Automatic Translation of Nominal Compounds from English to Hindi

by

Prashant Mathur, Soma Paul

in

International Conference on Natural Language Processing
(*ICON-2009*)

Report No: IIIT/TR/2009/219

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

Automatic Translation of English Nominal Compound in Hindi

Prashant Mathur
Language Technology Research Centre, IIIT Hyderabad
mathur@students.iiit.ac.in

Soma Paul
Language Technology Research Centre, IIIT Hyderabad
soma@iiit.ac.in

Abstract

English nominal compounds can be variously translated into Hindi. This paper presents an automatic translation system for translating English bigram nominal compound into Hindi. The method comprises of the following steps: (1) Translation template generation (2) Extraction of nominal compound from English corpus (3) Finding the appropriate sense of the components of the compound using WSD tool (4) Lexical substitution of the component nouns using Bi-Lingual Dictionary (5) Corpus Search using translation templates and Ranking of possible candidates. We have shown that the correct sense selection of the component nouns of a given nominal compound during the analysis stage significantly improves the performance of the system and makes the present work distinct from all the previous works done for automatic bilingual translation of Nominal compounds.

1.0 Introduction

Multiword nominal compound is a frequently occurring expression in English[1]. A two word nominal compound (henceforth NC) is a construct of two nouns, the rightmost noun being the head and the preceeding noun the modifier as found in 'cow milk', 'road condition', 'machine translation' and so on[2]. Rackow et al. (1992) has rightly observed that the two main issues in translating the source language NC *correctly* in 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 (b) becomes apparent when we examine a parallel corpus of English and Hindi that we have used for the present work. We have found that English nominal compounds can be translated in Hindi in the following varied ways:

a. As Nominal Compound
'Hindu texts' → *hindU SastroM*, 'milk production' → *dugdha utpAdana*

b. As Genitive Construction
'rice husk' → *cAval kI bhUsI*, 'room temperature'→ *kamare ke tApamAn*

c. As Adjective Noun Construction
'nature cure' → *prAkrtik cikitsA*, 'hill camel'→ '*pahARI UMT*'

The words *prAkrtik* and *pahARI* being adjectives derived from *prakriti* and *pAhAR* respectively.

d. As other syntactic phrase
wax work → *mom par ciwroM* 'work on wax',
body pain → *SarIr meM dard* 'pain in body'

e. As one word
Cow dung → *gobar*

[1] Tanaka and Baldwin (2004) 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.

[2] A nominal compound may be constituted of a more complex structure as 'customer satisfaction indices', 'social service department' and so on.

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

f. Others
Hand luggage → *haat meM le jaaye jaane vaale saamaan* 'luggage to be carried by hand'

However, no definite clue is available in the data that helps one in selecting the right construction type of Hindi for translating a given English NC. Tanaka and Baldwin (2004) observes that a translator or MT system attempting to translate a corpus will run across NCs with high frequency, but that each individual NN compound will occur only a few times (with around 45-60% occurring only once). The upshot of this for MT systems and translators is that NN compounds are too varied to be able to pre-compile in an exhaustive list of translated NN compounds. The system must be able to deal with novel NN compounds on the fly. Building an automatic translation system for nominal compounds from the source language (SL) English to the target language (TL) Hindi thus becomes a very challenging task in NLP. With Google translator we could achieve an accuracy of 45% with the same test data that we have used to evaluate our model. It could give a correct translation in 29% cases when a nominal compound remains a nominal compound in Hindi. When an NC is translated in genitive construction in Hindi, the translator could return the correct result 10% of cases. For other cases such as when NC translated as Adjective noun pair or as a single word, the performance of Google translator is poor.

This paper presents the architecture of a "Nominal Compound Translator" system that has been able to give an accuracy of 57% when tested on unseen gold standard test data. We limit our discussion to English two word nominal compounds in this paper. The approach adopted to build the system has a close resemblance to the approaches described in Bungum and Oepen (2009) for Norwaygian to English nominal compound translation and Tanaka and Baldwin (2004) (English to Japanese nominal compound and vice versa). All these works including the one described in this paper follow a *template based corpus search* approach. However, the present system distinctly differs from the aforementioned works for the analysis stage. Our system, unlike others, attempts to select the correct sense of nominal components by running a WSD system on the SL data. As a result of that the number of possible translation candidates to be searched in the target language corpus is significantly reduced. Translation of nominal compound combines the following subtasks: (1) Template Generation for Candidate Selection from target language Hindi (2) Extraction of NCs from English corpus (3) Finding relevant sense of the components of NCs. (4) Component Translation to Hindi using Bi-Lingual Dictionary (5) Corpus Search using templates and Ranking of possible candidates.

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.

2.0 Data

At the time of taking up the present project we made a preliminary study of NCs in English-Hindi parallel corpora in order to identify the distribution of various construct types which English NC 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 has nominal compound. The percentage of various translations is given in Table 1.

We have also come across some cases where an NC corresponds to a paraphrase construct for which we have not given a count in this table. There are .08% cases (see table 1) when an English NC becomes a single word form in Hindi. The single word form can either be a simple word as in ('cattle dung'→ *gobar*) or a compounded word such as 'blood pressure' → *raktacApa*, 'transition plan' → *parivartana-yojanA*.

Construction Type	No of occurrence	%
Nominal compound	3959	42.9
Genitive	1976	21.4
Adjective Noun phrase	557	.06
Single Word	766	.08
Transliterated Nominal Compound	1208	13.0

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

The above table records major translation types. There are 1208 cases (approximately 13%) where the English nominal compound is not translated but transliterated in Hindi. They are mostly technical terms, names of chemicals and so on.

The figure given in Table 1 is a report of the empirical study performed on English-Hindi parallel corpora. We prepare a set of *translation templates* that represents the construct types of Hindi (as in table 1). In section 4, we will discuss how these templates are used for searching possible translation in Hindi raw corpus. From table 1, we come to know that the frequency of English NC remaining as nominal compound in Hindi is the highest. The second highest construction is the genitive construct. Parallely we have performed a study with Hindi informants to find out how many cases an English nominal compound can legitimately be translated into a syntactic genitive construct even when it can have other more accurate translation. Our experiment shows that a nominal compound is well accepted as a genitive construct in Hindi in 59% of cases. This is an interesting finding which we have used in designing the heuristics of the present task.

3.0 Related Works

While working on the automatic translation of English nominal compound to Hindi, we came across works on two different approaches: a) transfer based approach (Rackow et al. (1992)) and b) corpus search based probabilistic approach (Bungum and Oepen (2009) (henceforth B&O), Tanaka and Baldwin (2004) (henceforth T&B)). Rackow et al. tried to set a mapping between the head noun of source language and target language in terms of some grammatical and semantic feature which helped them in selecting the right lexical item for the target language. The strategy adopted by both B&O and T&B has close similarity to ours as far as the template generation and the procedure of corpus search is concerned. First, they generate templates which represent various construct types of the target language and then search these templates in a huge corpus. The two works differ in using different strategy for ranking of the possible translated candidates that are found in the corpus. We have adopted the T&B proposal for ranking. T&B suggests ranking candidate translation based on target language distributional properties, essentially corpus frequency. They develop a measure called "interpolated CTQ (Corpus-based translation quality) metric" which extracts frequency counts from the target language corpus (for the details see section 4.4).

While working on source language side, both B&O and T&B disregard local contexts and does not attempt to identify the sense of nominal compound in the given context. They have, on the other hand, taken into account of all possible translations of the component nouns while performing the corpus search. In this way the number of search candidates has become many. We will discuss in section 4 that while translating a nominal compound we have tried to consider the meaning of that compound in the given context, that is, the sentence in which it has occurred. In this regard, our work becomes distinct from other works referred to in this section.

4.0 Preparation of Data and Approach

This section describes our procedure in details. The system is comprised of the

following stages: a) Preparation of data and template generation b) Determining sense of the component nouns in the given context, c) Lexical substitution using bilingual dictionary, d) corpus search using translation templates and e) Ranking of the possible candidates.

4.1 Preparation of Source Language Data

Two sets of language data are prepared for the work. The first set is a parallel corpus of around 50,000 sentences in which 9246 sentences have nominal compound. The source language sentences have been manually examined for nominal compounds and their correspondent translation is identified in the Hindi target language[3]. The second set of data consists of 7000 raw sentences of English on which we have run Tree-tagger[4] which is a POS tagger. The tagger not only gives part of speech of the words but also outputs the lemma for each word. The lemma is required in the later stage for searching the word in the wordnet. Sentences with nominal compounds are extracted from the tagged data and the nominal compounds are strictly restricted to be two consecutive noun construction type. We obtain 1584 sentences with distinct nominal compounds out of which 1000 sentences are randomly chosen for processing. These sentences are manually translated into Hindi and used half of it as development data and half of it as gold standard test data.

4.2 Generation of Translation Templates

One of the most important subtasks in this work is determining the translation templates. Each template is a possible translated construct type of English NC in Hindi. The parallel corpus data are inspected and generalized into translation templates. As shown in section 2, the two templates <E1[5] E2> → <H1 H2> and <E1 E2> → <H1 kA[6] H2> are the most frequent ones. The other interesting candidate is Adjective noun phrase in Hindi. Hindi has a rich derivational system for adjective formation. In this work we have identified till now 44 templates.

4.3 Sense Selection for Source Language NCs

The context determines the sense of a given English NC in a corpus. When the component nouns are taken independently, they might represent more than one sense. For each sense the English word might be translated into more than one Hindi equivalent word using English to Hindi bilingual dictionary. Let me explain the complexity of lexical substitution with data from the corpus. We came across the following sentences in the test data:

a. 'Millions of people in the border area need to feel safe again'

b. 'Road safety aims to reduce the harm (deaths, injuries, and property damage) resulting from crashes of road vehicles'

The nominal compound identified in sentence (a) and (b) are 'border area' and 'road safety' respectively. All four words can be used in more than one sense as given in 2nd column of table 2.

Word	No. of senses from Wordnet
Border	5
Area	6
Road	2
Safety	6

[3] In order to execute this task we have used a JAVA based interface "Sanchay" that has been developed in-house. Using an interface to do this task helped us to maintain consistency in work.

[4] We used Tree-Tagger (POS-Tagger) for tagging the corpus of 1.7M words. It gave an accuracy of 94%.

[5] E stands for English and H stands for Hindi

[6] kA is a genitive marker in Hindi. It has variants kI and ke. Therefore <H1 kA H2>, <H1 ke H2> and <H1 kI H2> form three translation candidates.

Table 2 : Number of Senses Listed in Wordnet

For each sense there exists a synset which consists of one or more semantically equivalent words in the wordnet. If we consider all words for all senses of the component nouns and attempt to translate all of them using a bilingual dictionary the number of translation candidates will be huge in number. Moreover we will be searching for those candidates that are not relevant for the English NC in the given context. In order to avoid the proliferation of data, we have chosen to use a WSD tool. We ran WordNet-SenseRelate (Peterson et al.) on our data for the purpose. This tool specifies the wordnet sense id for each noun component within NC as shown in table 3:

Word	Sense selected by WSD tool	Synset
Border	#1	<'boundary line', 'border', borderline', 'delimitation', 'mete'>
Area	#3	<'area', 'region'>
Road	#1	<'road', 'route'>
Safety	#2	<'safety', refuge'>

Table 3: Output of WSD tool

The third column of table 3 presents the synset associated with the sense selected by the WSD tool. Once the synsets are acquired in this process the translation for each word in the synset is obtained from a bilingual dictionary. Once we look into a bilingual dictionary, again we may come across many equivalents of a word which do not match to the sense id selected for that word. For example, the word 'border' (a member of the synset of 'border') has one equivalent *jhaalar* in the bilingual dictionary that is used in the domain of 'decoration' and not 'location'. We would like to discard such equivalents. Otherwise the whole attempt of using WSD tool on the source language side will be lost. The ideal situation would have been to have a mapping from the synset id of a word in English wordnet to the corresponding Hindi synset id in Hindi wordnet. Since that was not available to us, we have maintained the following strategy. We first acquire all possible translations for all the words within a synset from all possible dictionary resources. Then we take out those Hindi words which are common translations to all English words of a synset, if there is one. For example, we got the following translations for the two synsets <'road', 'route'> from bilingual dictionaries:

Word	Translation
Road	*path, maarg, saDak, raastaa*
Route	*maarg, saDak, raastaa*

Table 4: Translation using a bilingual dictionary

From table 4, we find out that *maarg, saDak, raastaa* are common translation for 'road' and 'route'. Once the Hindi equivalents are obtained they are used to frame the translation candidates which are searched in the corpus for a match. When common equivalent(s) is not found for all member words of a synset, we try for maximum number of member words for which a common translation is available. The worst case is when we do not find any common translation and that was rare in our experiment. For example, for the synset members of 'border' as well as 'safety' we have not come across any common Hindi equivalents. For such cases, we try out translations of all synset members one by one for generating the translation templates.

4.4 Corpus Search and Ranking Translation Candidates

We have performed the corpus search on a Hindi indexed corpus of 28 million words. For ranking, a reference ranking based on the frequency of occurrence of the translate candidates in full in the TL corpora is taken as baseline. To improve on the baseline, a stronger ranking measure is borrowed from Baldwin and Tanaka (2004). It rates a given translation candidate according to corpus evidence for both the fully specified translation and its parts in the context of the translation template in question. The

measure is called interpolated CTQ metric that extracts the frequency counts from the target language corpus in the following manner:

$$CTQ(w1^H, w2^H, t) = \alpha p(w1^H, w2^H, t) + \beta p(w1^H, t)p(w2^H, t)p(t)$$

where $\alpha p(w1^H, w2^H, t)$ is the probability of occurrence of template t with w1 and w2 as its instances and $\beta p(w1^H, t)p(w2^H, t)p(t)$ is the probability of occurrence of translation template t with w1 as its instance at one time multiplied by the probability of occurrence of translation template t with w2 as its instance at another time multiplied by the occurrence of translation template t. Naturally the first term will be given higher priority than the second term. The result presented in the next section will show that the incorporation of frequency of occurrence of $\beta p(w1^H, t)p(w2^H, t)$ has distinctly improved the recall in our system.

5.0 Result and Analysis

This section presents the result of our various experiments performed as part of translating automatically English NC to Hindi. The results show a distinct improvement in performance as we go from baseline ranking method to CTQ method of ranking. We have used three methods of lexical substitution for components of nominal compounds into Hindi equivalents and the result obtained for each method is presented at table 1 and table 2. As part of the first method we have not done any word sense disambiguation of the component words of source language NC; on the contrary we have straightaway used the bilingual dictionaries for substituting English NC components to all possible Hindi equivalents. For the second method, the first sense of wordnet for the components of the given English NC has been selected as default sense and all the members of synset of the first sense have been substituted using a bilingual dictionary. The motivation for this approach is two fold: a) a word occurs mostly in its default sense which is listed as the first sense in any lexicon; b) if the input word is not available in a bilingual dictionary for substitution, a synset gives us other equivalent words. This increases the robustness of the system. The third method is the one we have adopted for the present task – using a WSD tool on the source language NC and select the appropriate sense of the given word in that context. The purpose of trying out various methods for lexical substitution is for examining whether the usage of WSD tool brings in any improvement to the overall performance of the translator tool. The table below shows that it does. The pre-processed input that has been used for lexical substitution is not humanly analyzed data but is actually obtained as the output of Tree-Tagger that gives 94% accuracy and the WSD tool WordNet-SenseRelate that has produced 80% accurate case for nominal compound disambiguation[7]. The results of corpus search of the translation candidates are given in the following two tables. The baseline frequency model performs in the following:

Lexical substitution method	Recall	Precision
Only Bilingual dictionary	14.2%	50%
Wordnet 1st sense + Bilingual dictionary	24%	46.15%
Wordnet sense selection by WSD tool + Bilingual Dictionary	24.63%	53.68%

Table 5: Ranking using Baseline Frequency Model

[7] It is interesting to note that the accuracy reported for the WordNet-SenseRelate output on general data is 58%. When we tested the tool for nominal compound, it gave an accuracy of around 80% for the same.

With the use of CTQ measure metric, the accuracy of translation is distinctly improved as shown in the following table:

Lexical substitution method	Recall	Precision
Only Bilingual dictionary	19%	56.25%
Wordnet 1st sense + Bilingual dictionary	28%	54.1%
Wordnet sense selection by WSD tool + Bilingual Dictionary	28.50%	62.1%

Table 6: Ranking Using CTQ (Corpusbased Translation Quality)

The recall of this experiment was very low. In order to increase the coverage of translation, we have done the following study. We involved two informants to verify on the development data whether the compounds which were not found during corpus search can legitimately be translated as a genitive construct. We found that the heuristics is working for 59% cases. Therefore we incorporated this as a default translation case for our system. Whenever a corpus search for a translation candidate fails, we assign a genitive translation for that nominal compound. This results in a steep improvement in recall although the precision falls down a little. We ran the experiment on the output of 1st and 3rd lexical substitution methods. The result is reported in the following table:

Lexical Substitution Method	Recall	Precision
Bilingual Dictionary	24.86%	54%
WSD + CTQ	44.5%	57.04%

Table 7: Ranking after inclusion of default translation (X kA Y, X kI Y, X ke Y as templates)

6.0 Conclusion and Future Work

This paper describes the architecture of a template based translation system for translating English nominal compound into Hindi. We have observed that English nominal compounds can variously be translated into Hindi. However no clue is available to determine which type of Hindi constructs a given English nominal compound would be translated into. We have, therefore, adopted a corpus search approach that performs the search of candidate templates in a Hindi indexed corpus. While generating templates, we found out that adjectival templates are hard to generate because adjective formation from noun is a complex derivational process in Hindi. It does not only involve attaching an adjectival suffix on the noun but also many a time requires a change in the vowel of the stem. In the present work, we have performed poorly for adjective noun translation templates. The future work includes the correct generation of adjectival form from the modifier nouns so that correct templates for 'Adjective Noun' construct can be obtained. One advantage of this approach is that a translation if it exists in the corpus will never be missed. Therefore accuracy of translation will depends largely on the amount of target language data searched for the translation candidates.

7. References

George A. Miller. 1994. WORDNET: A Lexical Database for English. HLT.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*. Manchester, UK.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Pierrette Bouillon, Katharina Boesefeldt, and Graham Russell. 1994. Compound

nouns in a unification-based MT system. In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, Stuttgart, Germany.

Siddharth Patwardhan, Satanjeev Banerjee and Ted Pedersen. 2005. SenseRelate::TargetWord – A Generalized Framework for Word Sense Disambiguation. Proceedings of the ACL Interactive Poster and Demonstration Sessions, Ann Arbor, MI

Sparck Jones, K. 1983. "So what about parsing compound nouns?," in Automatic Natural Language Processing, K. Sparck Jones and Y. A. Wilks, eds., Ellis Horwood, Chichester, 164--168.

Su Nam Kim, Timothy Baldwin: Automatic Interpretation of Noun Compounds Using WordNet Similarity. IJCNLP 2005: 945-956

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

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by Machine of Complex Nominals: Getting it right. In Proceedings of the ACL04 Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain.

Tanaka, Takaaki and Timothy Baldwin. 2003b. Translation Selection for Japanese-English Noun-Noun Compounds. In Proceedings of Machine Translation Summit IX, New Orleans, LO, USA.

Ulrike Rackow, Ido Dagan, Ulrike Schwall. 1992. Automatic Translation of Noun Compounds. COLING 1992, 1249-1253

Zouhair Maalej, English-Arabic Machine Translation of Nominal Compounds, *in* Proceedings *of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*. Geneva: ISSCO, pp. 135–146, 1994.