Using Levins verb classification for preposition sense selection in English to Indian language MT

by

Samar Husain, Phani Chaitanya, Ganeshwar Rao Dulam, Tariq Khan, Dipti M Sharma

in

In Proceedings of the Conference on Language and Technology 2009 (CLT09), Lahore, Pakistan. 2009.

Report No: IIIT/TR/2009/3

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2009

Using Levin's verb classification for preposition sense selection in English to Indian language MT

Samar Husain, Phani Chaitanya, Ganeshwar Rao Dulam, Tariq Khan*, and Dipti M. Sharma
Language Technologies Research Centre, IIIT, Hyderabad, India.
samar@research.iiit.ac.in, {srpchaitanya, grdulam}@students.iiit.ac.in, safertariq@gmail.com, dipti@mail.iiit.ac.in

Abstract

The paper describes an initial attempt to use Levin's verb classes for the task of preposition sense selection in English to Indian language machine translation. Two language pairs have been selected to describe the approach, English-Hindi and English-Telugu. We exploit the correspondence of verb class's semantics vis-à-vis some prepositions it takes. Results obtained from the experiments show very high performance with an overall average accuracy of 93.6% for English-Telugu language pair. When combined with previous approach, the performance increased by 6.2%.

1. Introduction

It is well known that EFL learners find it hard to grasp the minute semantic intricacies which come into play while using prepositions. One can perhaps generalize this problem for other foreign language learners also [2]. It should then not come as a surprise that in automatic machine translation too, the task of translating a preposition from one language to another is difficult. One can treat automatic translation of preposition as a sense selection task. Framing this task in such a way makes it very different from the problem of word sense disambiguation. As many source language adposition senses generally collapse into a single target adposition, one does not need to always distinguish between the various senses. What is required then is correct source to target correspondence. Additionally, the pervasive use of adpositions in many languages makes it a crucial element during translation. It is also known that the elements that make up the lexical class of preposition are highly polysemous. Thus, inappropriate sense selection of an adposition during machine translation can have a negative impact on the quality of the translation, sometimes changing the semantics of the translated sentence drastically. This makes preposition sense selection (PSS) module a critical component of any reliable MT system.

Extensive body of work devoted to understanding English prepositions linguistically exists in the literature. Syntactically [11], [3], [26], and [25]; semantically by [27] and [28]; and the Pragmatic aspects by [4]. Prepositions hold a very important status in Cognitive linguistics, and have been studied in great detail both theoretically [16], [17], [15], [2] and from an MT perspective [32].

The work of automatically selecting the correct sense has also received considerable attention. [13] attempts to translate locative prepositions between English and French. It introduces the notion of 'representation of conceptualization' based in turn on [5]. The paper synthesizes this idea with the thesis of ideal meaning [8]. [30] has tried to resolve conceptual geographical prepositions using inference rule based on cognitive maps which people have of the external world. [7] uses knowledge representation formalism for preposition phrase (PP) interpretation.

Some studies pertain to systems which have been implemented for MT; [6] uses aligned parallel corpora to induce automatic rules by applying transformation-based learning. [1] makes use of contextual information to determine the meanings of 'over'. [31] uses a transfer rule based approach to translate locative PP-phrase, the approach uses the dependency relations marked as indices with individual word and a bilingual lexicon which has mapping between source and target lexical item (with indices). [24] looks at the semantics of the head noun of the reference object to get the lexical meaning of prepositions in an English-Bengali MT system. While [10] uses multiple contextual features of the parent phrase head of the PP and the PP

* CALTS, Hyderabad Central University, Hyderabad, India

noun complement for English-Hindi and English-Telugu MT.

In this paper we show how Levin's verb classes [18] can be used to select the appropriate sense of a preposition in the target language with very high accuracy. The method exploits the semantics of the verb classes and the kind of preposition they can potentially take. The approach suggests that one can without getting the semantics of the noun, the verb, or the preposition get a very robust system based only on the Levin's verb classification. Using various resources like WordNet [23], NomBank[1], PrepNet [28], etc. becomes unnecessary. Of course, the classification itself does implicitly consider the semantics of the verb, since the verb, owing to its semantics, subcategorizes for only certain arguments; but from a computational point of view one uses a single lexicon with no explicit semantic information. The final system will, apart from the verb class information, also need certain other syntactic information such as PP attachment and correct identification of the phrasal verb (if present). The proposed approach presumes that this information will be available. In Section 5 we describe how we go about getting such information. The paper does not claim to solve the problem of PSS in toto, on the contrary, the work proposed is meant to complement other methods.

The paper is arranged as follows; Section 2 briefly describes Levin's verb classification. In Section 3 we explain our approach. Section 4 lists some verb classes which lend themselves to the approach. We show the evaluation results in Section 5. Section 6 discusses the results and makes some general remarks on related issues. Finally, we conclude the paper in Section 7.

2. Levin's verb classification

Levin's verb classification is an elaborate attempt to investigate English verbs. The work is guided by the assumption that the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning [18]. In effect what this means is that the syntactic behavior of a verb is semantically defined. Drawing from earlier works dedicated to such an investigation, Levin has shown the correlations between the semantic and syntactic behavior of English verbs. The classification task makes use of wide variety of diathesis alternations characteristic of English verbs. A semantic class finally obtained adheres to or does not adhere to a set of alternations. We'll see in Section 3 that for our task too we make use of certain alternations as we go about finding a verb class suitable for PSS.

3. Approach

Verbs are generally thought to be at the core of a sentence in almost all the languages. Various linguistic theories have used this insight in varying ways to account for the syntactic and the semantic structure of a sentence. Since verb classes can be identified throughout language and are asserted to exist across languages as their basic meaning components can be applied cross-linguistically [12], one can therefore expect to see a possible adpositional correspondence between some language pairs as well. On the other hand, different conceptualization of same real-world situation lead to different selection of linguistic devices (such as adpositions) in different languages [32]; however what we are interested in is the correspondence between the adpositions in a language pair and not the difference in the conceptualization system. Such an investigation is out of the scope of this paper.

Certain verbs due to their intrinsic semantic properties constrain the type of elements which appear on the surface structure. This insight lies at the heart of the Levin's verb classes. The proposed scheme tries to exploit this characteristic for the PSS task. What is explored as part of this paper is the following:

(a) *Can we find a class of verbs which owing to its semantics constrains the kind of prepositions it takes?*

(b) *For such prepositions, is there a one-to-one correspondence on the target side?*

(c) *Does the chosen target adposition have a non-default/non-canonical usage?*

What this means for sense selection is that having found such a verb class every other potential parameter becomes invariant. All one needs is this verb class. Instead of looking at the prepositions directly, we discover them via the verb classes. We reject a verb class for which the target adposition takes a default sense; the selection of default sense would already be part of the baseline approach and can safely be left out.

There is no straightforward way to automatically select such verb classes. For the purpose of this paper we have tried to manually find some such classes

[1] http://nlp.cs.nyu.edu/meyers/NomBank.html

which satisfy the above three conditions. We search for such classes using some specific alternations. These alternations generally involve the participation of prepositions. Some of these are:

- *Prepositions drop alternation,*
- *With/against alternation,*
- *Through/with alternation,*
- *Locative inversion, etc.*

Apart from these alternations, the 'Comments' at the end of each verb class proved to be very useful. Note here that we could have possibly used VerbNet[2] for identifying the verb classes as well, but we opted for Levin's verb classification owing to its descriptive richness in the form of alternations and 'comments'. Of course, having done the analysis one could use the VerbNet lexicon as an alternative practical resource, since VerbNet more or less extends Levin's verb classes.

What makes the proposed approach attractive is that having identified the verb classes we need not worry whether we encounter a normal usage or a metaphorical usage of the verb. We know that even in its metaphorical usage, a verb tends to respect its basic syntactic requirements.

The work proposed can perhaps be automatically induced to a certain degree. Such an approach would require a large word aligned parallel corpus, unfortunately we do not have such a resource. In Section 6 we revisit this issue again and give some pointers towards the potential insights which our work can have for automatic induction of valid verb classes.

4. Verbs Classes

This section describes some of the verb classes chosen for English-Hindi and English-Telugu. Table 1 and Table 2 show such verb classes. The empty cells in the tables signify default sense. A '|' and '/' means alternative (these alternatives generally arise due to dialect variations), a '!' means negation. Note that the classes shown in the tables may not be the only ones which satisfy the three conditions mentioned in the previous section, there may be many more such verb classes. A more exhaustive study is needed to come up with a complete list. The present study only shows that the proposed approach works effectively.

4.1 English-Hindi MT

In this section we describe some of verb classes that can be used in the English-Hindi MT for PSS.

(a) Verbs of Contact by Impact: #18[3]

The verbs in this class describe moving one entity in order to bring it into contact with another entity [18]. Most of the verbs in the class will have the '*with/against*' alternation. Notice that it is the semantics of the class which constrains the sense of preposition. Table 1 shows the postposition for *with* and *against* on the Hindi side, which are different from their default usage.

(b) *Peer verbs: #30.3*

Peer verbs comprise a subset of verbs of perception. These verbs necessarily take a PP. Apart from *listen, check,* and *snoop*, all the other verbs can take '*at*'. When they do, the corresponding Hindi postposition gets fixed. This can be seen Table 1, the selected Hindi sense is not canonical.

(c) Judgment verbs: #33

The verbs in this class do not always take a PP, nevertheless a subset of this verb can be exploited. Specifically, when the 'negative verbs' such as *abuse, fine, punish,* etc. take a '*for*', it gets translated as *'ke kArana'* or alternatively *'kii vajaha sai'* in Hindi.

4.2 English-Telugu MT

This section describes some of the verb classes which lend themselves in the English-Telugu MT. Table 2 shows the result for all such verbs.

(a) *Marvel Verbs: #31.3*

Marvel verbs form a subset of the verbs of psychological state. All the verbs in this class are intransitive. They obligatory take various prepositions to express the stimulus/object of emotion.

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[3] The numbers after the colon reflect the original index number of a verb class in [18].

Table 1: Verb Classes for English-Hindi MT

(the number in the far left column are Levin's verb class index, the complete list along with actual names can be seen in Appendix)

	To	**With**	**Into**	**Against**	**At**	**For**
Default	ko	ke sAtha	andara ko	ke virudha	para	ke liye
11.5	taka		meM			
18		se (!18.3)	meM	se??		
30.3					kI tarapha	
33 (NEG)						ke kArana / kii vajaha sai
36.3		se				
37.5	se	se				
37.7	se	se				
22.1			meM			
22.3, 22.5	se	se	meM			
22.4	para					

Table 2: Verb Classes for English-Telugu MT

	At	***On***	***From***	***For***
Default	daggara	mIda	nuMci	kosaM
54.4 Price verbs	ku | ki			
13.5.1 Get verbs		na		
25.3 Illustrate verbs		paina		
31.3 Marvel verbs			toti	
39.6 Gorge verbs				varaku
51.1 Directed motion	ki|lo	na		

(b) *Get Verbs: #13.5.1*

The verbs of this class are part of a larger class of Get verbs. These verbs can take a benefactive argument using a PP *for*. However, *for* takes a default target sense and therefore we do not show the result for this preposition. Interestingly, when *on* appears as part of an optional PP it does get uniquely mapped in Telugu.

(c) *Price Verbs: #54.4*

This verb class is subsumed in Measure verbs. The verbs in this class describe an agent measuring the value of an attribute of an entity along a scale relevant to the characterization of that attribute. The verbs can optionally take a preposition to specify the measuring rate.

5. Experiment

In this section we evaluate the effect of certain verb classes for PSS in an English-Telugu MT. We later integrate the proposed approach with the approach proposed in [10] and show that there is a considerable improvement over previous published results. The tables specify the number of sentences used to evaluate the system. All the sentences were randomly selected from the ERDC[4] corpus.

We evaluate the performance of the proposed approach for 5 verb classes, namely, *marvel, price, gorge, illustrate* and *get.* Note that apart from the verb class information we will need certain other syntactic information such as PP attachment and correct identification of the phrasal verb (if present). Our implemented system presumes that this information

[4] Electronic Research and Development Centre, NOIDA, UP, India.

Table 3: Results for English-Telugu PSS

Verb Class	Preposition	Default Sense	Sense Identified	Number of sentences	Precision (%)
Price V.	At	daggara	ki	23	86.9
Marvel V.	From	nuMci	toti	45	97.7
Gorge V.	From	nuMci	varaku	34	91.1
Illustrate V.	On	mIda	paIna	49	95.7
Get V.	On	mIda	na	41	92.6
				Total = 192	Average = 93.6

will be provided as input. We get this information from the English language analyzer of the Shakti MT system[5]. To make sure that the mistakes made by the analyzer do not affect our system, we manually correct the errors of attachment, etc. Table 3 shows the result, for all the verb classes we get very high accuracy with an average overall performance of 93.6%. The results point towards the robustness of their use.

Table 4 shows the performance for *at, from,* and *on* using a modified rule base system proposed in [10]. The ~200 sentences for each of these prepositions were randomly selected from the ERDC corpus. Below we summarize this system. For more details, see [10];

(a) A raw sentence is fed to the Shakti MT system which performs various source language analysis, information such as PP attachment and correct identification of the phrasal verb (if present) is crucial.

(b) The output of step (a) is taken by the module which automatically constructs a six field tuple. The tuples reflect various contextual information. At this point only some fields such as, field 1 (source language preposition), field 2 (modified category) and field 4 (modifier category) are filled.

(c) This constructed tuple is then compared with the appropriate tuples present in the rule file. For this constructed tuple to satisfy the various constraints mentioned in the tuple with which it is compared, resources such as place filter, time filter, lists and WordNet are consulted automatically. The tuple which satisfies all the constraints is selected; the last field of this tuple contains the dictionary id of the sense.

(d) Output the selected sense.

[5] (http://shakti.iiit.ac.in). Note here that the proposed algorithm has been tested with Shakti version 0.83x which has still not been released. The released version is 0.73.

Table 4: Previous rule-based approach results

Prep	No. of Sentences	Precision (%)
At	199	73.8
From	197	82.2
On	199	78.4
	Total = 595	Average = 78.1

We combine the rule base with our verb classification information. Table 5 shows the result of the combined approach. The accuracy of all the prepositions increased using the combined approach; there is an increase of 4%, 4.7% and 9% for *at, from*, and *on* respectively. There is a 6.2% increase over the performance achieved by [10].

Table 5: Results from the combined approach

Prep	No. of Sentences	Precision (%)
At	222	77.8
From	231	86.9
On	289	87.4
	Total = 742	Average = 84.3

6. Discussion

The results show that using verb class information makes an effective strategy for PSS. It can complement other approaches and increase the baseline performance. The approach can systematically exploit the semantics of some verb classes to account for the various prepositions the verb can take.

The errors for the proposed approach were mainly due to *different sense of the verb*. Majority of errors creep into the system because the sense of the verb is unknown. Sometimes we might encounter polysemous senses of the selected verb. In such a case the target sense will generally be marked wrongly. To reduce such errors one could remove all the verbs which occur in more that one verb classes; but this would reduce the number of potential verbs which can be used for PSS task.

The approach can also prove to be effective from the machine learning perspective. It is interesting to note that the state of the art in statistical MT either

based on word alignment or phrase based techniques such as [14], [19], [20], [21], and [22] use very basic linguistic feature set. These features generally do not include conjoined features where the probability of the target word (such as postposition) is induced by considering both the source preposition and the verb or the noun which this preposition modifies. Use of conjoined features is generally avoided, as it can lead to sparsity problems. Nevertheless, for our task it is easy to see that to automatically induce the verb classes such as those discussed in this paper the use of conjoined features can prove to be very effective. The translation of the preposition on the target side is not only dependent on the lexical information (word-alignment) of the preposition but also on the verb (or noun) to which it attaches. We plan to explore this aspect in the future.

7. Conclusion and future direction

In this paper we proposed an approach to preposition sense selection in English to Indian language sense selection using Levin's verb classification. We motivated this approach as a system which can complement other techniques and which can provide a better baseline. We later showed some of those verb classes for English-Hindi and English-Telugu MT system. We evaluated the system and showed that the proposed approach gives very high performance consistently; the average precision for the verbs considered for English-Telugu PSS task came out to be 93.6%. We also combined our system with previous work and showed that there is an overall improvement of 6.2% over the previous state of the art for the prepositions considered. Finally, we discussed some of the causes which lead to errors in the implemented system.

8. References

[1] Yukiko Sasaki Alam. 2004. Decision Trees for Sense Disambiguation of Prepositions: Case of Over. In *HLT/NAACL-04.*
Available:http://www.aclweb.org/anthology-new/W/W04/W04-2608.pdf

[2] Marija M. Brala. 2000. Understanding and translating (spatial) prepositions: an exercise in cognitive semantics for lexicographic purposes. In *Working Papers in English and Applied Linguistics***,** Vol. 7. University of Cambridge Press. pp. 1 - 24.
Available:
http://www.rceal.cam.ac.uk/Publications/Working/Vol7/Brala.pdf

[3] Joseph Emonds. 1985. *A unified theory of syntactic categories.* Dordrecht: Foris.

[4] Gilles Fauconnier. 1994. *Mental spaces.* Cambridge: Cambridge University Press.

[5] M. Grimaud. 1988. Toponyms, Prepositions, and Cognitive Maps in English and French, *Journal of American Society of Geolinguistics.*

[6] Ebba Gustavii. 2005. Target Language Preposition Selection - an Experiment with Transformation-based Learning and Aligned Bilingual Data. In *Proceedings of EAMT 2005*.
Available:
http://stp.lingfil.uu.se/~ebbag/Target_Language_Preposition_Selection.pdf

[7] Sven Hartrumpf, Hermann Helbig and Rainer Osswald. 2005. Semantic Interpretation of Prepositions for NLP Applications. In *EACL 2006 Workshop: Third ACL-SIGSEM Workshop on Prepositions*.
Available: http://www.aclweb.org/anthology/W06-2105

[8] A. Herskovits. 1986. *Language and Spatial Cognition, An Interdisciplinary Study of Prepositions in English*. Cambridge University Press.

[9] Cliffort Hill. 1982. Up/down, front/back, left/right. A contrastive study of Hausa and English. In *Weissen born and Klein*, 13-42.

[10] Samar Husain, Dipti Misra Sharma and Manohar Reddy. 2007. Simple Preposition Correspondence: A problem in English to Indian language Machine Translation. *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, ACL. Prague, Czech Republic. 2007.*
Available: http://www.mt-archive.info/ACL-2007-Husain.pdf

[11] Ray Jackendoff. 1977. *The architecture of the language.* Cambridge, MA: MIT Press.

[12] Ray Jackendoff. 1990. *Semantic Structures.* MIT Press, Cambridge, Massachusetts.

[13] Nathalie Japkowicz and Janyce M. Wiebe. 1991. A System For Translating Locative Preposition From English Into French. In *Proc. Association for Computational Linguistics.*
Available: http://www.aclweb.org/anthology-new/P/P91/P91-1020.pdf

[14] Shankar Kumar, Franz Josef Och, Wolfgang Macherey. 2007. Improving Word Alignment with Bridge Languages. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pages 42-50.*
Available: http://www.aclweb.org/anthology-new/D/D07/D07-1005.pdf

[15] Ronald W. Langacker. 1987. *Foundations of Cognitive Grammer*. Stanford University Press. Stanford. California.

[16] George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

[17] George Lakoff and Mark Johnson. 1987. *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.

[18] Beth Levin. 1993. *English verb classes and alternations*. Chicago/London: The University of ChicagoPress.

[19] K. Ganchev, João Graça and B. Taskar. 2008. Better Alignments = Better Translations? *Association for Computational Linguistics (ACL)*. Columbus, Ohio, June 2008.
Available:
http://www.seas.upenn.edu/~kuzman/publications/ganchev_acl_2008_slides.pdf

[20] P. Liang, B. Taskar, and D. Klein. 2006. Alignment by Agreement. *Human Language Technology conference - North American chapter of the Association for Computational Linguistics (HLT-NAACL06)*, New York, June 2006.
Available: http://www.cs.berkeley.edu/~pliang/papers/alignment-naacl2006.pdf

[21] P. Liang, Alexandre Bouchard-Cote, D. Klein and B. Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. *Association for Computational Linguistics (ACL06)*, Sydney, Australia, July 2006.
Available:
http://www.seas.upenn.edu/~taskar/pubs/acl06.pdf

[22] B. Taskar, S. Lacoste-Julien, and D. Klein. 2005. A Discriminative Matching Approach to Word Alignment. *Empirical Methods in Natural Language Processing (EMNLP05)*, Vancouver, British Columbia, October 2005.
Available: http://www.seas.upenn.edu/~taskar/pubs/mmwa.ps

[23] George A. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography.*

[24] Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2005. Handling of Prepositions in English to Bengali Machine Translation. In *EACL 2006 Workshop: Third ACL-SIGSEM Workshop on Prepositions*.
Available: http://www.aclweb.org/anthology-new/W/W06/W06-2113.pdf

[25] Geoffrey Pullum and Rodney Huddleston. 2002. Prepositions and prepositional phrases. In *Huddleston and Pullum (eds.)*, 597-661.

[26] Gisa Rauh. 1993. On the grammar of lexical and nonlexical prepositions in English. In *Zelinskiy-Wibbelt (eds.)*, 99-150.

[27] Patrick Saint-Dizier and Gloria Vazquez. 2001. A compositional framework for prepositions. In *IWCS4, Tilburg, Springer*, lecture notes, p. 165-179.
Available: http://www.irit.fr/~Patrick.Saint-Dizier/publi_fichier/iwcs4.pdf

[28] Patrick Saint-Dizier. 2005. PrepNet: A framework for describing prepositions: Preliminary investigation results. In *Proc. of IWCS 6*, Tilburg.
Available: http://www.irit.fr/~Patrick.Saint-Dizier/publi_fichier/iwcs6rev.pdf

[29] Joseph M. Sopena, Agusti Loberas and Joan L. Moliner. 1998. A connectionist approach to prepositional phrase attachment for real world texts. In *COLING-ACL '98*, 1233-1237.
Available: http://www.aclweb.org/anthology-new/P/P98/P98-2201.pdf

[30] T. Tezuka, R. Lee, H. Takakura, and Y. Kambayashi. 2001. Web-Based Inference Rules for Processing Conceptual Geo-graphical Relationships. In *Proc. of the 2nd Int. Conf. on Web Information Systems Engineering, The 1st Int. Workshop on Web Geographical Information Systems*.
Available: http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/wise/TezukaLKT01

[31] Arturo Trujillo. 1992. Locations in the machine translation of prepositional phrases. In *Proc. TMI-92.*
Available: http://www.mt-archive.info/TMI-1992-Trujillo.pdf

[32] Cornelia Zelinskiy-Wibbelt. 1993. Interpreting and translating prepositions: A cognitively based formalism. In *Zelinskiy-Wibbelt (eds.)*, 351-390.

9. Appendix

11.5:	Drive verbs
13.5.1:	Get verbs
18:	Verbs of Contact by Impact
22.1:	Mix verbs
22.3:	Shake verbs
22.4:	Tape verbs
22.5:	Cling verbs
25.3:	Illustrate verbs
30.3:	Peer verbs
31.3:	Marvel verbs
33 (NEG):	Judgment verbs
36.3	Meet verbs
37.5:	Talk verbs
37.7:	Say verbs
37.9:	Advice verbs
39.6:	Gorge verbs
54.4	Price verbs
51.1:	Directed motion