Syntactic Construct : An Aid for translating English Nominal Compound into Hindi

Soma Paul
IIIT Hyderabad
soma@iiit.ac.in

Prashant Mathur
IIIT Hyderabad
mathur@research.iiit.ac.in

Sushant Kishore
IIIT Hyderabad
susanta@research.iiit.ac.in

Abstract

This paper illustrates a way of using paraphrasal interpretation of English nominal compound for translating them into Hindi. Input Nominal compound is first paraphrased automatically with the 8 prepositions as proposed by Lauer (1995) for the task. English prepositions have one-to-one mapping to post-position in Hindi. The English paraphrases are then translated into Hindi using the mapping schema. We have got an accuracy of 71% over a set of gold data of 250 Nominal Compound. The translation-strategy is motivated by the following observation: It is only 50% of the cases that English nominal compound is translated into nominal compound in Hindi. In other cases, they are translated into varied syntactic constructs. Among them the most frequent construction type is "Modifier + Postposition + Head". The translation module also attempts to determine when a compound is translated using paraphrase and when it is translated into a Nominal compound.

1 Introduction

Nominal Compounds are syntactically condensed constructs which have extensively been attempted to expand in order to unfold the meaning of the constructions. Currently there exist two different approaches in Computational Linguistics: (a) Labeling the semantics of compound with a set of abstract relations (Moldovan and Girju, 2003) (b) Paraphrasing the compound in terms of syntactic constructs. Paraphrasing, again, is done in three ways: (1) with prepositions ("war story" → "story about war") (Lauer 1995) (2) with verb+preposition nexus ("war story" → "story pertaining to war", "noise pollution" → "pollution caused by noise") (Finin 1980) (3) with Copula ("tuna fish" → "fish that is tuna") (Vanderwende,1995). Nominal compound (henceforth NC) is a frequently occurring construct in English[1]. A bigram or two word nominal compound, is a construct of two nouns, the rightmost noun being the head (H) and the preceding noun the modifier (M) as found in "cow milk", "road condition", "machine translation" and so on. Rackow et al. (1992) has rightly observed that the two main issues in translating the source language NC correctly into the target language involves (a) correctness in the choice of the appropriate target lexeme during lexical substitution and (b) correctness in the selection of the right target construct type. The issue stated in (a) involves correct selection of sense of the component words of NCs followed by substitution of source language word with that of target language that best fits for the selected sense (see Mathur and Paul 2009).

From the perspective of machine translation, the issue of selecting the right construct of target language becomes very significant because English NCs are translated into varied construct types in Hindi. This paper motivates the advantage of expanding English nominal compounds into "paraphrases with prepositions" for translating them into Hindi. The English NCs are paraphrased using Lauer's (1995) 8 prepositions. In many cases prepositions are semantically overloaded. For example, the NC "Hindu law" can be paraphrased as "law of Hindu". This paraphrase can mean "Law made by Hindu" (not for Hindu people alone though) or "Law meant for Hindu" (law can be made by anyone, not by the Hindus necessarily). Such resolution of meaning is not possible from "preposition paraphrase". The paper argues that this is not an issue from the point of view of trans-

[1] Kim and Baldwin (2005) reports that the BNC corpus (84 million words: Burnard (2000)) has 2.6% and the Reuters has (108M words: Rose et al. (2002)) 3.9% of bigram nominal compound.

Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, pages 32–38,
Los Angeles, California, June 2010.

lation at least. It is because the Hindi correspondent of "of", which is "*kA*", is equally ambiguous. The translation of "Hindu law" is "*hinduoM kA kAnUn*" and the construction can have both aforementioned interpretations. Human users can select the right interpretation in the given context. On the other hand, 'paraphrase with preposition' approach has the following advantages: (a) Annotation is simpler; (b) Learning is easier and (c) Data sparseness is less; (d) Most importantly, English prepositions have one to one Hindi postposition correspondents most of the times. Therefore we have chosen the strategy of "paraphrasing with prepositions" over other kind of paraphrasal approach for the task of translation. The paper explores the possibility of maintaining one to one correspondence of English-Hindi preposition-postpositions and examines the accuracy of translation. At this point it is worth mentioning that translation of English NC as NC as well as different syntactic constructs in Hindi is almost equal. Therefore the task of translating English NCs into Hindi is divided into two levels: (1) Paraphrases for an NC are searched in the web corpus, (2) An algorithm is devised to determine when the paraphrase is to be ignored and the source language NC to be translated as NC or transliterated in NC, and (3) English preposition is replaced by Hindi corresponding postposition. We have compared our result with that of google translation system on 250 that has been manually created.

The next section describes the data in some detail. In section 3, we review earlier works that have followed similar approaches as the present work. Our approach is described in section 4. Finally the result and analysis is discussed in section 5.

2 Data

We made a preliminary study of NCs in English-Hindi parallel corpora in order to identify the distribution of various construct types in Hindi which English NCs are aligned to. We took a parallel corpora of around 50,000 sentences in which we got 9246 sentences (i.e. 21% cases of the whole corpus) that have nominal compounds. We have found that English nominal compound can be translated into Hindi in the following varied ways:

1. As Nominal Compound

   "Hindu texts" → *hindU shAstroM*

   "milk production" → *dugdha utpAdana*

2. M + Postposition + H Construction

   "rice husk"→ *cAvala kI bhUsI*,

   "room temperature"→ *kamare ke tApamAna*

   "wax work"→ *mom par citroM*
   "work on wax"

   "body pain"→ *sharIra meM darda*
   "pain in body"

   English NCs are frequently translated into genitive[2] construct in Hindi. In English "of" is heavily overloaded(very ambiguous), so the genitives are in Hindi. The two other postpositions that we see in the above data are *par* "on" and *meM* "in/at" and they refer to location.

3. As Adjective Noun Construction

   "nature cure" → *prAkritika cikitsA*

   "hill camel" → *pahARI UMta*
   The words *prAkrtik* and *pahARI* being adjectives derived from *prakriti* and *pAhAR* respectively.

4. Single Word

   "cow dung" → *gobara*

The distribution of various translations is given below:

Construction Type	**No. of Occurrence**
Nominal Compound	3959
Genitive(of-*kA/ke/kI*)	1976
Purpose (for-*ke liye*)	22
Location (at/on-*par*)	34
Location (in-*meM*)	93
Adjective Noun Phrase	557
Single Word	766
Transliterated NC	1208

Table 1: Distribution of translations of English NC from English Hindi parallel corpora.

There are 8% cases (see table 1) when an English NC becomes a single word form in Hindi. For rest of the cases, they either remain as NC (translated 43% or transliterated 13%) or correspond to syntactic construct. When NC is translated as NC, they are mostly technical terms

[2]"of" corresponds to "kA/ke/kI", which are genitive markers in Hindi.

or proper names. Our data shows that there are around 40% cases when English NC is translated as various kinds of syntactic constructs such M + Postposition + H, Adj + H or longer paraphrases ("Hand luggage" → *hAth meM le jAne vAle sAmAn* "luggage to be carried by hand"). Out of these data, 70% cases are when English NC is translated into M[3] + postposition + H. Thus the translation of NC into postpositional construction is very common in Hindi.

For preparation of test data, we extracted nominal compound from BNC corpus (Burnard et al., 1995). BNC has varied amount of text ranging from newspaper article to letters, books etc. We extracted a sample of noun-noun bigrams from the corpus and manually translated them into Hindi.

In this paper, we propose an algorithm that determines when the syntactic paraphrase of English NC is to be considered for translation and when it is left for direct lexical substitution in Hindi.

3 Related Works

There exists no work which has attempted the approach that we will be discussing here for translating English NC into Hindi. From that perspective, the proposed approach is first of its kind to be attempted. However, paraphrasing English NCs is a widely studied issue. Scholars (Levi 1978; Finin 1980) agree there is a limited number of relations that occur with high frequency in noun compounds. However, the number and the level of abstraction of these frequently used semantic categories are not agreed upon. They can vary from a few prepositional paraphrases (Lauer, 1995) to hundreds and even thousands more specific semantic relations (Finin, 1980). Lauer (1995), for example, considers eight prepositional paraphrases as semantic classification categories: of, for, with, in, on, at, about, and from. According to this classification, the noun compound "bird sanctuary", for instance, can be classified both as "sanctuary of bird" and "sanctuary for bird".

The automatic interpretation of noun compounds is a difficult task for both unsupervised and supervised approaches. Currently, the best-performing NC interpretation methods in computational linguistics focus only on two-word noun compounds and rely either on rather ad-hoc, domain-specific, hand-coded semantic taxonomies, or on statistical models on large collections of unlabeled data.

The majority of corpus based statistical approaches to noun compound interpretation collects statistics on the occurrence frequency of the noun constituents and uses them in a probabilistic model (Resnik, 1993; Lauer, 1995; Lapata and Keller, 2004). Lauer (1995) was the first to devise and test an unsupervised probabilistic model for noun compound interpretation on Grolier encyclopedia, an 8 million word corpus, based on a set of 8 prepositional paraphrases. His probabilistic model computes the probability of a preposition p given a noun-noun pair *n1-n2* and finds the most likely prepositional paraphrase

$$p* = argmaxP(p|n1, n2) \quad (1)$$

However, as Lauer noticed, this model requires a very large training corpus to estimate these probabilities. Lapata and Keller (2004) showed that simple unsupervised models applied to the noun compound interpretation task perform significantly better when the n-gram frequencies are obtained from the web (accuracy of 55.71% on Altavista), rather than from a large standard corpus. Our approach also uses web as a corpus and examines frequency of various preposition paraphrases of a given NC. The next section describes our approach.

4 Approach

This section describes our procedure in details. The system is comprised of the following stages: (a) Web search of prepositional paraphrase for English NC; (b) mapping the English preposition to corresponding Hindi postposition; (c) Evaluation of correct paraphrasing on English side as well as evaluation of translation.

4.1 Paraphrase Selection for Translation

Based on the observation from English-Hindi parallel corpus data that we examined as part of this project, we have designed an algorithm to determine whether an English NC is to be translated as an analytic construct or retained as an NC in Hindi. We used Yahoo search engine to perform a simple frequency search for "M Preposition H" in web corpus for a given input NC. For example, the paraphrases obtained for the NC "finance minister" is given in table 2 and frequency of various paraphrases is shown in the second column:

[3] M: Modifier, H: Head

Paraphrase	Web Frequency
minister about finance	2
minister from finance	16
minister on finance	34300
minister for finance	1370000
minister with finance	43
minister by finance	20
minister to finance	508
minister in finance	335
minister at finance	64
minister of finance	5420000

Table 2: Frequency of Paraphrases for "finance minister" after Web search.

In the table we notice that the distribution is widely varied. For some paraphrase the count is very low (minister about finance) while the highest count is 5420000 for "minister of finance". The wide distribution is apparent even when the range is not that high as shown in following table:

Paraphrase	Web Frequency
agencies about welfare	1
agencies from welfare	16
agencies on welfare	64
agencies for welfare	707
agencies with welfare	34
agencies in welfare	299
agencies at welfare	0
agencies of welfare	92

Table 3: Frequency of Paraphrases for "welfare agencies" after Web search.

During our experiment we have come across three typical cases: (a) No paraphrase is available when searched; (b) Frequency counts of some paraphrases for a given NC is very low and (c) Frequency of a number of paraphrases cross a threshold limit. The threshold is set to be mean of all the frequencies of paraphrases. Each of such cases signifies something about the data and we build our translation heuristics based on these observations. When no paraphrase is found in web corpus for a given NC, we consider such NCs very close-knit constructions and translate them as nominal compound in Hindi. This generally happens when the NC is a proper noun or a technical term. Similarly when there exists a number of paraphrases each of those crossing the threshold limit, it indicates that the noun components of such NCs can occur in various contexts and we select the first 3 paraphrase as probable paraphrase of NCs. For example, the threshold value for the NC *finance minister* is: Threshold = 6825288/8 = 853161. The two paraphrases considered as probable paraphrase of this NC is are therefore "minister of finance" and "minister for finance". The remaining are ignored. When count of a paraphrase is less than the threshold, they are removed from the data. We presume that such low frequency does not convey any significance of paraphrase. On the contrary, they add to the noise for probability distribution. For example, all paraphrases of "antelope species" except "species of antelope" is very low as shown in Table 4. They are not therefore considered as probable paraphrases.

Paraphrase	Web Frequency
species about antelope	0
species from antelope	44
species on antelope	98
species for antelope	8
species with antelope	10
species in antelope	9
species at antelope	8
species of antelope	60600

Table 4: Frequency of Paraphrases for antelope species after Web search.

4.2 Mapping English Preposition to Hindi Post-position

The strategy of mapping English preposition to one Hindi post-position is a crucial one for the present task of translation. The decision is mainly motivated by a preliminary study of aligned parallel corpora of English and Hindi in which we have come across the distribution of Lauer's 8 prepositions as shown in table 5.

The table (Table 5) shows that English prepositions are mostly translated into one Hindi postposition except for a few cases such as "at", "with" and "for". The probability of "on" getting translating into "*ko*" and "of" into "*se*" is very less and therefore we are ignoring them in our mapping schema. The preposition "at" can be translated into "*meM*" and "*para*" and both postpositions in Hindi can refer to "location". However, the two prepositions "with" and "for" can be translated into two distinct relations as shown in Table 5. From our parallel corpus data, we therefore

Prep	Post-Pos	Sense	Prob.
of	*kA*	Possession	0.13
	ke	Possession	0.574
	kI	Possession	0.29
	se	Possession	0.002
from	*se*	Source	.999
at	*meM*	Location	0.748
	par	Location	.219
with	*se*	Instrument	0.628
	ke sAtha	Association	0.26
on	*par*	Loc./Theme	0.987
	ko	Theme	0.007
about	*ke bAre meM*	Subj.Matter	0.68
in	*meM*	Location	.999
for	*ke lie*	Beneficiary	0.72
	ke	Possession	0.27

Table 5: Mapping of English Preposition to Hindi postposition from alligned English-Hindi parallel corpora.

find that these prepositions are semantically overloaded from Hindi language perspective. The right sense and thereafter the right Hindi correspondent can be selected in the context. In the present task, we are selecting the mapping with higher probability. English Prepositions are mapped to one Hindi Post-position for all cases except for "at" and "about".

Preposition	Postposition
of	*kA/kI/ke*
on	*para*
for	*ke liye*
at	*para/meM*
in	*meM*
from	*se*
with	*ke sAtha*
about	*ke bAre meM* *ke viSaya meM* *ke sambaMdhi*

Table 6: Preposition-Postposition Mapping

Post-positions in Hindi can be multi-word as in "*ke bAre meM*", "*ke liye*" and so on. In the present paper we are translating the English preposition to the mostly probable postposition of Hindi. That does not mean that the preposition cannot be translated into any other postposition. However, we are taking the aforementioned stand as an preliminary experiment and further refinement in terms of selection of postposition will be done as future work. For the present study, lexical substitution of head noun and modifier noun are presumed to be correct.

5 Result and Analysis

In this section we will describe results of two steps that are involved in our work: (a) Selection of English preposition paraphrase for a given English NC; (b) Translation of English Preposition to Hindi Post-position.

For a given NC we used a brute force method to find the paraphrase structure. We used Lauer's prepositions (of, in, about, for, with, at, on, from, to, by) for prepositional paraphrasing. Web search is done on all paraphrases and frequency counts are retrieved. Mean frequency (*F*) is calculated using all frequencies retrieved. All those paraphrases that give frequency more than *F* are selected. We first tested the algorithm on 250 test data of our selection. The result of the top three paraphrases are given below :

Selection Technique	Precision
Top 1	61.6%
Top 2	67.20%
Top 3	71.6%

Table 7: Paraphrasing Accuracy

We have also tested the algorithm on Lauer's test data (first 218 compounds out 400 of NCs) and got the following results (Table 8). Each of the test data was marked with a preposition which best explained the relationship between two noun components. Lauer gives X for compounds which cannot be paraphrased by using prepositions For eg. *tuna fish*.

Prep	O_{Lauer}	O_{CI}	Percentage
Of	54	37	68.50%
For	42	20	47.62%
In	24	9	37.50%
On	6	2	33.33%

Table 8: Distribution of Preposition on Lauer test data of 218 NC

O_{Lauer} : Number of occurrence of each preposition in Lauer test data

O_{CI} : Number of correctly identified preposition by our method

In Table 9 we compare our result with that of Lauer's on his data. We gave the results with criteria: 1) only "N prep N" is considered. 2) Non-Prepositions (X) are also considered.

Case	Our Method	Lauer's
N-prep-N	43.67%	39.87%
All	42.2%	28.8%

Table 9: Comparison of our approach with Lauer's Approach

Now that we have paraphrased NCs, we attempt to translate the output into Hindi. We *assume* that we have the right lexical substitution. In this paper we have checked for the accuracy of the right Hindi construction selection.

For a given NC we got the paraphrase as "H prep M" or "MH". We use English preposition mapping as described in section 4.2 for translating NC in Hindi. For MH type compounds direct lexical substitution is tried out. We tested our approach on the gold data of 250 Nominal Compounds. We translate the same 250 NCs using google translation system in order to set up a baseline. Google Translator could translate the data with 68.8% accuracy.

Google returns only one translation which we evaluated against our test data. In our case, we have taken 3 top paraphrases as described in section 4.1 and translated them into Hindi by using the English preposition to Hindi postposition mapping schema. The following table presents the accuracy of the translation of the top three paraphrases

Case	Precision
Top 1	61.6%
Top 2	68.4%
Top 3	70.8%

Table 10: Translation Accuracy

In this work we have not considered the context of English NC while translating them into Hindi. Table 11 gives the accuracy of each post-position as translated from English preposition.

The other prepositions have occurred very less in number and therefore not given in the table.

Preposition	Post Position	Accuracy
Of	*kA/ke/kI*	94.3%
For	*ke liye*	72.2%
In	*meM*	42.9%

Table 11: Translation Accuracy for some individual prepositions

6 Conclusion and Future Work

This paper describes a preliminary approach for translating English nominal compound into Hindi using paraphrasing as a method of analysis of source data. The result of translation is encouraging as a first step towards this kind of work. This work finds out a useful application for the task of paraphrasing nominal compound using preposition. The next step of experiment includes the following tasks: (a) Designing the test data in such a way that all correspondents get equal representation in the data. (b) To examine if there are any other prepositions (besides Lauer's 8 preposition) which can be used for paraphrasing (c) To use context for translation.

References

Gildea. D. and Jurafsky. D. 2002. *Automatic labeling of semantic roles*, Computational Linguistics 28 (3), 245-288.

Lapata, M. and Keller, F. 2004. *The Web as a baseline: evaluating the performance of unsupervised Web-based models for a range of NLP tasks*. In: Proceedings of the Human Language Technology conference (HLT/NAACL), Boston, MA, pp. 121-128.

Lauer, M. 1995 *Designing statistical language learners: experiments on noun compounds*, Ph.D. Thesis, Macquarie University, Australia

Moldovan, D. and Girju, R. 2003 *Knowledge discovery from text* In: The Tutorial Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan.

Mathur, P. and Paul, S. 2009 *Automatic Translation of Nominal Compound into Hindi*. In: Proceedings of International Conference on Natural Language Processing (ICON), Hyderabad

Moldovan, D., Girju, R., Tatu, M., and Antohe, D. 2005 *On the semantics of noun compounds* Computer Speech & Language 19(4): 479-496

Girju, R. 2009 *The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study* Computational Linguistics 35(2): 185-228

Vanderwende, L. 1995 *The analysis of noun sequences using semantic information extracted from on-line dictionaries* Ph.D. Dissertation, Georgetown University.

Barker, K. and Szpakowicz, S. 1998 *Semi-automatic recognition of noun modifier relationships* In Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98) , pages 96-102, Montreal, Canada.

Finin, T.W. 1980 *The semantic interpretation of nominal compounds* In Proc. of the 1st Conference on Artificial Intelligence (AAAI-80), 1980.

Isabelle, P. 1984 *Another look at nominal compounds* In Proc. of the 10th International Conference on Computational Linguistics (COLING '84), Stanford, USA, 1984.

Kim, S.N. and Baldwin, T. 2005 *Automatic Interpretation of Noun Compounds Using WordNet Similarity* IJCNLP 2005:945-956

Rackow, U., Dagan, I. and Schwall, U. 1992 *Automatic translation of noun compounds* In Proc. of the 14th International Conference on Computational Linguistics (COLING '92), Nantes, Frances, 1992