in words which are bound to occur frequently in the test data and achieve maximum impact on the accuracy. In the following paragraphs, we explain our probabilistic cross linking model.

The model proposed by Khapra et al. (2009) is a deterministic model where the expected count for (Sense S, Marathi_Word W), *i.e.*, the number of times the word W appears in sense S is approximated by the count for the corresponding cross linked Hindi word. Such a model assumes that each Marathi word links to appropriate Hindi word(s) as identified manually by a lexicographer. Instead, **we propose a probabilistic model where a Marathi word can link to every word in the corresponding Hindi synset with some probability**. The expected count for (S, W) can then be estimated as:

$$E[\#(S,W)] = \sum_{h_i \in cross\_links} P(h_i|W,S) * \#(S,h_i) \quad (2)$$

where, $P(h_i|W,S)$ is the probability that the word h_i from the corresponding Hindi synset is the correct cross-linked word for the given Marathi word. For example, one of the senses of the Marathi word *maan* is {neck} *i.e.* *"the body part which connects the head to the rest of the body"*. The corresponding Hindi synset has 10 words {*gardan, gala, greeva, halak, kandhar and so on*}. Thus, using Equation (2), the expected count, $E[C(\{neck\}, maan)]$, is calculated as:

$$\begin{aligned} & E[\#(\{neck\}, maan)] = \\ & \quad P(gardan|maan, \{neck\}) * \#(\{neck\}, gardan) \\ & + P(gala|maan, \{neck\}) * \#(\{neck\}, gala) \\ & + P(greeva|maan, \{neck\}) * \#(\{neck\}, greeva) \\ & + \ldots \; so\ on\ for\ all\ words\ in\ the\ Hindi\ synset \end{aligned}$$

Instead of using a uniform probability distribution over the Hindi words we go by the empirical observation that some words in a synset are more representative of that sense than other words, *i.e. some words are more preferred while expressing that sense*. For example, out of the 10 words in

the Hindi synset only 2 words {*gardan, gala*} appeared in the corpus. We thus estimate the value of $P(h_i|W,S)$ empirically from the Hindi sense marked corpus by making the following independence assumption:

$$P(h_i|W,S) = P(h_i|S)$$

The rationale behind the above independence assumption becomes clear if we represent words and synsets using the Bayesian network of Figure 1. Here, the Hindi word h_i and the Marathi word W

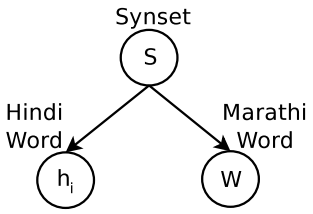

Figure 1: Bayesian network formed by a synset S and the constituent Hindi and Marathi words

are considered to be derived from the same parent concept S. In other words, they represent two different manifestations- one in Hindi and one in Marathi- of the same synset S. Given the above representation, it is easy to see that given the parent synset S, the Hindi word h_i is independent of the Marathi word W.

7 Optimal annotation using Selective Sampling

In the previous section we dealt with the question of optimal cross-linking. Now we take up the other dimension of this economic system, *viz.*, optimal use of annotated corpora for better accuracy. In other words, if an application demands higher accuracy for WSD and is willing to pay for some annotation then there should be a way of ensuring best possible accuracy at lowest possible cost. This can be done by including small amount of sense annotated data from the target language. The simplest strategy is to randomly annotate text from the target language and use it as training data. However, this strategy of random sampling may not be the most optimum in terms of cost. Instead, we propose a selective sampling strategy where the aim is to identify ***hard-to-disambiguate*** words from the target language and use them for training.

The algorithm proceeds as follows:

1. First, using the probabilistic cross linking model and aligned Wordnets we learn the parameters described in Section 4.

2. We then apply this scoring function on untagged examples (development set) from the target language and identify ***hard-to-disambiguate*** words *i.e.,* the words which were disambiguated with a very low confidence.

3. Training instances of these words are then injected into the training data and the parameters learnt from them are used instead of the projected parameters learnt from the source language corpus.

Thus, the selective sampling strategy ensures that we get maximum value for money by spending it on annotating only those words which would otherwise not have been disambiguated correctly. A random selection strategy, in contrast, might bring in words which were disambiguated correctly using only the projected parameters.

8 A measure for cost-benefit analysis

We need a measure for cost-benefit analysis based on the three dimensions of our economic system, *viz.*, annotation effort, lexicon creation effort and performance in terms of F-score. The first two dimensions can be fused into a single dimension by expressing the annotation effort and lexicon creation effort in terms of cost incurred. For example, we assume that the cost of annotating one word is k_a and the cost of cross-linking one word is k_c rupees. Further, we define a baseline and an upper bound for the F-score. In this case, the baseline would be the accuracy that can be obtained without spending any money on cross-linking and annotation in the target language. An upper bound could be the best F-score obtained using a large amount of annotated corpus in the target domain. Based on the above description, an ideal measure for cost-benefit analysis would assign a

1. reward depending on the improvement over the baseline performance.

2. penalty depending on the difference from the upper bound on performance.

3. reward inversely proportional to the cost in-

curred in terms of annotation effort and/or manual cross-linking.

Based on the above wish-list we propose a measure for cost-benefit analysis. Let,

$$MGB = Marginal\ Gain\ over\ Baseline\ (MGB)$$
$$= \frac{Performance(P) - Baseline(B)}{Cost(C)}$$
$$MDU = Marginal\ Drop\ from\ Upperbound\ (MDU)$$
$$= \frac{UpperBound(U) - Performance(P)}{Cost(C)}$$

then

$$CostBenefit(CB) = MGB - MDU$$

9 Experimental Setup

We used Hindi as the source language (S_L) and trained a WSD engine using Hindi sense tagged corpus. The parameters thus learnt were then projected using the *MultiDict* (refer section 3 and 4) to build a resource conscious Marathi (T_L) WSD engine. We used the same dataset as described in Khapra et al. (2009) for all our experiments. The data was collected from two domains, *viz.*, Tourism and Health. The data for Tourism domain was collected by manually translating English documents downloaded from Indian Tourism websites into Hindi and Marathi. Similarly, English documents for Health domain were obtained from two doctors and were manually translated into Hindi and Marathi. The Hindi and Marathi documents thus created were manually sense annotated by two lexicographers adept in Hindi and Marathi using the respective Wordnets as sense repositories. Table 2 summarizes some statistics about the corpora.

As for cross-linking, Hindi is used as the pivot language and words in Marathi synset are linked to the words in the corresponding Hindi synset. The total number of cross-links that were manually setup were 3600 for Tourism and 1800 for Health. The cost of cross-linking as well as sense annotating one word was taken to be 10 rupees. These costs were estimated based on quotations from lexicographers. However, these costs need to be taken as representative values only and may vary greatly depending on the availability of skilled bilingual lexicographers and skilled monolingual annotators.

Language	#of polysemous words		average degree of polysemy	
	Tourism	Health	Tourism	Health
Hindi	56845	30594	3.69	3.59
Marathi	34156	10337	3.41	3.60

Table 2: Number of polysemous words and average degree of polysemy.

10 Results

Tables 3 and 4 report the average 4-fold performance on Marathi Tourism and Health data using different proportions of available resources, *i.e.*, annotated corpora and manual cross-links. In each of these tables, along the rows, we increase the amount of Marathi sense annotated corpora from 0K to 6K. Similarly, along the columns we show the increase in the number of manual cross links (MCL) used. For example, the second column of Tables 3 and 4 reports the F-scores when probabilistic cross-linking (PCL) was used for all words (*i.e.*, no manual cross-links) and varying amounts of sense annotated corpora from Marathi were used. Similarly, the first row represents the case in which no sense annotated corpus from Marathi was used and varying amounts of manual cross-links were used.

We report three values in the tables, *viz.*, F-score (F), cost in terms of money (C) and the cost-benefit (CB) obtained by using x amount of annotated corpus and y amount of manual cross-links. The cost was estimated using the values given in section 9 (*i.e.*, 10 rupees for cross-linking or sense annotating one word). For calculating, the cost-benefit baseline was taken as the F-score obtained by using no cross-links and no annotated corpora *i.e.* 68.21% for Tourism and 67.28% for Health (see first F-score cell of Tables 3 and 4). Similarly the upper bound (F-scores obtained by training on entire Marathi sense marked corpus) for Tourism and Health were 83.16% and 80.67% respectively (see last row of Table 5).

Due to unavailability of large amount of tagged Health corpus, the injection size was varied from 0-to-4K only. In the other dimension, we varied the cross-links from 0 to 1/3rd to 2/3rd to full only

Selective Sampling	Only PCL			1/3 MCL			2/3 MCL			Full MCL		
	F	C	CB	F	C	CB	F	C	CB	F	C	CB
0K	68.21	0	-	72.08	12	-0.601	73.31	24	-0.198	73.34	36	-0.130
1K	71.18	10	-0.901	74.96	22	-0.066	77.58	34	0.111	77.73	46	0.089
2K	74.35	20	-0.134	76.96	32	0.080	**78.57**	**44**	**0.131**	79.23	56	0.127
3K	75.21	30	-0.032	77.78	42	0.100	78.68	54	0.111	79.8	66	0.125
4K	76.40	40	0.036	78.66	52	0.114	79.18	64	0.110	80.36	76	0.123
5K	77.04	50	0.054	78.51	62	0.091	79.60	74	0.106	80.46	86	0.111
6K	78.58	60	0.097	79.75	72	0.113	80.8	84	0.122	80.44	96	0.099

Table 3: F-Score (F) in %, Cost (C) in thousand rupees and Cost Benefit (CB) values using different amounts of sense annotated corpora and manual cross links in Tourism domain.

Selective Sampling	Only PCL			1/3 MCL			2/3 MCL			Full MCL		
	F	C	CB	F	C	CB	F	C	CB	F	C	CB
0K	67.28	0	-	71.39	6	-0.862	73.06	12	-0.153	73.34	18	-0.071
1K	72.51	10	-0.293	75.57	16	0.199	**77.41**	**22**	**0.312**	78.16	28	0.299
2K	75.64	20	0.167	77.29	26	0.255	78.13	32	0.260	78.63	38	0.245
3K	76.78	30	0.187	79.35	36	0.299	79.79	42	0.277	79.88	48	0.246
4K	77.42	40	0.172	79.59	46	0.244	80.54	52	0.253	80.15	58	0.213

Table 4: F-Score (F) in %, Cost (C) in thousand rupees and Cost Benefit (CB) values using different amounts of sense annotated corpora and manual cross links in Health domain.

Strategy	Tourism	Health
WFS	57.86	52.77
Only PCL	68.21	67.28
1/6 MCL	69.95	69.57
2/6 MCL	72.08	71.39
3/6 MCL	72.97	72.61
4/6 MCL	73.39	73.06
5/6 MCL	73.55	73.27
Full MCL	73.62	73.34
Upper Bound	83.16	80.67

Table 5: F-score (in %) obtained by using different amounts of manually cross linked words

Strategy	Size of target side annotated corpus						
	0K	1K	2K	3K	4K	5K	6K
Random + PCL	68.21	70.62	71.79	73.03	73.61	76.42	77.52
Random + MCL	73.34	75.32	75.89	76.79	76.83	78.91	80.87
Selective Sampling + PCL	68.21	71.18	74.35	75.21	76.40	77.04	78.58
Selective Sampling + MCL	73.34	77.73	79.23	79.8	79.8	80.46	80.44

Table 6: Comparing F-scores obtained using random sampling and selective sampling (Tourism)

Strategy	Size of target side annotated corpus						
	0K	1K	2K	3K	4K	5K	6K
Annotation + PCL	68.21	71.20	74.35	75.21	76.40	77.04	78.58
Only Annotation	57.86	62.32	64.84	66.86	68.89	69.64	71.82

Table 7: Comparing F-scores obtained using Only Annotation and Annotation + PCL(Tourism)

(refer to Tables 3 and 4). However, to give an idea about the soundness of probabilistic cross-linking we performed a separate set of experiments by varying the number of cross-links and using no sense annotated corpora. Table 5 summarizes these results and compares them with the baseline (Wordnet first sense) and skyline.

In Table 6 we compare our selective sampling strategy with random sampling when fully probabilistic cross-linking (PCL) is used and when fully manual cross-linking (MCL) is used. Here again, due to lack of space we report results only on Tourism domain. However, we would like to mention that similar experiments on Health domain showed that the results were indeed consistent.

Finally, in Table 7 we compare the accuracies obtained when certain amount of annotated corpus from Marathi is used alone, with the case when the same amount of annotated corpus is used in conjunction with probabilistic cross-linking. While calculating the results for the second row in Table 7, we found that the recall was very low due to the small size of injections. Hence, to ensure a fair comparison with our strategy (first row) we used the Wordnet first sense (WFS) for these recall errors (a typical practice in WSD literature).

11 Discussions

We make the following observations:

1. PCL v/s MCL: Table 5 shows that the probabilistic cross-linking model performs much better than the WFS (a typically reported baseline) and it comes very close to the performance of manual cross-linking. This establishes the soundness of the probabilistic model and suggests that with a little compromise in the accuracy, the model can be used as an approximation to save the cost of manual cross-linking. Further, in Table 7 we see that when PCL is used in conjunction with certain amount of annotated corpus we get up to 9% improvement in F-score as compared to the case when the same amount of annotated corpus is used alone. Thus, in the absence of skilled bilingual lexicographers, PCL can still be used to boost the accuracy obtained using annotated corpora.

2. Selective Sampling v/s Random Annotation: Table 6 shows the benefit of selective sampling over random annotation. This benefit is felt more when the amount of training data injected from Marathi is small. For example, when an annotated corpus of size 2K is used, selective sampling gives an advantage of 3% to 4% over random selection. Thus the marginal gain (*i.e.*, value for money) obtained by using selective sampling is more than that obtained by using random annotation.

3. Optimal cost-benefit: Finally, we address the main message of our work, *i.e.*, finding the best cost benefit. By referring to Tables 3 and 4, we see that the best value for money in Tourism domain is obtained by manually cross-linking 2/3rd of all corpus words and sense annotating 2K target words and in the Health domain it is obtained by manually cross-linking 2/3rd of all corpus words but sense annotating only 1K words. This suggests that striking a balance between cross-linking and annotation gives the best value for money. Further, we would like to highlight that our 3-factor economic model is able to capture these relations better than the 2-factor model of Khapra et al. (2010). As per their model the best F-score achieved using manual cross-linking for ALL words was 73.34% for both Tourism and Health domain at a cost of 36K and 18K respectively. On the other hand, using our model we obtain higher accuracies of 76.96% in the Tourism domain (using 1/3rd manual cross-links and 2K injection) at a lower total cost (32K rupees) and 75.57% in the Health domain (using only 1/3rd cross-linking and 1K injection) at a lower cost (16K rupees).

12 Conclusion

We reported experiments on multilingual WSD using different amounts of annotated corpora and manual cross-links. We showed that there exists some trade-off between the accuracy and *balancing* the cost of annotation and lexicon creation. In the absence of skilled bilingual lexicographers one can use a probabilistic cross-linking model and still obtain good accuracies. Also, while sense annotating a corpus, careful selection of words using selective sampling can give better marginal gain as compared to random sampling.

References

Agirre, Eneko and Oier Lopez de Lacalle. 2009. Supervised domain adaption for wsd. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 42–50, Morristown, NJ, USA. Association for Computational Linguistics.

Agirre, Eneko and German Rigau. 1996. Word sense disambiguation using conceptual density. In *In Proceedings of the 16th International Conference on Computational Linguistics (COLING)*.

Chan, Y.S., H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *In Proc. of ACL*.

Khapra, Mitesh M., Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 459–467, Singapore, August. Association for Computational Linguistics.

Khapra, Mitesh, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *5th International Conference on Global Wordnet (GWC2010)*.

Lee, Yoong Keok, Hwee Tou Ng, and Tee Kiah Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140.

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *In Proceedings of the 5th annual international conference on Systems documentation*.

Mihalcea, Rada. 2005. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *In Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP)*, pages 411–418.

Mohanty, Rajat, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma. 2008. Synset based multilingual dictionary: Insights, applications and challenges. In *Global Wordnet Conference*.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–47.

Veronis, Jean. 2004. Hyperlex: Lexical cartography for information retrieval. In *Computer Speech and Language*, pages 18(3):223–252.

Walker, D. and R. Amsler. 1986. The use of machine readable dictionaries in sublanguage analysis. In *In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press*, pages 69–83.