Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.
Fluency Enhancement

Applications to Machine Translation

By

Steve Lawrence Manion

2009
ABSTRACT

The quality of Machine Translation (MT) can often be poor due to it appearing incoherent and lacking in fluency. These problems consist of word ordering, awkward use of words and grammar, and translating text too literally. However we should not consider translations such as these failures until we have done our best to enhance their quality, or more simply, their fluency. In the same way various processes can be applied to touch up a photograph, various processes can also be applied to touch up a translation. This research outlines the improvement of MT quality through the application of Fluency Enhancement (FE), which is a process we have created that reforms and evaluates text to enhance its fluency.

We have tested our FE process on our own MT system which operates on what we call the SAM fundamentals, which are as follows: Simplicity - to be simple in design in order to be portable across different languages pairs, Adaptability - to compensate for the evolution of language, and Multiplicity - to determine a final set of translations from as many candidate translations as possible. Based on our research, the SAM fundamentals are the key to developing a successful MT system, and are what have piloted the success of our FE process.
PREFACE

The main objective of this research was to build a SAM based MT system that used our FE process to improve the quality (fluency) of its output. We have successfully completed our objective, however what was not expected was that we could also extract our FE process from the MT system, and decide whether it functioned as a built-in, or as an add-on capacity. Consequently, we also found that we could apply our FE process to other MT systems and language applications that are outside the scope of MT. Lastly we also found the performance our FE process improves if there is more linguistic data accessible to the MT system. Therefore the scope of this research has also been expanded to include methods of obtaining and structuring larger and more superior linguistic data. The objectives of this research have been broken down and discussed at the end of the first chapter of this thesis. Fig. 2 below illustrates the diverse applications of our FE process.

**FIG. 2. BRIEF DESCRIPTIONS OF HOW FE CAN BE USED IN MULTIPLE APPLICATIONS**
Two papers were published and presented through the following peer reviewed conferences:

**Self Learning Live Translation System (PowerPoint Presentation)**

The 3\textsuperscript{rd} International Conference on Convergence and hybrid Information Technology
Busan, South Korea, 11\textsuperscript{th} – 13\textsuperscript{th} November 2008

**Fluency Enhancement of Machine Translation (Poster Presentation)**

To be published in the ICSPCS proceedings / ISBN: 978-0-9756934-6-9
ICSPCS – The 2\textsuperscript{nd} International Conference on Signal Processing and Communication Systems
Gold Coast, Australia, 15\textsuperscript{th} – 17\textsuperscript{th} December 2008
Intellectual Property Statement

This research was conducted in collaboration with Kaleido K and Massey University. As the student (Steve L. Manion) belongs to Kaleido K, the research will also continue on after the submission of this thesis. Aspects of this research are sensitive; in particular anything related what is referred to as “Fluency Enhancement”. For this reason software code has not been provided in the publication of this research and only abstract diagrams have been used to illustrate software functionality.

Kaleido K is now in the process of commercializing this research. What is listed below may be of interest to those who wish to follow the activities of Kaleido K.

Kaleido K Patent

Kaleido K owns the pending patent 573943
Fluency Enhancement – A Process that Reforms and Evaluates Text to Enhance its Fluency

Kaleido K Website

http://www.kaleidok.com
Our MT system, FE technology and linguistic resources will be accessible here in the near future; we encourage you to visit, and join the mailing list in the mean time

Linguistic Data Consortium

http://www.ldc.upenn.edu/
The corpora developed with the web crawler built in this research is currently in negotiations to be published and distributed by the Linguistic Data Consortium

Kaleido K Language Community

The Facebook group which aids and contributes to the development of Kaleido K technology
ACKNOWLEDGEMENTS

Contributing Individuals / Institutions / Enterprises

Amal Punchihewa
Thesis Supervisor

Kaleido K
Developers of Bilingual Corpora / Korean to English Dictionary

Seul Hwa Lee / Wang Hyu Lee / Christina Manion / Benji Morgan
Website & Survey Translation / Survey Distribution

Massey University – Kathy Hamilton / Trish O’Grady / Gayle Leader
Thesis Administration / Funding and Support to attend ICCIT 08 and ICSPCS 08 Conferences

Project Resources

Google’s Web 1T 5-Gram Corpus – N-Gram Data used to obtain initial FE test results
Microsoft Server SQL 2005 – Used to store and index the N-Gram data
Windows Server 2003 – Required to run Microsoft Server SQL 2005
NetBeans IDE 6.0 – Used to construct FE, the MT system and web crawler in Java
Several Computers – Used for testing FE, the MT system and the web crawler
SpeedTest.com – Used for estimating internet speeds for web crawler calculations
# Table of Contents

1. **Introduction** ........................................................................................................... 1
   1.1 About the Author ........................................................................................................ 1
   1.2 Organization of Thesis .............................................................................................. 3
     1.2.1 Content Layout .................................................................................................... 3
     1.2.2 Appendices Layout ............................................................................................ 5
   1.3 What is Machine Translation? .................................................................................. 6
   1.4 Description of Problem ........................................................................................... 7
     1.4.1 Background ......................................................................................................... 7
     1.4.2 Machines & Language Translation ....................................................................... 10
     1.4.3 Language Models ................................................................................................ 11
     1.4.4 The Obstacles of Translating Language .............................................................. 14
   1.5 Proposed Solution .................................................................................................... 17
     1.5.1 Motivations ....................................................................................................... 17
     1.5.2 Our Machine Translation System’s Design Fundamentals ................................. 18
     1.5.3 Objectives .......................................................................................................... 21

2. **Academic Literature Review** .................................................................................. 23
   2.1 Methods of Translation ......................................................................................... 23
     2.1.1 Rule Based Machine Translation ....................................................................... 23
     2.1.2 Corpus Based Machine Translation ................................................................... 29
     2.1.3 Hybrid Machine Translation ............................................................................. 37
   2.2 Evaluation of Machine Translation ......................................................................... 40
     2.2.1 The BLEU Score ............................................................................................... 41
     2.2.2 BLEU Score Example 1 .................................................................................... 42
     2.2.3 BLEU Score Example 2 .................................................................................... 43
     2.2.4 Modified N-Gram Precision .............................................................................. 43
     2.2.5 Raising the Order of the N-Gram Precision ....................................................... 44
     2.2.6 Example of BLEU Score Calculation ................................................................. 45
     2.2.7 Alternative Evaluation Techniques .................................................................... 48
   2.3 Who leads Machine Translation? ........................................................................... 52
     2.3.1 SYSTRAN ......................................................................................................... 53
     2.3.2 Google ............................................................................................................. 54
     2.3.3 Carnegie Mellon ............................................................................................... 54
     2.3.4 Who is the leader? ............................................................................................. 55
   2.4 A Brief History of Machine Translation Methods .................................................. 56
     2.4.1 The Ignition of the Industry – (Pre 1954) ............................................................ 58
     2.4.2 Realizations & Reductions in Funding (1955 – 1966) ......................................... 59
     2.4.3 Dormant Times (1967 – 1976) ........................................................................... 60
     2.4.5 The Diverse Needs & Solutions of the Translation Industry (Post 1990) ........ 62
   2.5 Summarization of the Machine Translation Industry ............................................. 64

3. **Introduction to Fluency Enhancement** ................................................................... 68
   3.1 A Brief Outline ....................................................................................................... 68
## 6. CONCLUSION ................................................................................................................................... 130

6.1 MEETING THE OBJECTIVES .................................................................................................................... 130

6.2 ADVANTAGES & DISADVANTAGES .......................................................................................................... 132

6.3 FUTURE WORK ................................................................................................................................... 133

6.4 CLOSING STATEMENT & FINAL THOUGHTS ............................................................................................... 135

## 7. APPENDICES ..................................................................................................................................... 141

7.1 APPENDIX A ...................................................................................................................................... 141


7.2 APPENDIX B ....................................................................................................................................... 132

7.2.1 English Market Research Survey ................................................................................................ 132

7.2.2 Korean Market Research Survey ................................................................................................ 139

7.2.3 Raw Data of Market Research Survey ....................................................................................... 146

7.3 APPENDIX C ....................................................................................................................................... 154

7.3.1 Abbreviations ............................................................................................................................ 154

7.3.2 Technical Terms ......................................................................................................................... 155

7.3.3 References ................................................................................................................................... 156
LIST OF FIGURES

FIG. 1. THE SAM FUNDAMENTALS .................................................................................................................. 2
FIG. 2. BRIEF DESCRIPTIONS OF HOW FE CAN BE USED IN MULTIPLE APPLICATIONS ........................................... 3
FIG. 3. ICSPCS POSTER PRESENTATION PHOTOS .............................................................................................. 4
FIG. 4. KALEIDO K LANGUAGE COMMUNITY ...................................................................................................... 1
FIG. 5. NEW ZEALAND ........................................................................................................................................ 2
FIG. 6. THE RESULT OF MACHINE TRANSLATION MISCONCEPTIONS ....................................................................... 8
FIG. 7. REPRESENTATION OF THE CURRENT RELIEF MACHINE TRANSLATION PROVIDES ............................................ 9
FIG. 8. THE 19TH CENTURY AMERICAN ICON OF EASE[1] ............................................................................. 12
FIG. 9. THE VISUAL (AND OTHER) INPUTS MACHINE TRANSLATION REQUIRES TO SOLVE LANGUAGE AMBIGUITIES .... 15
FIG. 10. RESEARCH OBJECTIVES ...................................................................................................................... 21
FIG. 11. DICTIONARY REQUIREMENT MODEL FOR DIRECT MACHINE TRANSLATION .............................................. 24
FIG. 12. DICTIONARY REQUIREMENT MODEL FOR INTERLINGUAL MACHINE TRANSLATION ..................................... 25
FIG. 13. BRIDGING DISTANT LANGUAGE PAIRS USING MULTIPLE INTERLINGUAL LANGUAGES .......................... 26
FIG. 14. DICTIONARY REQUIREMENT MODEL FOR TRANSFER BASED MACHINE TRANSLATION ............................ 27
FIG. 15. INCREASE IN DICTIONARIES REQUIRED FOR RULE BASED MACHINE TRANSLATION METHODS ......... 28
FIG. 16. AN EXAMPLE OF GENERAL STATISTICAL MACHINE TRANSLATION ARCHITECTURE ............................... 31
FIG. 17. EXPECTATION MAXIMIZATION – PARAMETER ESTIMATION FROM THE CONNECTED CORPUS (FRENCH TO ENGLISH) . 32
FIG. 18. WORD ALIGNMENT INDUCED PHRASES (SPANISH TO ENGLISH) ......................................................... 33
FIG. 19. ENGLISH TO JAPANESE SYNTAX TREE ................................................................................................. 34
FIG. 20. BEAM WORD GRAPH ............................................................................................................................ 35
FIG. 21. BLEU SCORES ON UNDER INCREASING AMOUNTS OF TRAINING DATA FOR PORTAGE SMT ALONE AND SYSTRAN MT WITH PORTAGE APE[8]. .......................................................... 37
FIG. 22. ARCHITECTURE FOR A MULTI-ENGINE MACHINE TRANSLATION driven by a Statistical Machine Translation Decoder[9] ................................................................................................................................. 38
FIG. 23. DISTINGUISHING HUMAN FROM MACHINE ......................................................................................... 44
FIG. 24. MACHINE TRANSLATION EVALUATION OF ADEQUACY CORRELATION[12] .................................................... 50
FIG. 25. DEVELOPMENT FLOW OF THE MACHINE TRANSLATION INDUSTRY WITH CHRONOLOGICAL ORDER OF KEY EVENTS ... 57
FIG. 26. ADVANTAGES AND DISADVANTAGES OF EACH MT APPROACH ................................................................. 64
FIG. 27. THE VAUQUOIS TRIANGLE .................................................................................................................. 65
FIG. 28. DEVELOPMENT STEPS OF EACH MACHINE TRANSLATION APPROACH [17] ............................................ 66
FIG. 29. THE FLUENCY ENHANCEMENT PROCESS ............................................................................................... 69
FIG. 30. THE INFLUENCE OF PACKAGING AND LOCALIZATION PRE-PROCESSING ON FLUENCY ENHANCEMENT .......... 71
FIG. 31. REORDERING ........................................................................................................................................ 73
FIG. 32. SYNONYM SWAPPING .......................................................................................................................... 73
FIG. 33. PART OF SPEECH ADDITION ..................................................................................................................... 73
FIG. 34. PART OF SPEECH REMOVAL ................................................................................................................... 74
FIG. 35. MORPHING .............................................................................................................................................. 74
FIG. 36. PARAPHRASING .................................................................................................................................... 74
FIG. 37. SPELLING ................................................................................................................................................. 74
FIG. 38. LOCALIZING ............................................................................................................................................ 75
FIG. 39. PUNCTUATING ....................................................................................................................................... 75
FIG. 40. THE FLUENCY ENHANCEMENT PROCESS .................................................................................................. 75
FIG. 41. THE PACKAGER – INPUTS & OUTPUTS .................................................................................................... 78
FIG. 42. THE PACKAGER – INTERNAL OPERATIONS ............................................................................................ 79
LIST OF TABLES

TABLE 1. DIRECT MACHINE TRANSLATION FROM GERMAN TO ENGLISH ............................................................ 23
TABLE 2. DIRECT MACHINE TRANSLATION FROM JAPANESE TO ENGLISH .......................................................... 24
TABLE 3. EXAMPLE BASED TRANSLATION FROM JAPANESE TO ENGLISH ............................................................. 29
TABLE 4. EXAMPLE BASED TRANSLATION FROM ENGLISH TO GERMAN USING GENERALIZATION TAGS .................... 30
TABLE 5. UNTRANSLATED TOKENS (EXCLUDING NUMBERS AND PUNCTUATIONS) IN OUTPUT FOR NEWS COMMENTARY TASK (GERMAN TO ENGLISH) FROM DIFFERENT MT SYSTEMS[9] ................................................................................... 39
TABLE 6. CALCULATED MODIFIED N-GRAM PRECISIONS ..................................................................................... 47
TABLE 7. GRAPHICAL REPRESENTATION OF F-MEASURE [12] ................................................................................ 49
TABLE 8. MACHINE TRANSLATION EVALUATION OF FLUENCY CORRELATION[12] ...................................................... 50
TABLE 9. NIST 2008 BLEU-4 RESULTS FOR ARABIC TO ENGLISH ........................................................................... 52
TABLE 10. NIST 2008 BLEU-4 RESULTS FOR CHINESE TO ENGLISH ......................................................................... 52
TABLE 11. NIST 2008 BLEU-4 RESULTS FOR ENGLISH TO CHINESE ......................................................................... 53
TABLE 12. NIST 2008 BLEU-4 RESULTS FOR URDU TO ENGLISH ............................................................................. 53
TABLE 13. SAMPLE N-GRAM DATA .................................................................................................................... 70
TABLE 14. EVALUATION DATA (PRESENCE INDICATION) .......................................................................................... 76
TABLE 15. FIRST SENTENCE TRANSLATION COMPARISON ...................................................................................... 104
TABLE 16. SECOND SENTENCE TRANSLATION COMPARISON .................................................................................. 105
TABLE 17. THIRD SENTENCE TRANSLATION COMPARISON ..................................................................................... 106
TABLE 18. TOP 25 FLUENCY ENHANCED TRANSLATIONS OF THE FIRST SENTENCE ...................................................... 108
TABLE 19. ADVANTAGES AND DISADVANTAGES TO FE TECHNOLOGY ....................................................................... 132

LIST OF EQUATIONS

(1) THE EQUATION USED TO CALCULATE A MODIFIED N-GRAM PRECISION ............................................................... 43
(2) THE EQUATION USED TO CALCULATE A BLEU SCORE ............................................................................................ 45
(3) THE EQUATION SET USED TO CALCULATE THE BREVITY PENALTY ............................................................................. 45
(4) EXAMPLE BREVITY PENALTY CALCULATION ......................................................................................................... 46
(5) EXAMPLE BLEU SCORE CALCULATION ........................................................................................................... 47
(6) NUMBER OF TOKENS NEEDED TO BE FOUND PER SECOND ....................................................................................... 90
(7) AVERAGE SIZE OF ENGLISH TOKEN (UNIGRAM) ................................................................................................... 90
(8) REQUIRED DATA ACQUISITION SPEED ................................................................................................................ 90
(9) AVERAGE DIFFERENCE BETWEEN THE ENHANCED AND THE ORIGINAL ACCURACY OF EACH TRANSLATION ........................ 99
(10) EXPECTED AMOUNT OF SENTENCES TO BE PROCESSED PER MONTH .................................................................... 122