Machine Translation Accuracy by Focusing on Gender Issues for English to Hindi Translation

Beenu Yadav

Department of Computer Science and Engineering, SCRIET, CCS UNIVERSITY, Meerut, India
beenu.yadav87[at]gmail.com

Abstract: ***In English-Hindi machine translation, it is an important task to determine the gender of names correctly. When a sentence having only one subject is given as a pronoun and is composed of two or more clauses, then it is important to determine the gender of the proper noun for correct translation of the given sentence. In order to overcome these issues numerous methods have been proposed. In this particular research paper we focus on the gender identification issues from English to Hindi translation. This paper proposes statistical method for the given problem by focusing on Hindu names. Proposed method yields the improvement in accuracy. This means that the proposed solution is more plausible for the gender classification of the Hindu names.***

Keywords: statistical method, machine translation, named-entity recognition, natural language processing

1. Introduction

Machine translation systems are developed continuously still it has various problems. Machine translators produce relatively good results between languages English and German which share characteristics. But translation between two dissimilar languages such as English and Hindi is not good enough.

Hindi language is different from European languages such as English, German and French etc. In Hindi language, the omission of subject is frequently occurred but in European languages, subjects are not omitted. When a sentence is composed of multiple clauses, it is important to identify correct gender of the subject in English-Hindi machine translation. The goal of this paper is to determine gender of the subject.More generally, this task can be expanded into a classification of normal English names. It is a totally independent issue to know whether a subject is shared by several clauses or not. Hence, in this paper, we focus on the gender classification of English names[5].

Figure 1.1 shows an error in the result of the Google translator related with our topic. For example when we write a sentence “Sita is going,” As we can see the outcome is incorrect, because the Google translator couldn’t identify Sita’s gender.

In order to overcome this problem, two tasks have to be solved. The first task is named-entity recognition in which each proper noun representing person name should be correctly identified as a name. This is a well-known problem in natural language processing community and has been deeply studied for last few years. Recent work in this task results very high performance. The second task is to identify the correct gender of the recognized names. When the gender of the name is correctly identified, only then the proper pronoun can be generated in the machine translation systems. To solve this problem, a probabilistic method is used in this paper, in which probabilistic information on human names is gathered from open domain web pages. Since the named entity recognition is independent topic from the determination of the gender for a name and has been widely studied by various researchers, this paper focuses only on gender issues [3].

This paper is organised as follows. Section I Introduction. Section II Literature review on named entity recognition and pronoun generation in English. Section III proposes how the gender of English names is classified using a statistical method. Section IV gives an overall description how the proposed method is used in English-Hindi machine translation. Section V gives the experimental results. Finally, Section VI draws conclusions.

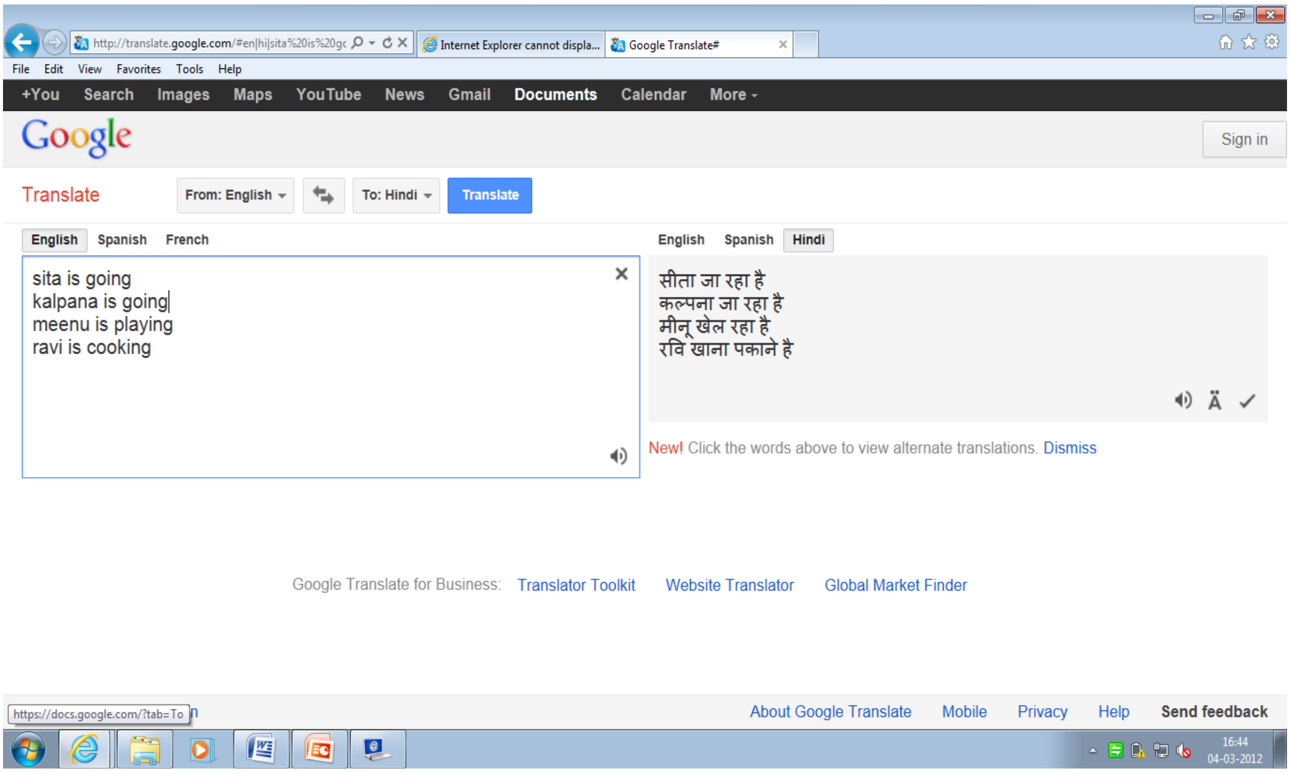
2. Literature Review

The named-entity recognition task has been implemented by two types of methods. The first one use regular-expression like dictionariesand patterns is known as rule-based methods. When the large dictionaries contain the number of entities and the patterns are extracted from a large scale corpus [7]. Rule-based method provides high performance but the cost of constructing large dictionaries is very high and the extraction of rule is also difficult.

The second method is to use the statistical information for this task [6]. In this method we collect the statistical knowledge and then determine the category of the entity using knowledge. Currently research is being done into statistical machine translation and example based machine translation [2].

3. Classification of Gender of English Names

From the point of view of machine translation process, the task of classifying the gender for a given English name is a kind of classification task. For example, for a given name n , the task is to determine gender g € G = {male,female} . For training set A = {(n_1,g_1), (n_2,g_2),.......,(n_n,g_n)} contain a pair of name and its gender, the task is to find function f : n →G. The function f can have the conditional probability and various forms. When the conditional probability is used for f, Fig. 1.1 An example shows the mistake which the Google translator makes


http://translate.google.com/#en|hi|sita%20is%20gc
Internet Explorer cannot displa...
Google Translate#
File Edit View Favorites Tools Help
+You Search Images Maps YouTube News Gmail Documents Calendar More
Google
Sign in
Translate
From: English
To: Hindi
Translate
English Spanish French
English Spanish Hindi
Google Translate for Business: Translator Toolkit Website Translator Global Market Finder
https://docs.google.com/?tab=To
About Google Translate Mobile Privacy Help Send feedback
16:44
04-03-2012

sita is going
kalpana is going
meenu is playing
ravi is cooking

सीता जा रहा है
कल्पना जा रहा है
मीनू खेल रहा है
रवि खाना पकाने है

New! Click the words above to view alternate translations. Dismiss

We focus on to identify the gender of a newly entered name n’□ as follows.

$$g^* = \underset{y€\{male,female\}}{\operatorname{argmax}} \; D(y|x_1\, n’, \Theta), \qquad (1)$$

Where distribution D, ϴ is a parameter for distribution D and it is calculated from the training set/data A.
$n = (x_1, x_2, x_3, \ldots\ldots, x_m)$, for each syllable within a name is x and the maximum length of a name is m.
Therefore, the probability of the name n is:

$$P(n) = P(x_1, x_2, \ldots\ldots\ldots, x_m)$$
$$= P(x_1)P(x_2|x_1)\ldots\ldots\ldots P(x_m|x_1, \ldots\ldots, x_{m-1}) \qquad (2)$$

It is rewritten as

$$P(x_j|x_1, \ldots\ldots, x_{j-1}) = \frac{C(x_1, \ldots\ldots, x_j)}{C(x_1, \ldots\ldots\ldots, x_{j-1})} \qquad (3)$$

Where counting function is C.

A name is considering as a sequence of syllables and considers it as an n-gram model. Means long name n implies high dimensionality in this model. In equation (3) the probability is to be zero in high dimensionality. In equation (2). It leads the value of P (n) zero. Hence, a as result it is impossible to distinguish the gender of a given name n using probability. In order to overcome this problem, weadopts the Back-off model in this paper by Katz [4].

4. Overall Description

For determining the gender of English names, we use a subsystem for generating pronouns in English-Hindi machine translation system [1]. The English–Hindi machine translation system consists of five steps: The given source text, morphological analysis, syntactic analysis, semantic and contextual generation, target language generation. Thus, proposed method is used in the target sentence generation step.

For the given source text , it first identify the named entities by named-entity recognizer. For each named-entity found, if it is a person, then try to identify the gender by the proposed method. After identifying the gender, pronoun is generated.

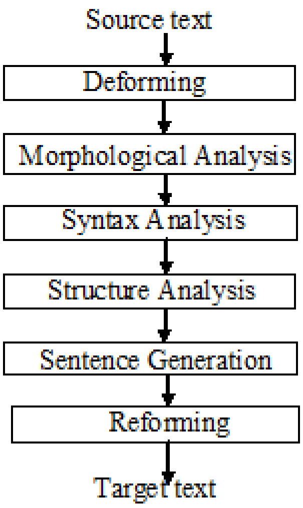


Figure 2: Machine translation process

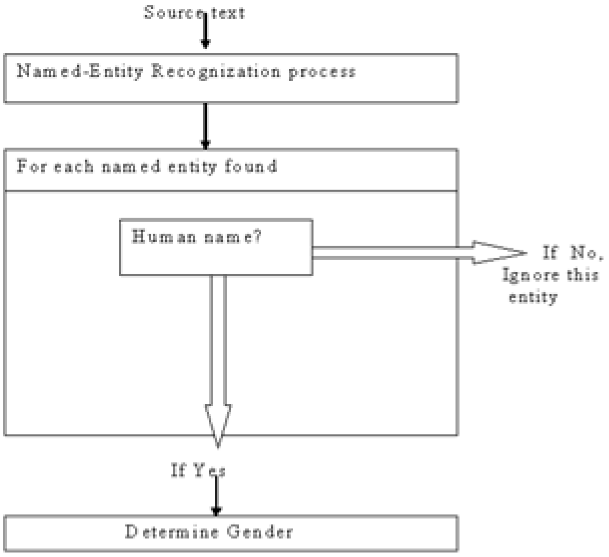

Source text
Named-Entity Recognization process
For each named entity found
Human name?
If No, Ignore this entity
If Yes
Determine Gender

Figure 3: The procedure for generating an appropriate gender for a human name.

In English-Hindi machine translation, the process of pronoun generation for proper nouns. The algorithm for generating an appropriate pronoun P for a human name N with its gender G.

```
Algorithm: Determine pronoun of the given sentence
Input     : A given name N and its gender G
Output    : A pronoun P
[step 1]    num = identify the number of name N
[step 2]    if num = plural then P = they
[step 3]    else if G = female then P = she
[step 4]    else P = he.
```

5. Conclusion

In this paper, we have proposed a method for determining the gender of English names for English-Hindi machine translation. The proposed solution is base on the statistics computed with a collection of English names. The proposed method give 90% of accuracy, while we achieves only 62.11% accuracy by database looking-up. That is, the proposed method improves 27.89% of accuracy over the simple looking-up. It implies that the proposed method is plausible for gender classification of English names in English-Hindi machine translation systems.

6. Future Work

As a future work, we will work on common name problems to improve the accuracy. Our work is limited for only human names.

References

[1] Argamon, S.,Koppel, M., J. And Shimoni, A. R. 2003. Gender, genre, and wrs.iting style in formal written text.

[2] E-S Chung, Y-G Hwang, and M-G Jung, "Korean Named Entity Recognition Using HMM and Co-Training Model", In proceedings of the 6th International Workshop on Information Retrieval with Asian languages, 2003.

[3] Bird, S., Klein, E., Loper, E. Et al. 2009. NLTK; Natural Language Processing Toolkit.

[4] K.-N.Kim, Y.-H. Yoon, J.-Y. Seo, and H,-S. Kim, "Named Entity Recognition Using Acyclic Weighted Digraphs: A Semi- Supervised Statistical Method", Lecture Notes in Computer Science, 2005.

[5] Hee-Geun Yoon, Seong-Bae Park, Yong- Jing Han and Sang- Jo Lee "Determining Gender of Korean Names with Context",In proceeding of the International Conference on Advanced Language Processing and Web Information Technology,2008.

[6] Anke Frank, Christiane Hoffmann, Maria Strobel "Gender Issues in Machine Translation".

[7] C.-N. Seon, Y.-J. Ko, J. Kim, and J.-Y. Seo, "Named Entity Recognition Using Machine Learning Methods and Pattern-Recognition Rules", In proceedings of the 6th Natural Language Processing Pacific Rim Symposium, 2001.