English-Persian Phrase-based Statistical Machine Translation: Enhanced Models, Search and Training

A THESIS PRESENTED IN FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER ENGINEERING

at Massey University
Albany (Auckland), New Zealand

Mahsa Mohaghegh

2012

ABSTRACT

Machine translation (MT), as applied to natural language processing, has undergone substantial development over the past sixty years. While there are a number of different approaches to MT, there has been increasing interest in statistical machine translation (SMT) as the preferred approach to MT. Advances in computational power, together with the exploration of new methods and algorithms have enabled a general improvement in the output quality in a number of systems for various language pairs using this approach. However, there is a significant lack of research work in the area of English/Persian SMT, mainly due to the scarcity of data for this language pair, and the shortage of fundamental resources such as large-scale bilingual corpora. Several research studies have been published on work in the area of machine translation involving the Persian language; however, results producing fluent, usable output are rare.

This thesis shows how SMT was implemented with this language pair for the first time, and how we created a cutting-edge hybrid SMT system capable of delivering high-quality translation output.

We present the development of what is currently the largest English/Persian parallel corpus, constructed using a web crawler to source usable online data, together with the concatenation of existing parallel corpora. As yet another contribution of the research, we propose an improved hybrid corpus alignment method involving sentence length-based and word correspondence-based models to align words, phrases and sentences in the corpus. We also show the impact that the corpus domain can have on the translation output, and the necessity to consider domains of both bilingual and monolingual corpora where they are included in the training and language models.

Two open-source toolkits, Moses and Joshua, were modified to work with the Persian language, and their behaviour and performance results were compared to determine which performed better when implemented with the Persian language.

We present our work in designing, testing, and implementing a novel, three-level Transfer-based automatic post-editing (APE) component based on grammatical rules, which operates by analysing, parsing, and POS-tagging the output, and implements functions as transformers which perform corrections to the text, from lexical

transformation to complex syntactical rearrangement. We show that rule-based approaches to the task of post-editing are superior to the commonly-used statistical models, since they incorporate linguistic knowledge, and are strong in terms of syntax, morphology, and structural semantics – qualities which are very desirable when performing grammatical correction and syntactical restructuring.

We implement independent manual evaluation as well as standard automatic techniques, in order to assess more accurately the translation output. This evaluation shows that the use of the APE component is able to improve translation output significantly, that is, by at least 25%, resulting in high-quality translation output.

Our system performs well by using a combination of the capabilities of two main MT approaches – SMT and RBMT – in different areas of the system as a whole. SMT provides the main system with consistent, mathematical-based translation, and the Transfer-based algorithm in the APE component operates with comprehensive linguistic rules in order to improve incorrect sentences, and fine-tune translation output. This results in a robust, state-of-the-art system, which noticeably exceeds other currently available solutions for this language pair.

ACKNOWLEDGEMENTS

In the name of God who owns soul and wisdom. These are the best attributes of God.
*~ **Ferdowsi** (935 – 1020)*

Firstly, I would like to express my gratitude to my advisors Professor Dr.Hossein Sarrafzadeh, and Dr. Rezaul Hasan. This thesis would not have been possible without their support, advice, and valuable ideas and suggestions.

I would also like to thank my committee member Dr. Tom Moir for his constant support and valuable advice and critical comments. His advice and patience was much appreciated.

I wish to extend my heartfelt thanks to my parents, whose continuous encouragement lightens my path into higher education.

At this point, I would like to express my everlasting gratitude to my best friend for his patience, support, and love. This thesis would not have been possible without Mike's continuous encouragement and patience. Thank you for reminding me that there are things more important than this work, and for standing next to me during this time.

I also want to thank my best friend Dr. Mandana Arzaghi, who pointed out early that there are life and opportunities also outside of Iran. During these last years, geographical distances were against seeing more each of each other, but you were always there for me as a true friend and a good example of perseverance and success. I know you will always be there as a good influence for me.

I would also like to thank Mehdi Mohammadi who has been working at the University of Shikh Bahaee for the many fruitful discussions and joint experiments related to Automatic Post-Editing and Hybrid translation approaches.

My PhD studies turned out to be an unforgettable experience, mostly thanks to the support from my colleagues and friends. I was lucky to be part of a great group of scientifically ambitious, intelligent, and industrious researchers at Massey e-Centre.

Lastly, I offer my regards and appreciation to all of those who have supported me in many aspects during the course of the research, my friends, family, colleagues, reviewers, etc, and anyone else not mentioned here.

Dedicated to my parents

DECLARATION

The author declares that this is her own work except where due acknowledgement has been given. It is being submitted for a PhD in Engineering to Massey University, New Zealand.

This Thesis describes the research carried out by the author at the School of Engineering, Massey University, New Zealand, from June 2008 to December 2012, supervised by Dr Rezaul Hasan.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACL	Complement Clause of Adjective
ADV	Adverb
ADVC	Adverbial Complement of Verb
AJCONJ	Conjunction of Adjective
AJPP	Prepositional Complement of Adjective
AJUCL	Adjunct Clause
AOL	America Online
APE	Automatic Post Editing
APOSTMOD	Adjective Post-Modifier
APP	Apposition
APREMOD	Adjective Pre-Modifier
ASR	Automatic Speech Recognition
AVCONJ	Conjunction of Adverb
BLEU	Bilingual Evaluation Understudy
CAT	Computer Assisted Translation
CFGs	Synchronous Context-Free Grammars
CNW	Canada Newswire
COMPPP	Comparative Preposition
C-STAR	International Consortium for Research on Speech Translation
DARPA	Defence Advance Research Project Agency
DG	Dependency Grammar
EBMT	Example-Based Machine Translation
EGIU	English Grammar in Use
ELRA	European Language Resource Association
EM	Expectation–Maximization Algorithm
En	English
ENC	Enclitic Non-Verbal Element
Fa	Farsi
FAMT	Fully Automatic Machine Translation
FLDB	Farsi Linguistic Database
FST	Finite State Transducer
FTD	US Air Force's Foreign Technology Division
GLP	Gross Language Product
HAMT	Human-Assisted Machine Translation
Hiero	Hierarchical
HMT	Hybrid Machine Translation
IBM	International Business Machines Corporation
IEEE	Institute of Electrical and Electronics Engineers, Inc.
IR	Information Retrieval
IRNA	Iranian News Agency
IWSLT	International Workshop on Spoken Language Translation
LG	Link Grammar
LV	Linking Verb
LVP	Light Verb Particle
MAHT	Machine-Assisted Human Translation
MAP	Maximum A-Posteriori
MERT	Minimum Error Rate Training
MESU	Measure
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MLE	Maximum Likelihood Estimation
MOS	Mosnad

MOZ	Ezafe Dependent
MPEC	Modern Persian-English corpus
MST	Parser Maximum Spanning Tree Parser
MT	Machine Translation
NADV	Adverb of Noun
NCL	Clause of Noun
NCONJ	Conjunction of Noun
NE	Non-Verbal Element of Infinitive
NEZ	Ezafe Complement of Adjective
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NPEC	News Persian English Corpus
NPOSTMOD	Post-Modifier of Noun
NPP	Preposition 0f Noun
NPREMOD	Pre-Modifier of Noun
NPRT	Particle of Infinitive
NSPEC	News Subtitle Persian-English Corpus
NVE	Non-Verbal Element
ODJ	Object
ODJ2	Second Object
OOV	Out of Vocabulary
PAHO	Pan American Health Organization
PARCL	Participle Clause
PART	Interrogative Particle
PB	Phrase Based
PCONJ	Conjunction of Preposition
PCTS	Parallel Corpus Test Set
PeEn-SMT	Persian-English Statistical Machine Translation
PEN	Parallel English-Persian News Corpus
POS	Part of Speech
POSDEP	Post-Dependent
PPL	Perplexity Threshold
PREDEP	Pre-Dependent
PREM	Pre-Modifier
PRO	Predicate
PROG	Progressive Auxiliary
PUNC	Punctuation Mark
RBMT	Rule-Based Machine Translation
RHS	Right Hand Side
ROOT	Root
SAMT	Syntax Augmented Machine Translation
SBJ	Subject
SCFG	Stochastic Context-Free Grammar
SDL	Scalable Enterprise Translation Server
SDL	SDL Language Weaver
SMT	Statistical Machine Translation
SOV	Subject-Object-Verb
SRILM	Sri Language Model
SVO	Subject–Verb-Object
TAM	Tamiz
TEP	Tehran English-Persian Corpus
TER	Translation Error Rate Te
TER	Translation Error Rate
TMC	Tehran Monolingual Corpus
U	Unicode

UN	United Nations
UTIRE	University of Tehran Information Retrieval Evaluation System
VCL	Complement Clause of Verb
VCONJ	Conjunction of Verb
VPP	Prepositional Complement of Verb
VPRT	Verb Particle
WER	Word Error Rate
WSD	Word Sense Disambiguation

LIST OF PUBLICATIONS

1. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, Mehdi Mohammadi, ***GRAFIX: Automated Rule-Based Post Editing System to Improve English-Persian SMT*** (Short paper- COLING 2012, the 24th International Conference on Computational Linguistics. Mumbai, India, December 2012)
   http://aclweb.org/anthology-new/C/C12/C12-2085.pdf

2. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, ***A Hierarchical Phrase-Based Model for English-Persian Statistical Machine Translation*** (Full Paper - Innovations 12, 8th International Conference on Innovations in Information Technology. AL AIN ,UAE, March 2012)
   http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6207733

3. **Mahsa Mohaghegh**, ***Advancements in English-Persian Hierarchical Statistical Machine Translation.*** (Short paper - NZCSRSC New Zealand Computer Science Research Student Conference. Otago, April 2012)
   https://sites.google.com/a/nzcsrsc.ac.nz/nzcsrsc2012/programme/posters

4. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, Tom Moir, ***Improving Persian-English Statistical Machine Translation: Experiments in Domain Adaptation*** (Full paper – *In Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, IJCNLP 2011, pages 9–15,Chiang Mai, Thailand, November 2011)
   http://www.aclweb.org/anthology/W/W11/W11-3002.pdf

5. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, ***An Overview of the Challenges and Progress in PeEn-SMT: First Large Scale Persian-English SMT System*** (Full Paper - Innovations 11, 7th International Conference on Innovations in Information Technology. Abu Dhabi, April 2011)
   http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5893841

6. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, ***The Impact of Domain for Language Model in the PeEn-SMT: First Large Scale Persian-English SMT System***. (Short paper-NZCSRSC New Zealand Computer Science Research Student Conference. Palmerston North, April 2011)

   https://sites.google.com/a/maori.geek.nz/nzcsrsc2011/papers/paper-abstracts#TOC-Mahsa-Mohaghegh-The-Impact-of-Domain-for-Language-Model-in-the-PeEn-SMT:-First-Large-Scale-Persian-English-SMT-System-

7. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, ***Multilingual Information Service System for Tourists*** (Poster – *NZBio Conference*, Auckland, March 2011)

8. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh, Tom Moir**.** ***Improved Language Modelling for English-Persian Statistical Machine Translation.*** (Full Paper - *In Proceedings of SSST-4 Workshop at COLING-2010*, Beijing, China, August 2010)

http://www.aclweb.org/anthology-new/W/W10/W10-3810.pdf

9. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh. ***Performance Evaluation of Statistical English-Persian Machine Translation.*** (Full Paper – JADT2010 – *10th International Conference on Statistical Analysis of Textual Data.* Sapienza, University of Rome, June 2010)

   http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1091-1100_114-Sarrafzadeh.pdf

10. **Mahsa Mohaghegh**. ***A Statistical Approach to English-Persian Machine Translation.*** (Short Paper – NZCSRSC – *New Zealand Computer Science Research Student Conference.* Wellington, April 2010)

    http://ecs.victoria.ac.nz/Events/NZCSRSC2010/Papers

11. **Mahsa Mohaghegh**, Tom Moir, Abdolhossein Sarrafzadeh. ***A Statistical Approach to English-Persian Machine Translation.*** (Poster – *NZBio Conference*, Auckland, March 2010)

    http://www.academia.edu/238736/A_Statistical_Approach_to_English-Persian_Machine_Translation

12. **Mahsa Mohaghegh**, Abdolhossein Sarrafzadeh. ***Analysis of the Effect of Data Variation in a Statistical English-Persian Machine Translation.*** (Full Paper - Innovations '09 – *6th International Conference on Innovations in Information Technology*. Dubai, December 2009.

    http://dl.acm.org/citation.cfm?id=1802285
    http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05413782

13. **Mahsa Mohaghegh**, Tom Moir. ***First English-Persian Machine Translation.*** (Poster – NZPGC – *New Zealand Postgraduate Conference*. Wellington, November 2009)

    http://www.academia.edu/238733/the_first_english-persian_statistical_machine_translation