IMPROVING PERFORMANCE OF HINDI TO ENGLISH PBMT USING EFFICIENT PHRASE MANAGEMENT TECHNIQUES

SUSMITA GUPTA

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2018

IMPROVING PERFORMANCE OF HINDI TO ENGLISH PBMT USING EFFICIENT PHRASE MANAGEMENT TECHNIQUES

by

SUSMITA GUPTA

Department of Mathematics

Submitted
in fulfillment of the requirements of the degree of doctor of philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2018

CERTIFICATE

This is to certify that the thesis entitled ***Improving Performance of Hindi to English PBMT Using Efficient Phrase Management Techniques*** submitted by ***Mrs. Susmita Gupta*** to the Indian Institute of Technology Delhi, for the award of the Degree of the **Doctor of Philosophy**, is a record of the original bona fide research work carried out by him under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Prof. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology Delhi

ACKNOWLEDGEMENTS

I am deeply indebted to *Mother Saraswati*; the Goddess of Knowledge. Due to *Her* divine grace I have been able to carry out this research work.

This thesis would not have been possible without the support of many people. It is impossible to acknowledge all of them here, but there are few people who deserve a special mention.

I am forever indebted to my Ph.D. supervisor **Professor Niladri Chatterjee** for teaching me basics of the research. I am grateful for his unfailing support throughout this work. He was great in letting me do what I wanted, and encouraged me when the things were not going fine. He has been extremely patient with my failings, and his trust in my abilities helped me overcome many difficulties.

I thank IIT Delhi for providing me necessary facilities to carry out my research work. I thank Head of Department of Mathematics, DRC Chairperson and my SRC members for their help at various stages of my research work.

I am thankful to manager of Cadence Design Systems Ms Reenee Raizada Tayal, for granting permission to join Ph.D. in IIT Delhi. I am especially thankful to my colleagues for their constant support during my research work.

Most importantly I would like to thank many people in my personal life for giving their unwavering support throughout, especially during this research work. No word can be ever enough to thank my family for their wholehearted support for carrying out research work without bothering for the day-to-day problems. Finally, I thank to all my elders for their blessings for successfully carrying out this research work.

Last but not least, I am very much thankful to the anonymous examiners and reviewers for their valuable comments and suggestions for improvising this thesis.

Susmita Gupta

ABSTRACT

Despite the arrival of Neural Machine Translation (NMT) systems traditional Phrase-based Statistical Machine Translation (PBSMT) systems are still relevant particularly for the resource poor languages. The demand on huge volume of data that is an essential requirement for NMT systems still leads researchers on Machine Translation involving such languages to use resort to PBSMT as the preferred translation paradigm. The focus of the present research is to improve the quality of PBSMT systems by using Natural Language based techniques. The schemes developed have been tested for Hindi to English MT. Translations involving resource-poor languages (e.g. Hindi) are often constrained by limited parallel corpora which is an essential knowledge source for any SMT system. To overcome this limitation, we have prescribed four different techniques at different stages of the PBSMT system. The four schemes are:

1. Efficient Pruning of the Phrases. The scheme proposed is aimed at pruning the phrases that are not pertinent for the translation task at hand. This ensures good quality translation at a much smaller time.

2. Dynamic Phrase Table Generation. The scheme proposed here uses a Phrase Table created dynamically, for a given the input sentence. This helps in generating only the useful phrases for the given translation task.

3. Improved Similarity Measurement of Source Language Phrases with special attention to Hindi. The scheme proposed helps in finding similar phrases which can help in generating translation for the given input sentence.

4. Developing Efficient Recombination Rules. As the translation for a given input is generated phrase by phrase, it is important that the generated translations follow the linguistic rules of target language. The proposed scheme aims at achieving this goal.

The combined effect of the proposed scheme is measured and improvement in translation quality and/or the time requirement while maintaining the quality is achieved. The proposed scheme should be suitable for translation involving Indian languages other than Hindi as well.

सार

तंत्रिका मशीन अनुवाद (एनएमटी) सिस्टम के आगमन के बावजूद परंपरागत वाक्यांश आधारित सांख्यिकीय मशीन अनुवाद (पीबीएसएमटी) सिस्टम अभी भी संसाधन गरीब भाषाओं के लिए प्रासंगिक हैं। एनएमटी सिस्टम के लिए डेटा (तथ्य) की भारी मात्रा में मांग की आवश्यक अभी भी मशीन अनुवाद पर शोधकर्ताओं को दूसरे तरीको पर ले जाती है, जिसमें पीबीएसएमटी का उपयोग पसंदीदा अनुवाद प्रतिमान के रूप में करने के लिए किया जाता है। वर्तमान शोध का ध्यान प्राकृतिक भाषा आधारित तकनीकों का उपयोग करके पीबीएसएमटी सिस्टम की गुणवत्ता में सुधार करना है। विकसित योजनाओं का हिंदी से अंग्रेजी एमटी के लिए परीक्षण किया गया है। संसाधन-गरीब भाषाओं (जैसे हिंदी) को शामिल करने वाले अनुवाद अक्सर समान समांतर निगम द्वारा बाधित होते हैं जो कि किसी भी एसएमटी प्रणाली के लिए एक आवश्यक ज्ञान स्रोत है। इस सीमा को दूर करने के लिए, हमने पीबीएसएमटी प्रणाली के विभिन्न चरणों में चार अलग-अलग तकनीकों को निर्धारित किया है। चार योजनाएं हैं:

1. वाक्यांशों का कुशल कटौती। प्रस्तावित योजना का उद्देश्य उन वाक्यांशों को छेड़छाड़ करना है जो अनुवाद कार्य के लिए प्रासंगिक नहीं हैं। यह बहुत कम समय पर अच्छी गुणवत्ता वाले अनुवाद सुनिश्चित करता है।

2. गतिशील वाक्यांश तालिका जनरेशन। यहां प्रस्तावित योजना इनपुट वाक्य के लिए, गतिशील रूप से बनाई गई वाक्यांश तालिका का उपयोग करती है। यह दिए गए अनुवाद कार्य के लिए केवल उपयोगी वाक्यांश उत्पन्न करने में मदद करता है।

3. हिंदी पर विशेष ध्यान देने के साथ स्रोत भाषा वाक्यांशों के बेहतर समानता मापन। प्रस्तावित योजना समान वाक्यांश खोजने में मदद करती है जो दिए गए इनपुट वाक्य के लिए अनुवाद उत्पन्न करने में मदद कर सकती हैं।

4. कुशल पुनर्संरचना नियम विकसित करना। चूंकि दिए गए इनपुट के लिए अनुवाद वाक्यांश द्वारा वाक्यांश उत्पन्न होता है, इसलिए यह महत्वपूर्ण है कि जेनरेट किए गए अनुवाद लक्ष्य भाषा के भाषाई नियमों का पालन करें। प्रस्तावित योजना का लक्ष्य इस लक्ष्य को प्राप्त करना है।

प्रस्तावित योजना के संयुक्त प्रभाव को गुणवत्ता बनाए रखने के दौरान अनुवाद गुणवत्ता में सुधार और / या समय की आवश्यकता में सुधार किया जाता है। प्रस्तावित योजना हिंदी के अलावा अन्य भाषाओं में अनुवाद के लिए उपयुक्त होना चाहिए।

Table of Contents

List of Figures

List of Tables