Thesis For
Master of Engineering
In
Information & Telecommunications Engineering

Massey University
Palmerston North, New Zealand

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.

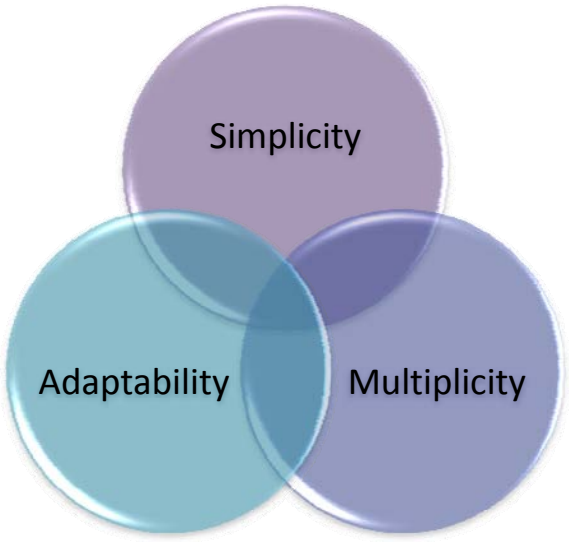


FIG. 1. THE SAM FUNDAMENTALS

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.

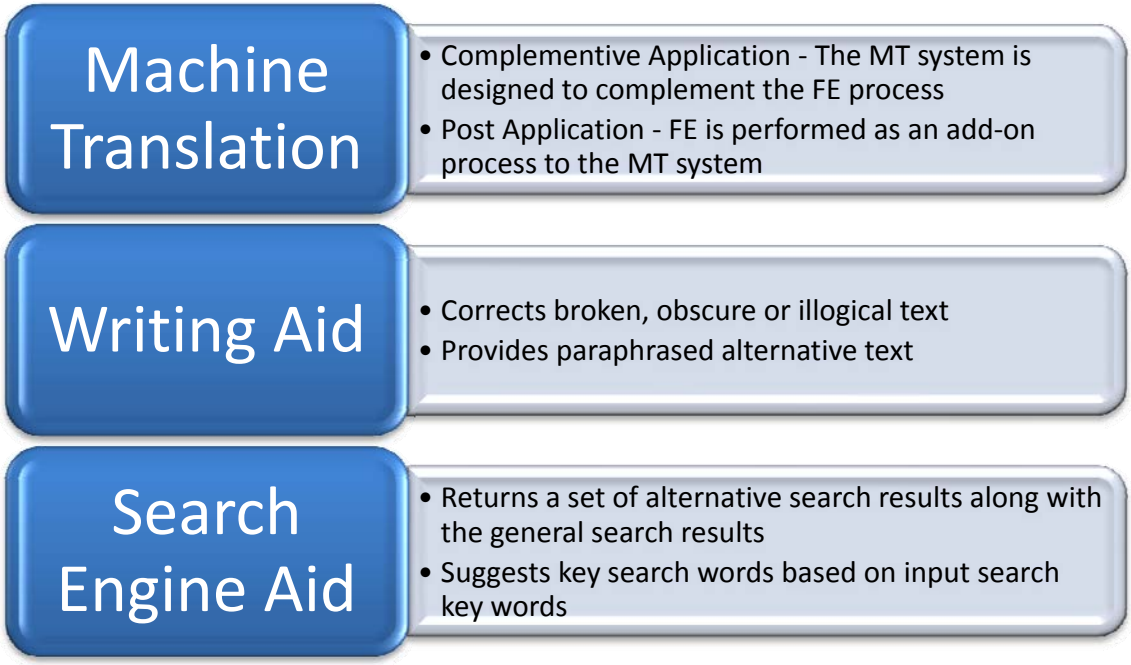


FIG. 2. BRIEF DESCRIPTIONS OF HOW FE CAN BE USED IN MULTIPLE APPLICATIONS

CONFERENCES

Two papers were published and presented through the following peer reviewed conferences:

Self Learning Live Translation System (PowerPoint Presentation)

Pages 631-639 of ICCIT proceedings / ISBN: 978-0-7695-3407-7

The 3rd International Conference on Convergence and hybrid Information Technology

Busan, South Korea, 11th – 13th November 2008

Fluency Enhancement of Machine Translation (Poster Presentation)

To be published in the ICSPCS proceedings / ISBN: 978-0-9756934-6-9

ICSPCS – The 2nd International Conference on Signal Processing and Communication Systems

Gold Coast, Australia, 15th – 17th December 2008

FIG. 3. ICSPCS POSTER PRESENTATION PHOTOS

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

http://www.facebook.com/home.php?#/group.php?gid=24918989484
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
2.4.4 Commercialization & Recovery of the Industry (1977 – 1989) 61
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

3.2 THE SUB-PROCESSES OF THE FLUENCY ENHANCEMENT PROCESS 69
3.3 STATISTICAL INFLUENCE 70
3.3.1 Web Sourced N-Grams 70
3.3.2 Massive Text Corpora 70
3.4 SPECIFIC DETAILS OF THE SUB-PROCESSES 71
3.4.1 Application Based Pre-processing 71
3.4.2 Reformation Process 72
3.4.3 Evaluation Process 76
3.4.4 Application Based Post Processing 77

4. TRANSLATION SYSTEM DESIGN 78

4.1 TRANSLATION PROCESS 78
4.1.1 Data Model of the Translation System 78
4.1.2 Application of Fluency Enhancement 84
4.1.3 Fluency Enhancement Techniques 86
4.2 DATA COLLECTION 87
4.2.1 Test Data 87
4.2.2 Expansion of Data 88
4.2.3 The Web Crawler 92
4.3 IMPLEMENTATION 96

5. RESULTS & DISCUSSION 98

5.1 FLUENCY ENHANCEMENT RESULTS 98
5.2 COMPARING DIFFERENT APPROACHES 103
5.2.1 Comparison of Competing Online Machine Translation Solutions 103
5.2.2 Example Application of Fluency Enhancement 107
5.3 WEBCRAWLER RESULTS 110
5.4 MARKET RESEARCH RESULTS 113
5.4.1 The Survey 113
5.4.2 What do you think of this tool? 116
5.4.3 How often do you think you would use this tool? 117
5.4.4 Do you think your friends would use this tool? 118
5.4.5 If you used this tool, when entering text into it, would you attempt to write in English, your native language or both? 119
5.4.6 If you used this tool, how would you like to access it? 120
5.4.7 If you were to pay for this tool, would you prefer to pay based on the length of time used or per sentence improvement? 121
5.4.8 For each payment method, how much would you expect to pay? 122
5.4.9 Have you purchased, seen or heard of a tool that restructures and improves your English writing such as this one? 123
5.4.10 Except English, are there any other languages this tool could help you improve your writing in? 124
5.4.11 What is your age group? 125
5.4.12 What is your native language? 126
5.4.13 What is your occupation? 127
5.4.14 In your everyday life, how often do you need to use English? 128

5.4.15 Survey Summary 129

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
7.1.1 Entrant Details of the NIST 2008 Machine Translation Evaluation[20] 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

FIG. 43. THE INFLUENCE OF PACKAGING AND LOCALIZATION ON FE....79
FIG. 44. THE SEEDER – INPUTS AND OUTPUTS....80
FIG. 45. SEEDER – INTERNAL OPERATIONS....81
FIG. 46. 4-1 LEVEL N-GRAM LINK....82
FIG. 47. LINKER – INPUTS AND OUTPUTS....82
FIG. 48. LINKER – INTERNAL OPERATIONS....83
FIG. 49. PRIORITIZATION REGIME FOR LESS LINKS FIRST WITH A PENTAGRAM ORDER SPAN....84
FIG. 50. ADDING OF A NEW MAP LEVEL ON EXHAUSTION OF ANOTHER....85
FIG. 51. BLOBBING TO ACHIEVE FLUENCY ENHANCEMENT....86
FIG. 52. SPEED TEST'S INFORMATION THE INTERNET SPEED OF KOREAN ISPS....91
FIG. 53. THE SPEED OF OUR INTERNET CONNECTION TO A SERVER IN SEOUL FROM SUWON....91
FIG. 54. THE INTERFACE OF OUR WEBCRAWLER BEFORE STARTING A SEARCH....92
FIG. 55. THE SEARCH SCOPE PANE OF OUR WEBCRAWLER INTERFACE....93
FIG. 56. THE LOCALES & EXTENSIONS PANE OF OUR WEBCRAWLER INTERFACE....94
FIG. 57. THE KEY SEARCH WORDS PANE OF OUR WEBCRAWLER INTERFACE....95
FIG. 58. IMPLEMENTATION – A BUILT IN UPDATE MECHANISM....96
FIG. 59. FUTURE EVALUATION RESULTS TO BE POSTED....98
FIG. 60. EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN CASTING RADIUS IS INCREASED FOR INDIVIDUAL CLAUSES IN SEED TRANSLATIONS 1 & 2....100
FIG. 61. AVERAGE EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN CASTING RADIUS IS INCREASED FOR SEED TRANSLATIONS 1 & 2....101
FIG. 62. EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN N-GRAM ORDER SPAN IS INCREASED FOR INDIVIDUAL CASTING RADII....102
FIG. 63. COMPARISON OF COMPETING ONLINE MACHINE TRANSLATION SOLUTIONS....103
FIG. 64. SEED TRANSLATION POSSIBILITIES FOR FIRST SENTENCE TRANSLATION....107
FIG. 65. FLUENCY ENHANCEMENT USED AS A WRITING TOOL....109
FIG. 66. THE WEBCRAWLER DURING EXECUTION....110
FIG. 67. A CLOSE UP OF THE WEBCRAWLER'S PERFORMANCE....111
FIG. 68. TEST RESULTS OF OUR WEB CRAWLER....111
FIG. 69. AN ABSTRACT ILLUSTRATION OF THE FLUENCY ENHANCEMENT TOOL....115
FIG. 70. WHAT EACH TIER THINKS OF THE FLUENCY ENHANCEMENT TOOL....116
FIG. 71. HOW OFTEN EACH TIER WOULD USE THE FLUENCY ENHANCEMENT TOOL....117
FIG. 72. WHETHER EACH TIER'S FRIENDS WOULD USE THE FLUENCY ENHANCEMENT TOOL....118
FIG. 73. PREFERRED INPUT LANGUAGES FOR EACH TIER....119
FIG. 74. PREFERRED ACCESS METHODS....120
FIG. 75. PREFERRED PAYMENT METHODS....121
FIG. 76. PAYMENT EXPECTATIONS FOR EACH PAYMENT METHOD....122
FIG. 77. AWARENESS OF SIMILAR PRODUCTS....123
FIG. 78. OTHER POPULAR LANGUAGES THE FLUENCY ENHANCEMENT TOOL....124
FIG. 79. BREAKDOWN OF AGE GROUPS SURVEYED....125
FIG. 80. NATIVE LANGUAGES OF THOSE SURVEYED....126
FIG. 81. OCCUPATIONS OF THOSE SURVEYED....127
FIG. 82. EVERYDAY NECESSITY FOR ENGLISH....128
FIG. 83. LEARNING CURVES FOR CONFUSION SET DISAMBIGUATION....133
FIG. 84. THE VAUQUOIS TRIANGLE....135
FIG. 85. OBJECT PROFILING TO OVERCOME SEMANTIC AMBIGUITIES....136

FIG. 86. APPLICATION OF OBJECT ORIENTATED PROGRAMMING THEORY TO NATURAL LANGUAGE....138
FIG. 87. IMPLEMENTATION OF A DICTIONARY WITH A SEMANTIC FRAMEWORK....140

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

1. INTRODUCTION

1.1 ABOUT THE AUTHOR

My name is Steve Lawrence Manion, and I have been a language enthusiast for as long as I can remember and have always enjoyed building things. I speak English, Japanese and Korean fluently and understand the basics of both German and Chinese. I have read and researched extensively on linguistics and translation, and am intrigued by the nature of language. I am also a programmer, so of course I often try to realize my ideas by developing them in the form of algorithms.

Two years ago I formed a company in New Zealand called Kaleido K, whether I am working towards a qualification or not, the fruits of my research will always be nurtured and developed there. After several years of work on the intricate solution of MT, Kaleido K will be releasing a beta version of its MT system in the near future. 200 people were surveyed on the topic of this MT system and about 180 people belong to the Kaleido K Facebook group. Survey participants who were further interested in this research also gave their emails, so we now have a beta tester base altogether of approximately 350 people to help provide feedback on the MT system we will release. If you are interested in becoming a beta tester yourself, please visit my company site, www.kaleidok.com. The first version of the Kaleido K website is already active; here users can join and contribute to the development of our technology.

FIG. 4. KALEIDO K LANGUAGE COMMUNITY

In regard to this research I have two key goals in mind. The first is to get my MT system up and running, and to make it publicly accessible online in the near future, so it may be ready to compete in the Open MT Evaluation. This event is the closest thing to what would be the MT Olympics, in which all the major MT developers test their MT systems against each other. While the event is competitive, the main purpose of it is to bring developers together to discuss the issues of MT, so they can collaborate together. To be able to successfully compete in this event is to be officially acknowledged as a developer in the MT industry. Further more if a good result can be obtained at this event, this will demonstrate that the research achieved by Kaleido K is indeed very significant to the development of the MT industry.

FIG. 5. NEW ZEALAND

A second goal of mine is to raise the profile of the New Zealand MT industry. New Zealand is geographically far off from the rest of the world, and doesn't feel the burdening effects of the lack of translation resources to the extent that the rest of the world does. However there are a lot of talented programmers and linguists in New Zealand, and to raise public awareness and get more people working on the solution to MT would be great. There are to date, no New Zealand MT systems that have participated in the Open MT Evaluations, New Zealand should definitely get more involved in the MT industry, especially as the demand for translation resources is ever growing. Even if we don't require MT so much ourselves, there is no reason why we can't develop and distribute the technology to the rest of the world which is hungry for it.

1.2 ORGANIZATION OF THESIS

This thesis will educate you on the benefits the growing potential of our research. We have meticulously investigated the problem of MT and have devised our own process which borrows ideas from other methods and includes some inventive steps of our own. As previously mentioned we call this process Fluency Enhancement (FE), a process to reform and evaluate text to improve its fluency. This process can be applied to several areas of computational linguistics, speech processing and so forth but we will keep strictly to the topic of MT. With this in mind, here is a preview of this thesis's structure.

1.2.1 Content Layout

Chapter 1: Introduction

Naturally we will start with our introduction, in which we define what MT is and the issues that surround it, then put forward our proposed solution. The purpose of the introduction is to state clearly, the purpose of conducting such research, and also to act as a bouncing board onto the academic literature review, which will go into much further detail in regard to some of the subjects brought up in the introduction.

Chapter 2: Academic Literature Review

The academic literature review provides all the necessary information to bring the reader up to speed with the MT industry so they can comprehend how our FE process can contribute significantly to MT development. The following aspects will be covered in this order: a guide to MT approaches and methods, how MT technology is evaluated, a brief history of MT and lastly a summarization of the MT industry and its future direction.

Chapter 3: Introduction to Fluency Enhancement

Next we present the final preparation chapter which will highlight what exactly FE is. It is described in a very abstract sense without getting into too much detail so the reader can understand how it is applied to an MT system, and also how it can also have applications outside the scope of MT as well.

Chapter 4: Machine Translation System Design

After prepping you with what FE is, it is then time to show in detail how we have applied it to our own MT system, and review the design of our MT system as well. At this deeper level of detail, we can outline the several aspects we can alter about FE to tweak its performance for each individual application. Also documented in this chapter is the collection of linguistic data. For this we have created a web crawler which is capable of constructing more superior corpora to those used in this research. Also we discuss the implementation of the MT system here too and reflect how it honours the SAM fundamentals referred to earlier.

Chapter 5: Results & Discussion

Our results and discussion section includes results for our MT system using FE, the web crawler, and the market analysis of our FE tool. The results of our MT system using FE are based on the sample sentences used in chapter 4 which explain FE. The web crawler results demonstrate its performance on different computers and how it meets our requirements. Lastly for the market research results, each question addressed to the market is individually graphed and assessed accordingly to put together an overall market profile. Industry comparable results were not able to be acquired by the submission deadline of this thesis; however they will be available via the Kaleido K website soon for those who are interested or happen to read this thesis in the future.

Chapter 6: Conclusion

Lastly we will summarize in our conclusions and recap on the important points of this research and how we managed to achieve our objectives. Here the advantages and disadvantages of FE will also be examined.

1.2.2 Appendices Layout

Appendix A – Elaborated Data

Listed in this appendix are elaborated forms of data that are not in the middle of the thesis due to possible disruptions to the flow of the thesis.

Appendix B – Market Research Resources and Data

This appendix is dedicated to the market research section of this thesis. Here the survey used for survey is provided in both English and Korean. Also the raw data that was used to obtain the graphs and figures in the market research section can be found here.

Appendix C – Abbreviations, Technical Terms & References

All abbreviations, technical terms that belong to the MT industry, references and supplementary reading material are all denoted in this appendix. Each reference is indicated with a number in [square brackets], and the source of each reference can be found with each corresponding number in this appendix.

1.3 WHAT IS MACHINE TRANSLATION?

In simple terms, MT is the process of a machine (these days a computer) executing the translation of a source input language into a target output language. In recent times, the field of MT also includes tools that aid translation, collect data for the translation process, and even help evaluate the quality of MT. MT that calculates translations from vast amounts of data also falls under the study of Computational Linguistics (CL), which is a sub category of Artificial Intelligence (AI). While the history of MT has seen it rise and fall in popularity, it is now firmly one of the most important areas of research development with the world's increasing rate of globalization and the need for adequate communication growing more acute. There are three general types of MT that are particularly in demand, these are explained briefly below.

Machine Translation for Dissemination

This form of MT is used when the translations need to be of publishable quality, which is a must for multinational corporations that translate a lot of their documentation. In this case, often the MT system produces translation drafts, in which translators edit before publishing. This form of MT greatly aids translators and helps them churn through a larger volume of documents in a shorter amount of time.

Machine Translation for Assimilation

This is the form of MT that is commonly found on the Internet, and it provides translations in real time providing a result which could convey the general meaning of the text, but is not always intelligible. Basically it aids people to quickly understand the general meaning of whatever text they place into the MT system; however these systems are generally not designed to handle the broad range of language that is usually input into them.

Machine Translation for Communication

This form of MT is used for conversing over the internet, and needs to perform in real time like in MT for Assimilation, which can also be used for this purpose. The difference between MT for Communication is that it focuses on handling language that is frequently used in conversation, thus it is a more appropriate form of MT to be used in emails and chat rooms.

1.4 DESCRIPTION OF PROBLEM

1.4.1 Background

MT is a problem that has been investigated off and on since the 1950s to this very day, with momentum towards MT research continuing to grow. Various approaches, each with their own set of methods have been developed. Even hybridizations of these approaches to better reap and mitigate respectively the advantages and disadvantages of each approach have attracted research focus as well. Yet despite all of this, there is not a single MT system that is capable of consistently delivering translations that are indistinguishable from human translators. The more the problem is investigated, the more its complexity appears to be further underestimated than previously predicted. An underlying problem of MT is that many developers and end users of MT technology also are not able to think in a multilingual context, and because of this they harbour several misconceptions about the nature of language.

For developers, unintentionally when they design MT systems, they have misconceptions about all languages based on their native language. They jump to conclusions and make misguided assumptions, which are then reflected on the MT systems they design. These misconceptions usually become apparent to the developer when they then try to apply their MT system to different language pairs, and they discover that the new language pairs challenge their MT system in a way they did not design it to be challenged. In making this point however, we do not dismiss the argument that there may be some benefits to designing entirely different MT systems for different language pairs. However we do believe that it is good design practice to make the most of reusability, and because MT developers may lack the ability to think in a multilingual context, they jeopardize the portability (reusability) of their MT systems across different language pairs and accidently blind themselves to envisioning a much more universal MT solution that is technically of a higher order than anything we know of today. MT is such a complex problem, it is easy for one to bury oneself in it and not be able to see a simpler solution out of it.

Secondly end users of MT systems may fail to understand the complexity of language and have a belief that there is a perfect translation for every expression in every language. Thus the MT industry does not really have a positive public image. It's not that end users reject MT; it's that they are disappointed by it through their own misconceptions. However this is not entirely their

fault as many end users are not multilingual, thus have a poor understanding of the complexities of language. But an even larger problem that compounds with this is the promises of MT developers and distributers. In order to be competitive in the MT market, they overstate the capabilities of their product, which further fuels this misconception of the perfect translation that end users may have. In this combination of MT developers exaggerating the success of their MT system and end users putting too much faith and trust into these MT systems, misguided, humorous and sometimes culturally offensive translations have spread throughout the world. Fig. 6 below illustrates a victim of having too much faith in MT capabilities.

FIG. 6. THE RESULT OF MACHINE TRANSLATION MISCONCEPTIONS

In some cases, developers and distributors are not necessarily overstating the capabilities of their MT products either, but they do need to explain how they evaluated their product in more detail. Even though there are metric units (such as BLEU which will be explained later) to evaluate the success of MT systems, the results are almost always not comparable to each other. Different MT systems may be better suited to different translation applications, based on the data they draw from. Also often MT evaluation results are not obtained in real time; many competitors in open MT evaluations dedicate a single computer to each translation and may let it run for days before an answer is found, which is probably not what the end user would prefer. Some evaluation standards, even favour different MT approaches more than others, which further makes it difficult to identify the most appropriate MT product for the end user. More on this will be discussed later.

The problem of inadequate MT quality is unavoidable; translation resources are spread like butter over too much bread, and no one has figured out how to supply more butter at the rate it is needed. The world is globalizing at an increasing rate. Military, political, economic, social activities and so forth continue to further drive the demand for any form of translation to new heights. In many situations, even though the quality that MT can provide is not on par with a human, it is adequate enough to convey the general idea, and despite its shortcomings, the fact that MT can provide any relief at all to the demand for translation is very fortunate, particularly in this day in age with the frenzy of communication occurring over the internet between people from all over the world.

All this said the bottom line of this problem is that MT needs to be further improved to bridge the quality gap between human and machine. The further this gap is closed, the more MT systems will be able to provide a level of quality that is acceptable to a broader range of users, which will in turn relieve the demand on translation resources. Fig. 7 below illustrates approximately the current relief MT is able to provide *without* having any post editing being *vital* to the translation process.

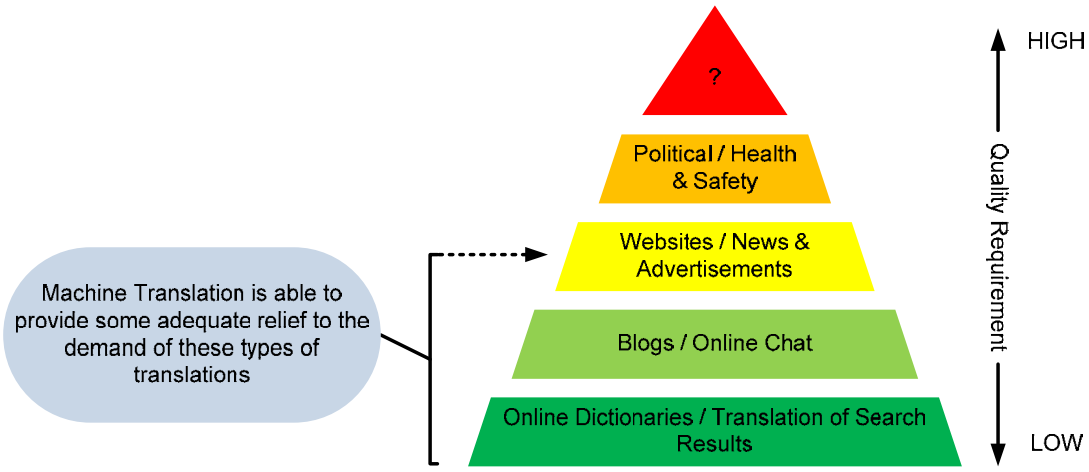


FIG. 7. REPRESENTATION OF THE CURRENT RELIEF MACHINE TRANSLATION PROVIDES

The goal of this research is to design an MT approach that can lift the bar in MT quality, and effectively lift the height of our dotted arrow in Fig. 7 above. The following sections of our introduction will cover in specific detail the obstacles that occur when translation language with a machine that our MT system must overcome.

1.4.2 Machines & Language Translation

Can a machine learn a language? We argue that a machine is not able to unless it has a level of AI that allows it to have its own conscience. In this case the machine would not be just a passive user of language, but an active user and contributor to the evolution of language by being creative with its rules and constructs.

Without this level of AI, what we do believe a machine is capable of doing is observing large enough quantities of linguistic data to learn how to mimic a language's patterned structure and swatch in and out these patterns to almost flawlessly replicate any desired text in this target language. Perhaps you may disagree with this, but consider how you yourself learnt your native language, and how language continues to evolve. When you were young, you were exposed to your native language, you learnt to mimic the common patterns, and if you have kept well socially connected to the world then you have learnt to keep up with the ever evolving patterns of your spoken languages.

You may argue that you are not mimicking and simply following the official grammatical rules and spelling of your spoken language that you have learnt. But the main reason we have grammatical rules for language is so we can teach and pass on language to others; the grammatical rules themselves are almost always broken and changed, and also can have any number of exceptions due to the cultural and social climate of where the language is spoken. In short, grammatical rules are made to simplify the complex nature of language; in essence they are a big blanketing rule of thumb but in no way control the evolution of language.

On the following page we will now strengthen our argument with two different language models that favour either a human or a machine.

1.4.3 Language Models

Ideal Language Model (ILM)

If you followed grammatical rules that were constant and without exceptions then you would be simply following the rules. If languages had defined constant rules then achieving high quality machine translation would be a much easier task, since you could simply code all the rules of all languages into the MT system and you would be finished. Constant rules would disable the collecting of rule exceptions over time. If new expressions were needed, the creation of new grammatical structures and words to capacitate further evolution of language would be required. However over time the collection of rules and words rather than rule exceptions would occur. This is fine for a machine but would become a headache for the human race to remember such a vast library of grammar and words. The ILM is ideal for MT systems but not for humans, thus that's why it is not used.

Realistic Language Model (RLM)

Unfortunately the grammatical rules and evolution of language are largely controlled by people, and the ILM described previously is traded in for the RLM. People create new grammar and spelling as they see fit, in order to express themselves the way they wish to. Thus new spellings are formed and grammatical rules are broken and the exceptions to them all flourish. Considering all this, if you are following rules that are constantly open to change and variation, then technically, this is more like mimicking, since you are not able to rely on the rule holding true in every single case, you have to copy what others say in order for your speaking to be comprehendible and aligned to that of others.

The RLM is what MT must deal with, and consequently this is why the problem of MT grows so complex. A machine really needs to learn how to mimic a language, just like a human does, in order to communicate effectively in that language. Preferable to the ILM, if a machine follows the rules word for word it will fail, and soon be speaking language that is out of date as time goes on as well. It is a well known fact that humans are an exceptional species when it comes to mimicking what is exposed to their senses; this is why the RLM is in fact the human's ILM, as long as we are well socially connected, we can quickly adapt to the evolution of language. There is some irony to this though, for a human when it comes to learning a foreign language, mimicking the evolution of

that respective language is a much harder task. You must understand the culture and social climate of the country where that language is spoken to try understand the origins of each evolutional step. Consider the expression “As easy as pie”; what is this expression supposed to mean for a non-native speaker? In New Zealand, a pie is generally not described as easy, but as delicious or hot so it doesn’t seem logical to describe it as easy; however it makes perfect sense to the native English speakers. The following information is an excerpt taken from an online dictionary known as the Phrase Finder and gives specific details of the expression’s origin:

FIG. 8. THE 19TH CENTURY AMERICAN ICON OF EASE[1]

The easiness of pie comes with the eating. At least, that was the view in 19th century America, where this phrase was coined. There are various mid 19th century US citations that, whilst not using 'as easy as pie' verbatim, do point to 'pie' being used to denote pleasantry and ease. 'Pie' in this sense is archetypally American, as American as apple pie in fact.

The earliest example of the actual phrase 'as easy as pie' that can be found comes from the Rhode Island newspaper *The Newport Mercury*, June 1887 - in a comic story about two down and outs in New York:

"You see veuever I goes I takes away mit me a silverspoon or a knife or somethings, an' I gets two or three dollars for them. It's easy as pie. Vy don't you try it?"

Pie seems to rank right up there with cake in the US lexicon of ease and pleasantry - 'a piece of cake', 'take the cake', 'cake-walk' are all American phrases from the 19th century [1].

Thus according to the excerpt on the previous page we can conclude the expression originated from America in the 19th century and spread throughout the most the English speaking countries including New Zealand a long time ago. Anyone learning English today would have to learn this expression from its context of use, in order to use it correctly, which again is technically mimicking, as it is highly likely that most native speakers have no knowledge of the origins of this expression.

Considering the situation just explained, it is now clear that the RLM can also be unfavourable for humans as well. Not every human has a knack for learning a foreign language, and often a human's childhood plays a big part in their understanding of language. In fact a child that has not learnt a language by their teenage years will never have the chance to develop any decent language skills for the rest of their adult life. There have been several documented cases of this in feral children, the most famous of them being a boy called Victor from France [2]. So while humans are of course excellent at mimicking, in terms of learning a foreign language, only a select few have the ability to become distinguished translators. Language is very complicated and it may take several years for a translator to mimic a language like a native speaker, so it is easy to understand how the world ends up scarce of translation resources. In light of this situation, a machine is not as elegant as a human at mimicking a language, however they can learn how to in a much shorter period of time. Because there are so many translations required by the world, and it takes such a long time to train up a human translator, improving the quality of MT to relieve the acute demand is critical and unquestionably the more cost and time effective solution.

In summary, the RLM complements the imagination of humans in the creation of new language; however it also causes communication fumbles between humans who speak different languages. For MT, learning any language is like learning a foreign language for a human, though humans have the upper hand because they are better at mimicking something of such a complex nature. While machines have speed on their side, they are further hindered by the fact they prefer a logical yes or no answer to everything and the RLM is filled with gray areas of maybe yes or maybe no. It should now be clear that when designing an MT system that can learn to mimic and replicate a target language on par with a human, the aspects of the RLM that trip up machines must be somehow resolved. As a result this will improve the quality of MT.

1.4.4 The Obstacles of Translating Language

A lot of people have a naïve understanding of translation. Watching a subtitled movie they may complain of the subtitles not making sense, and failing to convey the punch line of an in-movie joke and so forth. But people who come to this conclusion most likely do not understand the complexities of language. The translator may in some part be at fault, but the crux of the problem is that languages are simply different. These people most probably believe that there is a perfect translation for everything into every single language. When people begin to learn their first foreign language, for a while they usually demand how to say this and that from their teachers and peers. After a while they begin to accept the truth that there is not a completely perfect and identical in meaning translation for everything they want to say. Once this truth is realized, they can usually learn to have new found respect for the subtitles and MT systems that appeared so ridiculous during the time they were monolingual. We will now review some of the obstacles of language translation.

Mechanical Differences

You need to have realistic expectations. There is not a perfect translation for every piece of text. Every language is mechanically different, thus they are not as interchangeable as some people may believe. Each language has its own benefits and limitations.

For example English enforces the use of upper and lower case letters, which can be frustrating for programmers if they are subject to case sensitive errors whilst programming. However on the contrary English can be easily used to make acronyms with upper case letters, which is a convenient way of expressing many technical terms used in several industries. As for another example, the Korean language, as well as many other languages can be written using no spaces, and further more it uses less characters than most western languages when it comes to writing a sentence of equivalent meaning, so you can say more and save space in text messages and on blog sites where your reply may have a character cap. From these two examples, you can probably identify some unique mechanical differences in your own native language too. The compounded effects of these various mechanical differences usually cause translations to be unintelligible for any number of reasons.

Different Words & Expressions

Different languages have different words and expressions. An MT system often struggles to identify when and how to use these different words and expressions. A machine translator is at a disadvantage because it usually only has text and sound as data input, however a human translator has sight, knowledge, taste and smell and other such data input which further help the human translator come up with a more sensible translation in a broader scope of contexts.

There are several ambiguous words and expressions that can only be translated properly if the context is completely understood; also sometimes to understand the context more text or speech is required, in which case an MT system may lack the resources to do so. Such words are polysemantic and can have several meanings. Consider seeing a sign in a field that says "The footpath across the field is free – but the bull may charge!" from context you can easily understand the bull will not be charging admission, but may charge towards you and chase you. Also consider this sentence, "The brick was dropped on the table and it broke", once again you would need a visual input to understand what *it* refers to in order to understand if the brick or the table broke. Both of these examples explain how the degree of contextual input an MT system lacks, correlates to the potential degree of ambiguity that may arise. When translations appear illogical, what we have described here is usually the cause of the problem.

FIG. 9. THE VISUAL (AND OTHER) INPUTS MACHINE TRANSLATION REQUIRES TO SOLVE LANGUAGE AMBIGUITIES

Language Defines How You Think

Language defines how you think and express yourself. As you are limited to the words and expressions of the language you speak, then you are likely to describe a situation relatively different than someone who speaks another language, since you are both dealing with the different limitations and benefits of the respective language you speak.

This was demonstrated in the UK at Bristol University. English, Japanese and Turkish speakers witnessed a carton where Sylvester the cat swallows a bowling ball and rolls down a hill. Each person was then asked to describe with hand gestures how the cat moved. Japanese and Turkish speakers dissected the action showing its path (moving down) and manner (rolling) with separate hand gestures. Contrarily the English speakers combined the actions path (moving down) and manner (rolling) with one single hand gesture. This is because English has the phrasal verb "roll down" which describes the path and manner of the action happening together, whereas Japanese and Turkish do not, and two different verbs are used in conjunction to describe the same event. In identifying this aspect of language, Dr Kita stated: "My research suggests that speakers of different languages generate different spatial images of the same event in a way that matches the expressive possibilities of their particular language. In other words, language influences spatial thinking at the moment of speaking. If people express themselves differently in each language then their thought process is also different [3]." This supports our argument that language defines how you think, which identifies another hurdle for MT to overcome. In translating Japanese or Turkish to English, how can an MT system translate the literal meaning of "he rolls as he descends" to the more intelligible translation of "he rolls down"? When translations appear too literal in meaning, what we have described here is usually the cause of the problem.

The Ultimate Effects of these Obstacles

So in making these few points, it should now be clear that these obstacles incur a margin of inaccuracy for any translation. If a translation appears to be perfect in conveying the exact same meaning, it is in part, because the two languages had some strong coincidental mechanical and linguistic similarities which placed no real burdening obstacles for the subject translation. However in any case, an MT system should be designed so it can deal with any of the potential obstacles previously described and more.

1.5 PROPOSED SOLUTION

1.5.1 Motivations

The motivations of this research are very much in plain sight if you consider the everyday lives of people in this world. On so many levels the world needs to communicate, and in several situations people want to be particular about how they want to express themselves. We ourselves (the writers of this research) are users of MT, and what is offered out there still does not suffice our needs. When I am translating a document, I may use an MT system, but I only receive one result, and the chances of it reading fluently are also unlikely, so if I am not quite confident in using the target language, then there is not a lot I can do with the translation result since I am incapable of doing any adequate post editing. MT systems need to change; often their one glove fits all solution falls short for many users. MT systems tell the end user how to use it, when it should be the end user calling the shots. Not only should they be able to choose translation direction, they should also be able to choose the topic, the number of translation possibilities required, and even be able to configure the MT system to how they like it. Let them use it and configure it so they get the results they are happy with, and so they are also in some part accountable when they get bad results.

It's time to educate the public more about MT and provide them with more user configurable tools to do their own translating. No more copying and pasting text to be translated and taking it to a particular website where you have to again from scratch manually set up the input and output languages to push out only one translation. It's time to arm the end users with MT tools, such as browser tool bars and plug-ins that can identify the text encoding on a given website. Have these tools be customizable to remember previous settings and have designated hot keys. Have the tools provide more than one translation answer and also perhaps a self assessment feature on its confidence in each of its translations and so forth. Perhaps some companies have already taken this initiative, but we believe it needs to become common practice to offer end users such options and control. While our motivations are mainly to improve the quality of MT, we also want to change the face of how MT is perceived and used by the public.

1.5.2 Our Machine Translation System's Design Fundamentals

Each MT system is built upon some important fundamentals that usually dictate the actual success of the MT system. Some fundamentals are commonly shared whereas others are unique and can be solely responsible for an MT system doing better or worse than another MT system. In this section we will explore the three key fundamentals that our MT system is built on, and identify the significance and the influence they have on the development of our MT system's design.

Simplicity

It is important to design the simplest MT system possible so it can be used to translate any language pair in any direction. To achieve simplicity, coding rules to guide the behaviour of the MT system needs to be kept to a minimum. As soon as rules are made to deal with a particular issue that exists in a particular language, then new rules are also created to deal with the exceptions of the initial rule. This can continue until the MT system then becomes too complex and rigid in design. Our stance is not to throw out rules altogether, as Rule Based Machine Translation can still provide reasonable results for some language pairs and ensure that the rules of syntax are adhered to. However the implementation of rules has to be done in a very uniform and abstract way so the MT system remains stable and maintainable for the developer.

For example most MT systems have a syntax parser. When building a syntax parser, make sure it is able to analyse syntax by drawing from comparisons in a database, avoid techniques where syntax comparisons are hardcoded into the MT system's algorithm, and in this way the parser can be easily converted to parse syntax in other languages too. Furthermore, if the parser comes across strange uses of language, let it automatically denote the strange use as an exception, and if the strange use of language continues to appear let it be brought to the developer's attention or even automatically formed into a syntactic rule. Wherever possible, simplify the processes of MT. When implementing rules, do not rely on the rules themselves to achieve translation, rely on the fact that the rules will be broken and design accordingly.

Adaptability

Self learning is crucial for an MT system, no individual or group of people can possibly keep up with the manual task of updating an MT system to constantly align with the evolution of language. This mechanism of self update must be a built in feature of the MT system; otherwise the MT system is not able to adapt and will decrease in accuracy the more language evolves. Self learning needs to be implemented in the way of a feedback loop, so the more it is interacted with, the more it can train itself in accuracy. As for one adaptability measure, for our MT system, we believe that each text that is put through the MT system should be regarded as a popular translation. To improve the performance of the MT system, each popular translation should be collected and analysed for reoccurring patterns (with respect to privacy – never published and discarded after analysis). Text that appears to be a popular translation can be given special attention, and more suiting and varied translations of these texts can be derived to further tweak performance. These popular translations actually form their own individual corpus, and this corpus can also be used to help aid translation in the opposite direction as well. More on this will be discussed later in our MT system's implementation.

To get results for this thesis we have used a large English text corpus prepared by Google [4] and Kaleido K's English / Korean bilingual dictionary. However we want to obtain our own data that is broader, larger and more multilingual. So we have built our very own web crawler that is in the process of constructing superior corpora. The corpora we are constructing will also be more up to date than the Google corpus that was constructed in 2006 and it will also be able to be split up in terms of topic (i.e. medical linguistic data can be extracted from the corpora), so in all respects our own corpora will do a better job of upholding our Adaptability fundamental.

Multiplicity

It is important that an MT system explores as many translations as possible in order to acquire the most likely translations. A key benefit of this is finding the most natural translation for the target text. Furthermore translations can be processed in respect to different fields of expertise such as language used in *law* or *medical science* for example. In reality there can be endless translations for a given piece of text, so a simple calculation to derive a single translation will not do. We need to produce and analyze numerous translations and then find the most popular and well understood translations amongst all these translations in order to achieve a final translation that is fluent and cohesive and that is free of awkward use of words or mechanical problems. What can be of concern is the extra processing involved that can be detrimental to real time performance. This can be constrained though if pruning is done during the translation process. Pruning is where unlikely translations are prematurely terminated before they see out their final translation result; this is very effective because with this method a large number of translations that are destined to be failures are no longer processed. When you are juggling thousands of possible results this makes a big difference, and you can focus on the handful of translations that really do have potential.

Determining whether a translation should be prematurely terminated, is a matter of doing a quick pass through the Google corpus, to statistically understand if the words in the translation have much chance of occurring in the same sentence. A cut off point usually needs to be decided for this. Depending on how much accuracy versus real time performance the end user wants, a certain amount of finalist translations can be determined and further worked on, while the others are prematurely terminated. For example “Where are you?” would be chosen as a finalist while “Where do you exist?” would not likely be chosen. The importance of multiplicity will be clarified in more detail later when the design approach of our MT system is explained.

1.5.3 Objectives

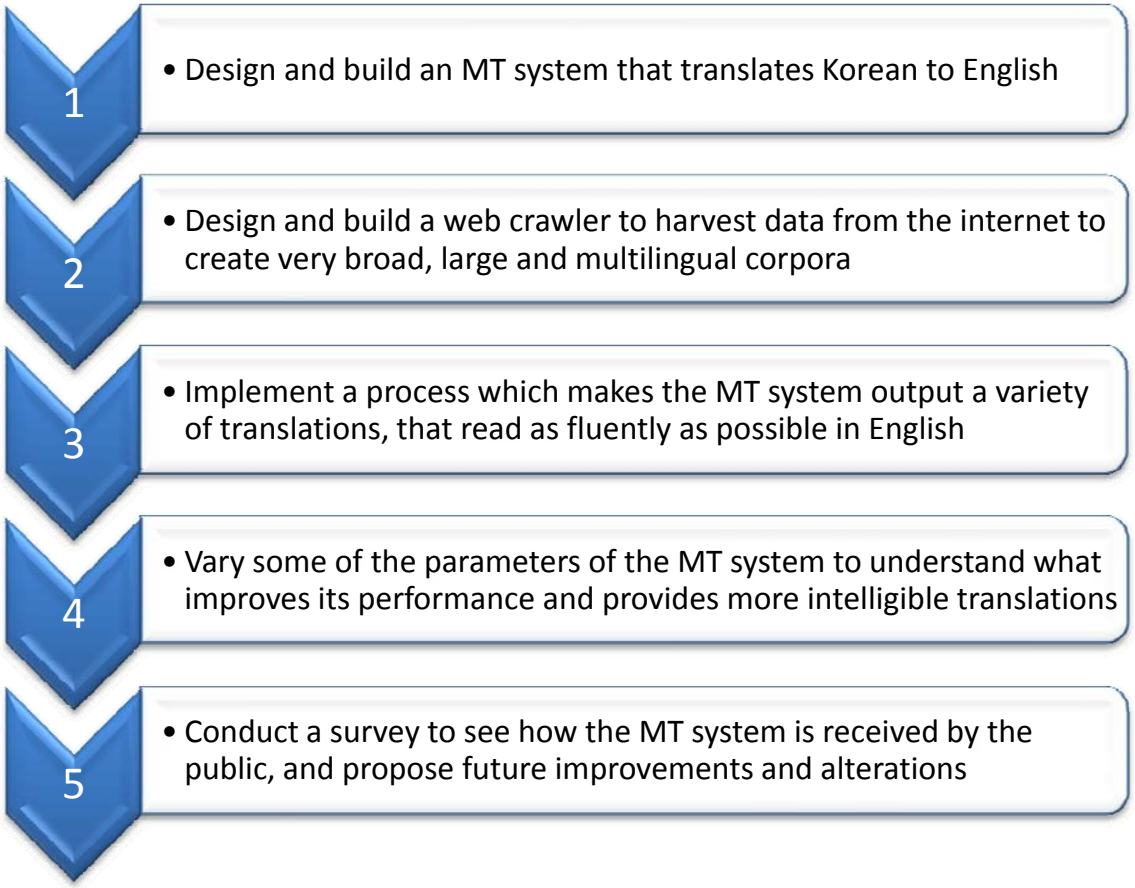


FIG. 10. RESEARCH OBJECTIVES

Our key objectives can be seen in Fig. 10 above. We will now explain each objective in a little more detail:

Objective 1: The MT system must translate Korean text into English text. This must occur in real time and be able to output at least some result for any piece of Korean text input into the system.

Objective 2: The web crawler must be deployed across multiple computers and communicate / add linguistic data to the Network Attached Storage (NAS). The web crawler must be programmed to identify and obtain quality linguistic data. Web pages that provide a quality source for linguistic data should be indexed in a database in the NAS, and these pages can be searched again with new

versions of corpora we develop. The corpora must be superior to the corpus we have used in this research provided by Google. In this regard there will be more linguistic data in more languages, the linguistic data will have longer and broader N-Grams, and the linguistic data will be also categorized by topic, so for example only medical linguistic data can be extracted from the corpora if necessary.

Objective 3: Putting together an MT system is the easy part, getting it to make intelligible translations is the difficult part. A process must be layered over the MT process that ensures the MT system doesn't just translate text, but ensures that it is somehow intelligible to a human. The concept of fluency is important, what might look like the correct translation to a machine may need a lot of post editing for a human to make sense of it. Thus our objective is to ensure the human has to do very little post editing after they see the final output. Thus a process which produces a variety of translations that are paraphrased versions of each other and that read more fluently must be devised. (In other words, Fluency Enhancement)

Objective 4: The process described in the previous objective needs to accept parameters which may allow the performance of the process to be tweaked in order to get the best performance for each unique application it may be used for. (Even outside the scope of MT). The parameters could also be indirectly chosen by the end user if not coded into the MT system. For example the FE process could draw from different linguistic data, if the end user knew the content of the translation was related to medical science; then our process could be bias to medical linguistic data when attempting to enhance the fluency of the translation.

Objective 5: Finally we need to put forward the concept of our FE process integrated into a tool to the market and get some feedback on what they think. Enough people need to be surveyed so we can obtain a good idea of the market profile and identify what we need to do to transport our FE process from just a novel idea to something that is commercially distributable.

In our conclusion we will again address these objectives and explain how and to what extent we were able to successfully complete them.

2. ACADEMIC LITERATURE REVIEW

2.1 METHODS OF TRANSLATION

2.1.1 Rule Based Machine Translation

Rule Based Machine Translation (RBMT) as its name suggests, uses *rules* to overcome the syntactic and semantic barriers of language translation. RBMT methods work well for languages that do not have abundant linguistic resources, and also can provide translations that are more syntactically correct since the syntax rules of each language can be coded accordingly. The Rule Based approach comprises of 3 methods, known as Direct MT, Interlingual MT and Transfer Based MT. Each method has its own advantages and disadvantages which we will discuss in the following text and later compare them all together at the end of this section.

Direct Machine Translation

Direct MT is a very flat and unstructured method of MT. As one might guess, a translation consists of locating each word in the source text and translating it into the target text *directly* word for word without making any efforts to translate any syntactic or semantic elements of the text. The end result is a translation that is a list of words that may not appear to have any grammatical or logical relationship to each other. In some cases this approach can perform well if the languages are not too distant from each other, for example German to English, which can be seen below.

TABLE 1. DIRECT MACHINE TRANSLATION FROM GERMAN TO ENGLISH

GERMAN to ENGLISH (Close Language Pair)										
Ich habe einen Bruder und er ist 18 Jahre alt.										
Ich	habe	einen	Bruder	und	er	ist	18	Jahre	alt	.
I	have	a	brother	and	he	is	18	years	old	.
I have a brother and he is 18 years old.										

"I have a brother and he is 18 years old." is basically a perfect translation. It is not uncommon to obtain a reasonably intelligible translation from German to English using this method of

translation. However take a look at this next example in which we tried to translate text of similar meaning from Japanese to English which are languages that are very distant from each other.

TABLE 2. DIRECT MACHINE TRANSLATION FROM JAPANESE TO ENGLISH

JAPANESE to ENGLISH (Distant Language Pair)							
私は弟がいて、彼が１８歳です。							
私は	弟が	いて	、	彼が	１８	歳	です	。
I	brother	there is	,	he	18	years	is	.
I brother there is, he 18 years is.							

The above result is complete garble, but we can still to an extent understand the idea behind the translation, this is because we can post edit the translation in our head with human intelligence. However for much more complex text using Direct MT, the meaning could be beyond any post editing measures leaving the end user guessing.

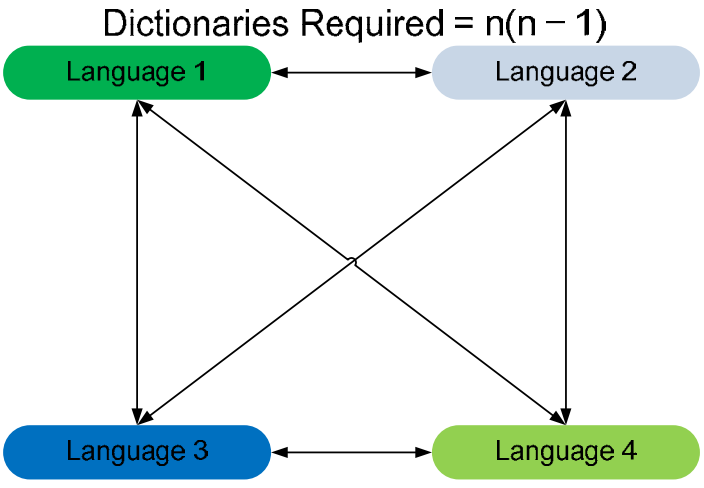


FIG. 11. DICTIONARY REQUIREMENT MODEL FOR DIRECT MACHINE TRANSLATION

For each language pair, a dictionary is required for each direction of translation. Direct MT requires $n(n-1)$ dictionaries where n is the number of languages present in a given MT system. Fig. 11 illustrates an example of this concept, notice how for 4 languages, $4(4-1) = 12$ dictionaries are required. The dictionary cost for Direct MT can be high as it increases in a quadratic nature with the increase in languages added to the MT system.

Interlingual Machine Translation

The Interlingual MT method uses a universal (*interlingual*) language that works as a medium between languages in the translation process. One major advantage of this approach is the amount of dictionaries required to make large multilingual MT systems is reduced as only *2n* dictionaries are required. This advantage is noticed more significantly as more languages are added to the MT system; this is because the increase in dictionaries required is linear rather than quadratic. Notice in Fig. 12, only *2n = 8* dictionaries are required, rather than 12 dictionaries which is what the previous method Direct MT required.

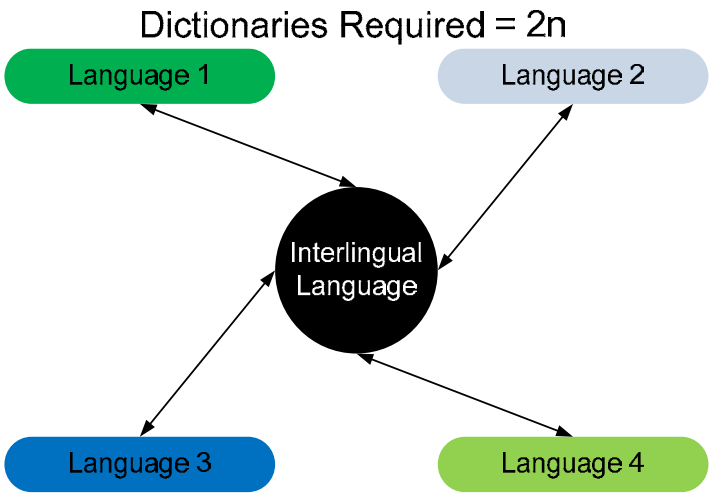


FIG. 12. DICTIONARY REQUIREMENT MODEL FOR INTERLINGUAL MACHINE TRANSLATION

There is also a disadvantage that grows as more languages are added to the MT system; this being the interlingual language must accommodate all the syntactic and semantic complexities of all the added languages. This is a very difficult task and some languages are very distant from each other in terms of syntactic and semantic elements. One measure to counter this problem is to have more than one interlingual language, which helps bridge the gap between more distant languages. This solution can also adopt an official language to be one of the interlingual languages. For example Russian can be used as an interlingual language to help bridge the gap between English and Ukrainian [5].

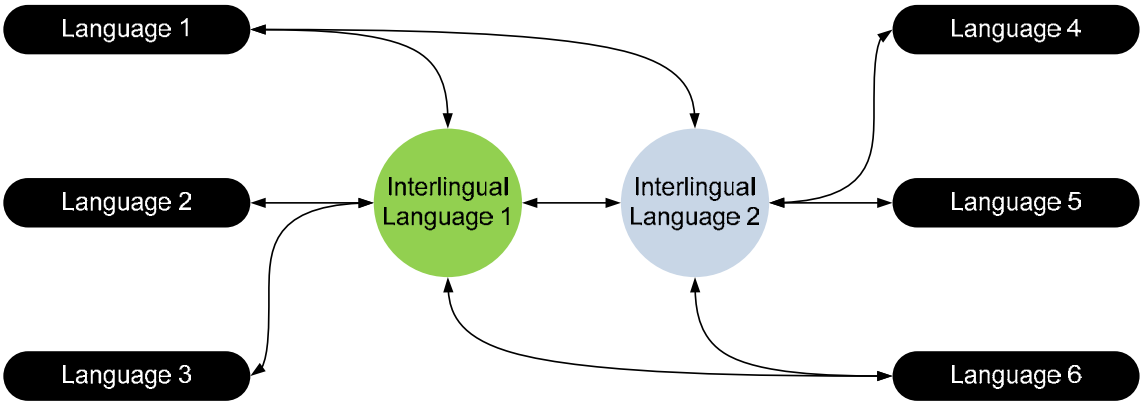


FIG. 13. BRIDGING DISTANT LANGUAGE PAIRS USING MULTIPLE INTERLINGUAL LANGUAGES

Fig. 13 illustrates how two interlingual languages could be applied to an Interlingual MT system. It is difficult to know whether using multiple interlingual languages can increase or decrease the amount of dictionaries required in the MT system. It depends entirely on the relative distances between the language pairs. For example some language pairs may only need to be linked through one interlingual language, where as other language pairs may need to be linked through more. Also the paths in which each translation process will take through the interlingual language pairs can have influence too; as there is no point in building dictionaries for paths that translation processes will never traverse through.

Lastly while the Interlingual MT method can occasionally perform well in competition with other MT methods, the fact that interlingual languages result in a reduction of dictionaries serves as a double sided sword. This is because all of the interlingual languages need to be updated as the syntactics and semantics of each language evolve over time. Updating these interlingual languages to reflect the changes in all of the individual languages can grow into a mammoth task. One needs to devise an effective auto updating system that is also safe in not being able to harm the output quality of the MT system as it updates. Even if this is feasible, the design and implementation of this feature may also be a difficult task in itself. If an official language is used as the interlingual language, then rather than the official language needing to be updated, the issue is shifted to keeping the translation processes and data up to date. This is usually a more favourable option, particularly for MT systems that heavily rely on linguistic data, as then the issue is the simpler task of ensuring the linguistic data is recent.

Transfer Based Machine Translation

Just like Interlingual MT, Transfer Based MT makes use of interlingual languages, however Transfer Based MT is different because the interlingual languages are dependent on the languages in the MT system, whereas the interlingual languages used for Interlingual MT are independent and can be used with any languages. Since the Transfer Based MT's interlingual languages are dependent on the amount of languages used in the MT system, one of the primary benefits of Interlingual MT of only needing 2*n* dictionaries is lost. Transfer Based MT uses a special interlingual language for every single language pair. This incurs a dictionary cost of $2n(n-1)$, which is not only now quadratic, but is also twice that of Direct MT's dictionary cost. In Fig. 14, $2 \times 4 \times (4-1) = 24$ dictionaries are required for the MT system. This is very high, but Transfer Based MT can prove to be more accurate than Interlingual MT since it has specifically designed interlingual languages for each language pair. In this light, whether to use either Transfer Based MT or Interlingual MT is usually a trade off between accuracy or dictionary reduction costs being more important.

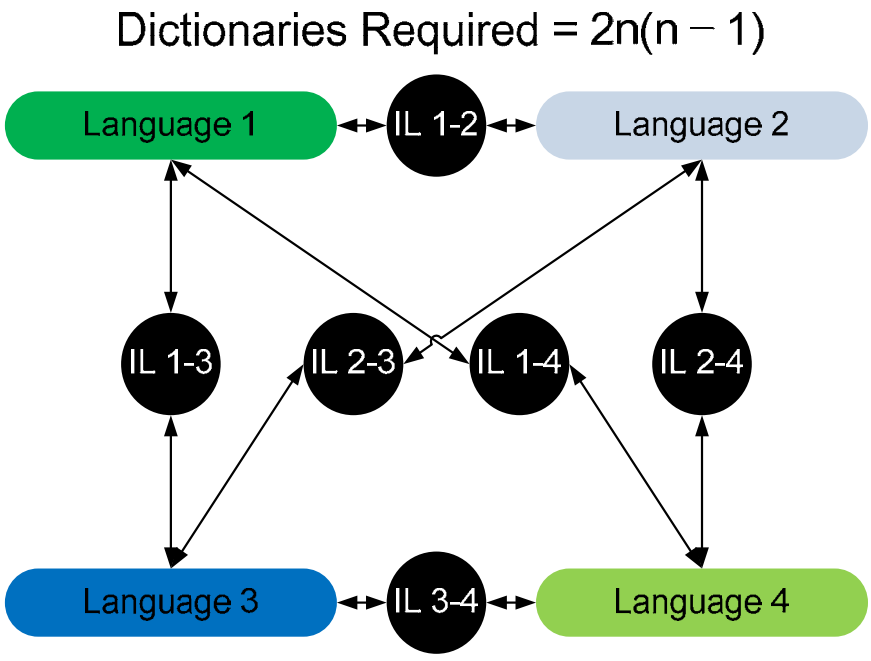


FIG. 14. DICTIONARY REQUIREMENT MODEL FOR TRANSFER BASED MACHINE TRANSLATION

Preferred Applications for Rule Based MT Methods

Fig. 15 demonstrates how the dictionary costs for each RBMT method increase as more languages are added to an MT system. As can be seen, Interlingual MT is far less expensive with its linear increase as opposed to Direct MT and Transfer Based MT which have dictionary costs that increase in a quadratic nature.

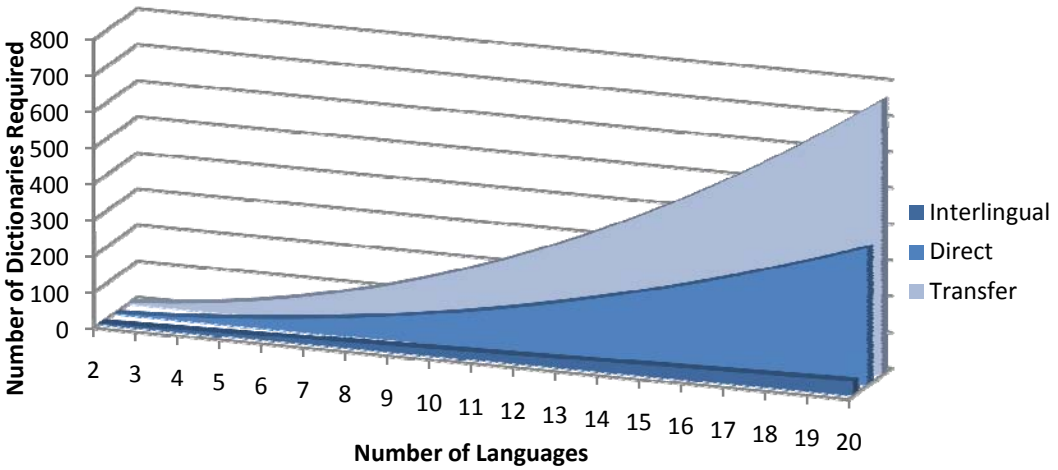


FIG. 15. INCREASE IN DICTIONARIES REQUIRED FOR RULE BASED MACHINE TRANSLATION METHODS

Each RBMT method has a purpose that it is best suited to. Direct MT is best used as a mechanical dictionary, and if used for full text MT then it should only be used for languages that are relatively close in distance. Interlingual MT is best used for large multilingual MT systems that need dictionary costs to be at a minimum. It can also be useful for languages that are close in distance and possibly do not have any linguistic resources available in the first place. For example Interlingual MT would work well for an MT system that translated the many European languages in Europe. Transfer Based MT is best used for MT systems that are not concerned with dictionary costs and need improved accuracy, especially for language pairs that may be distant. This is the best Rule Based MT method for an MT system that would encompass many languages that are significantly diverse from each other providing the engineer has the time and resources to put together such a potentially complicated MT system.

2.1.2 Corpus Based Machine Translation

Corpus Based Machine Translation (CBMT) as its name suggests, uses sheer volumes of text data (*corpora*) to overcome the syntactic and semantic barriers of language translation. CBMT methods work well for languages that have abundant linguistic resources, and also it can provide output that is more semantically and fluently correct since it can compare translations to real translated text to verify intelligibility. The Corpus Based approach comprises of 2 methods, known as Example Based MT and Statistical MT. Each method has its own advantages and disadvantages which we will discuss in the following text and then we will compare them against each other at the end of this section.

Example Based Machine Translation

Example Based Machine Translation (EBMT) is very popular for applications such as those that aid tourists. The situation the translation is required for is usually predictable. For example the tourist may want to order food. With several phrases stored in memory relating to ordering food, and plugging in and out the food items and counters where appropriate, an EBMT system can provide fairly good results such as those in Table 3. However this method may struggle to come up with a sensible translation for anything that needs to be translated outside the topic scope of ordering food, it is completely dependent on how broad the phrasal coverage is of the corpora used to power the EBMT system.

TABLE 3. EXAMPLE BASED TRANSLATION FROM JAPANESE TO ENGLISH

<X>を<Y>杯ください。	
コーヒーを 3 杯ください。	ビールを 2 杯ください。
3 coffees please.	**2 beers** please.

While EBMT is usually the ideal choice for MT use in a predictable situation, some MT developers have also used it to cover a very broad range of topics so it can cope with a wider range of translations. Take a look at this example taken from research conducted by the School of Computer Science at Carnegie Mellon. EBMT requires huge amounts of translated text in order to perform well. In this study they decided to generalize words, with tags such as <PERSON>, <CITY> and <DATE>. Generalization of words enables EBMT to make better use of the bilingual corpora it

uses to derive translations. Take a look at the text translated from English to German in Table 4. If we have some translated text of similar structure in our EBMT system, we can translate (*by example*) any other text that has this same generalized structure [6].

TABLE 4. EXAMPLE BASED TRANSLATION FROM ENGLISH TO GERMAN USING GENERALIZATION TAGS

<PERSON> was in <CITY> on <DATE>. **<PERSON> war am <DATE> in <CITY>.**	
John Hancock was in **Philadelphia** on **July 4th**.	**Joe Bloggs** was in **Moscow** on **October 10th**.
John Hancock war am **4. Juli** in **Philadelphia**.	**Joe Bloggs** war am **10. Oktober** in **Moskau**.

As can be seen, generalized EBMT can provide competitive results, however just like all CBMT methods, its success is largely dependent on the size of the bilingual corpora available to it. Quality bilingual text is not always easy to come by, and collecting up a sizable amount in order to encompass the infinite amount possible translations that could be put through the MT system is most probably impossible. Thus improvements such as generalizations and other pattern matching measures are important in order to make the most of the bilingual corpora available to the EBMT system.

Statistical Machine Translation

Statistical Machine Translation (SMT) is relatively new and its use in MT development has spread widely. It works by statistically analysing bilingual corpora to guide the decision making process in the MT system. As can be seen in Fig. 16 below, the general process can be broken down into smaller processes. This example is from the University of Southern California, firstly the text is translated as individual pieces (by the translation model) into broken pieces of English text that have a probability *P(s|e)* of being the correct correlating translation. Then the broken English text is analyzed and concatenated together into longer pieces of text that have a probability *P(e)* of being a good translation. Using the Translation Model and the Language Model, the Decoding Algorithm's job is to find translation *e* that maximizes *P(e)* × *P(s|e),* which will be the chosen output for the SMT system [7].

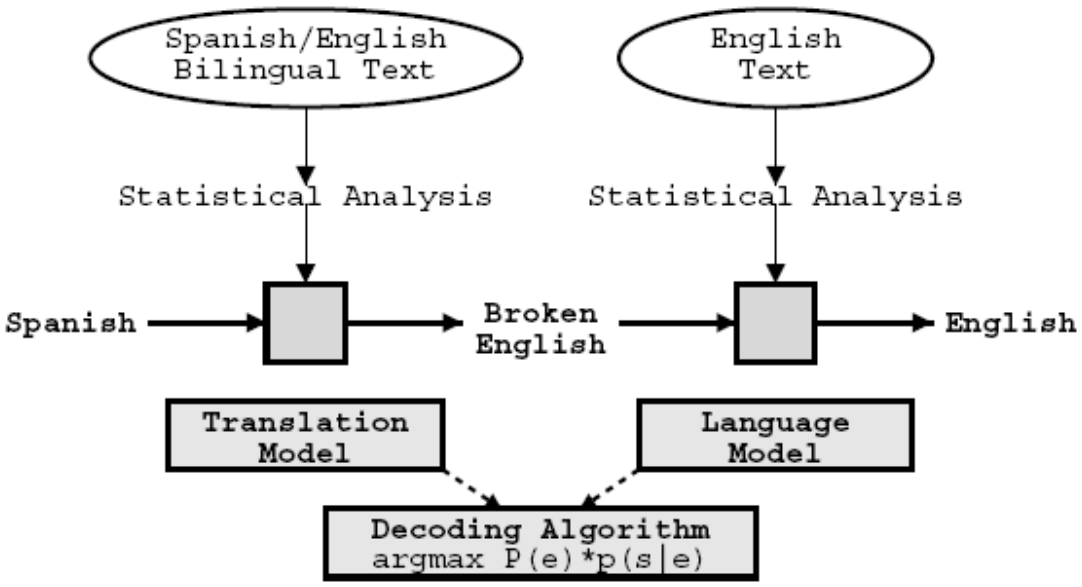


FIG. 16. AN EXAMPLE OF GENERAL STATISTICAL MACHINE TRANSLATION ARCHITECTURE

Now you have a general idea of how an SMT system works, we will briefly explain how some of the processes can be executed. Please note that there are several ways to execute these subprocesses however we will only cover a few to demonstrate the possibilities.

Expectation Maximization

One way of achieving a translation model is to use an Expectation Maximization algorithm, which ploughs through the bilingual corpora and identifies the probabilities of word and phrase matches. Fig. 17 demonstrates with a small example of how this is done. The process continues to increment over the data and makes educated guesses as to which words correlate to each other in each language. So for our example below, the first increment would isolate the French word "la" as being the English equivalent of "the". On the second, third and forth increment, through a process of elimination the French words "maison", "fleur" and "bleu" can be extracted to be the English equivalents of "house", "flower" and "blue" respectively. In this manner we can eventually derive all the possible equivalents and obtain probabilities for each of them. This is not limited to only being performed on separate words, but it also can be performed on phrases as well.

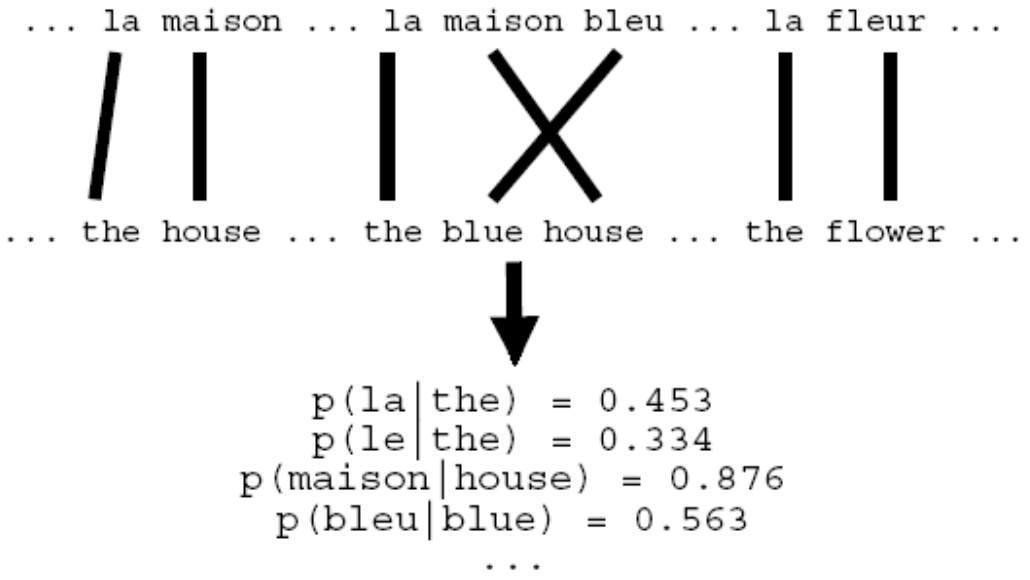
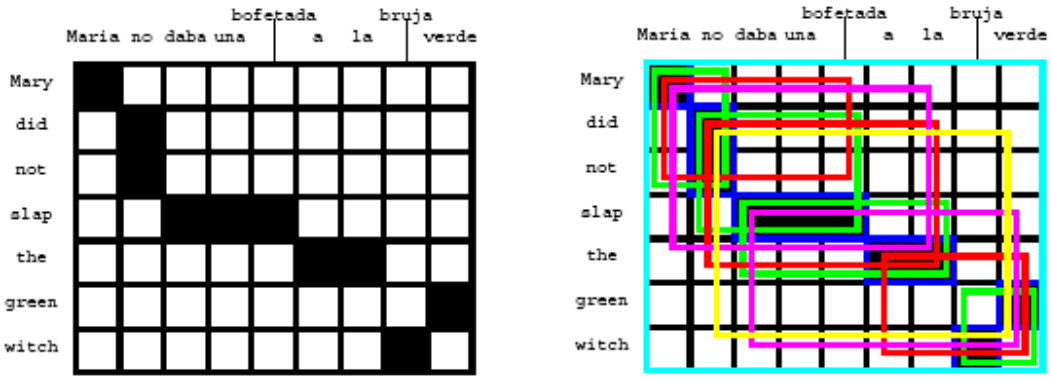


FIG. 17. EXPECTATION MAXIMIZATION – PARAMETER ESTIMATION FROM THE CONNECTED CORPUS (FRENCH TO ENGLISH)

Word Alignment Induced Phrases

Next we must understand what to do with these probabilities. One technique is to use Word Alignment Induced Phrases, which is where you form a grid like the one seen below in Fig. 18. A phrase alignment has to contain all alignment points for all words it covers [7]. Notice where the squares are dark, this is where the words match each other in meaning fully or partially. Refer to the coloured squares and to the broken text represented underneath. For each increment, we are able to link more of the squares together, and put together more meaningful pieces of text, until finally the whole grid is grouped in the teal blue (largest) box, which marks the finished translation. Notice how this method was able to reorder the words "green witch" correctly, as the equivalent Spanish words appear the other way around.



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch), (Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch), (Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

FIG. 18. WORD ALIGNMENT INDUCED PHRASES (SPANISH TO ENGLISH)

Syntax Trees

Another popular method is using Syntax Trees. Look carefully at Fig. 19 below, notice how the sentence is formed into a tree like structure that represents the grammatical structure of the text using tags for the parts of speech [7]. These words are then reordered to where they would syntactically most likely be in the target language of Japanese. Following this the necessary Japanese parts of speech are inserted if required, and then lastly the English words are translated.

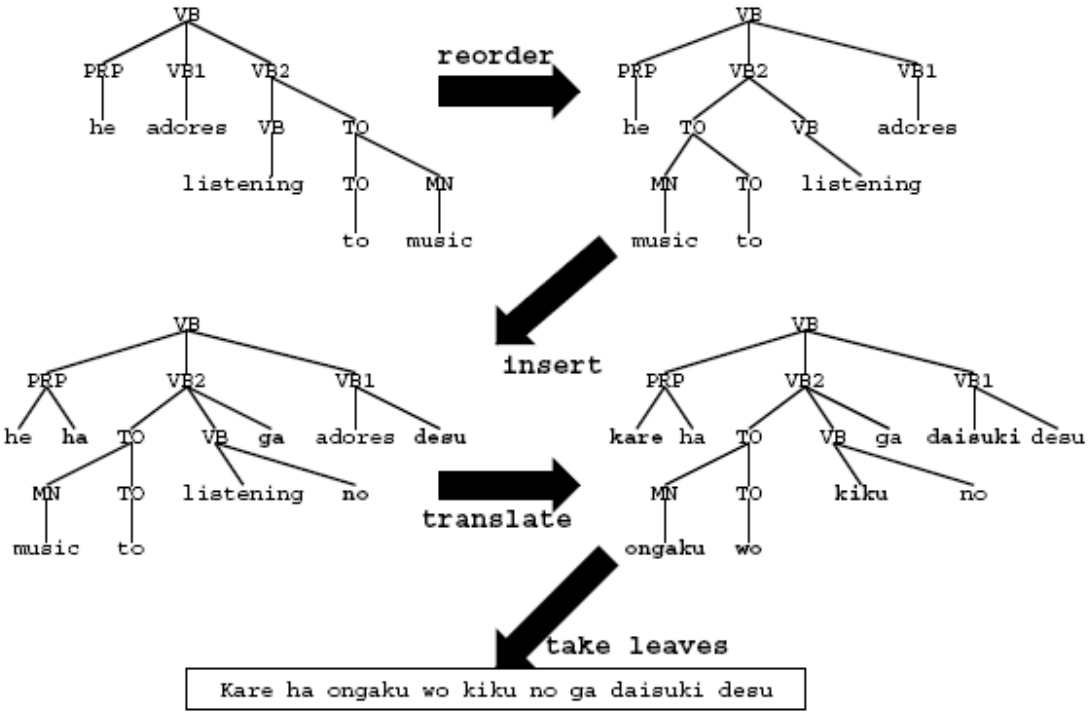


FIG. 19. ENGLISH TO JAPANESE SYNTAX TREE

Word Beaming

Now we will review a technique used in the language model. What you see below is our already translated sentence, but now using what is called a Beam Word Graph, we can probabilistically ensure that our translation reads well in English. Words are chosen, and then probabilities of other words occurring after these chosen words are weighted. If there is a strong enough probability, the path continues to beam across, until no more significant probabilities remain. The path that is able to travel the furthest distance across is most likely the best reading translation according to our language model. Fig. 20 below demonstrates this. "Mary did not slap the green witch" turned out to be the most probable answer [7].

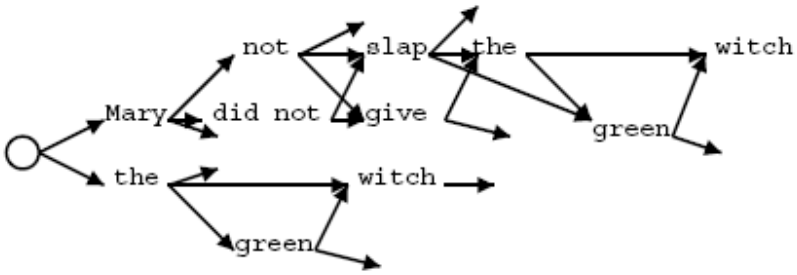


FIG. 20. BEAM WORD GRAPH

Preferred Applications for Corpus Based MT Methods

Both CBMT methods have applications that they are best suited to. EBMT as you may have guessed is best suited to applications where an MT system may have physical restrictions on how much storage it can have. In these cases, EBMT is the best option. EBMT is usually used in predictable situations or by using generalizations in which for both corpus size can be reduced. Portable / hand held MT systems most probably use EBMT since such devices do not have such large amounts of storage. The Google provided corpus once indexed into a database was approximately 300GB in size. Add on top of that Kaleido K's bilingual dictionary and other miscellaneous memory costs, it is easy to see why EBMT is the best choice with its ability to reduce corpus size.

SMT is best used when there are no physical limits on processing power or storage space. SMT is also capable of putting out a translation no matter how unique the original translation was. Thus a broader range of topics can be translated by an SMT system. However SMT usually has a few drawbacks. If the language is not very widely spoken, obtaining enough linguistic data for the SMT system to be effective may not be possible. Secondly, SMT usually does not make efforts to uphold grammatical rules, and relies on statistics to ensure correct grammar is adhered to; however occasionally translations are grammatically sound but completely illogical. EBMT can usually avoid this, because it is drawing from human translated examples, and merely substituting appropriate words.

In short usually EBMT excels in logical translations and reduction of processing resources required. On the other hand SMT usually excels in being able to tackle a diverse variety of translations and doesn't require any syntactic analysis, just the checking of translation fluency.

2.1.3 Hybrid Machine Translation

The two primary approaches to MT both have similar and different merits; however neither is able to outperform the other in all aspects of MT. Due to this fact, many MT systems being developed today are actually a hybrid of the two approaches RBMT and CBMT in order to acquire the best of both approaches and mitigate their weaknesses. Overtime 2 general methods have formed under the umbrella of the Hybrid MT approach, which are Sequential MT and Multi-Engine MT.

Sequential Hybrid Machine Translation

Sequential Hybrid Machine Translation (SHMT) uses both approaches, but in a sequential manner, with either one of the approaches completing the translation first, then the other filling the role of post editing the translation. The National Research Council of Canada has published their success with SHMT, using SMT as an Automatic Post Editing (APE) process to follow the RBMT process [8]. PORTAGE and SYSTRAN were the chosen MT systems for each approach respectively. The results obtained from the SHMT system surpassed results obtained with these same SMT and RBMT systems alone. Below in Fig. 21 are the results that were obtained, as you can see, despite the SMT system catching up in accuracy eventually, the SHMT had a much faster learning curve as the number of training sentences increased. This demonstrates that the SHMT system is able to perform just as well as SMT with much less linguistic data. RBMT systems are usually the ideal choice for language pairs with limited linguistic data, but this SHMT system has also proven itself to be a competitive solution with its synergizing of both MT approaches.

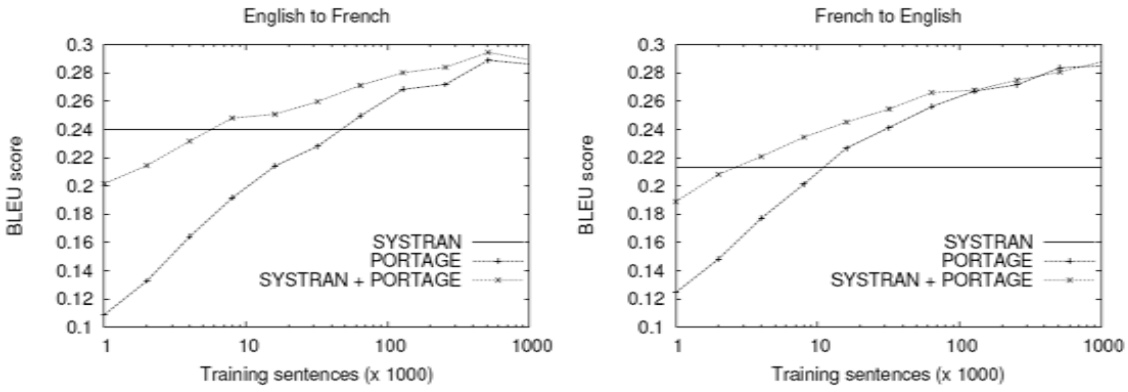


FIG. 21. BLEU SCORES ON UNDER INCREASING AMOUNTS OF TRAINING DATA FOR PORTAGE SMT ALONE AND SYSTRAN MT WITH PORTAGE APE [8].

Multi-Engine Machine Translation

Multi-Engine Machine Translation (MEMT) also uses both RBMT and SMT approaches, and which one to use at a given time is usually based on a selection criterion. An example of a selection criterion could be: When there are a lot of linguistic resources available, use SMT, and when there are not a lot of linguistic resources available, then use RBMT. MEMT can also operate without a selection criterion and simply use all methods of both RBMT and SMT all at once, then simply output what it calculates to be the most sensible result and even possibly perform some post editing. These are just common set ups for MEMT, in fact the way in which an MEMT system is programmed to run is really up to the imagination of its creator. An architectural example of one can be observed below in Fig. 22, which is an MEMT system created by researchers from a group of universities and research institutes in Germany [9].

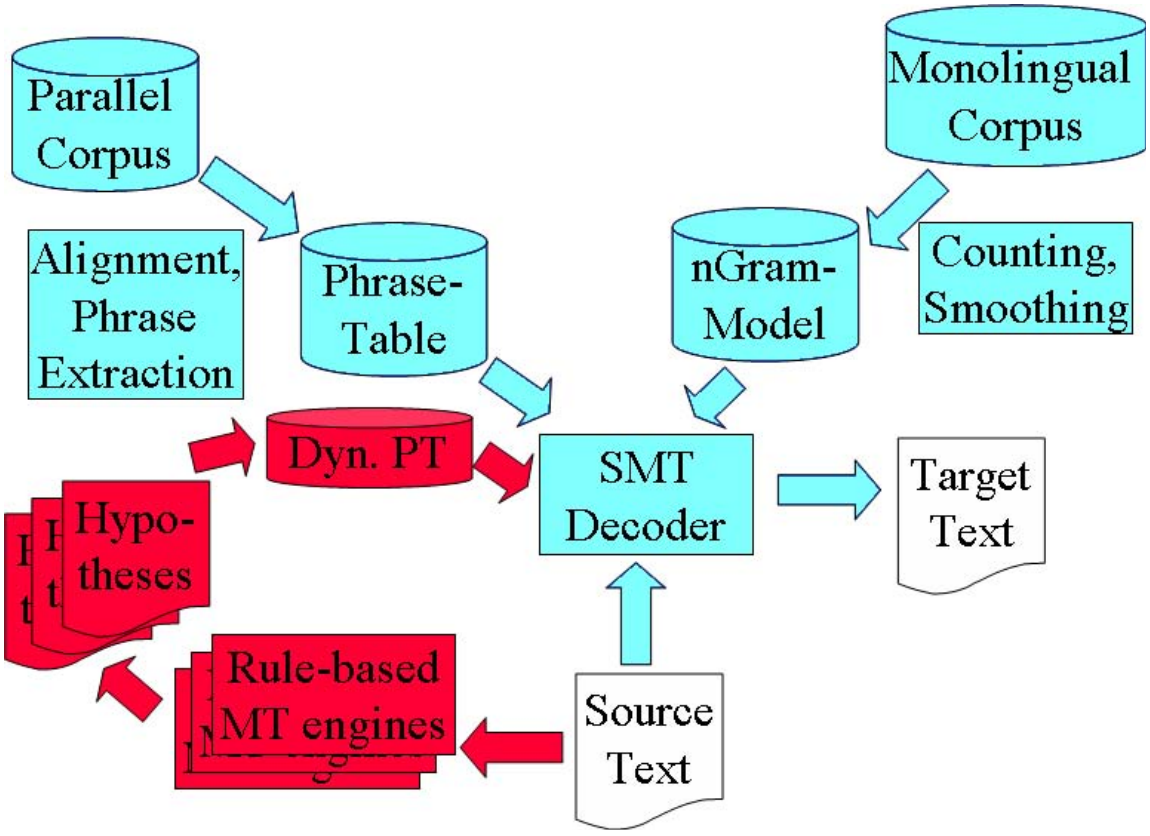


FIG. 22. ARCHITECTURE FOR A MULTI-ENGINE MACHINE TRANSLATION DRIVEN BY A STATISTICAL MACHINE TRANSLATION DECODER[9]

This MEMT system has 3 engines, 2 SMT engines (in teal blue / lightly shaded) and 1 RBMT engine (in red / darkly shaded). All 3 engines attempt to translate the source text into the target language. Each engine attempt is feed to the SMT decoder, which then attempts to render each translation so it is statistically more fluent, which may be harmful to linguistically correct translations formed by the RBMT engine. Finally the most probable translation is output as the final translation. Here are the results below in Table 5 obtained by this MEMT system, when compared to 2 other pure RBMT systems and 1 pure SMT system. Table 5 indicates how often the different MT systems were unable to translate a word, so the less words an MT system is unable to translate, the better its performance is. The Reference Base is the perfect score, as these include words that should not be translated (such as acronyms i.e. USA). As you can see, the MEMT system is easily able to obtain a better result, since at least one of its translation engines are capable of solving most ambiguities that it can encounter.

TABLE 5. UNTRANSLATED TOKENS (EXCLUDING NUMBERS AND PUNCTUATIONS) IN OUTPUT FOR NEWS COMMENTARY TASK (GERMAN TO ENGLISH) FROM DIFFERENT MT SYSTEMS[9]

MT Systems	Number of Tokens Unable to be Translated
Reference Base	2091 (4.21%)
RBMT 1	3886 (7.02%)
RBMT 2	3508 (6.30%)
SMT	3976 (7.91%)
Hybrid (MEMT)	2425 (5.59%)

2.2 EVALUATION OF MACHINE TRANSLATION

Now we have discussed the various MT approaches commonly used, it's now time to cover how all of the MT methods of these approaches are evaluated. Human evaluation of MT can be time consuming and expensive, and it is also difficult to make human evaluations comparable with each other since each evaluation is subject to each human's ability to objectively judge a good translation. Hence a quicker, more inexpensive and automatic method of MT evaluation is required. One method of MT evaluation known as BLEU has grown to be widely popular as it was one of the first automated MT evaluation methods that correlated well with human judgement. BLEU stands for Bilingual Evaluation Understudy. While other more recent MT evaluation methods may correlate slightly better to human judgement, BLEU is still the most widely adopted and understood evaluation standard used to evaluate MT today. Therefore in the next few pages we will go into further detail regarding BLEU, and then we will compare it to other industry standard methods of MT evaluation and discuss the advantages and disadvantages of each.

2.2.1 The BLEU Score

To explain in simple terms, a BLEU score is calculated by firstly acquiring 3 or 4 reference translations of the source text by highly distinguished human translators of that respective language pair. These translations are then labelled Reference 1, Reference 2 and Reference 3 respectively. Secondly each MT system is then commanded to translate the source text, and these translations are called the Candidate translations. Hence there are Candidate 1, Candidate 2 and so forth for each respective MT system. Each candidate translation is then compared word for word to the 3 reference translations; the closer each translation matches word for word, the higher the BLEU score. 0.00 meaning not one word matched between a candidate translation and the reference translations, and 1.00 meaning that a candidate translation was able to match at least one reference translation word for word.

In order to pan out a well rounded BLEU score, of course many (usually 1000) translations are executed in order to obtain an overall BLEU score for each MT system. This counters the problem where in some translation instances they may be very accurate however because these translations do not directly match with the reference translations they acquire a bad BLEU score. Multiple BLEU score calculations should help find a well rounded BLEU score that illustrates well the MT system's correlation to human judgement. Secondly to avoid long candidate translations acquiring a better score because they simply have more words to match with the reference translations, a Brevity Penalty is also included into the BLEU score calculation. Candidate translations that are much longer than the closest matching reference translation are penalized for being too long, as a result lowering their BLEU scores. Lastly one final detail to mention is that BLEU scores calculated without the same number of reference translations, input text and linguistic data to draw from should not be directly compared to each other as the scores were obviously calculated under different testing conditions.

The following examples of how a BLEU score is calculated is directly taken from one of IBM's officially published papers on BLEU [10]. We will use these examples to highlight the fundamentals of calculating a BLEU score, and a little later we will go through an actual calculation to show the reader how to calculate a BLEU score by himself/herself.

2.2.2 BLEU Score Example 1

Candidates (Machine Produced)

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

References (Human Produced)

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

In the above example, Candidate 1 shares the most words in common with the 3 Reference translations. Take the phrase "It is a guide to action" for example; this appears exactly, or in a relatively similar form in all 3 Reference translations. This is what achieves a higher BLEU score.

2.2.3 BLEU Score Example 2

Candidate (Machine Produced)

Candidate 1: the the the the the the the.

References (Human Produced)

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

2.2.4 Modified N-Gram Precision

We can use example 2 above to demonstrate some of the basic mathematics involved in the calculation of a BLEU score. Candidate 1 is a dummy translation and is used here to make obvious the general calculation of a BLEU score. Despite it being a meaningless translation, it still can achieve a *modified unigram precision* of $\frac{2}{7}$. A modified N-Gram precision is calculated by using formula (1).

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{N-Gram \in C} Count_{clip}(N-Gram)}{\sum_{C \in \{Candidates\}} \sum_{N-Gram \in C} Count(N-Gram)}$$

(1) THE EQUATION USED TO CALCULATE A MODIFIED N-GRAM PRECISION

Calculations are done according to N-Grams. A *unigram* measurement searches for single word matches, a *bigram* measurement searches for two consecutive word matches and a *trigram* measurement searches for three consecutive word matches and so forth. Later we will outline some of the problems with BLEU, because such a senseless translation such as example 2 can in some cases achieve a modest BLEU score, it is often stated that BLEU can favour SMT over RBMT at times. Because of this, the BLEU score should not be considered as the ultimate indicator of success for any group of MT systems.

2.2.5 Raising the Order of the N-Gram Precision

As the modified N-Gram calculation is calculated to a higher order, the overall N-Gram precision decreases and Fig. 23 below demonstrates this. The dark blue represents human translations, and the light blue represents MT.

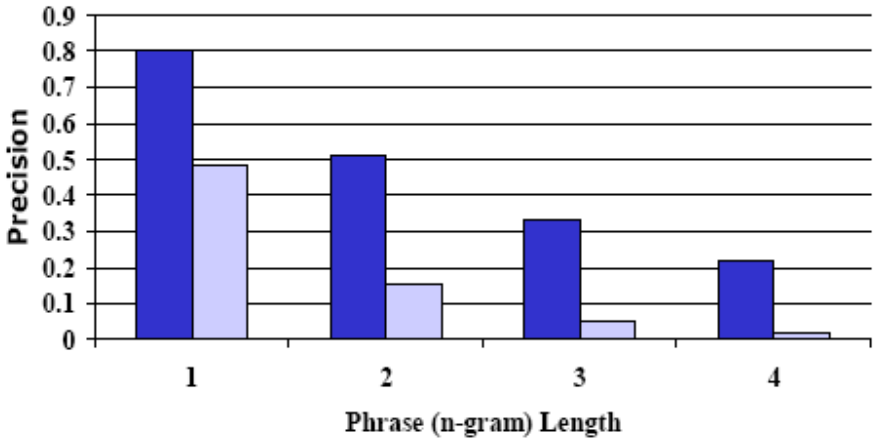


FIG. 23. DISTINGUISHING HUMAN FROM MACHINE

As can be seen, it is not hard to achieve high *modified unigram precision*; however it is very difficult to achieve a high *modified quadgram precision*. An actual BLEU score is not taken from any individual modified precision calculation, but rather from the degree of logarithmic decay that occurs between each modified N-Gram precision calculation. BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified N-Gram precisions [10]. An N-Gram length of up to 3 or 4 N-Grams is standard for BLEU calculations.

2.2.6 Example of BLEU Score Calculation

Now the fundamentals of BLEU have been discussed, it is time to go step by step through an actual BLEU score calculation. Equation (2) is the official equation used to calculate a BLEU score.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

(2) THE EQUATION USED TO CALCULATE A BLEU SCORE

Breaking Equation (2) down into its individual parameters p_n is the Modified N-Gram Precision for each value of *n*, as shown earlier in Equation (1) on page 43, in which the geometric mean of $p_1, p_2 \ldots p_n$ can be obtained. As for the other parameters, in most cases where 4 reference translations are used *N = 4* and $w_n = {}^{1}/_{N}$. Lastly BP is the Brevity Penalty, and is calculated with Equation (3).

$$\text{BP} = \begin{cases} 1 & \text{if } c > \\ e^{(1-{}^{r}/_{c})} & \text{if } c \leq r \end{cases}$$

(3) THE EQUATION SET USED TO CALCULATE THE BREVITY PENALTY

In Equation (3) *c* is the length of the candidate translation and *r* is the effective reference translation length.

Now we have explained the required equations used to obtain a BLEU score, let us calculate the BLEU score for this candidate translation against the following 4 reference translations. This example was obtained from lecture slides at Carnegie Mellon University's Language Technology Institute [11].

Candidate (Machine Produced)

Candidate 1: the gunman was shot dead by police.

References (Human Produced)

Reference 1: the gunman was shot to death by the police.

Reference 2: the gunman was shot to death by the police.

Reference 3: police killed the gunman.

Reference 4: the gunman was shot to dead by police.

In this example *r = 9* and *c = 8*, thus *c < r*, so the Brevity Penalty for this example is:

$$BP = e^{(1-9/8)} = 0.8825$$

(4) EXAMPLE BREVITY PENALTY CALCULATION

TABLE 6. CALCULATED MODIFIED N-GRAM PRECISIONS

p_1	p_2	p_3	p_4
the	the gunman	the gunman was	the gunman was shot
gunman	gunman was	gunman was shot	gunman was shot dead
was	was shot	was shot dead	was shot dead by
shot	shot dead	shot dead by	~~shot dead by police~~
dead	dead by	~~dead by police~~	~~dead by police .~~
by	~~by police~~	~~by police .~~	
police	police .		
.			
8/8 = 1.0	6/7 = 0.86	4/6 = 0.67	3/5 = 0.6

Now take a look at Table 6, this illustrates how the Modified N-Gram Precisions are calculated. This makes it easier to understand what is happening Equation (1). Shaded N-Grams such as "the gunman" and "was shot dead by" occurred at least once in the reference translations, however the crossed out N-Grams such as "dead by police" and "by police" did not occur in any of the reference translations. Therefore a Modified N-Gram Precision is simply a ratio illustrating how many N-Grams occurred for each order of N-Grams. From these Modified N-Gram Precisions, we can then obtain the geometric mean, and then by plugging in all our other calculated parameters into Equation (2) we can calculate the BLEU score, which can be seen below in Equation (5).

$$\text{BLEU} = 0.8825 \times \sqrt[4]{(1.0 \times 0.86 \times 0.67 \times 0.6)} = 0.68$$

(5) EXAMPLE BLEU SCORE CALCULATION

So our example candidate sentence "The gunman was shot dead by police." was able to acquire a BLEU score of 0.68. A BLEU score calculation is rather simple after you have seen it done once. We have actually created an algorithm that automatically calculates a BLEU score for an MT system. This will be accessible on the Kaleido K website shortly. Simply upload your candidate and reference translations accordingly in Microsoft Excel file format, click the submit button and a BLEU score for your MT system is provided for you.

2.2.7 Alternative Evaluation Techniques

NIST

After the success of BLEU and its wide adoption, the inadequacies of BLEU came to surface. NIST in many respects tries to rectify these inadequacies and improve the correlation between human and machine evaluation. In the calculation of a BLEU score, Modified N-Gram Precisions are given the same weight ($w_n = {}^{1}/_{N}$). This means if larger order N-Grams are not able to be found in the reference translations, the BLEU score can be dragged down significantly. A reasonable amount of reference translations can cancel out this negative effect; however this should not be relied upon. In calculating the geometric mean NIST is able to rectify this, by giving more weight to higher order Modified N-Gram Precisions when making matches to the references translations, and less weight to lower order Modified N-Gram Precisions. Secondly, BLEU is not able to differentiate between N-Grams which are either rich or not rich in content. NIST gives more importance to N-Grams such as “machine translation” as opposed to “at the”, in turn rewarding translations that provide more content (adequacy) as well as fluency.

F-Measure

In papers concerning F-Measure, it has been stated to outshine both BLEU and NIST. One strength that F-Measure has over BLEU and NIST is that the length of the candidate translation and the reference translations is irrelevant. Therefore the chance of the Brevity Penalty that BLEU and NIST share attributing negatively towards a well rounded translation is negated. A sound MT method should evaluate both adequacy (inclusion of meaningful content) and fluency (it sounding natural). F-Measure only rewards when both adequacy and fluency are simultaneously present, thus it makes a good MT evaluation method for correlating well with human judgement. Take a look at Table 7 to understand this idea. This graphically represents what F-Measure uses known as Maximum Match Size (MMS). On the left is the candidate translation and on the bottom is the reference translation, graphically MMS is an attempt to find the squares that hold the largest aligned matches as possible whilst avoiding repetition at the same time.

TABLE 7. GRAPHICAL REPRESENTATION OF F-MEASURE [12]

E					¤					
D				¤						
C			¤							¤
I									¤	
A	¤							¤		
B		¤					¤			
C			¤							¤
H										
	A	B	C	D	E	F	B	A	I	C

Without getting technical, F-Measure rewards candidate translations that manage to fill more squares such as in the above graphical representation. Thus the best candidate translation is the translation that can fill the most squares in the grid.

The two graphs below are taken from a paper written on F-Measure. Here we can see how BLEU, NIST and F-Measure compared against each other in respect to Adequacy and Fluency. These results were acquired by using Spearman correlation on a Chinese corpus using a single reference. As you can see F-Measure has better correlation that NIST, and NIST has better correlation than BLEU, which aligns with what we have stated so far.

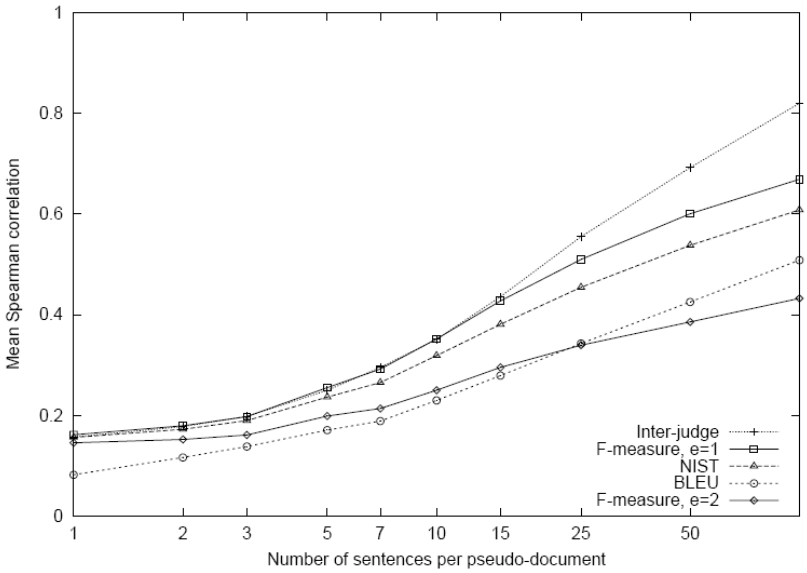


FIG. 24. MACHINE TRANSLATION EVALUATION OF ADEQUACY CORRELATION [12]

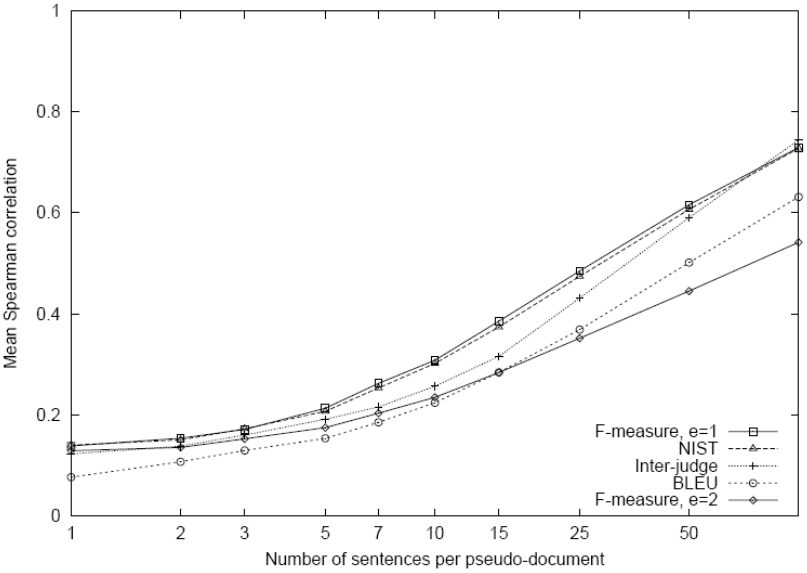


TABLE 8. MACHINE TRANSLATION EVALUATION OF FLUENCY CORRELATION [12]

METEOR

METEOR was developed at Carnegie Mellon University, and works in a similar fashion to BLEU but it only requires one or two reference translations and only needs to compare unigrams. It is able to do this through WordNet which is a freely available large net of words that are linked together by their synonymous definitions. When comparing the candidate translation to the reference translations, WordNet enables synonyms and words with identical stems to also be valid matches against the reference translations. As mentioned before, BLEU counters the negative effect of well rounded translations being overlooked by using a larger number of reference translations; however this can be costly on a large scale, so in this respect METEOR can be a better alternative method of evaluation as it only requires one or two reference translations due to its use of WordNet. METEOR also has a reordering penalty which is based on how much the candidate translation must be shuffled around to match one of the reference translations.

2.3 WHO LEADS MACHINE TRANSLATION?

Now a general understanding of the BLEU score has been established, we can discuss the success of different approaches and the companies that use them. These companies have all obtained a BLEU score on their respective MT systems, so an understanding of the BLEU score calculation method can help one appreciate the level of success that has been achieved. These are the scores for the National Institute of Standards and Technology (NIST) Open MT Evaluation 2008. The NIST entrants are documented in more detail in reference to their respective NIST IDs in Appendix A.

TABLE 9. NIST 2008 BLEU-4 RESULTS FOR ARABIC TO ENGLISH

ARABIC to ENGLISH	
Constrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.4557
IBM-UMD	0.4525
IBM	0.4507
BBN	0.434
LIUM	0.4298
ISI-LW	0.4248
CUED	0.4238
SRI	0.4229
Edinburgh	0.4029
UMD	0.3906
UPC	0.3743
Columbia	0.374
NTT	0.3671
CMUEBMT	0.3481
QMUL	0.3308
SAKHR	0.3133
UPC.LSI	0.3021
BASISTECH	0.2529
AUC	0.1415
Unconstrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.4772
IBM	0.4717
Apptek	0.4483
CMU-SMT	0.4312

TABLE 10. NIST 2008 BLEU-4 RESULTS FOR CHINESE TO ENGLISH

CHINESE to ENGLISH	
Constrained Data Track	
NIST ID	**BLEU-4 Score**
MSR-NRC-SRI	0.3089
BBN	0.3059
ISI-LW	0.3041
Google	0.2999
MSR-MSRA	0.2901
SRI	0.2697
Edinburgh	0.2608
SU	0.2547
UMD	0.2506
NTT	0.2469
NRC	0.2458
CASIA	0.2407
NICT-ATR	0.2269
ICT	0.2258
JHU-UMD	0.2111
XMU	0.1979
HITIRLab	0.1866
HKUST	0.1678
ISCAS	0.1569
NTHU	0.0393
Unconstrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.3195
CMU-SMT	0.2597
NRC-SYSTRAN	0.2523
UKA	0.2406
CMUXfer	0.131
BJUT	0.0735

TABLE 11. NIST 2008 BLEU-4 RESULTS FOR ENGLISH TO CHINESE

ENGLISH to CHINESE	
Constrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.4142
MSRA	0.4099
ISI-LW	0.3857
NICT-ATR	0.3438
HITIRLab	0.3225
ICT	0.3176
CMUEBMT	0.2738
XMU	0.2502
UMD	0.1982
Unconstrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.471
BJUT	0.2765

TABLE 12. NIST 2008 BLEU-4 RESULTS FOR URDU TO ENGLISH

URDU to ENGLISH	
Constrained Data Track	
NIST ID	**BLEU-4 Score**
Google	0.2281
BBN	0.2028
IBM	0.2026
ISI-LW	0.1983
UMD	0.1829
MITLLAFRL	0.1666
UPC	0.1614
Columbia	0.1459
Edinburgh	0.1456
NTT	0.1394
QMUL	0.1338
CMU-XFER	0.1016

2.3.1 SYSTRAN

SYSTRAN is the MT system behind Yahoo and AltaVista's Babel Fish. The origins of SYSTRAN technology began to form after the success of the Georgetown experiment (discussed in the next section) and the funding that followed in 1968. Since then they have provided translation solutions for the United States Department of Defence and the European Commission [13]. They cover a wide spectrum of languages, but have mainly dealt with the languages pairs of English/French and English/Russian, especially as their work during their earlier years was influenced by motivations of the Cold War.

The roots of SYSTRAN's methods are very influenced by their earlier work, which means they are a leader in RBMT technology. Therefore their methods reflect the traditional ways of MT which is also reflected in their recent BLEU scores which do not rank as highly as companies that use CBMT methods in their approach. They were able to achieve a BLEU score of 0.2523 for Chinese to English with an unconstrained data track, which placed them 3rd in this category. This is not actually a bad score since as we mentioned earlier, at times BLEU tends to favour SMT. SYSTRAN could close the BLEU gap more perhaps if they used their existing RBMT technology and focused more on MEMT systems.

2.3.2 Google

Google used SYSTRAN's MT technology until a few years ago when they decided to design their own MT system. This was obviously a good decision on their part when you consider Google's BLEU scores which were 0.4557 and 0.2999 for Arabic to English and Chinese to English respectively. In fact Google's own design achieved very competitive BLEU scores across the board at the NIST Open MT 2008 Evaluation.

They use a CBMT approach, with SMT as the method, and not only that, they also often use English as an interlingual language to reduce dictionary/data costs where they lack bilingual corpora for relatively unpopular language pairs. So for example when you translate Vietnamese to Polish, Vietnamese is first translated to English, then to Polish. Google has huge amounts of data and computing power at their disposal with the added advantage of having a huge presence on the internet, where bilingual corpora can be obtained and users can access and test the translator. Users can also directly contribute to the bilingual corpora. That means Google is obtaining free tweaking to their MT system so that the quality of its output is further refined the more it is used [14].

2.3.3 Carnegie Mellon

Carnegie Mellon University has published a lot of MT based research. In the NIST Open MT 2008 they entered 3 different MT systems, two of which were named CMUEBMT and CMUSMT, which we can safely assume were an EBMT system and a SMT system respectively. Overall they achieved a very mixed bag of results. With their SMT system, for the Arabic to English unlimited data track, they achieved quite a respectable score of 0.4312 in which the highest score was 0.4772 achieved by Google. With their EBMT system, for the English to Chinese constrained data track, they achieved a midrange score of 0.2738 in which the highest score was 0.4142 again achieved by Google. Other than that, their other scores did not really rank highly. Overall Carnegie Mellon University uses a variety of MT methods. Perhaps we can expect a successful hybrid MT system from them in the future as they continue to explore multiple MT methods simultaneously.

2.3.4 Who is the leader?

In our opinion, it is still too early to tell who the leader of MT is. For 60 years the ultimate solution to MT has always appeared to be around the corner. It's like climbing a mountain with the peak always appearing in site, but you only reach that peak to find that there is another peak beyond it. What we can say about the race to solve MT is that it's still open to anybody, especially since in this age we share information with each other so often. Every time someone comes up with a slightly better solution, it is quickly learned by the rest of the MT development community. Especially because competitors of the Open MT Evaluations each year (the MT Olympics) are obligated share their secrets of success with others at the follow up work shop each year. The development of MT progress in the future is likely to be steady, and success will ultimately be achieved by contributions from all researchers from all areas of the MT development community.

2.4 A BRIEF HISTORY OF MACHINE TRANSLATION METHODS

The history of MT development has quite distinct periods. In each period the researchers and developers of MT methods get inspired off into another direction. Usually the new direction of MT methods is perpetuated by the limitations of previous methods and the promising prospects of more recent methods. In this brief history we will not go into specific details about all the translation feats achieved and MT systems designed, but rather highlight the key events and researchers that have had significant influence in the MT industry to the extent that they shifted the direction of the industry's development.

Before reading over the brief history, take a look at Fig. 25 on the following page. We have put together this illustration to help the reader quickly understand the key events in MT history and the general flow of MT development over time. This illustration should help the reader more easily digest the more specific details of MT history as they read on.

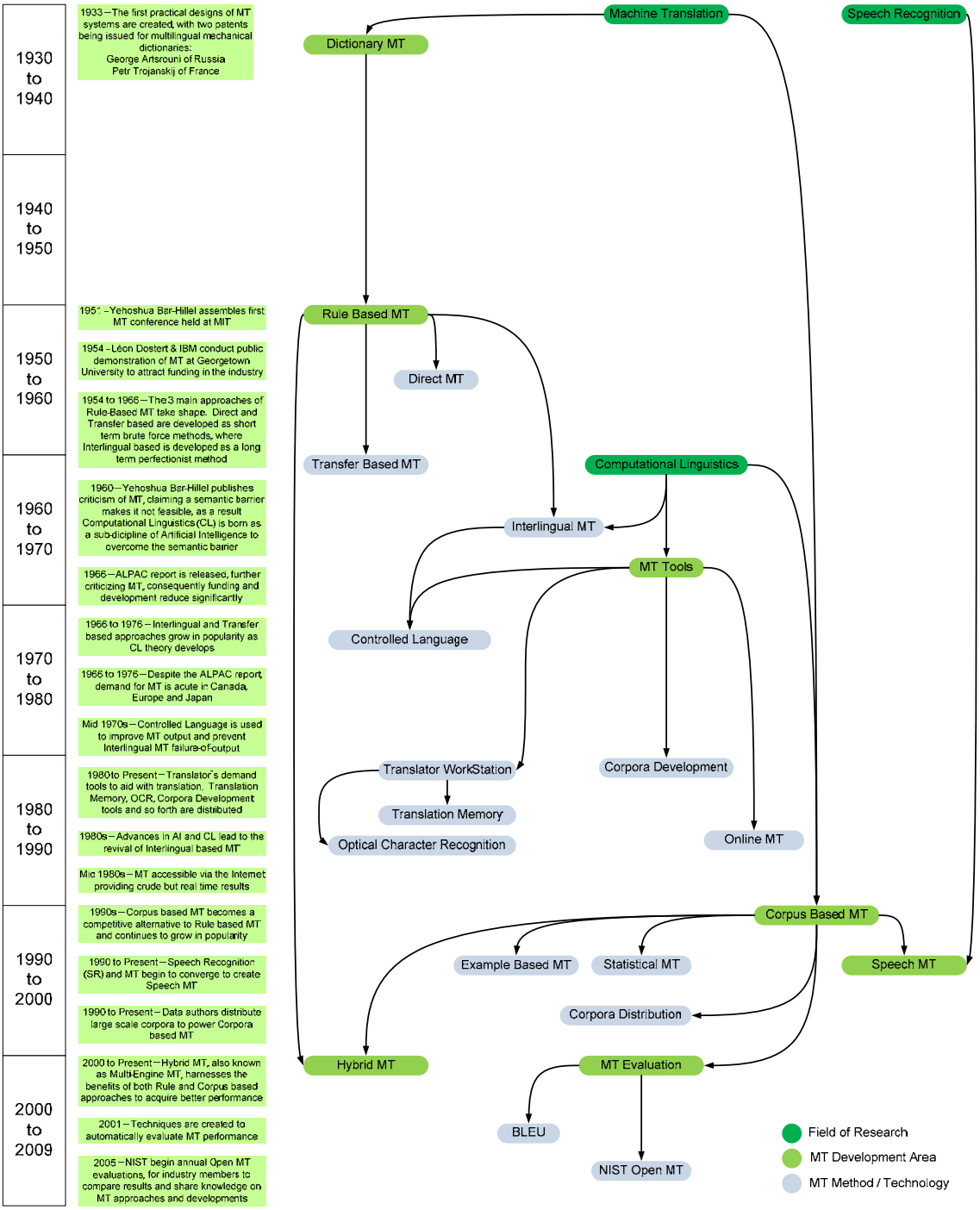


FIG. 25. DEVELOPMENT FLOW OF THE MACHINE TRANSLATION INDUSTRY WITH CHRONOLOGICAL ORDER OF KEY EVENTS

2.4.1 The Ignition of the Industry – (Pre 1954)

Over time researchers from different parts of the world came across the idea that the process of translation could be automated, and this is exactly what happened in the early days of the MT industry which can be recorded all the way back to the 17th century in which its possibilities were in various ways documented. In general, there were a lot of rudimentary translation devices and proposed ideas that were not interconnected in any way. In terms of who came up with a practical design for an MT system first, we could use MT patents as an indicator. The first two were issued in 1933, one to George Artsrouni of Russia, and one to Petr Trojanskij of France, who both had invented a type of multilingual mechanical dictionary.

Researchers remained largely unaccompanied in their endeavours until Yehoshua Bar-Hillel of Massachusetts Institute of Technology (MIT) organized the first MT conference at MIT in 1952 which almost everyone who had something to do with the MT industry was present. The most important aspect about this conference was the shared view that funding in the industry was required in much larger amounts. Léon Dostert of Georgetown University suggested the best way to do this was to get the industry some publicity as it had largely remained in the shadows up to this point in time.

So on January 7th 1954, Léon Dostert carried out his proposal at Georgetown University and collaborated with IBM using the IBM 701 mainframe computer. A carefully selected sample of 49 Russian sentences was translated into English, using a very restricted vocabulary of 250 words and just 6 grammar rules [15]. The media stated it was a major breakthrough and many proclaimed the MT process would be mastered within 3 to 5 years, flooding heavy investment into the MT industry as Léon Dostert and many others had hoped. Investment was not limited to the US; Russia and Europe were also motivated to increase research efforts in the MT industry. Motivated by the Cold War, the US and Russia were particularly focused on developing MT for Russian/English and English/Russian respectively.

2.4.2 Realizations & Reductions in Funding (1955 – 1966)

Now funding was abundant researchers quickly set out to work on a solution to MT. As progress was made it became quite clear to many that achieving seamless MT was quite an ambitious goal. Researchers became divided on how to solve the problem, with some wanting to design MT that used brute force to plough through dictionaries of words to obtain translations (which got results in the short term), while others wanting to design MT that was more elegant and based on language theory (which would get results in the long term). With development occurring down these two avenues, the three main genres of what would later be known as RBMT began to form. Direct MT was the adopted genre for brute force enthusiasts while Interlingual and Transfer Based MT were the adopted genre for the linguistic theory enthusiasts. Direct MT was the dominant method of this period, and MT systems based on this method were dubbed as the first generation of MT systems.

In 1960, Yehoshua Bar-Hillel, the enthusiast who put together the first MT conference in 1952, felt disillusioned and published his criticism of MT. His argument was that MT could not penetrate the semantic barrier that existed. His argument carried much weight, and he used a pen as an example to back up his point. A pen can have at least two meanings (a container for animals/children or a writing tool). In the sentence "The box was in the pen" we know that only the first meaning is plausible; the second meaning is excluded by our knowledge of the normal sizes of (writing) pens and boxes [15]. He carried on to point out that MT could not solve these problems without real world knowledge.

Of course with any funding, results are expected, however the RBMT methods the researchers were using were not capable of delivering on promises and funding into the industry steadily began to cease. By 1966 the ALPAC (Automatic Language Processing Advisory Committee) report was released exposing how the MT methods of that time fell short of fulfilling promises. The report largely agreed with Yehoshua Bar-Hillel's convictions and it stated that MT was slower, less accurate and twice as expensive as human translation and that "there is no immediate or predictable prospect of useful machine translation", which further stumped the development of the industry, particularly in the US [15].

2.4.3 Dormant Times (1967 – 1976)

While the demand for MT may have disappeared in the US, the demand for it remained in regions which unlike the US had persistent language barriers, such as Canada, Europe and Japan.
Canada as a bilingual nation required much bidirectional English/French translation in the government sector, while Europe with so many language barriers between each country needed even more multilingual MT solutions, and lastly Japan required much Japanese/English MT, particularly for the translation of published research. Thus development in the industry still took place, mostly outside of the US.

Even though the ALPAC report temporarily stumped MT development, it also forced it to take a new direction. The previous Direct MT methods that had been used were getting abandoned and Interlingual and Transfer based methods were being adopted. These methods if designed to do so, could partially account for real world knowledge, facilitating a primitive form of Artificial Intelligence (AI). Consequently the use of such methods started the beginnings of Computational Linguistics (CL), which is a sub discipline of AI. The use of CL in MT marks the second generation of MT systems.

2.4.4 Commercialization & Recovery of the Industry (1977 – 1989)

From the mid 70s MT began to recover from the crippling effects of the ALPAC report and had now been enjoying steady commercialization for about a decade. The quality of translations had also improved a lot for two particular reasons, the translations that were required were of a very specific nature, thus the use of a specific dictionary improved results, also companies such as Xerox had made efforts to use controlled language in translation, which is where only specific language constructs can be used as input into the MT system; this significantly removed the need for post editing after translations [15].

The mid 80s saw Interlingual techniques further rise in popularity with further advances in AI and computing technology. Interlingual MT methods also were being designed hand in hand with reversibility in mind, making MT systems fundamentally bidirectional. What were also being demanded in the market were MT tools to assist translators in increasing their productivity. This kicked off in the 80s, but by the 90s they were in big demand and a collection of tools together would be packaged and distributed as Translation Workstations. These included such tools as dictionary and terminology managers, translation memories and optical character recognition. The translators were not only demanding MT tools, so were learners of languages. Japan saw a huge development in hand held translation dictionaries. In France the Systran MT system was made accessible online via the Minitel network. Following this in the 90s, Systran along with Compuserve would make MT fully accessible on the internet. Systran's MT system went under the well known service Babelfish and eventually made available all their various language pairs.

2.4.5 The Diverse Needs & Solutions of the Translation Industry (Post 1990)

The 90s would see in the 3rd generation of MT systems which use a Corpus Based approach. Leading up to the 90s large amounts of linguistic data were not abundant and MT was predominately Rule Based, which required hours of skilled linguist's time to program the brain of MT systems. CBMT didn't require hours of cumbersome linguistic programming so was welcomed with open arms; it is in fact still the height of research interest in MT development today. The rise of CBMT popularity was partially fuelled by the explosion of data available all over the internet, which began to expand at an astonishing rate.

Since the internet has made it easier for the world to communicate more, the need for online instant translation has also increased, thus the amount of sites that offer this kind of service have popped up all over the internet. Secondly the internet has also seen the MT development community communicate a lot more. Developers now collaborate more on research and also share and distribute more linguistic resources with each other. The automated evaluation of MT systems using such evaluation schemes as BLEU in recent times has also become common place. This provides a feedback loop for the MT development community, and further helps them gauge the success of their various approaches to MT. A good example of this is the annual NIST Open MT Evaluation in which all the researchers meet together, test their MT systems against each other, discuss among themselves how they achieved their results and participate in MT workshops as well. This annual inter-collaboration and evaluation of MT, along with the many other similar conferences that are held each year, have sped up MT development considerably. This can be seen if you observe the progress of the BLEU score results of each consecutive Open MT Evaluation.

Since the world is globalizing more, MT has also spread across several different platforms and made its way into several aspects of everyday life. There are now many hand held MT devices, that range from character recognition pens to aid translators, to a variety translation applications offered by telecommunication companies to aid their mobile subscribers. MT usually requires heavy processing, but with improved hardware specifications both on mobile and stationary devices, MT is becoming more accessible than ever. We ourselves have previously developed our

own mobile MT system in 2006 and published the paper “Interactive Translation of Japanese to Korean” [16]. In our mobile MT system translations were executed using a mobile phone as the input/output device, and the translation could be processed at another location on a server and sent back to the mobile device. This is just a typical example of how diverse MT solutions are becoming in this day and age.

MT is not only crossing over to different platforms and devices; different MT related technologies are also converging together. Most of these other technologies fall under the umbrella of AI and are related to how humans communicate with the world and their environment. As we discussed in the previous chapter, MT systems find it hard to compete with human translators because they have only text as an input. So in order to close the gap, MT systems also require all the input data (speech, visual, smell, touch, knowledge) that a human has at its disposable when attempting a translation. A good example of technology convergence is the integration of speech recognition and synthesis technologies used with MT. Often they are being added on the front and back end of MT systems so the input and output of the MT systems can become completely audio-based.

To summarize, the history of MT has seen it continue to develop, and in the process become a more diverse and broader area of research. MT no longer just deals with the automated translation of text, it also deals with aspects of its convergence with other technologies, the development of more mobile and accessible solutions, the evaluation of its progress, the construction of the resources it uses and more. Most importantly the demand for MT technology continues to grow relentlessly, so we can expect the future history of MT to be a very exciting one.

2.5 SUMMARIZATION OF THE MACHINE TRANSLATION INDUSTRY

While Corpus Based approaches are considered the new way to solve MT, Rule Based approaches still offer benefits that are not able to be replaced, so neither approach can completely conquer the other. Fig. 26 below identifies the main advantages and disadvantages of each approach.

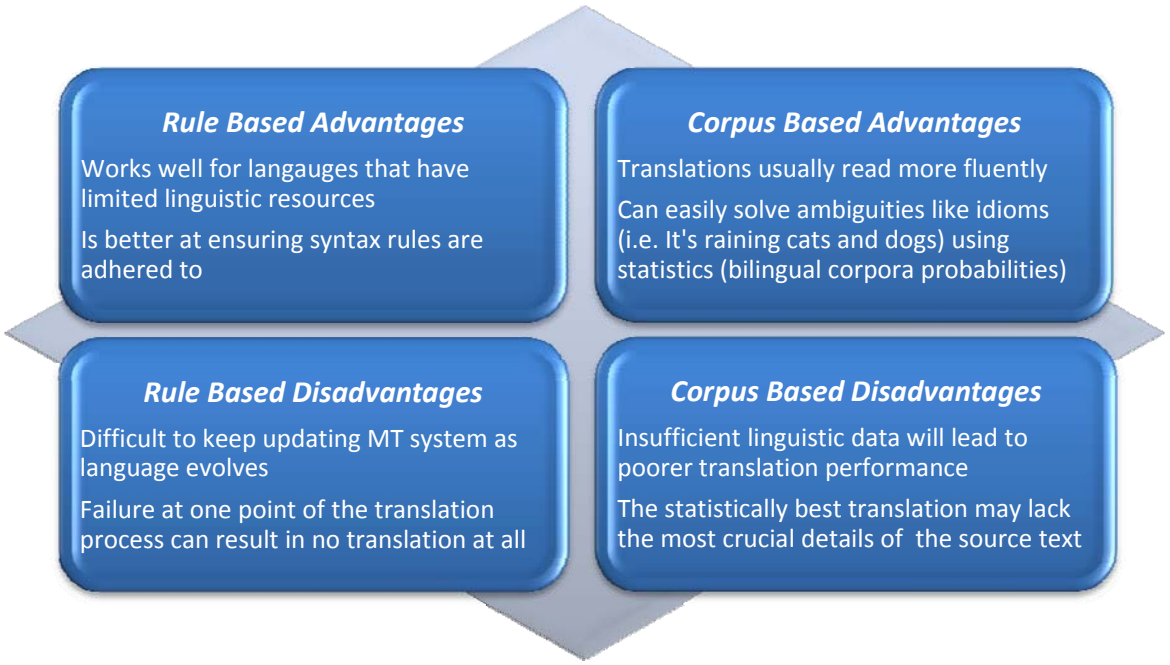


FIG. 26. ADVANTAGES AND DISADVANTAGES OF EACH MT APPROACH

For both approaches, RBMT and CBMT, their development can be described using the Vauquois Triangle as seen in Fig. 27. Development first starts at the bottom of the triangle where very basic methods are used to execute the specified approach. As the demand for more articulate translations increases, the respective methods used for each approach also grow in complexity.

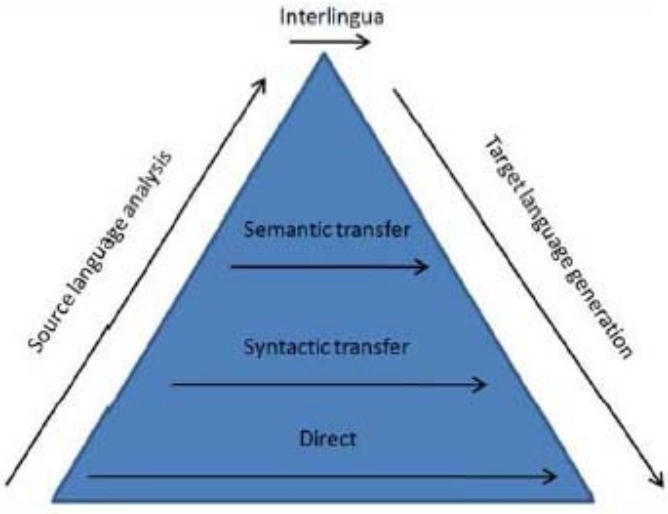


FIG. 27. THE VAUQUOIS TRIANGLE

RBMT started out from mechanical dictionaries, but later in order to overcome syntactic and semantic language barriers, Interlingual languages were also made use of. CBMT is has also developed in a similar fashion, it started out with primarily word to word translation using simple statistical models. Now for CBMT better statistical models are designed and the probabilities of not only words but phrasal equivalents are derived and used.

In Fig. 28 we can observe the development steps of each MT approach and by referring to the white (clear) and yellow (shaded) arrows, the areas of Ongoing Activity and Current Primary Focus Areas can be observed respectively. Notice how despite RBMT being the older approach has by no means been abandoned, and still much research is taking place into this approach. As for CBMT, the true potential of this approach has still not been realized, and heavy research efforts are being placed on Phrase and Syntax Based Models. What is not completely obvious here is because RBMT and CBMT are often used together in MEMT configurations, this is why research using both approaches still prevails. Notice the last staircase in Fig. 28, this is there to demonstrate that perhaps there are still more approaches to MT that we have not discovered and developed yet.

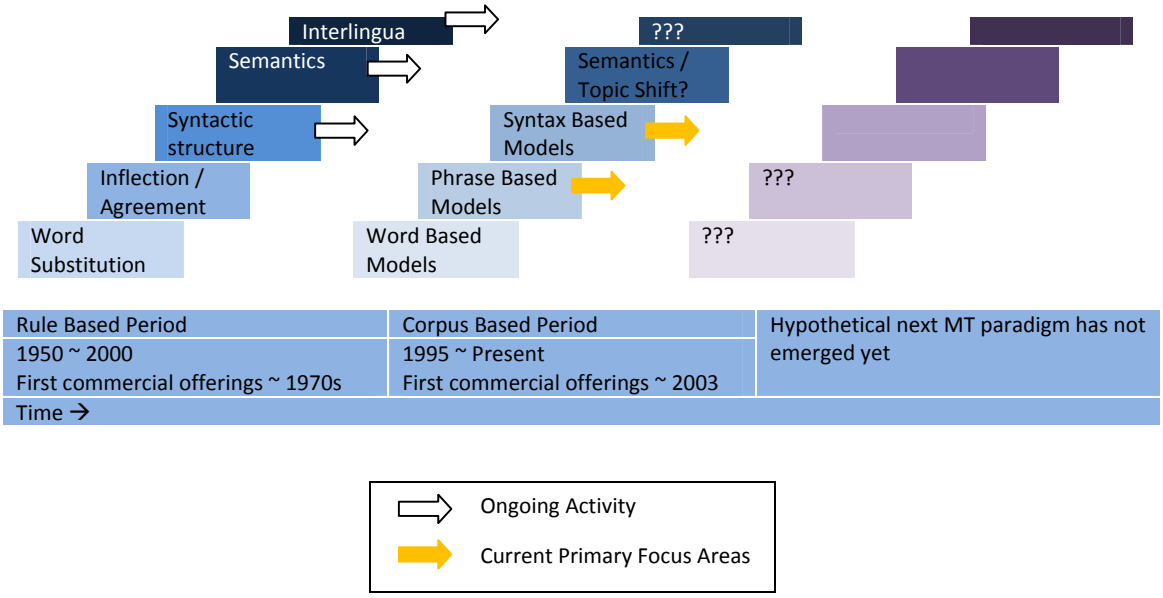


FIG. 28. DEVELOPMENT STEPS OF EACH MACHINE TRANSLATION APPROACH [17]

Another problem with the MT industry is the methods of evaluation. While BLEU is good because it is a fast and automated process, it has been pointed out from several members of the MT development community that it can favour CBMT. In recent evaluations RBMT has lagged behind CBMT, however to a human when looking at the candidate translations, that gap is not as big as the BLEU scores are illustrating it to be. Often RBMT provides more syntactically correct results, and CBMT can mix words up a bit and does not put as much importance on syntax. Despite this, CBMT still does well because it at least contains similar words to the reference translations in the candidate translation. Two things have been done to compensate this problem. Firstly RBMT developers are looking more into designing MEMT configurations to close the apparent gap BLEU illustrates. Secondly BLEU, and other similar scoring methods are annually reviewed and new evaluation methods are even being designed to ensure there is a fairer comparison between CBMT and RBMT in future evaluations. NIST, F-Measure and METEOR that were described earlier are good examples of new MT evaluation methods that correlate better with human judgement.

3. INTRODUCTION TO FLUENCY ENHANCEMENT

The purpose of this chapter is to explain in a very abstract manner what Fluency Enhancement (FE) is. In the following chapter we will then explain how we implemented FE into our MT system to improve translation quality.

3.1 A BRIEF OUTLINE

FE is a novel process that enhances the fluency of *text* through measures of reformation and evaluation. *Text* that undergoes this *Fluency Enhancement Process* is text that fails to read fluently because it is produced by a machine or an inarticulate person of language.

The scope of FE extends to various applications. Applications of the FE process are not limited to, but include improving the quality of MT, broadening search engine results, naturalizing text produced by an inarticulate person of language and more.

3.2 THE SUB-PROCESSES OF THE FLUENCY ENHANCEMENT PROCESS

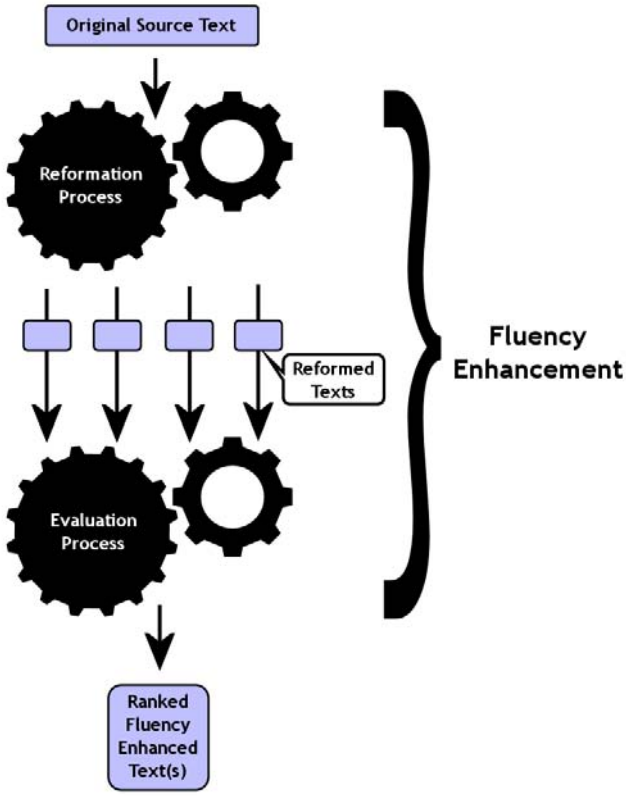


FIG. 29. THE FLUENCY ENHANCEMENT PROCESS

1. Depending on the type of application FE is used for, the original source text may need to undergo some specific pre-processing. *(Application Based Pre-processing)*
2. The original source text is reformed into numerous other texts that are relatively similar in meaning and have a higher level of fluency. *(Reformation Process)*
3. Each reformed text is evaluated and ranked in terms of fluency. *(Evaluation Process)*
4. Depending on the type of application FE is used for, one or more of the reformed texts are output in terms of their degree of fluency according to the output requirements of the application. *(Application Based Post Processing)*

3.3 STATISTICAL INFLUENCE

3.3.1 Web Sourced N-Grams

Several of the operations involved in FE are powered by the following statistical data in Table 13. The table holds N-Grams which are sequences of words; one word is a unigram, a two word sequence is a bigram and so forth. The statistical data in Table 13 represents how often each N-Gram appears on the internet. This data will be referred to herein forth as the *N-Gram Data*.

TABLE 13. SAMPLE N-GRAM DATA

N-Gram	Frequency
several different ideas in	76
several different ideas may	27
several different ideas of	52
several different ideas on	50
several different ideas some	16
several different ideas that	47
several different ideas to	65
several different ideas identities ,	37
several different ideas identities .	67

3.3.2 Massive Text Corpora

Several of the operations of FE are also powered by massive multilingual corpora. The statistical relationships that can be obtained from the massive multilingual corpora aid decision logic when it comes to obtaining similar text in identical or different languages. This data will be referred to herein forth as the *Corpora Data*.

3.4 SPECIFIC DETAILS OF THE SUB-PROCESSES

3.4.1 Application Based Pre-processing

FE can lend itself to several different applications. For some of these applications the text must undergo pre-processing before the general FE Process takes place. A typical instance of this is MT, which requires the text to be translated before it undergoes FE. As an example, consider this pre-processing measure described in Fig. 30:

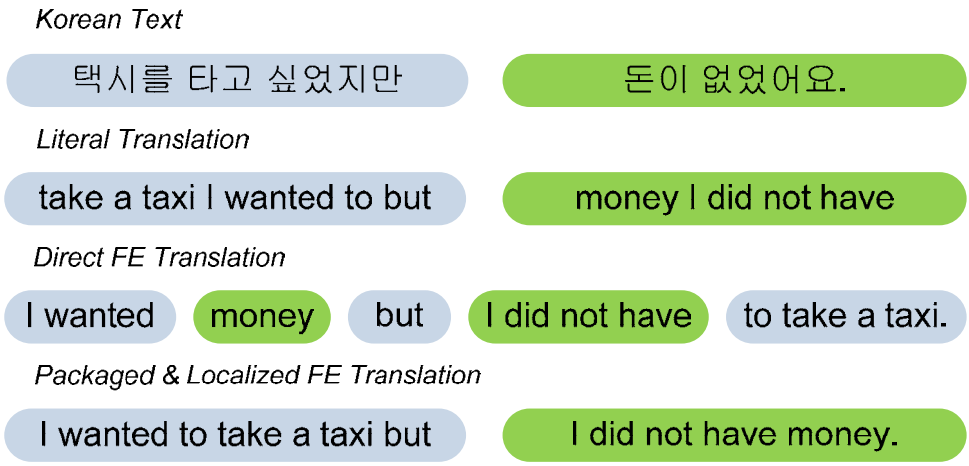


FIG. 30. THE INFLUENCE OF PACKAGING AND LOCALIZATION PRE-PROCESSING ON FLUENCY ENHANCEMENT

Before this Korean text undergoes FE, it can be put through the pre-processing measure of packaging and localization. This pre-processing measure ensures that the original meaning of each clause is packaged and kept localized, so when FE takes place, the clause boundaries do not spill over into each other mixing up words from different clauses. In the above example, notice how this pre-processing measure is able to preserve the original meaning of the text, yet still enhance its fluency. This is one of many typical pre-processing techniques that can take place to complement FE in its desired application.

3.4.2 Reformation Process

Main Operations

After any pre-processing stage, the original source text is then reformed into numerous texts that are relatively similar in meaning and have a higher probability of being more fluent. To achieve the higher level of fluency these are some, but not all of the main enhancement operations that are carried out in order to reform the original source text:

Reordering – Words are reordered into different sequences

Synonym Swapping – Words are swapped out with synonyms that would more likely occur in the given context

Part of Speech Addition and Removal – Specific parts of speech such as articles, prepositions, demonstratives and pronouns are either added or removed

Morphing – Words are morphed into their equivalent variations

Paraphrasing – Idioms and obscure text are replaced with equivalent paraphrases

Spelling – Incorrectly spelt words are spelt correctly

Localizing – Words that have various spellings in different regions of the world are spelt correctly for each respective region

Punctuating – Various punctuation is added where statistically suggested

These operations are carried out in a synergetic and controlled manner to ensure real time performance and only good reformations are kept. The operations are powered by the N-Gram and Corpora data. An operation will only be performed, providing there is enough supporting statistical data that indicates the operation will be beneficial to the reformation process.

Reordering

Reordering helps solve problems where the order of words disrupts the fluency of a sentence. Some languages, in particular English, rely on the order of words to convey the true meaning of the text. Reordering is best left as one of the last operations to help polish the fluency and preserve the original meaning of the text. Reordering is powered by the N-Gram Data.

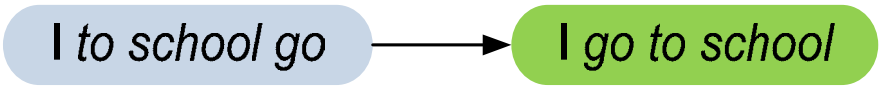


FIG. 31. REORDERING

Synonym Swapping

Synonym swapping first analyses the words around the target word to get a feel for the context of the sentence. Then it picks a synonym accordingly using the Corpora Data. Synonym swapping solves the problem of words being used in an incorrect context.

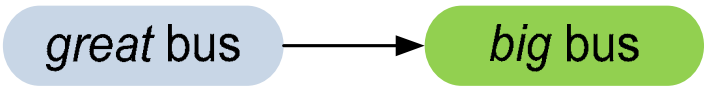


FIG. 32. SYNONYM SWAPPING

Part of Speech Interchanging

When text is not fluent, misuse of parts of speech is a huge contributor. Using the N-Gram Data, common parts of speech that come before or after particular words can be identified. This helps guide the reformation process in adding and removing parts of speech where appropriate.

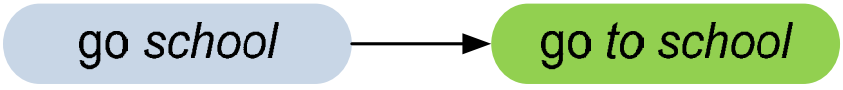


FIG. 33. PART OF SPEECH ADDITION

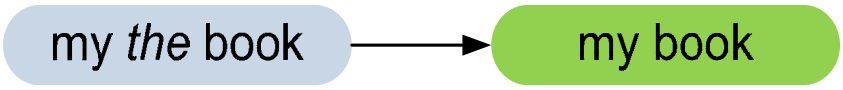


FIG. 34. PART OF SPEECH REMOVAL

Morphing

Sometimes there are odd rules or specific exceptions in language. These can be spotted and rectified accordingly using the N-Gram Data.

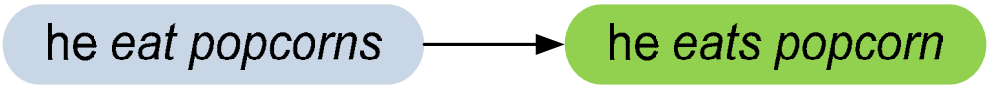


FIG. 35. MORPHING

Paraphrasing

Paraphrasing is great way help idioms and strangely worded text fit more fluently into text. Paraphrasing is powered by the Corpora Data.

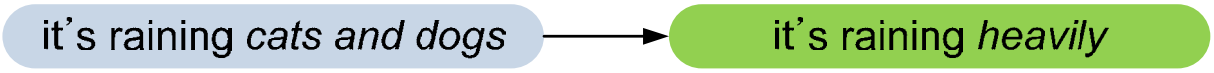


FIG. 36. PARAPHRASING

Spelling

Using N-Gram Data the spelling of words can be corrected.

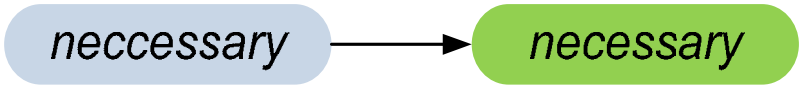


FIG. 37. SPELLING

Localizing

Various words have different spellings in different regions of the world. If we know the region where the text came from, we can perform the localizing operation during the reformation process.

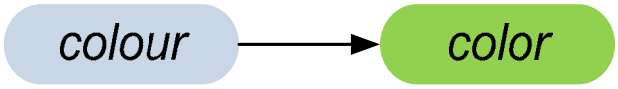


FIG. 38. LOCALIZING

Punctuating

Using the N-Gram Data punctuation can be added where it is statistically suggested.

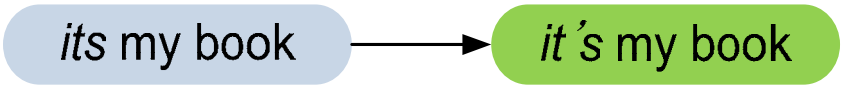


FIG. 39. PUNCTUATING

Here in Fig. 40 is a simple example of a reformation using some of the operations previously described.

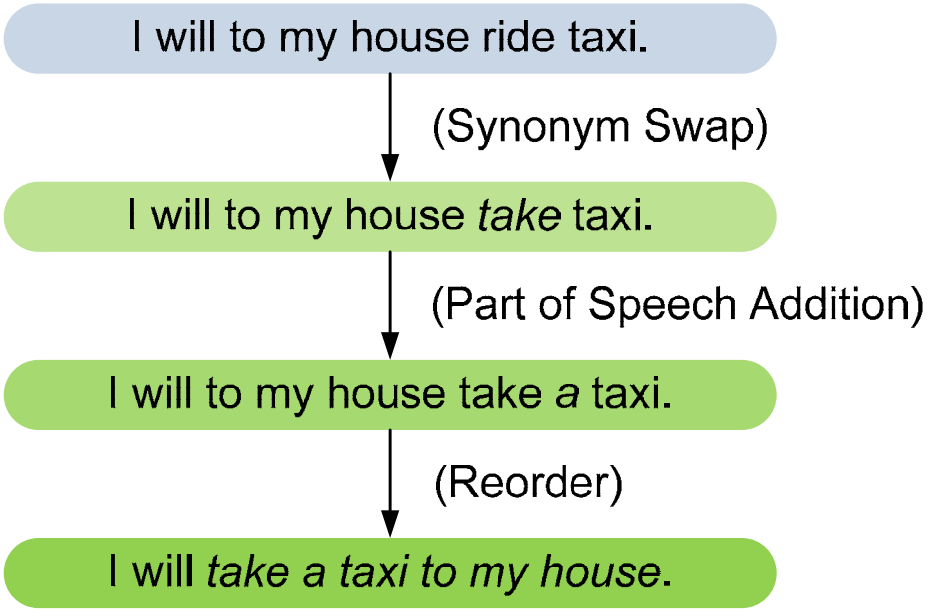


FIG. 40. THE FLUENCY ENHANCEMENT PROCESS

3.4.3 Evaluation Process

The evaluation process provides a numerical score of how fluent texts are. Thus this evaluation process is used to compare the fluency of each reformed text, and to provide the best output(s) accordingly while omitting failed reformations. To evaluate, each reformed text is now mixed up into every single N-Gram combination possible in terms of word sequence order. Table 14 below indicates whether the word combinations are in fact listed in our N-Gram Data. Here it is obvious that the reformed text ("the book is here") is more Fluency Enhanced as it is present 7 times in the N-Gram Data whereas the original source text ("the book exists here") is not present at all. In general, reformed texts that have greater and more significant presence in the N-Gram Data are chosen as the final output.

TABLE 14. EVALUATION DATA (PRESENCE INDICATION)

Evaluation of Original Source Text	Presence in N-Gram Data	Evaluation of Reformed Text	Presence in N-Gram Data
the book exists here	No	the book is here	Yes
the book here exists	No	the book here is	Yes
the here book exists	No	the here book is	No
the here exists book	No	the here is book	No
the exists book here	No	the is book here	No
the exists here book	No	the is here book	No
book the exists here	No	book the is here	No
book the here exists	No	book the here is	No
here the book exists	No	here the book is	Yes
here the exists book	No	here the is book	No
exists the book here	No	is the book here	Yes
exists the here book	No	is the here book	No
book exists the here	No	book is the here	No
book here the exists	No	book here the is	No
here book the exists	No	here book the is	No
here exists the book	No	here is the book	Yes
exists book the here	No	is book the here	No
exists here the book	No	is here the book	No
book exists here the	No	book is here the	Yes
book here exists the	No	book here is the	Yes
here book exists the	No	here book is the	No
here exists book the	No	here is book the	No
exists book here the	No	is book here the	No
exists here book the	No	is here book the	No

This evaluation covered here is kept very simple for the reader to understand the general idea. However there are many other factors to calculate such as varying text lengths, the actual frequencies of how often each N-Gram is present, analyzing texts of greater word lengths than the maximum N-Gram length and so forth.

3.4.4 Application Based Post Processing

By itself FE is not very useful; it must serve an application in order for it to be beneficial to a user. Depending on the application, different outputs may be required from our FE process. MT often requires one final answer, where as if it is used for a writing tool or an alternative search solution add-on, then several answers may be required. Below are some of the applications FE can serve.

Machine Translation Application (Post Process)

FE is performed on the output of an MT system to further improve the quality of MT. The quality of a translation can be touched up in the same way you can touch up the quality of an image.

Machine Translation Application (Complemented Process)

If the design of the MT system complements FE using several pre-processing measures in the process of each translation, then FE can act as the main method of translation. Even Dictionary Based MT produces reasonable results when it internally complements the FE process.

Search Engine Application

Search Engine results are influenced by what key search words a user places into the search engine. Using FE, variations of the key search words can also be obtained to provide *alternative search results* to be returned along with the main search results. The user may not put in the best keywords for the search results they want, so FE can ensure a broader range of results to increase the chance of pleasing the search engine end user.

Writing Aid Application

FE can also be used as a post processing tool to improve the fluency of broken, obscure or illogical text that may be produced from a person who is inarticulate in language. Thus helping people produce more fluent and comprehendible text when writing or speaking.

4. TRANSLATION SYSTEM DESIGN

4.1 TRANSLATION PROCESS

4.1.1 Data Model of the Translation System

As previously mentioned, to maximize the effectiveness of FE, there are certain design features that need to be included in the translation process. Depending on the structure of the MT system, implementing the necessary design features may or may not be possible. Our subject MT system is divided into three parts in which the first two aid FE and the final part executes FE. So when designing an MT system that may make use of FE, similar design features such as the ones discussed here can be implemented.

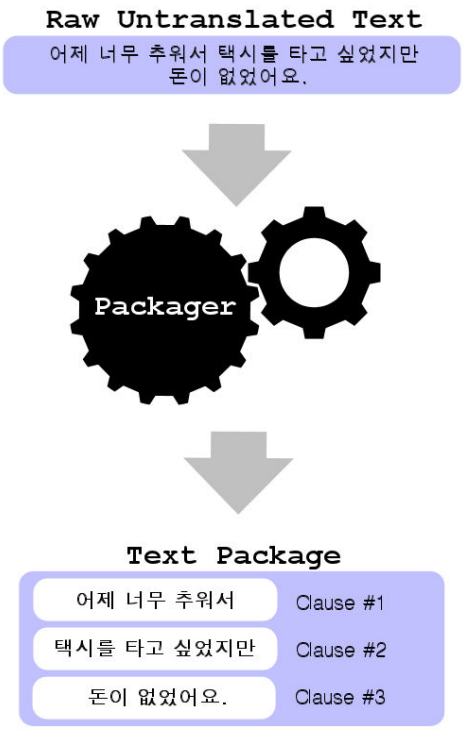


FIG. 41. THE PACKAGER – INPUTS & OUTPUTS

As shown in Fig. 41, the yet to be translated text is read into the MT system, and formed into a *Text Package* by the *Packager*. The purpose of the *Packager* is to identify sentence boundaries, and furthermore divide sentences into individual clauses within the source text so it is then broken

down into smaller word sequences to be translated. The *Packager* is designed to identify sentence boundaries and clauses using a directory of unique identifiers for each respective language. Thus the *Packager* does not need to be redesigned for each language, and only needs to be informed of what language it must work with, Fig. 42 illustrates this process.

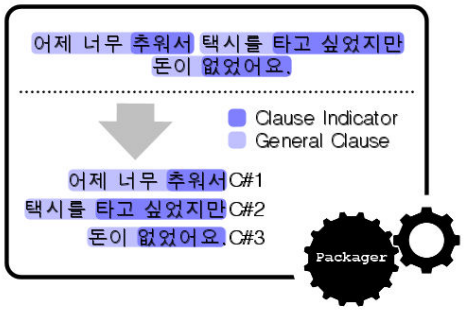


FIG. 42. THE PACKAGER – INTERNAL OPERATIONS

The term clause is a loose term here, and the *Packager* does not always break the source text down into what linguists would define as true clauses. The purpose of breaking the source text down like this is to ensure translations are localized when it comes to executing FE. If packaging is not done, and the source text is translated as is, then FE can link words from over different clause and sentence boundaries. In effect packaging is an effort to retain the original meaning and context of the source text so it is not lost during translation. Failing to do so lets the final translation acquire a whole new meaning that is unlikely to align with that of the original source text. Once again this situation is illustrated in Fig. 43.

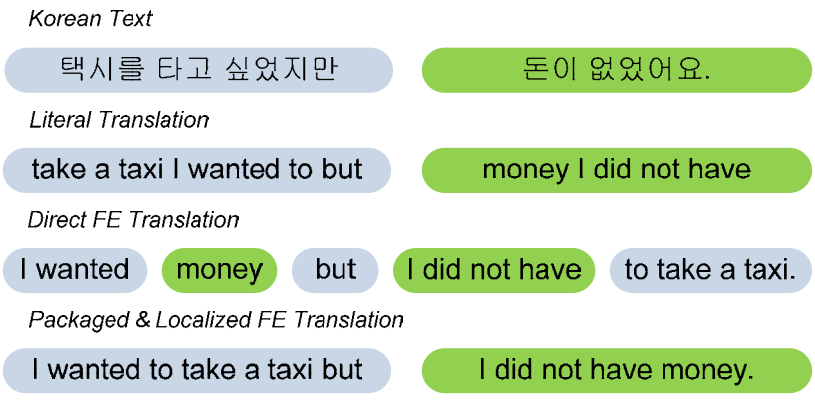


FIG. 43. THE INFLUENCE OF PACKAGING AND LOCALIZATION ON FE

Now the *Translation Seeds* need to be created. Numerous *Translation Seeds* are generated from each *Text Package* using the *Seeder* which can be seen in Fig. 44. Each *Translation Seed* the *Seeder* creates has a unique translation for each clause, and the difference between two *Translation Seeds* can be as small as only one word. In reality, humans can come up with several translations for the same text, so when building an MT system, why not do the same and explore all possible translations.

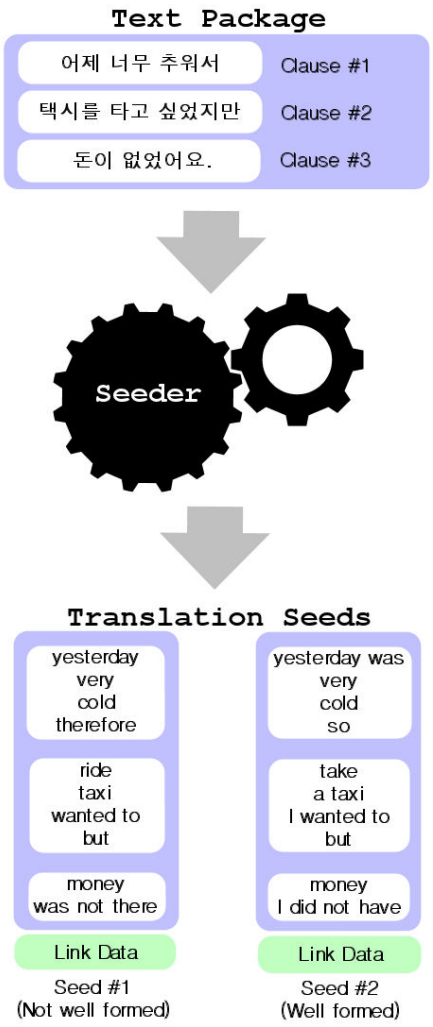


FIG. 44. THE SEEDER – INPUTS AND OUTPUTS

If we subject lots of translations to FE, and we calculate how successful FE is for each one, we can identify the quality of each respective translation. If we only enhance the fluency for only one translation, there is a small possibility FE will have little or even a negative effect, so the final

translation will of course be a gamble. Selecting the highest quality FE translations out of many is quite a reliable method of avoiding this problem, thus we must produce plenty of *Translation Seeds*. Fig. 45 illustrates the internal operations of the *Seeder*, take a look at the big green bubble on the bottom left; there are the bigram frequencies for each word combination in each of the *Translation Seeds*. Notice how *Translation Seed #2* (solid pink line) has higher frequencies of occurrence; this is because it is a better translation, thus is a better candidate for FE. Translations are obtained in our MT system by obtaining the most popular translations of each N-Gram which are collected into lists. Various N-Gram combinations from these lists are formed to make translations (the solid pink / dotted gray lines below represent the *Translation Seeds*). These popular translations are obtained from the expansive bilingual text corpora prepared by Kaleido K.

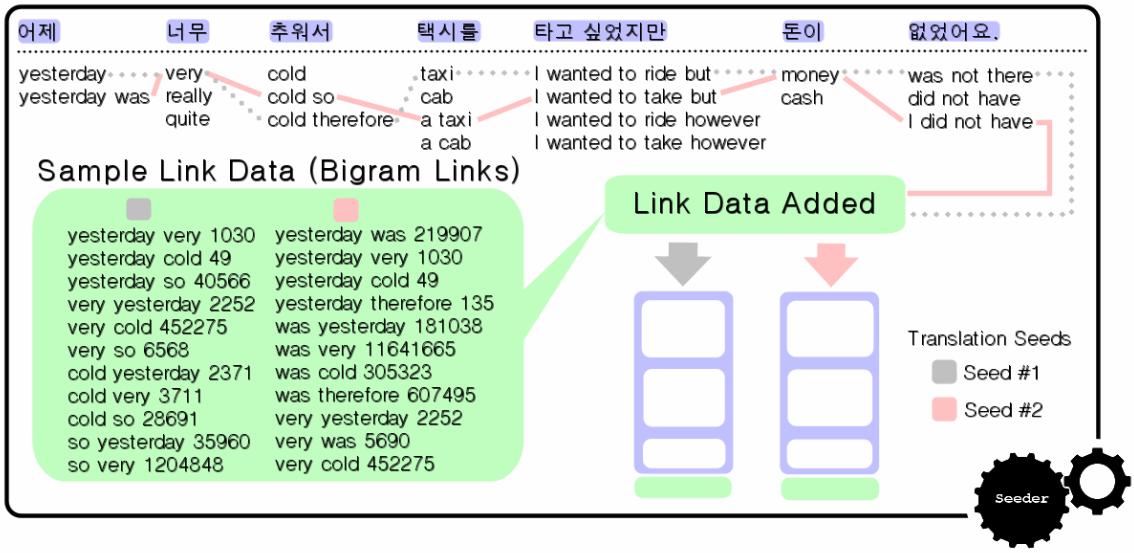


FIG. 45. SEEDER – INTERNAL OPERATIONS

The combinations of N-Grams could be countless which could drag out the translation process; however parameter limitations can help avoid this. Firstly the N-Grams must be relatively high in probability; secondly a limit on the amount of *Translation Seeds* created can be placed. Also *Translation Seeds* are formed by first using the most popular translated N-Grams, so we can be sure when placing a limit on the amount of *Translation Seeds* produced, that the *Translation Seeds* will be the most well formed ones. To aid the linking process which occurs in the next step, *Link Data* is also added to the *Translation Seed*. The *Link Data* holds statistical information about the likelihood of any sequence of words occurring after another sequence of words.

Now moving on to Fig. 47, the *Linker's* job is to perform FE. Within each clause, all combinations of the words are searched within the Google corpus. For every valid N-Gram match, a *Link* is created, and assigned a *Level Value* according to its composition as can be seen in Fig. 46. As the Google corpus contains N-Grams of up to an order of five words, then *Link Level Values* can range anywhere from single unigram to unigram *Links* (1-1), to pure pentagram *Links* (5). Each *Link* is also assigned a particular strength value, based on how often that *Link* appears in the Google corpus. Also some *Links* have more significance than other *Links* according to the type of FE application which further still has influence. All factors taken into account, a *Link Structure Score* (LSS) is calculated for each translation, and this indicates the likelihood of each *Translation Seed* being the highest quality FE translation. Ultimately the *Translation Seeds* with the highest LSSs are chosen as the final translation output, as they the highest quality FE translations out of all the FE translations according to the LSSs.

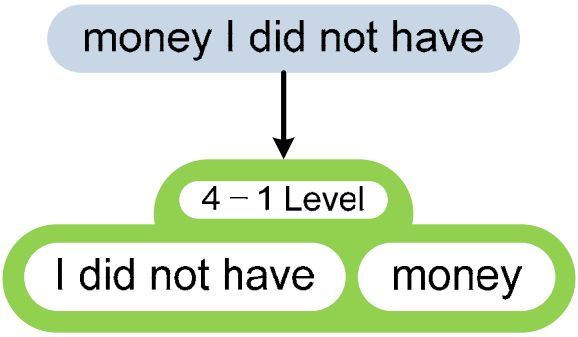


FIG. 46. 4-1 LEVEL N-GRAM LINK

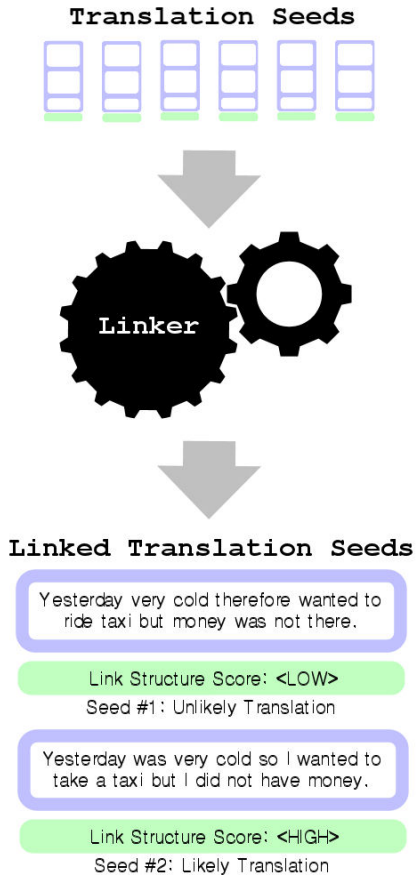


FIG. 47. LINKER – INPUTS AND OUTPUTS

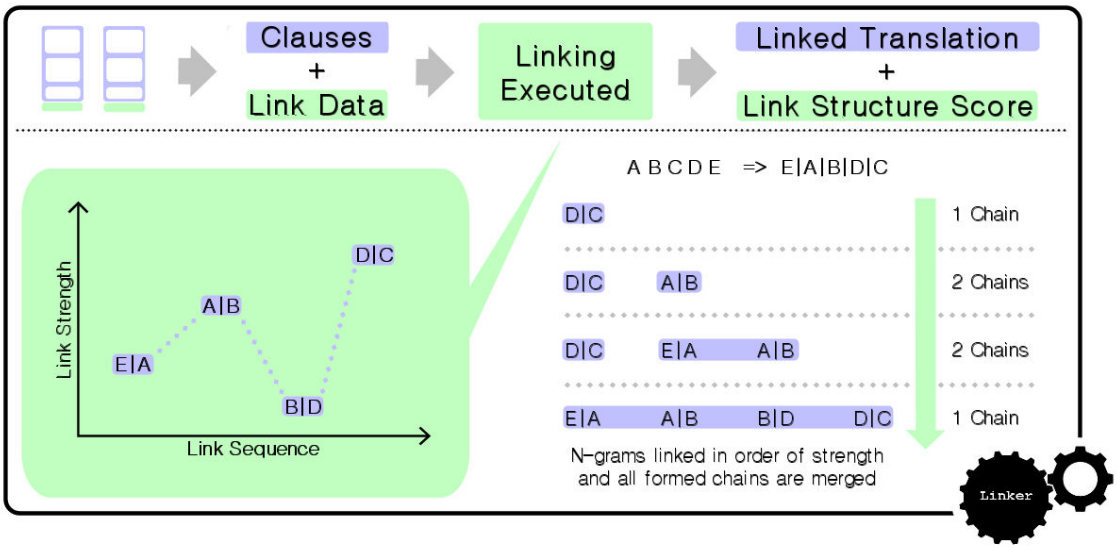


FIG. 48. LINKER – INTERNAL OPERATIONS

Fig. 48 above illustrates the internal operations of the *Linker*, and even though it is quite abstract it demonstrates how a basic FE process takes place. In our example above, originally the translation of the source clause is A|B|C|D|E. Now according the *Link Data*, D|C is the link with the highest probability, so the reforming starts there. Next A|B is found. Following that E|A is found, in which A already belongs to A|B so we now have E|A|B. The process is continued until we have the new reformation E|A|B|D|C which is the translation undergone FE. The FE process is actually much more complicated than just described, with many various application techniques and parameters used to tweak its performance. In the following section this will be covered, however for now the above example helps you understand what the final output of our MT system is.

4.1.2 Application of Fluency Enhancement

The application of FE techniques involves four factors denoted as follows, the *Prioritization Regime* (PR), the *Casting Radius* (CR), the *N-Gram Order Span* (NOS) and finally the calculation of the LSS. These factors each contribute uniquely to the application of FE, and later our results will identify the optimum range of values to achieve good results. Let's explain the four factors a little further.

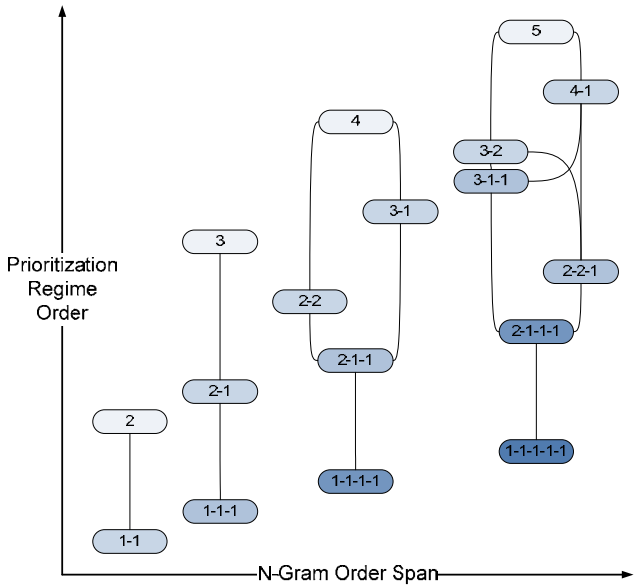


FIG. 49. PRIORITIZATION REGIME FOR LESS LINKS FIRST WITH A PENTAGRAM ORDER SPAN

The purpose of the PR is to sort higher priority *Links* from lower priority *Links*. Some typical PRs are *Less Links First*, *Size First* or *More Links First*. (Less Links First is illustrated above in Fig. 49) *Links* occur in sets, for example [3-1-1, 1-3-1, 1-1-3] and each set is named after the leading *Link* in the set, and each set is mapped out into what is known as a *Word Frequency Map Level* (WFML). All the WFMLs included together make up the *Word Frequency Map* (WFM). A WFM is generated for each translated clause that occurs in a *Seed Translation* and is also known as the *Link Data*. A WFM ranks each *Link* combination in terms of how frequently it occurs in the N-Gram corpus.

Secondly the CR is how far to search either side of the *Map Level Base* (MLB). For example if the MLB is at position *b* in the PR, and the CR is *c*, then WFMLs from positions *0* to *(b + c)* will be searched for *Links*. If a *Link* from *b* to *c* is chosen for FE, then the MLB is incremented forward, and following this a new WFML is generated to comply with the CR. The purpose behind this is to help

higher order *Links* have a larger chance of being selected than lower order *Links* which are selected much more easily due to their higher frequencies of occurrence.

Thirdly is the NOS. This considers such aspects as, the limitation on how many *Sub Links* a *Link* can have and what is the maximum size of an *N-Gram Link*. Lastly, there is the calculation of the LSS. There are several ways to calculate the significance of each *Link* and how it is merged with other *Links*. It is very important to develop a balanced weighting system so high quality FE translations can be separated from low quality FE translations easily. More simply, the LSS calculation calibrates the effectiveness of FE performance. The following is a walk through example of a typical application setup for FE.

For arguments sake, let's use *Less Links First* for the PR. This PR focuses on finding the Links that are large in size and do not make use of several *Sub Links* to achieve size. Let's use a CR of 4, and an NOS of up to a pentagram that can consist of up to 5 *Sub Links*. So given our chosen application, initially *Links* would be searched in the following order 5,4-1, 4, 3-2, 3-1-1, with 5 being the MLB *Link* and 3-1-1 being the CR *Edge Link* due to the CR being 4. *Links* are searched and found from these WFMLs, until a more fluent rearrangement of the original translation has been found. However if for instance, one WFML was exhausted, and no more *Links* existed at that WFML, a new WFML needs to be created in order to comply with the CR. So as seen in Fig. 50 the WFML 3-1 would be created if one of the previous WFMLs were exhausted. (Note WFMLs are referred to as *Map Levels* in Fig. 50)

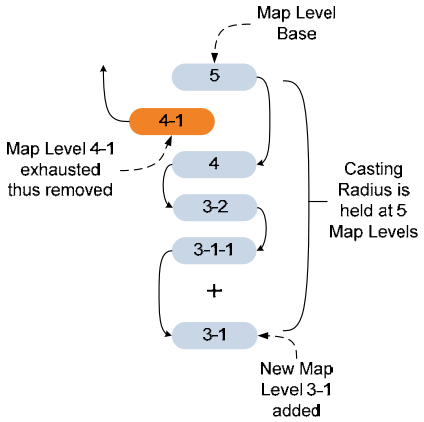


FIG. 50. ADDING OF A NEW MAP LEVEL ON EXHAUSTION OF ANOTHER

4.1.3 Fluency Enhancement Techniques

Blobbing

Let's introduce one FE technique we have dubbed *Blobbing* as can be seen in Fig. 51. Basically the most common *Link* is found and removed from the WFM, and this is used as the primary blob. From there we look for the second most common *Link*. If this *Link* contains an identical word sequence to the first *Link*, then the two *Links* are blobbed together. Also in the case that the *Link* found completely contains a whole new sequence of different words, then this can be used to start up a secondary blob. Eventually all the *Links* are blobbed together to form one final blob.

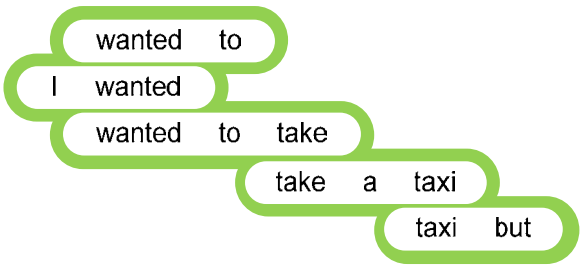


FIG. 51. BLOBBING TO ACHIEVE FLUENCY ENHANCEMENT

Another thing to note is that after all the blobbing has taken place we then need to calculate the LSS. Calculation of the LSS can be done in several ways for several different FE techniques. First of all, analyzing the frequencies of each *Link* is a must. After that further LSS derivations can be done by considering the FE technique used. Let's look at blobbing in Fig. 51, we could take into account the depth of the blob being 5, we could make some assumptions about the highest priority *Link* found, we could also reward instances of repetition such as the word "wanted" appearing 3 times. Defining the exact way to calculate an LSS is not in the scope of this thesis, and also as you will see in the results section, we can still discern the quality of an FE translation regardless.

4.2 DATA COLLECTION

4.2.1 Test Data

The N-Gram corpus used to produce the results in this thesis is provided by Google and is known as the "Web 1T 5-gram Version 1". It is enough data to produce results that prove the usefulness of our MT system that implements FE; however we still require a corpus that has larger, broader, more multilingual N-Grams that are also more up to date.

The Web 1T 5-gram Version 1 was produced 3 years ago in 2006. N-Grams up to the order of 5 are listed, but no further, and the data has also been pruned back, so that words occurring less than 200 times were labelled as <UNK>, and N-Grams appearing less than 40 times were removed. While we can understand the data was pruned and only listed up to 5 grams to rule out strange text and constrain the size of the corpus, we still need data that exceeds these limits. Our results also demonstrate this need, as the less the data is pruned and the broader and larger it is, we can acquire more accurate results. While Google have achieved putting together quite a large corpus, we are now in the process of producing an even bigger, broader and more up to date corpus. It should be finished mid this year, and we also intend to publish our corpus to make it available to the public through the Linguistic Data Consortium. Currently we are corresponding with them in regard to a good way of formatting our linguistic data. In fact one of the things we are discussing is to have our corpus be separable into several sub corpora that contain text related to specific areas of expertise (i.e. medical science, law).

4.2.2 Expansion of Data

Let's start with introducing the Google's Web 1T 5-gram corpus [4]. As listed in Google's official blog and at the Linguistic Data Consortium, here are the quoted facts:

Introduction

This data set, contributed by Google Inc., contains English word N-Grams and their observed frequency counts. The length of the N-Grams ranges from unigrams (single words) to five-grams. We expect this data will be useful for statistical language modelling, e.g., for machine translation or speech recognition, as well as for other uses.

Source Data

The N-Gram counts were generated from approximately 1 trillion word tokens of text from publicly accessible Web pages.

Character Encoding

The input encoding of documents was automatically detected, and all text was converted to UTF8.

Tokenization

The data was tokenized in a manner similar to the tokenization of the Wall Street Journal portion of the Penn Treebank [18]. Notable exceptions include the following:

- Hyphenated words are usually separated, and hyphenated numbers usually form one token
- Sequences of numbers separated by slashes (e.g. in dates) form one token
- Sequences that look like URLs or email addresses form one token

Data Sizes

```
File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:    1,024,908,267,229
Number of sentences:    95,119,665,584
Number of unigrams:         13,588,391
Number of bigrams:         314,843,401
Number of trigrams:        977,069,902
Number of fourgrams:     1,313,818,354
Number of fivegrams:     1,176,470,663
```

Sample Data

The following is an example of the 3-gram data contained this corpus:

```
ceramics collectables collectibles    55
ceramics collectables fine    130
ceramics collected by  52
ceramics collectible pottery  50
ceramics collectibles cooking  45
ceramics collection ,  144
ceramics collection .  247
ceramics collection    120
ceramics collection and       43
ceramics collection at 52
```

The following is an example of the 4-gram data in this corpus:

```
serve as the incoming  92
serve as the incubator 99
serve as the independent      794
serve as the index    223
serve as the indication       72
serve as the indicator 120
serve as the indicators       45
serve as the indispensable    111
serve as the indispensible    40
serve as the individual       234
```

Let's do some calculations to find out what it would take to acquire a larger corpus than this to suffice the needs of our MT system and improve the FE it uses. Firstly Google found 1,024,908,267,229 tokens. Let's assume they don't have the internet cached on a whole of lot of servers and assume they had to find the data from scratch (like we must do), and let's assume they took 1 month to do this. Then we must find this many tokens per second:

$$\frac{1{,}024{,}908{,}267{,}229}{(60 \times 60 \times 24 \times 30)} \cong 395{,}412 \text{ tokens per second}$$

(6) NUMBER OF TOKENS NEEDED TO BE FOUND PER SECOND

Secondly we need to understand how much data in megabytes this is per second. So by looking at the Unicode encoded text file of the Google unigrams we can calculate this. The file with the unigram frequencies removed (leaving only the words / tokens) is 133MB in size. The file itself contains 13,588,391 tokens, thus each text token in the English language on average is:

$$\frac{133{,}000{,}000}{13{,}588{,}391} = 9.788 \text{ bytes} \cong 9.8 \text{ bytes}$$

(7) AVERAGE SIZE OF ENGLISH TOKEN (UNIGRAM)

Therefore the amount of data needed to be acquired per second is:

$$395{,}412 \times 9.788 = 3870292 \text{ bytes} \cong 3.87 \text{ Megabytes per second}$$

(8) REQUIRED DATA ACQUISITION SPEED

A majority of this research has been conducted in South Korea, which is fortunate due to the fast internet that is available to Koreans. In fact Korea Telecom is our ISP and according to Fig. 52 they provide the fastest internet speeds in Korea.

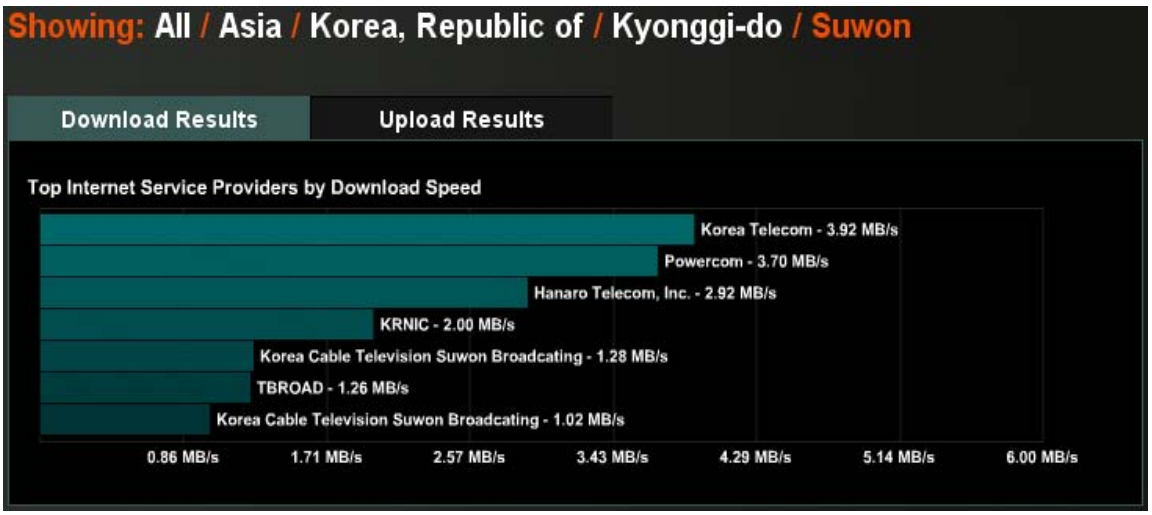


FIG. 52. SPEED TEST'S INFORMATION THE INTERNET SPEED OF KOREAN ISPS

In testing our internet connection with Speed Test we got the following result of 11.48MB/s as seen in Fig. 53, which demonstrates we should be able to easily acquire data at the speed of 3.17MB/s. In fact with the right amount of computing power and resources, theoretically we can achieve obtaining over three times the amount of data in the space of a month.

FIG. 53. THE SPEED OF OUR INTERNET CONNECTION TO A SERVER IN SEOUL FROM SUWON

4.2.3 The Web Crawler

Now our goals for data acquisition have been explained, let's review the software we have created which will help us achieve these goals. Firstly even though our web crawler needs to churn through a lot of linguistic data per second, we still need to be concerned with data quality. Thus we have also designed a lot of interface accessible and built in controls which ensure that the linguistic data we collect, meets a certain criterion so it is beneficial rather than detrimental to our MT system. Below in Fig. 54 is a snapshot of our web crawler just before it is about to engage in crawling the web.

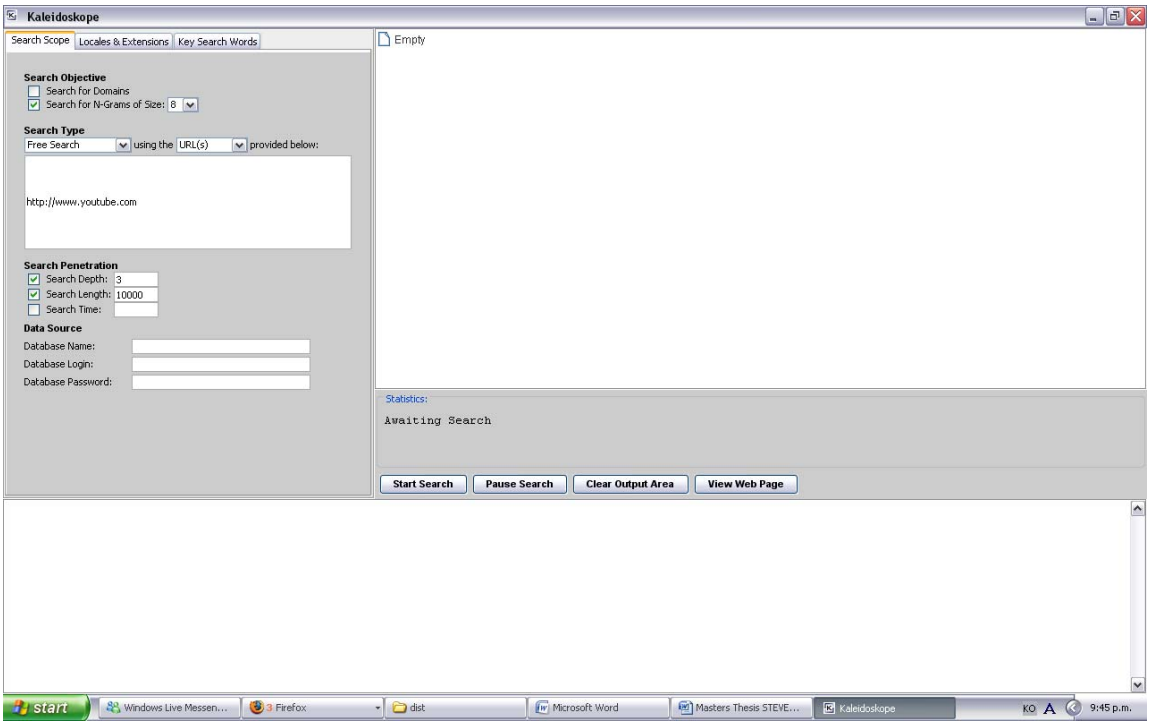


FIG. 54. THE INTERFACE OF OUR WEBCRAWLER BEFORE STARTING A SEARCH

On the left pane, we can set the search parameters and constraints. On the right pane appears a tree hierarchy of the pages searched, and just below that is the web crawler control bar. Finally the bottom pane is the output message area; all updates about the search are output here.

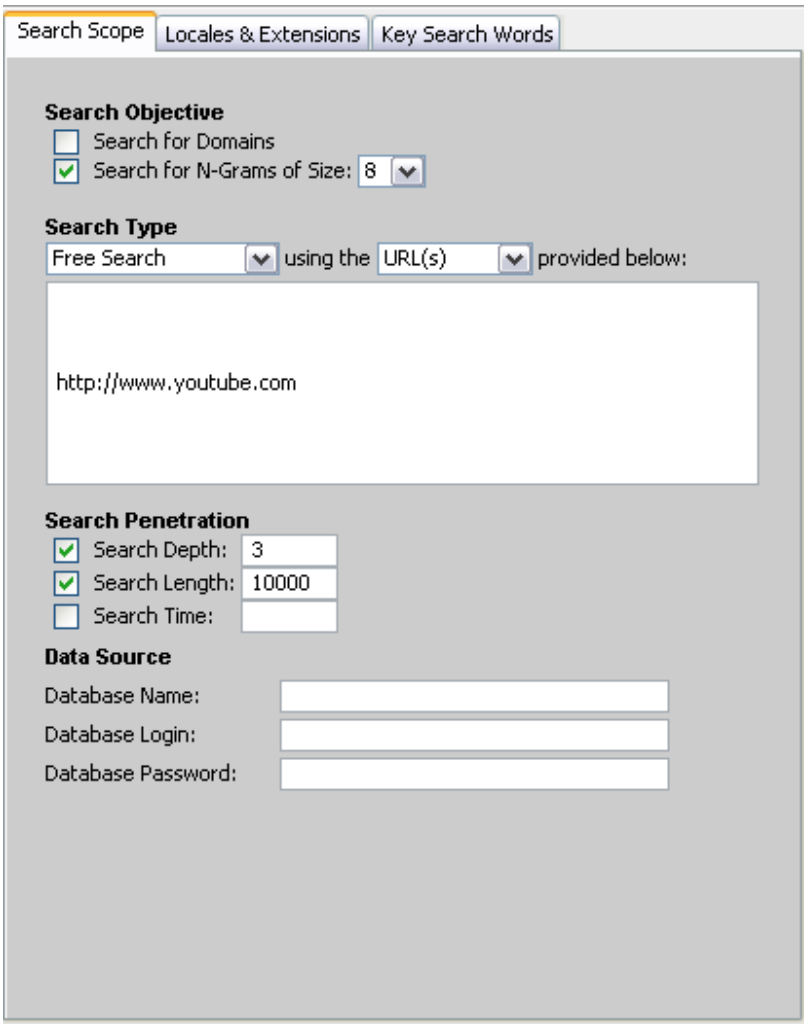


FIG. 55. THE SEARCH SCOPE PANE OF OUR WEBCRAWLER INTERFACE

Here in the Search Scope pane, we can set the starting parameters and constraints. As can be seen, we can select the size of N-Grams we wish to search, the type of search and the URL to start from. Furthermore we can decide the extent of the search with the search penetration parameters of search depth, length and time. Lastly we also need to identify a data source to store the captured N-Grams and the profiles of the web pages they came from.

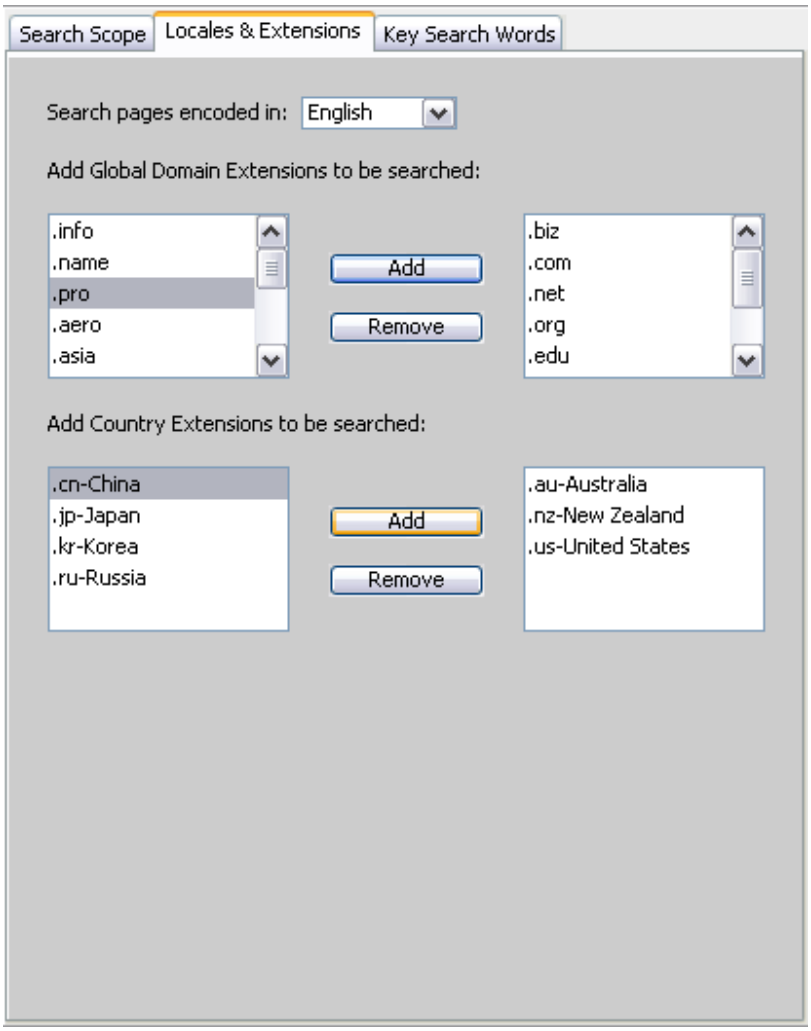


FIG. 56. THE LOCALES & EXTENSIONS PANE OF OUR WEBCRAWLER INTERFACE

In the Locales and Extensions tab, we can specify the language we are searching N-Grams for and the specific countries and web site extensions to get the N-Gram data from. This helps avoid getting N-Grams from different languages all mixed up, and also if we specify such things as countries, we can get localized linguistic data. For example we can get N-Grams for British and American English by searching ".us" and ".uk" sites individually.

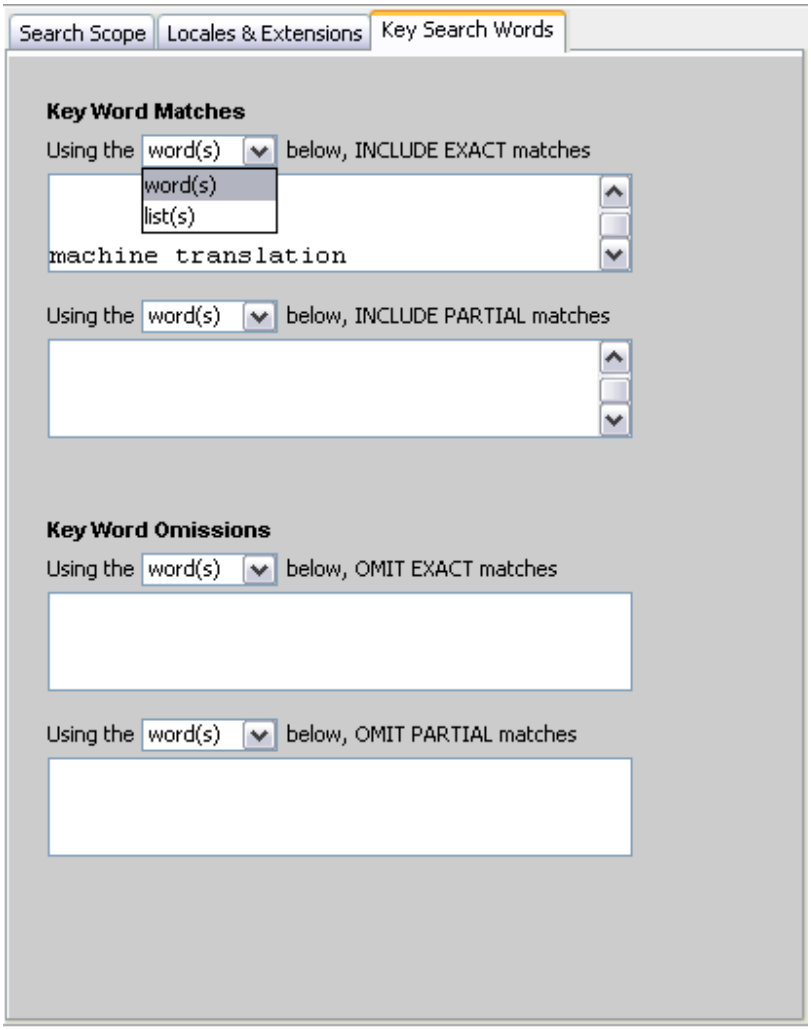


FIG. 57. THE KEY SEARCH WORDS PANE OF OUR WEBCRAWLER INTERFACE

Lastly we have the Key Search Word tab, in which we can use to acquire specific N-Gram data. For example if we wanted medical N-Gram data, we could supply the web crawler with a set of key search words that were related to medical science. If any of these medical words appear in the meta tags of the webpage, then the page can be used to get medical N-Gram data.

4.3 IMPLEMENTATION

The implementation of our MT system can be viewed in Fig. 58. The implementation ensures our MT system can adapt to the evolution of language and provide specific translations for different areas of expertise. The MT system will use a corpus formed by text data collected through the translation process, storing previous translations to complete a text corpus for each respective language. So all translations that occur from Korean to English or any other language will help build up the Korean text corpus that will aid translation from another language into Korean.

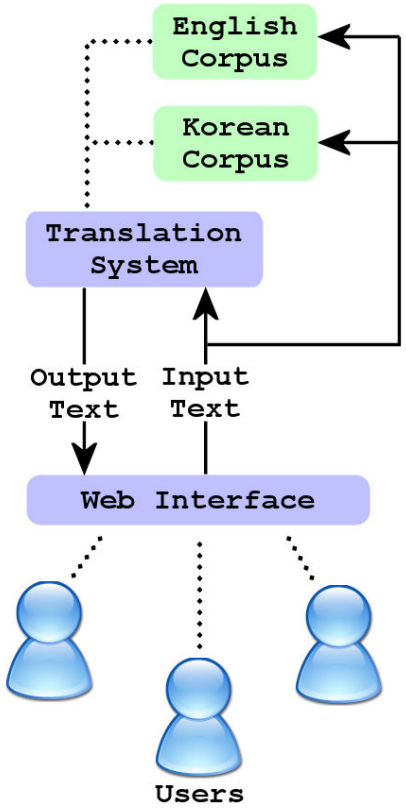


FIG. 58. IMPLEMENTATION – A BUILT IN UPDATE MECHANISM

This is a key benefit over other systems which rely on bilingual corpora made by collecting text from various sources on the internet and translated documents. How much relevance can a corpus like this have to everyday translations? It would be much better if the corpora were built up with text that people actually chose to translate. Even though you can argue that bilingual corpora were made with translated documents and therefore were also made up of translated

text that people desire to have translated, it still does not give you a measurement of how much the desire to have that text translated is. For example a sentence such as "Can you help me?" would be far more in demand for a translation than a one-off news headline or a specific clause in a terms and conditions agreement. So in short, if we use popular text to be translated, we can generate popular and agreed upon translations, thus improving accuracy.

Moreover when using the web interface, if users require a translation that uses language specific to an area of expertise then they can select that topic from a list. So when their text to be translated is stored in the database, it can be tagged as *medical* or *law* and so forth, and in return a translation can also be put together from a corpus in the target language that has text data tagged with *medical* or *law*. This can be optional to the user, but overtime it should yield better results and grow in popularity, as the user will naturally be encouraged to tag their translations in order to get better results. This idea also relates back to our earlier proposal of making corpora separable into several sub corpora to improve accuracy, as the linguistic data an MT system can be quite influential on the quality of its output.

5. RESULTS & DISCUSSION

5.1 FLUENCY ENHANCEMENT RESULTS

The design and construction of a full scale translation system is an enormous task. The translation system that we have built, clearly demonstrates that FE can output quality results, however the results can range from perfect to disastrous. Disastrous results are not failures, as they also read very fluently, but the problem is it is fluent garble. The core problem of the instability is due to the fact that the FE algorithm still needs more maturing, so there is a reasonable reduction in error margins. The results in the following pages highlight important aspects of FE and how they improve translations. However there are not comparable industry standard results. This has deliberately been avoided for the time being, however an industry standard test using BLEU, will be conducted at a later date, and the results posted on the Kaleido K website. These results should be available not long after the submission of this thesis at the URL below:

<<< http://www.kaleidok.com/eng/currentresults.html >>>

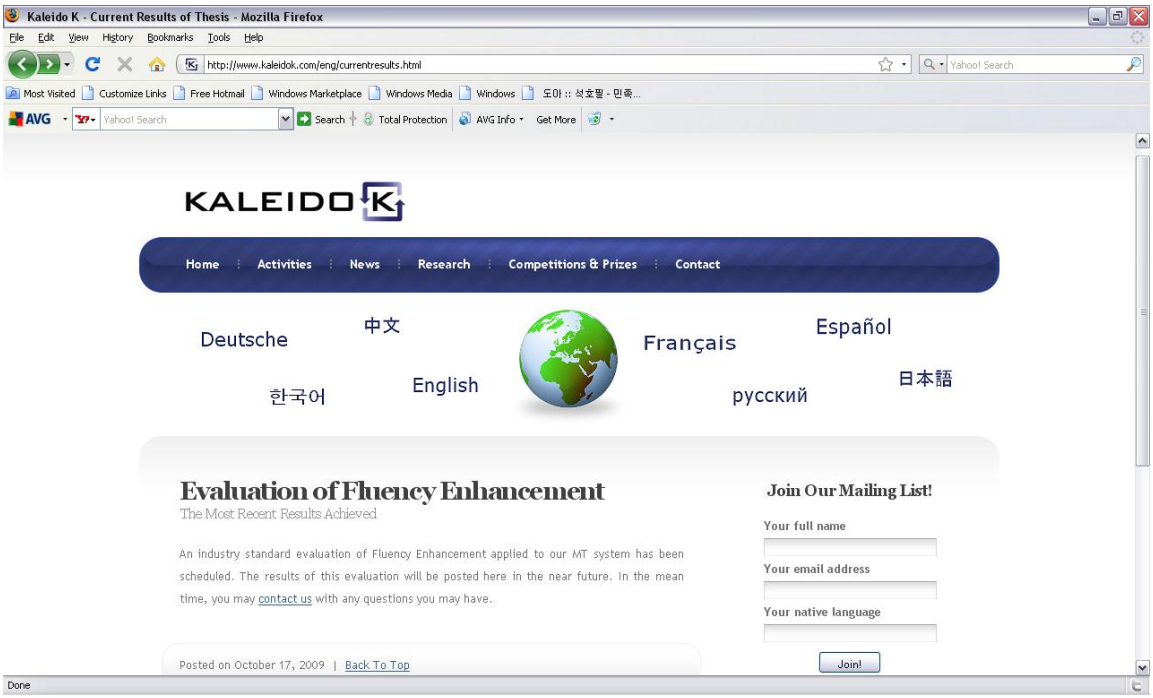


FIG. 59. FUTURE EVALUATION RESULTS TO BE POSTED

In the MT industry it is standard to use BLEU to evaluate the accuracy of an MT system [10]. However in this thesis a different measurement will be used that is more fitting to comprehend the effectiveness of FE. In all of the following results, we are interested in the *Fluency Enhancement Difference*. Basically, we find the average calculated difference between the enhanced and the original accuracy of each translation to discover the degree of improvement FE is able to achieve overall. Equation (9) represents this.

$$\Delta FE_{\,Average}(\%) = \sum_{i=1}^{T} \left(\frac{A_{Enhanced_i} - A_{Original_i}}{T} \right) \times 100$$

(9) AVERAGE DIFFERENCE BETWEEN THE ENHANCED AND THE ORIGINAL ACCURACY OF EACH TRANSLATION

The amount of translations *T* is decided by how many possible word sequences can be derived from the set of words present in the *Translation Seed*. The MT system will attempt FE on every single word sequence whether it is reasonably accurate or not accurate at all. For our tests, accuracy is defined by checking whether each word in each sequence has the correct neighbouring words in order to produce a sensible and fluent translation. Naturally if all the neighbours are all correct for each word, then the translation is considered 100% correct. It should be noted that this measurement is not an ideal measurement of translation accuracy. Nonetheless it is suitable because it allows easy measurement for comparing the improvement of translations against each other, which is what we are trying to understand here.

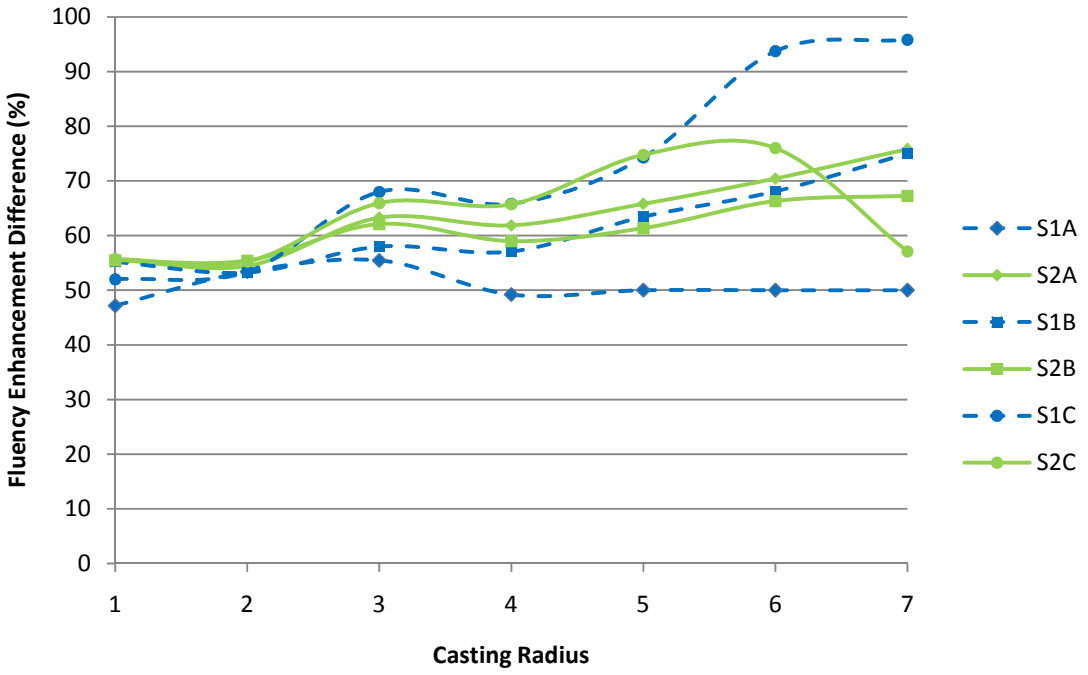


FIG. 60. EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN CASTING RADIUS IS INCREASED FOR INDIVIDUAL CLAUSES IN SEED TRANSLATIONS 1 & 2

Our first result in Fig. 60 demonstrates how increasing the CR improves FE. Our sample sentences from our *Seed Translations* were used to obtain the data above and are divided into three clauses A, B and C respectively. Clauses of *Seed Translation 1* are the dashed blue lines while clauses of *Seed Translation 2* are the solid green lines. Increasing of the CR improves results however they also become more irregular. *Seed Translation 2* is fairly consistent in its enhancement increase, whereas *Seed Translation 1* is more subject to these irregularities. This already suggests that *Seed Translation 2* will provide a more stable and reliable FE translation. Increasing the CR and observing which *Seed Translations* maintain their stability is a good way to determine a well rounded FE translation.

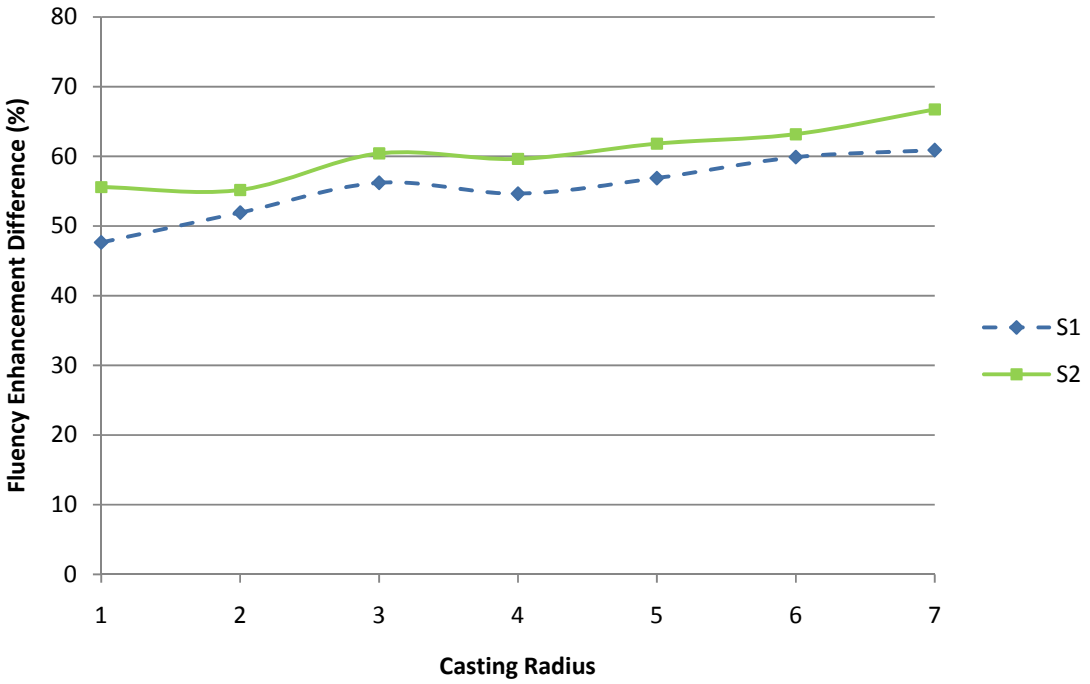


FIG. 61. AVERAGE EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN CASTING RADIUS IS INCREASED FOR SEED TRANSLATIONS 1 & 2

Next is Fig. 61 which is the combined averages of Fig. 60 for both *Seed Translations* This portrays a more important picture, notice how over all values of the CR, *Seed Translation 2* is able to achieve better FE than *Seed Translation 1*. This is because a more sensible translation can be derived from *Seed Translation 2*, so without calculating the LSS, we can already see which FE translation is a higher quality candidate for the final translation. This is a good time to point out that the heavy analysis that took place to produce these results can be bypassed by calculating a LSS. A LSS is used to predict this result and save time to ensure translations occur in real time, which is a vital goal for the success of any MT system.

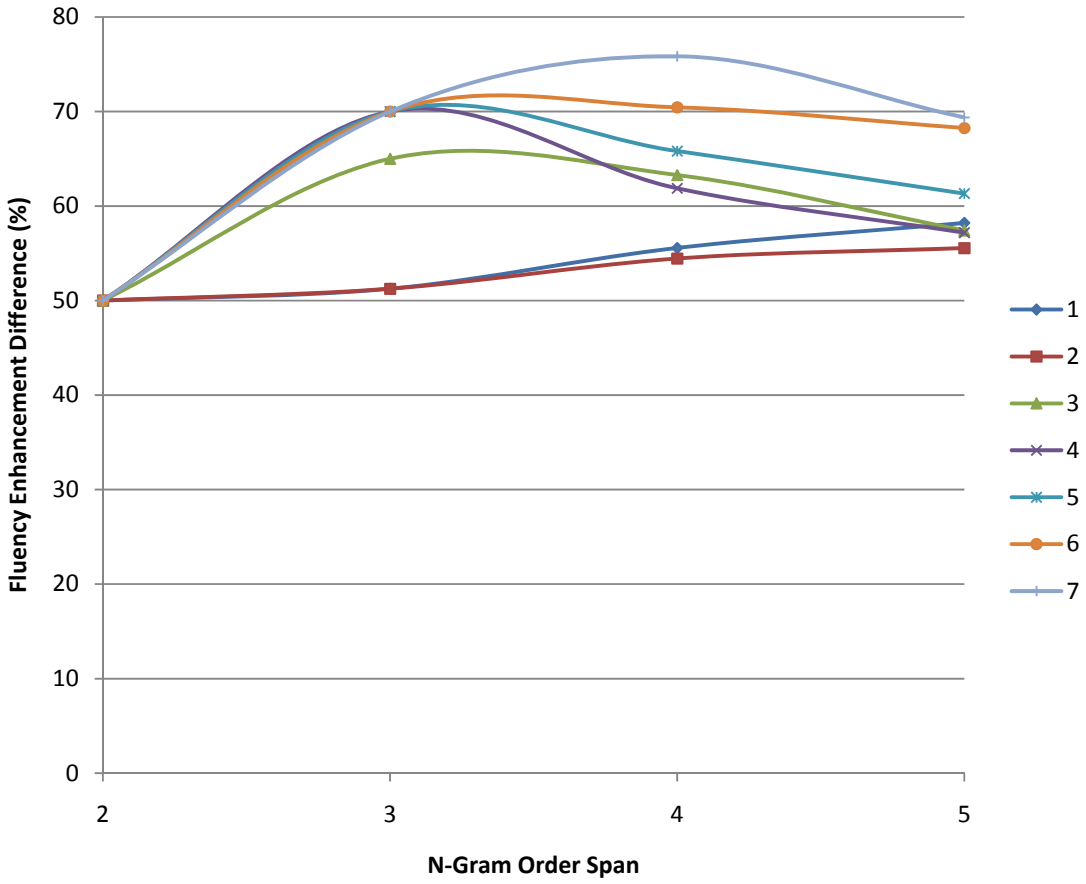


FIG. 62. EFFECT ON FLUENCY ENHANCEMENT DIFFERENCE WHEN N-GRAM ORDER SPAN IS INCREASED FOR INDIVIDUAL CASTING RADII

In Fig. 62 we are concerned with the FE Difference when increasing the NOS over a series of different CR values. Increasing the NOS is beneficial until an order of 5 is reached and the benefits drop back. This suggests that 3 or 4 is the optimum NOS value when used with higher CR values. In future testing with a corpus that has higher order N-Grams, we shall find out if increasing the CR in proportion with the NOS will actually see further improved FE translations.

5.2 COMPARING DIFFERENT APPROACHES

5.2.1 Comparison of Competing Online Machine Translation Solutions

In Fig. 63 below, we have a paragraph of 3 Korean sentences that we have translated into English using a human translator, and the 3 popular online MT systems Google, SYSTRAN and WorldLingo. Judging by the results, we would like to point out that Google appears to use a CBMT approach, whereas SYSTRAN and WorldLingo appear to use a RBMT approach. See if you can identify by yourself how different MT approaches produce different results. For example Google is able to get the first translation perfectly correct, however it flops on the following two translations because it is unable to parse some of the Korean words. On the contrary notice how SYSTRAN's and WorldLingo's translations fail to convey the meaning of the first translation, but are able make much more intelligible translations for the next two translations. Please note that these results were gathered in February 2009, and the purpose of comparing them here is not to compare the success of any company's MT solution, as much more than 3 translations would be needed to do that. The purpose of this is to identify how results can differ for both CBMT and RBMT approaches.

FIG. 63. COMPARISON OF COMPETING ONLINE MACHINE TRANSLATION SOLUTIONS

TABLE 15. FIRST SENTENCE TRANSLATION COMPARISON

	Everyone likes to get compliments.		
Google	Everyone likes to **receive** compliments.	In this case, Google obtained an excellent translation. The only difference is **receive** was used instead of **get** which is completely fine.	***Excellent***
SYSTRAN	The fact that **anyone** is **praised.**	The meaning of the original sentence has been completely lost. However the use of the words **anyone** and **praised** do indicate the sentence is about praising people.	***Inadequate***
WorldLingo	The fact that **anyone** it is **praised.**	A very similar result to SYSTRAN here. Once again the meaning of the original sentence has been completely lost. However the use of the words **anyone** and **praised** do indicate the sentence is about praising people.	***Inadequate***

From looking at Google's result here, the first sentence translation demonstrates how well CBMT can work, since if what needs to be translated is quite a common thing to say, there is likely to be linguistic data related to it in the corpus, ensuring a good translation. The RBMT systems both failed to produce an adequate translation for the first sentence, and can only give an idea of what is being discussed.

TABLE 16. SECOND SENTENCE TRANSLATION COMPARISON

	Compliments don't just make people happy.		
Google	I just **do not** jotahjige gibunman of **people.**	Google was unable to parse some words, so it translated their pronunciation. However since there were some words that were not able to be identified the translation suffered greatly. **Do not** and **people** appear, giving the translation some credibility.	***Inadequate***
SYSTRAN	Is **not** the fact that the **praise does** to **make** get **better only feeling of the person** only.	The meaning of the original sentence has been kept mostly intact, and if we focus on the words in bold, and assume we place **not** between **does** and **make**, it reads even better. **Praise does not make better only the feeling of the person.**	***Adequate***
WorldLingo	**The praise only feeling of the** jar **person** is **not to do** to **get better.**	This translation has the makings of a potentially good translation, but the ordering of words is too scrambled and the use of the word **jar** doesn't make a lot of sense. Notice how we can rearrange the order of the words in bold to make a translation that may have been adequate. **To do praise not only feeling of the person get better.**	***Inadequate***

For the second sentence translation, Google's CBMT approach failed to parse some Korean words, which in this case meant no sensible translation could be found. Not having enough linguistic data available to come up with a sensible translation is an issue for CBMT systems. Korean is a phonetic language in which the characters are often morphed in several different ways to indicate the use of different grammar. To get around this Google should perhaps use a more robust parser, or simply find more linguistic data to add to the corpus. On the other hand we see that the SYSTRAN's RBMT system was able to come up with a slightly intelligible answer, which emphasizes the reality that RBMT systems often perform better on language pairs that have limited linguistic resources.

TABLE 17. THIRD SENTENCE TRANSLATION COMPARISON

	Something that you say meaninglessly could completely change someone's whole life.		
Google	**Inadvertently change a** word **to** the **words** of the songdurijjae **life**.	Google was again unable to parse a word and translated its pronunciation. However the translation does contain words that are relevant to the true translation, and when rearrange can make a little bit of sense. **Words to inadvertently change a life.**	***Inadequate***
SYSTRAN	The end of **one single word changing a life unintentionally**, all does.	The translation sounds a little awkward, but the original meaning of the source text can be understood if the beginning and the end of the translation is ignored. **One single word changing a life unintentionally.**	***Adequate***
WorldLingo	The end of **inadvertent** nose **one single word does a life changes all.**	This translation sounds even more awkward, particularly with the irrelevant use of the word **nose.** However once again, if we ignore particular words, we can understand the gist of the translation. **Inadvertent one single word does a life changes all.** This translation could possibly be adequate; however the SYSTRAN translation is much better in comparison.	***Not Quite Adequate***

Finally with our last sentence translation, we see that the RBMT systems were able to do a little better. However the CBMT approach also had an answer that even though was not adequate still contained a bulk of the words required to make a good translation. Here we would like to point out that this is why BLEU can often favour CBMT systems over RBMT systems. A CBMT result will often contain the correct words needed for a translation even if it fails, however a RBMT result usually fails because the wrong words or grammar were chosen, so there are no similarities between the correct translation and the RBMT result. So essentially, this means RBMT gets punished more for failing, and rewarded less for its sometimes more syntactically correct answers such as the ones that have been seen here.

5.2.2 Example Application of Fluency Enhancement

FIG. 64. SEED TRANSLATION POSSIBILITIES FOR FIRST SENTENCE TRANSLATION

Now we would like to demonstrate an example of our approach to solving such translations by applying FE to our MT system. Above in Fig. 64 you can see the possible translations for the packaged and localized text. If every possible combination above is tried, that gives us 660 *Seed Translations* to work with.

TABLE 18. TOP 25 FLUENCY ENHANCED TRANSLATIONS OF THE FIRST SENTENCE

Rank	TRANSLATION	Structural Integrity Ratio	Link Structure Score
1	Everyone likes to be praised.	4	63
2	Everyone loves to get a compliment.	3.5	145
3	Everyone likes to get a compliment.	3.5	112
4	Everyone likes to be complimented.	3	2011
5	Everyone loves to be praised.	3	1161
6	Everyone loves to be complimented.	3	1103
7	Everyone has a weakness for being praised.	2.75	4954
8	Everyone is fond of getting a compliment.	2.75	149
9	Anyone likes to get a compliment.	2.67	2955
10	Everyone is fond of being praised.	2.67	273
11	Everyone is fond of being complimented.	2.67	273
12	Everyone is fond of getting compliments.	2.67	100
13	Everyone has a weakness for being complimented.	2.55	4954
14	Anyone has a weakness for being praised.	2.5	4614
15	Everyone loves to receive praise.	2.5	1039
16	Everyone loves to get praised.	2.5	964
17	Everyone loves to get complimented.	2.5	964
18	Everyone loves to get compliments.	2.5	964
19	Everyone has a fancy for being praised.	2.5	877
20	Anyone has a fancy for being praised.	2.5	877
21	Everyone likes to receive praise.	2.5	715
22	Everyone likes to get praised.	2.5	609
23	Everyone likes to get complimented.	2.5	609
24	Everyone likes to get compliments.	2.5	609
25	Anyone likes to be praised.	2.5	310

Above in Table 18 we have the top 25 FE results. Out of the 660 *Seed Translations*, 635 were found. Each was given a LSS and then ranked in tiers of its structural integrity ratio (SIR). The SIR is a measurement that ensures longer translations are not ranked more highly than more intelligible translations. Take a look at the translation ranked 13th, it has a higher LSS but it is obviously not the best translation. The translation ranked 1st has a SIR of 4, this is because both "Everyone likes to be", "likes to be praised" and "to be praised ." were all found in our corpus of N-Grams, proving the structure of the translation is well formed.

The purpose of Fig. 65 is to demonstrate how FE can obtain many alternative suggestions from the same source text. As we mentioned earlier this paraphrasing capability of FE means it can have several more applications outside the scope of MT. Of course for MT, the single translation of "Everyone likes to be praised." may be sufficient, however if we were to apply FE to a writing tool then perhaps we would want to give the writer as many ways as possible to express themselves. Perhaps the writer needs to specifically use the word "compliment", or the writer wants a more unique phrase such as "Everyone has a weakness for being complimented." and so forth.

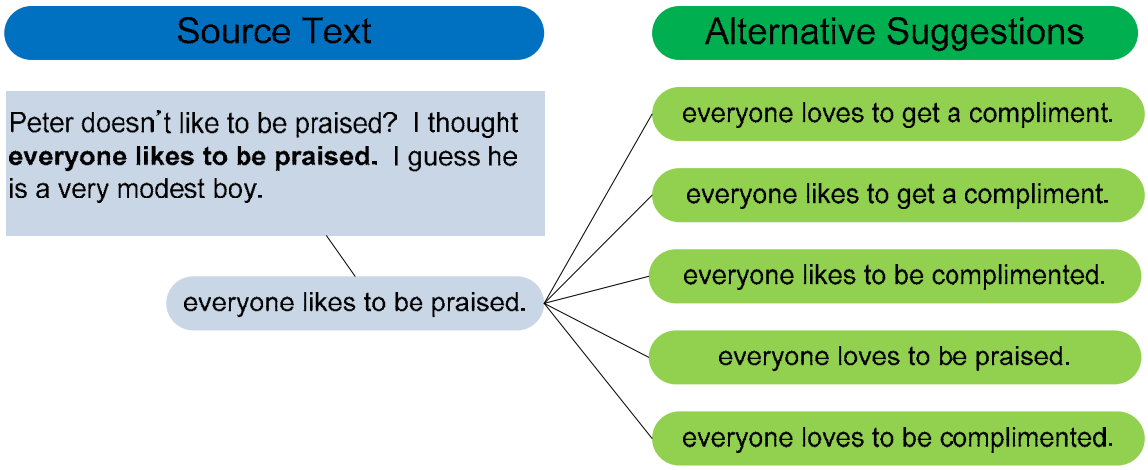


FIG. 65. FLUENCY ENHANCEMENT USED AS A WRITING TOOL

Likewise if we needed to provide alternative search results for a search engine, both the words "praised" and "complimented" could be search simultaneously. The possibilities for the paraphrasing capabilities are endless, subject to the imagination of the designer implementing FE. Look at how many alternative ways we can express the same sentence in Fig. 65.

5.3 WEBCRAWLER RESULTS

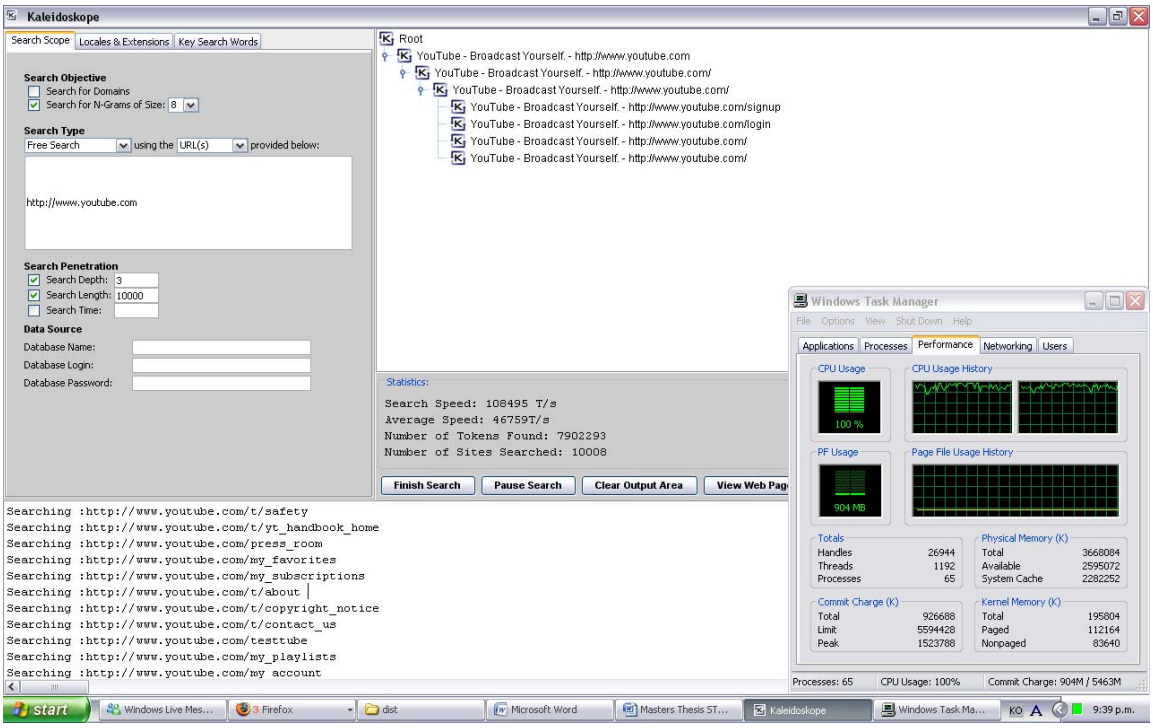


FIG. 66. THE WEBCRAWLER DURING EXECUTION

In Fig. 66 we have a snap shot of our web crawler during execution. The web crawler is multithreaded, and contains a queen spider that searches through the web, and delegates work to the many worker spiders. Each worker spider has its own thread, and contributes to the spider web, which holds all the captured data, and keeps the worker spiders informed of which pages have already been visited. The purpose of multithreading is to make the most of our CPU cycles to ensure the web crawler is crawling as fast as it possibly can.

As can be seen in Fig. 67, the CPU usage is at 100% and the current and average search speeds are 108,485 and 46,759 tokens per second respectively. If you can recall from the previous chapter, 395,412 tokens per second is our target. So 8 to 9 instances of our web crawler operating on different machines through the same router with one internet connection should mean we can achieve our desired speed, and have our new N-Gram corpus completed within a month.

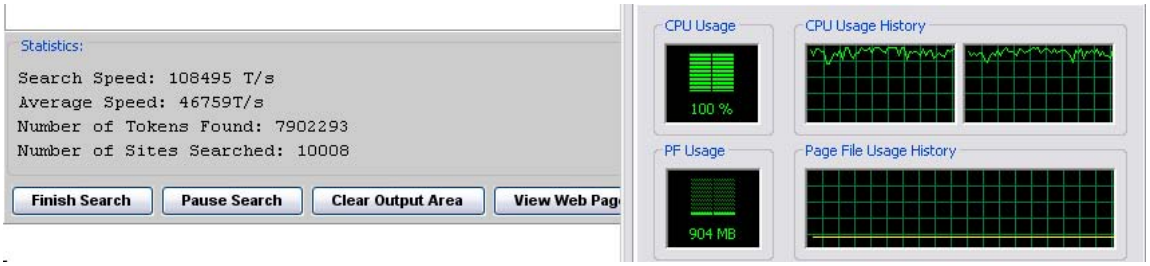


FIG. 67. A CLOSE UP OF THE WEBCRAWLER'S PERFORMANCE

We then tried running the web crawler on another computer (that had lower specifications) at the same time as we ran it on the main computer, with both computers connected through the same router using one internet connection. The results can be seen in Fig. 68 below.

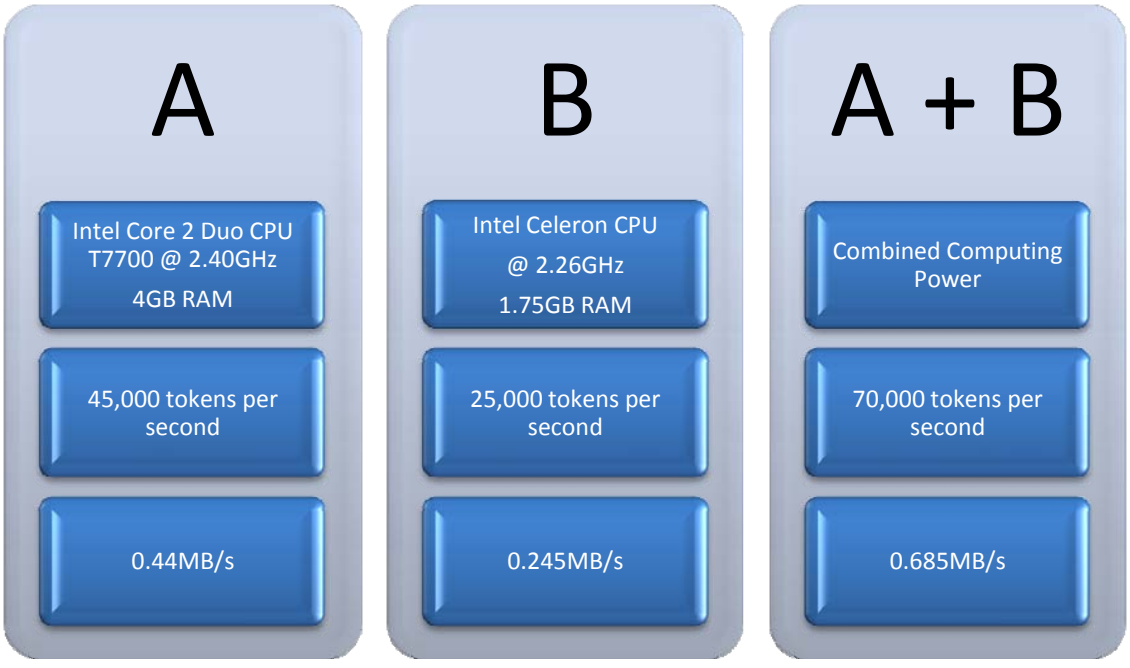


FIG. 68. TEST RESULTS OF OUR WEB CRAWLER

Our target speed is 3.87MB/s and our internet connection and router are easily able to pull in around 11MB/s, so we definitely have the bandwidth required in order to reach our target speed. As Fig. 68 above confirms, the more computers I plug into the router that are running our web crawler, the more speed we can accumulate. The web crawler almost works twice as fast on computer A, this is most probably because the processor on computer A is an Inter Core 2 Duo, as opposed to the single core CPU on computer B. So the issue now is purchasing enough computing power and connecting it all up to the same internet connection

Even though our web crawler has demonstrated already that it can get significant results, it still needs more development to ensure the following issues are accounted for:

Bad latency/bandwidth to remote servers – getting reasonable response times
Malicious pages – avoiding spam/spider traps
Politeness – not hitting a server too often
Duplication – avoiding site mirrors/duplicate pages
Distribution – be designed to run on multiple distributed machines
Scalability – crawl speed can be increased with more machines
Efficiency – uses all processing and network resources available
Quality – can identify and keep a list of high quality pages
Repetition – continue to repeat visit pages that change often (i.e. news)
Extensible – be able to handle several data formats, and know which to avoid
Language – to only search pages of a specified language

Some of the above issues have already been addressed and taken care of but thorough testing still needs to be done to ensure proper functionality. The main issues now are avoiding duplication on the web, politeness, acquiring a profile collection of quality sites to mine data from and avoiding malicious pages.

5.4 MARKET RESEARCH RESULTS

5.4.1 The Survey

We conducted a survey to get a better understanding of the potential market that FE applications could acquire. We surveyed at 5 different universities in South Korea and New Zealand and got a sample base of 200 people. The universities surveyed were:

- Kyunghee University, Suwon, South Korea
- Sungkyunkwan University, Seoul, South Korea
- Canterbury University, Christchurch, New Zealand
- Chonbuk University, Jeonju, South Korea
- Ajou University, Suwon, South Korea

193 of the 200 people surveyed, were non-native speakers of English, so FE could definitely help them. The survey was split into two sections; the first section was a 5 minute English test, which involved correcting English sentences in the appropriate places. This is something that would come naturally to a native speaker, but may be difficult to identify for a non-native speaker. The purpose of this part is to split those surveyed into two different groups, a higher tier group that have good English proficiency and a lower tier group that are not so proficient. Doing this allows us to see clearer if our target market is people who have mastered English well, or those who are just beginners.

When delivering the survey to each person, they were first asked “Do you speak English?” with the survey also being explained in English. This helped filter out people who did not speak English, as the FE tool would likely be useless for them. However we had both English and Korean versions of the survey, and often the actual issued survey was the Korean version; this ensured the survey participants understood how to complete the survey. On the following page is part 1 of the survey written in English for the purpose of this thesis. For the raw data of this market research refer to Appendix B.

MARKET RESEARCH SURVEY: PART 1

In the sentences below, find the mistakes and fix them.

Some sentences do not have any mistakes

Sometimes, you might need to delete or add to correct the sentence

A point is given for finding a mistake; a second point is given for fixing the mistake

The number of mistakes to find and fix is not for you to know

Lastly, this survey *must* be completed by you, *without* the help of any person or language resource.

1. Would you like some chocolate desert?

2. Often my father and I go to fishing at the lake.

3. I am in my final year of university and it is very difficult.

4. I will graduate this summer and my parents will come to the my ceremoney.

5. Why is your mother and father not able to come your graduation party?

6. After student graduate, they usually sell their books to second hand bookshop.

7. I am not sure how I will find a job next year.

8. I am so exciting about going to Mexico next weekend with john.

The maximum score obtainable was 20; however the average for non-native speakers was 5.59, which means on average non-native speakers could identify 2 or 3 mistakes and were able to correct them. The lower tier group is those who got 0 to 10 as their score, which was a majority of those surveyed. The higher tier group is those who got 11 to 20 as their score, and this group was considerably smaller. The average score for native speakers was 17.5; demonstrating FE may possibly be useful for native speakers as well.

In part 2 of the survey, this figure and explanation (written in basic English) were given to help the survey participant get an idea of how FE could be helpful.

MARKET RESEARCH SURVEY: PART 2

Please read about the tool below. Then answer the questions that follow:

A tool to improve your written English

The tool reads the English sentence you wrote. Then it makes many similar sentences that have the same meaning. The new sentences sound more natural and may better express what you want to say. The tool doesn't just fix grammar; it can also completely change the sentence so it sounds like a native speaker said it. This tool can also read your native language, and give you many new English sentences to use in your English writing. The picture below shows how the tool works.

Example:

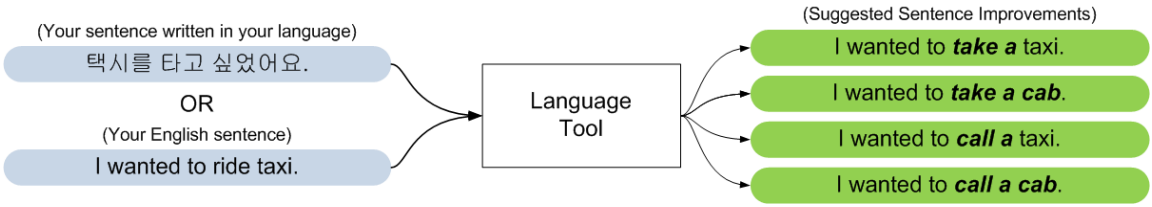


FIG. 69. AN ABSTRACT ILLUSTRATION OF THE FLUENCY ENHANCEMENT TOOL

Following this they were issued with questions about the FE tool described above. The following pages document the results of each question.

5.4.2 What do you think of this tool?

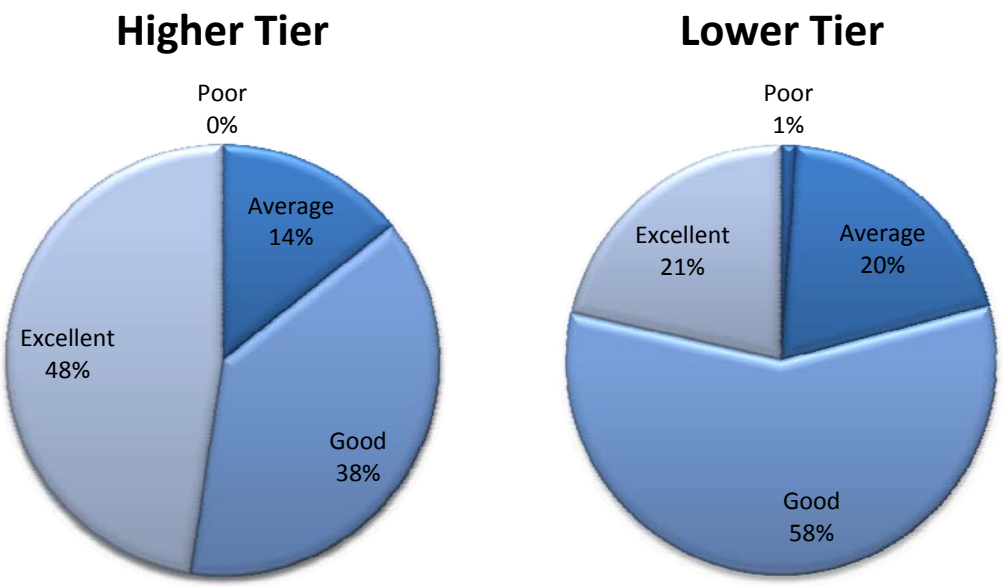


FIG. 70. WHAT EACH TIER THINKS OF THE FLUENCY ENHANCEMENT TOOL

Considering Fig. 70 above, the higher tier group can see how the FE tool can benefit them; however the lower tier group does not share the same enthusiasm. Notice how half of the higher tier group think it is excellent, as opposed to only 21% of the lower tier group agreeing with this, in fact most of the lower tier group consider the FE tool as just good, but a few people even thought it was a bad idea. Our first survey question already demonstrates the importance of dividing groups into two tiers, as English ability can alter the perception of our FE tool.

5.4.3 How often do you think you would use this tool?

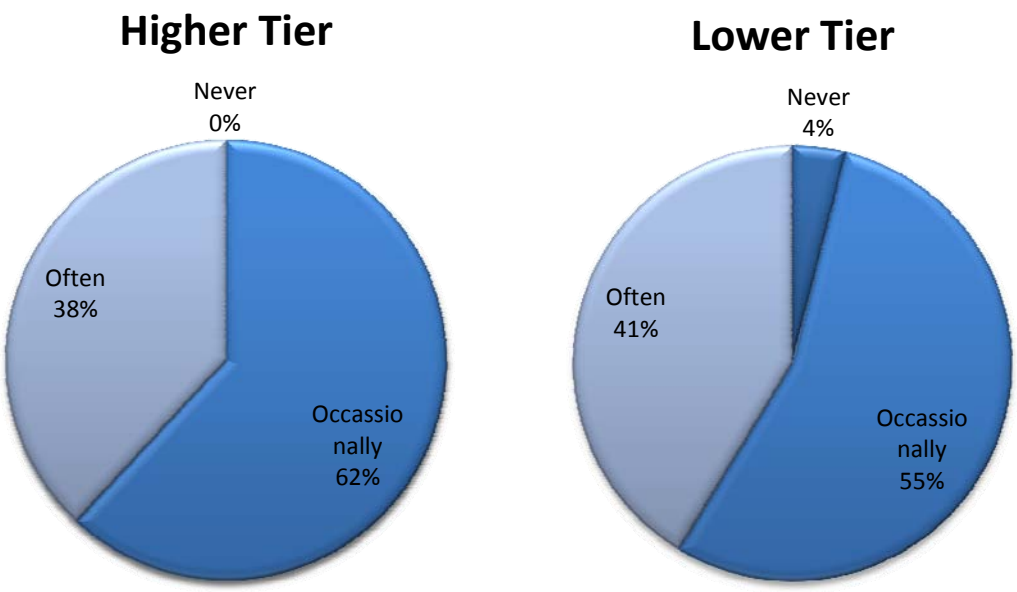


FIG. 71. HOW OFTEN EACH TIER WOULD USE THE FLUENCY ENHANCEMENT TOOL

As can be seen in Fig. 71, between the higher and lower tier groups, they both feel they would need to use this tool to approximately the same degree. As most of the surveying took place in South Korea, this represents the pressing need to understand English at academic institutions.

5.4.4 Do you think your friends would use this tool?

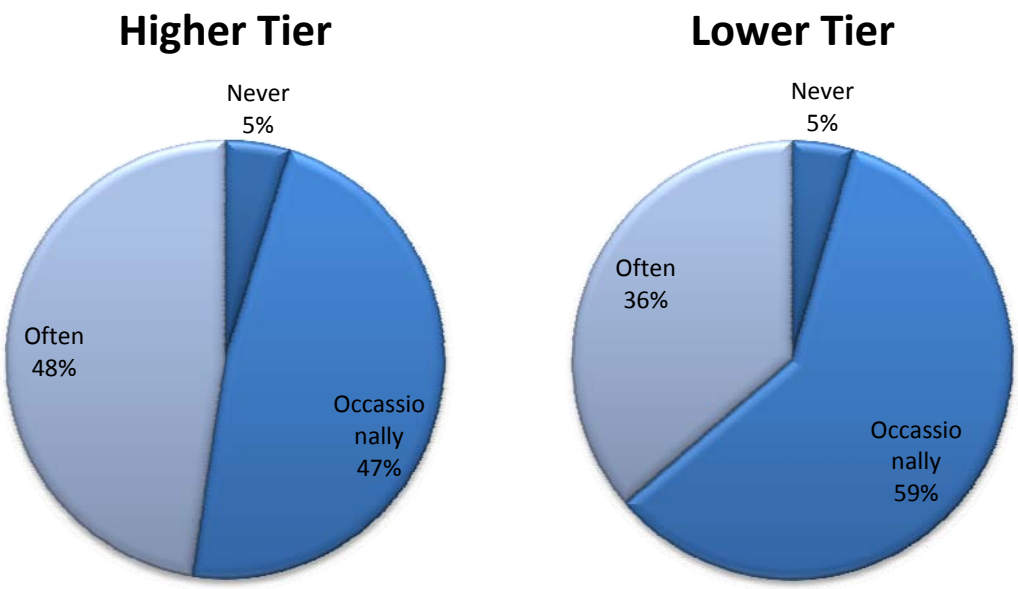


FIG. 72. WHETHER EACH TIER'S FRIENDS WOULD USE THE FLUENCY ENHANCEMENT TOOL

This question made the survey participants think if their peers could use the FE tool. 48% of the higher tier group felt their peers would use it often, as opposed to 36% of the lower tier group. Even though it is a small difference, it shows that people with better English proficiency, are likely to have a friend with better English proficiency too. We can make this assumption based on the first question "What do you think of this tool?", since this questioned demonstrated that more proficient English speakers felt more positively that they could make use of the tool.

5.4.5 If you used this tool, when entering text into it, would you attempt to write in English, your native language or both?

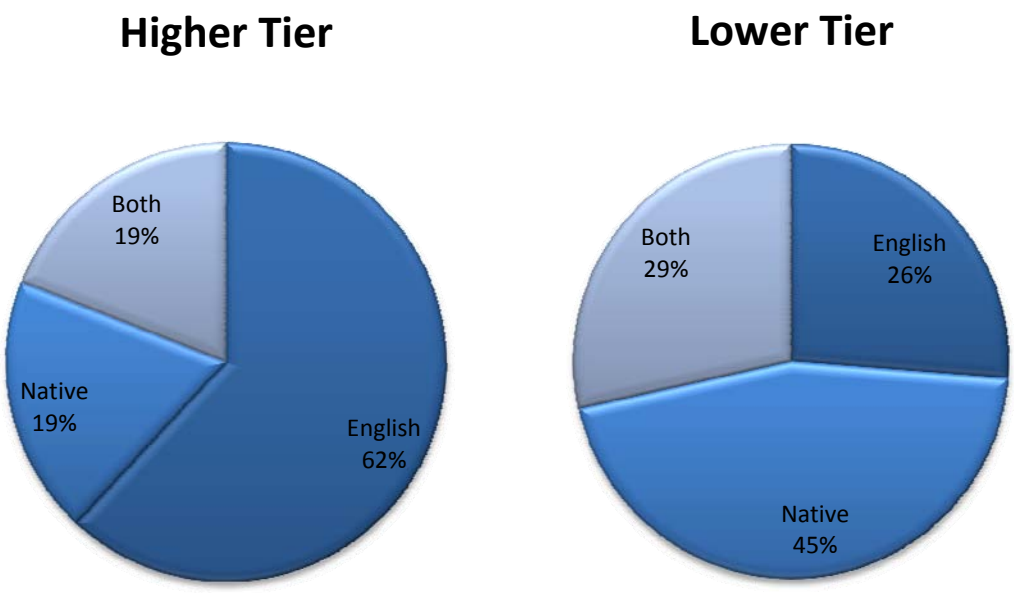


FIG. 73. PREFERRED INPUT LANGUAGES FOR EACH TIER

The higher tier group would attempt more to input English into the FE tool, this is because they can see the value of doing this, since it rectifies their broken English and helps them understand how to improve it. Being able to input your native language to get the English equivalent is more of an add-on of feature, which is an application of FE to MT. Because this add-on feature was mentioned, a lot of the lower tier group viewed the FE tool as only a translation tool, and also because they were not as proficient in English, most probably wanted to input in their native language into the FE tool only.

5.4.6 If you used this tool, how would you like to access it?

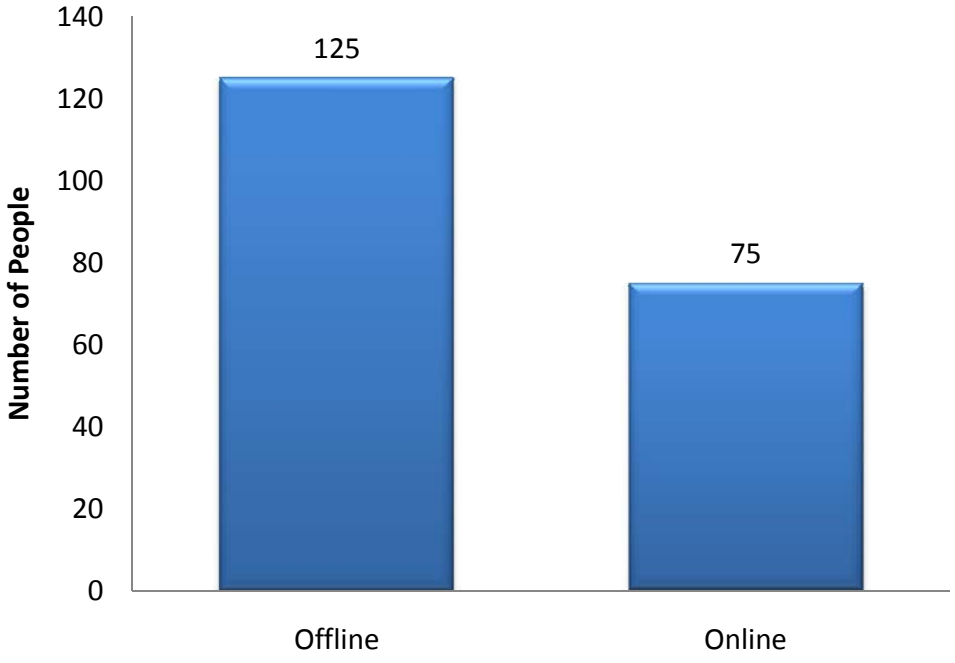


FIG. 74. PREFERRED ACCESS METHODS

Most wanted to access the FE tool offline, by downloading it and installing it on their computer. While this may be the preferred choice, it is in fact the more difficult option to implement. Since FE relies on a huge database that will most likely exceed the hard drive space of many users, and secondly will constantly need to be updated. The best offline option could be to install some client software, so the user could access the FE tool on their desktop, in which the computer needs to be connected to the internet to access the linguistic database. We think this alternative measure would be sufficient for most of the offline crowd. Since according to the comments many wrote, the convenience of having the FE tool on their desktop is one of the main reasons they preferred the offline option.

5.4.7 If you were to pay for this tool, would you prefer to pay based on the length of time used or per sentence improvement?

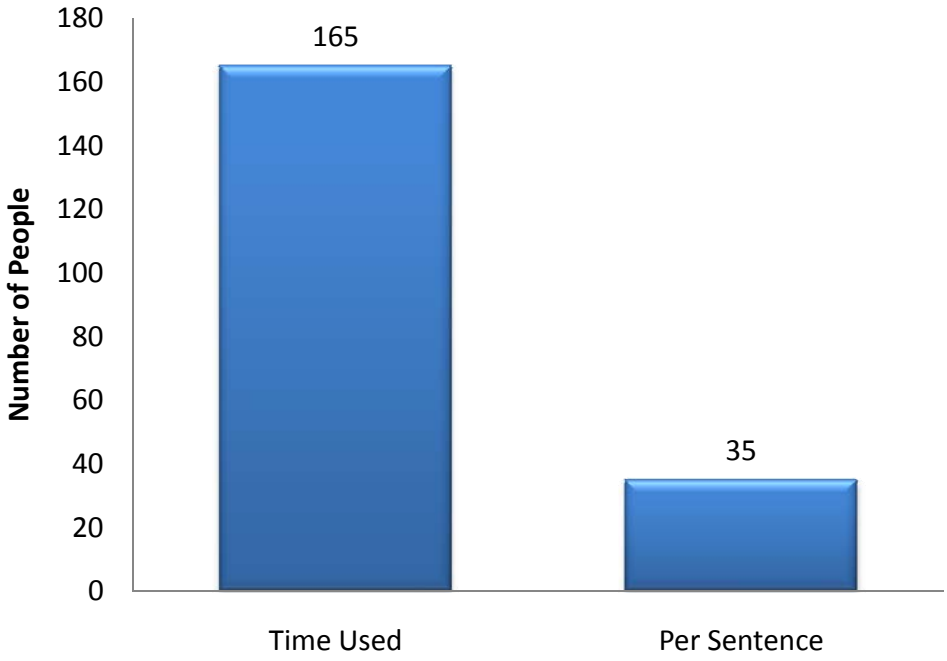


FIG. 75. PREFERRED PAYMENT METHODS

Firstly, let us clearly explain the two choices. Time Used was considered to be 1 month of unlimited use of the FE tool. Per Sentence was considered to be 200 sentences that could be up to 40 words in length. Most people didn't want to have limitations, or have to think in advance how much they would need to use the FE tool so they preferred to use the FE tool based on Time Used rather than Per Sentence.

5.4.8 For each payment method, how much would you expect to pay?

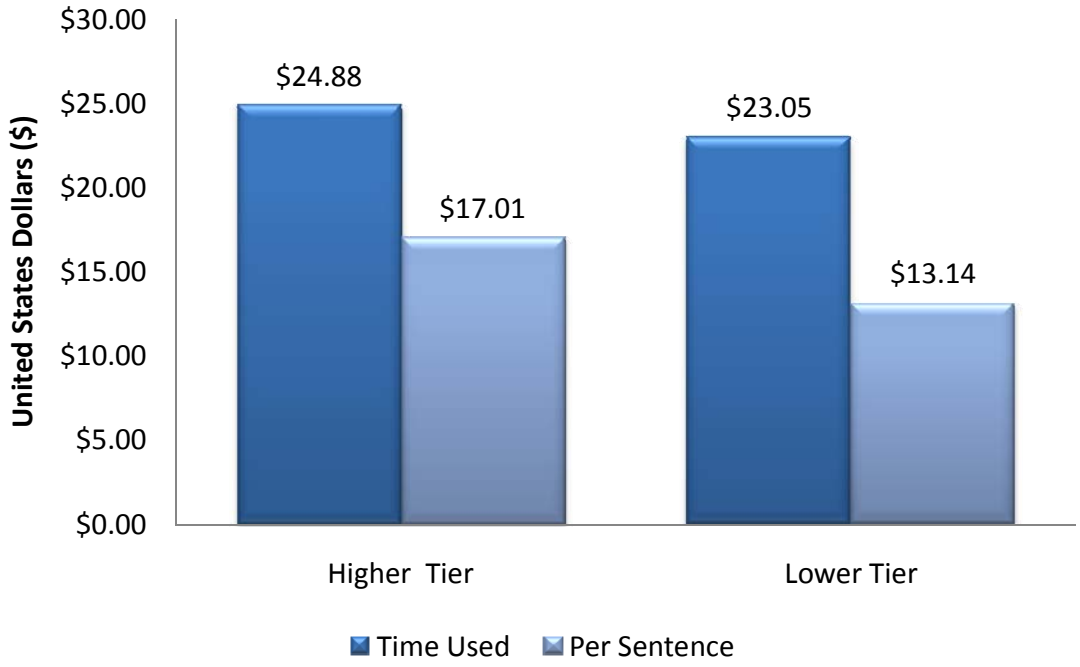


FIG. 76. PAYMENT EXPECTATIONS FOR EACH PAYMENT METHOD

As can be seen in Fig. 76, the higher tier group is willing to pay slightly more for the FE tool. For 200 sentences each group is willing to pay a certain amount, so from this we can calculate how many sentences they would expect to process with the tool in one month using Equation (10).

$$\frac{\$(\text{Time Used}) \times 200}{\$(\text{Per Sentence})} = \text{Sentences Process Per Month}$$

(10) EXPECTED AMOUNT OF SENTENCES TO BE PROCESSED PER MONTH

After doing some calculations, on a monthly basis the higher tier group expects to process 293 sentences and the lower tier group expects to process 351 sentences respectively per month. Thus we can conclude that the higher tier group expects to pay more and use it less, whereas the lower tier expects to pay less and use it more.

5.4.9 Have you purchased, seen or heard of a tool that restructures and improves your English writing such as this one?

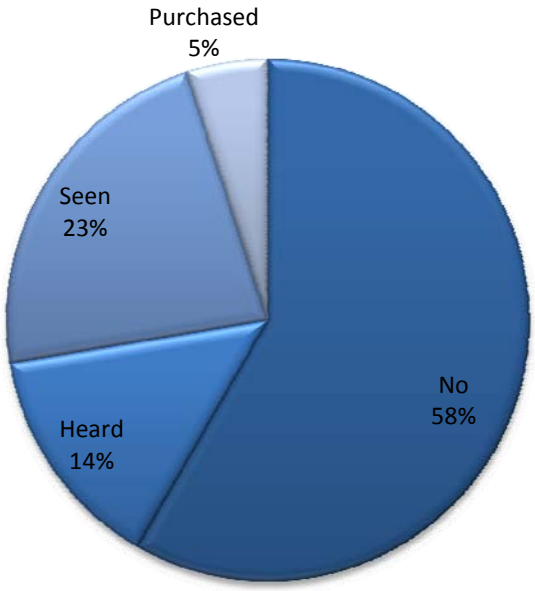


FIG. 77. AWARENESS OF SIMILAR PRODUCTS

This question was interesting, because most who answered it, either said "No" or went on to explain exactly what it was they had seen, heard or purchased. For each category the answers were relatively similar. For "Seen", they almost always mentioned Google Translate; however this only proved they saw the FE tool as just a translator, which it is not. Google Translate has no such tool to date that provides multiple paraphrased outputs of text that may also be translated if the user wishes. For "Heard", they generally mentioned hearing about such an idea from friends. For "Purchased", they usually mentioned some software that had failed them completely in the past, which made them comment sceptically about the FE tool described. As mentioned earlier in Chapter 1, many developers in the MT industry overstate the abilities of their products.

5.4.10 Except English, are there any other languages this tool could help you improve your writing in?

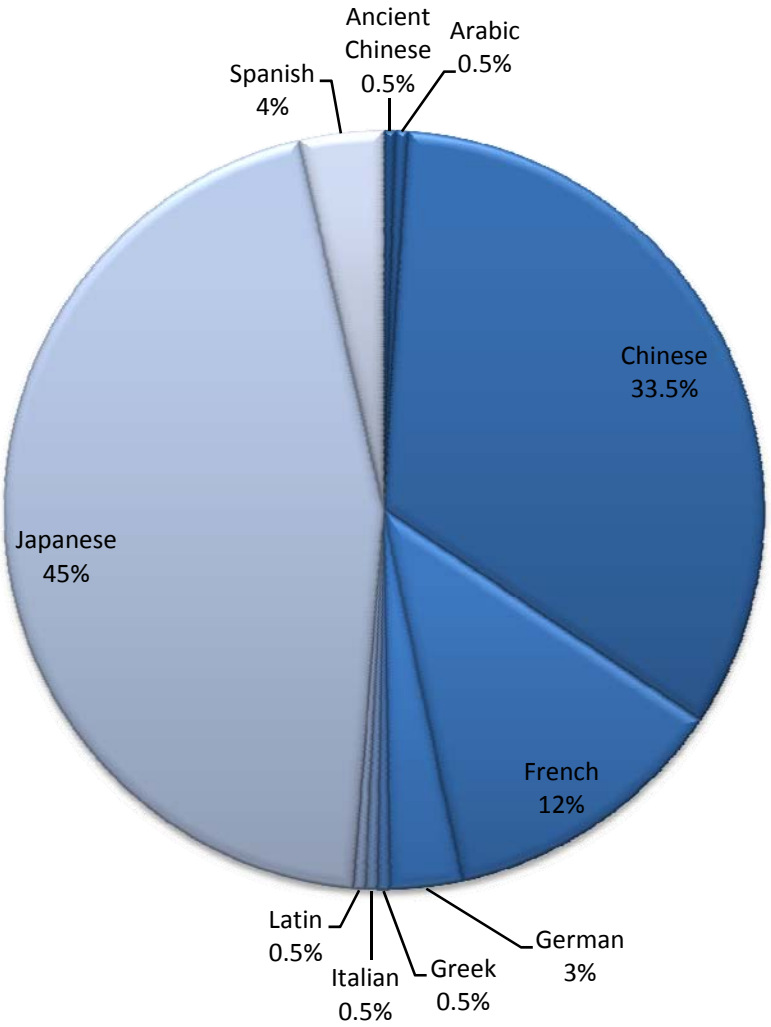


FIG. 78. OTHER POPULAR LANGUAGES THE FLUENCY ENHANCEMENT TOOL

Japanese and Chinese were by far the most dominant languages that users wished to be aided with. However we have to consider the fact that most of the survey participants were from Korea, so naturally these two languages were the most popular due to Korea's geographic proximity to these countries. Following that, other European languages were also of interest, in particular French.

5.4.11 What is your age group?

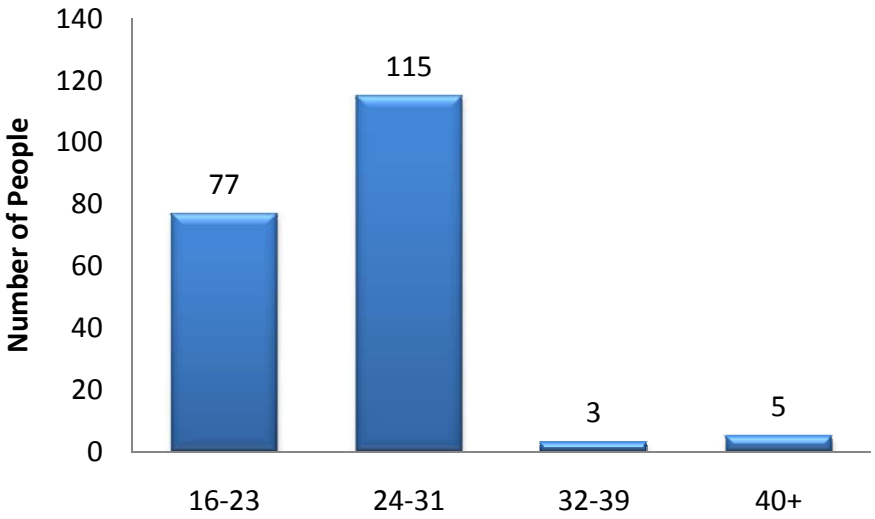


FIG. 79. BREAKDOWN OF AGE GROUPS SURVEYED

Most of the survey participants were of a younger age. This was not intentional, as many senior citizens were also asked to complete the survey, but in failing to understand or replying "No" to the question "Do you speak English?" they could not participate. This perhaps illustrates how the younger generation of Korea has a much better command of the English language that the previous generation.

5.4.12 What is your native language?

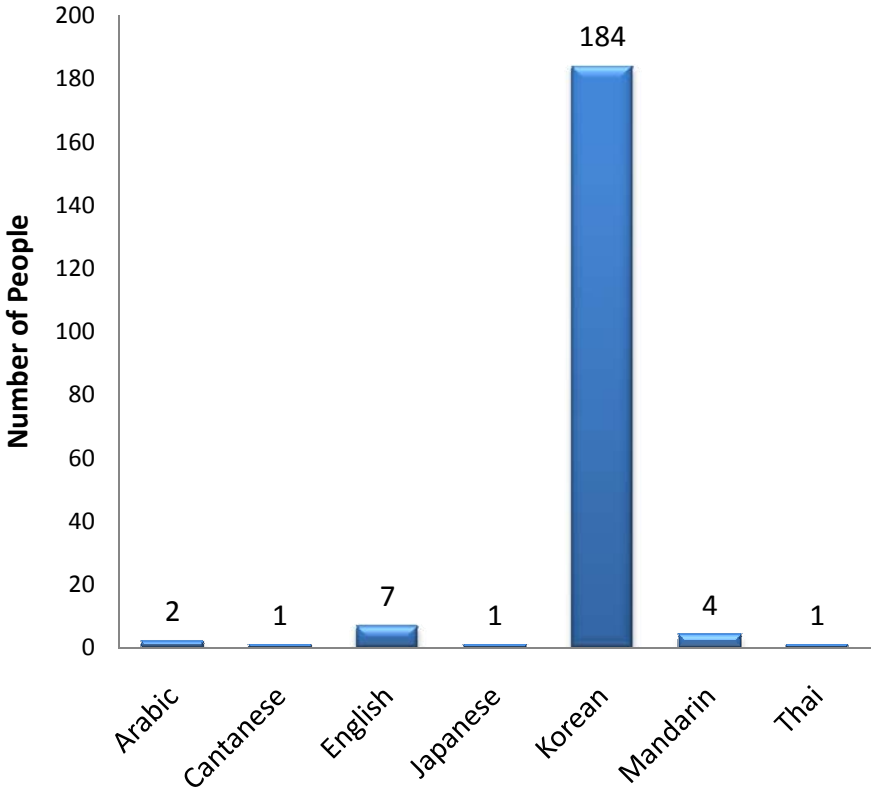


FIG. 80. NATIVE LANGUAGES OF THOSE SURVEYED

Most of the surveying was conducted at Korean universities, thus most of the survey participants were Korean. Seven native speakers were surveyed, who lived in South Korea. The purpose of surveying some native speakers as well was to understand how large the gap is between native and non native speakers of English. As we mentioned earlier, part 1 of our survey was able to identify that there was a reasonable gap in English ability between native speakers (average score of 17.5) and non native speakers (average score of 5.59) in Korea.

5.4.13 What is your occupation?

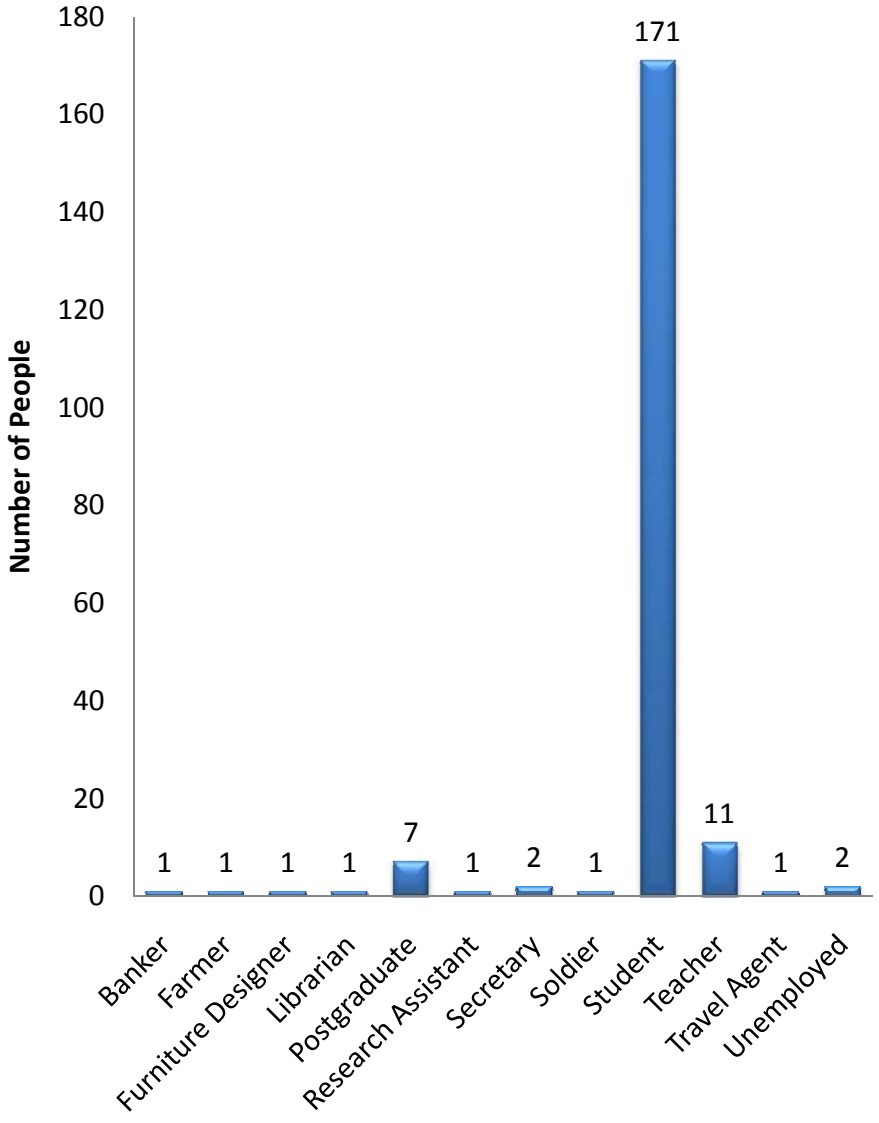


FIG. 81. OCCUPATIONS OF THOSE SURVEYED

Since the survey was conducted at universities, it is no surprise that most of the survey participants were students. Students were willing survey participants as they are young, and most likely require English in their everyday life for academic reasons.

5.4.14 In your everyday life, how often do you need to use English?

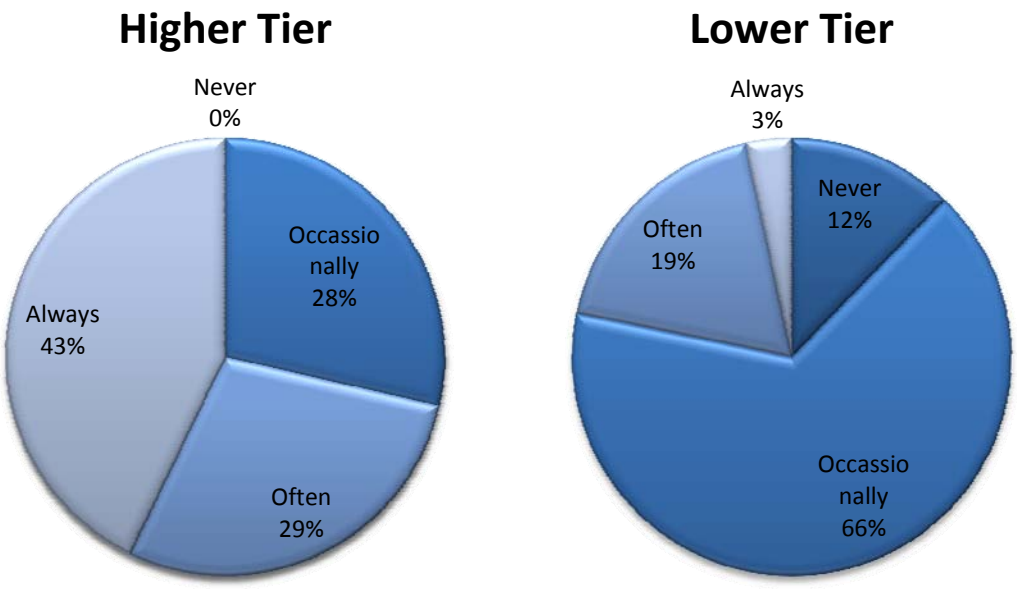


FIG. 82. EVERYDAY NECESSITY FOR ENGLISH

As seen in Fig. 82, the higher tier groups used English considerably more than those in the lower tier group. This is perhaps why those in the lower tier group struggled to do well in the first part of the survey. From this, we can assume that the higher tier group will make more use of our FE tool since they have a more pressing need for English.

5.4.15 Survey Summary

From the survey questions we have put forward to the survey participants, we can now build a profile of who is the target market based on our results. The higher tier group believed they would use the FE tool more often, input English into the FE tool more and also they expected to pay more. Most participants preferred offline time-based use of the FE tool and expected to at least translate approximately 300 or more sentences per month. They also desired the FE tool to be capable of aiding them with Japanese, Chinese and French as well. We are not able to determine exactly the main occupation of users since we only surveyed at universities (thus biased towards students), however we can assume that most users would be under the age of 30.

Target Market Profile:

Candidate:

- Uses English often
- Relatively high proficiency
- Requires Japanese / Chinese / French capable FE tool as well
- Under 30 years old
- Possibly a Korean student

FE Tool Use:

- Use in English
- Offline
- Time-based use (Approximately 300 sentences or more)
- $25 USD per month

6. CONCLUSION

6.1 MEETING THE OBJECTIVES

Objective 1: Machine Translation System Design & Construction

Our first objective was to build an MT system that translates Korean to English based on the SAM fundamentals – Simplicity, Adaptability and Multiplicity. We were able to successfully complete this objective to a reasonable extent. While our MT system was capable of achieving translations, the MT algorithm and the linguistic data it feeds upon could still be much more matured. This is why the publication via the internet of comparable industry standard results has been postponed until the near future when the MT system has been more matured.

Objective 2: Web Crawler Design & Construction

Our second objective was to build a web crawler so we could power our MT system with vast amounts of linguistic data. We have succeeded in building a web crawler, and through testing we have established that it should be able to obtain large amounts of quality linguistic data in a reasonable amount of time. However the web crawler still has room for improvement, and with a little more tweaking, we can finish our collection of linguistic data to further power our MT system.

Objective 3: Implementation of the Fluency Enhancement Process

Our third objective is the key objective of this research, which is the development of our FE process. Not only did we successfully design the FE process, we were also able to prove through our results that it really improves the quality (fluency) of translation output, especially when the subject MT system can access large amounts of linguistic data and has internal processes that complement the FE process.

Objective 4: Tweaking & Understanding the Fluency Enhancement Process

When designing the FE process, we tried to make the algorithm as dynamic as possible, so we had a range of input parameters, giving us a control panel to work from in tweaking the performance of the FE process applied to our MT system. This turned out to be a great idea, because from altering the input parameters we learnt about the weaknesses of our MT system. For example through increasing the values of the CR and the NOS, we found that increasing them together improved the quality of results (Refer to Fig. 62 on page 102). However we were restricted in increasing the values to a certain point because the Google N-Gram corpus only held N-Grams up to 5 tokens in length. Thus we realized we needed more N-Grams that were longer in length to further improve the FE process, so we then designed our web crawler to have a control parameter to search for larger N-Grams accordingly. We have and still are improving the FE process and this feedback loop of testing and tweaking is fundamental in doing this.

Objective 5: Develop a Market Profile for the Fluency Enhancement Tool

Our last objective was to test the market and understand if the FE process we had designed came across as a useful concept, and whether it also had some commercial potential. We put forward our idea in the form of an FE tool that functioned as a writing aid that had translation capabilities. The results of our market research were fairly positive, proving our FE tool has some commercial potential.

Overall we were able to successfully meet our objectives; however there is still room to work towards them in a broader scope. Our MT system, FE process and linguistic data still need to undergo further development, and further understanding of the market can also help ensure the design of our FE tool has some commercial viability. In fact it would be wise to redefine these objectives upon further research conducted after the submission of this thesis. Redefined objectives should also take into account the future work suggested later in the following pages.

6.2 ADVANTAGES & DISADVANTAGES

Table 19 has a list of advantages and disadvantages of our FE technology. These advantages and disadvantages are what exist now for FE in its relatively early stages of development. Further research into FE technology should see the disadvantages mitigated or even eradicated and the advantages substantially improved.

TABLE 19. ADVANTAGES AND DISADVANTAGES TO FE TECHNOLOGY

Advantages	Disadvantages
Most likely an FE translation will produce a better translation	There is a small possibility that the quality of the original translation will be degraded
Use of an LSS improves performance allowing FE to take place in real time	Ideal FE is not always able to be performed in real time
FE functionality can exist in the capacity of an MT system add-on	A pre-existing MT system may not be able to take full advantage of FE
The FE algorithm has portable functionality across any language that it has a corpus for	A reliable and expansive corpus needs to be acquired to achieve reasonable quality FE

From Table 19 we can conclude that it is advantageous to use FE when you can design the MT system to complement the FE process, an appropriate configuration to calculate the LSS is used and there is a large amount of linguistic data available for it to improve the MT system's output to a reasonable degree. If the MT system already has a well established and rigid design, with little linguistic resources and the FE process must execute as an add-on process, then the FE process will not improve MT system's output significantly.

6.3 FUTURE WORK

Fortunately future work on FE will in fact continue to occur. After the submission of this thesis Kaleido K will continue to endeavour with the development of FE technology and the linguistic resources required to power it. Here are some future goals of Kaleido K in regard the development of FE technology.

Superior Corpora

The quality, size and nature of corpora directly influence the performance of an MT system. The more data you feed an MT system the further it increases in accuracy. Research by Microsoft confirms this in Fig. 83. Notice however that the positive returns diminish as more data are added [19]. Thus we must continue to improve our corpora. Better corpora can be constructed with a better web crawler, so improving our web crawler is the key to achieving this goal.

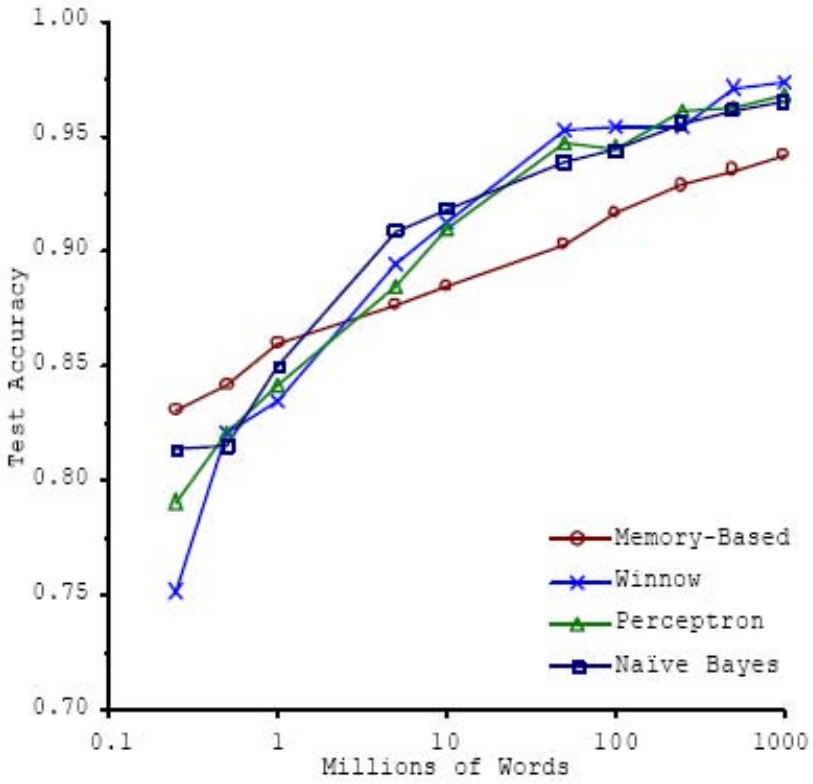


FIG. 83. LEARNING CURVES FOR CONFUSION SET DISAMBIGUATION

Optimization / Presets

More optimized applications of FE will further improve the quality of FE output, involving more suitable PRs, larger values for the CR and the NOS, and using more efficient and accurate LSS calculations. For different applications the way to optimize FE is likely to differ, thus we can derive some optimization presets that a user of FE can simply apply without having to understand how things such as the CR, NOS, PR, LSS and so forth interact with each other and influence the final output.

Real Time Performance

Our algorithm is usually able to perform FE in the time frame of under a minute. However if we want our MT system to ever be accessible via the internet, it of course has to perform much faster (basically within a browser window refresh). Thus efforts need to be made to increase the speed of real time performance. Recently we increased the speed of the FE algorithm by 300%, simply by creating a cache for previously searched N-Grams, meaning the database did not need to be accessed as often. Continued efforts such as these to ensure the FE process executes more elegantly should see the FE algorithm able to perform in real time.

Complementation

When internal processes of an MT system complement the FE process, the quality of translation output improves. Thus it is in our best interests to further understand how other processes in an MT system can complement FE and redesign them to optimize complementation. Redesigning the FE process so it can also be complemented by other technologies and resources external to the MT system should also not be overlooked. For example we could redesign the FE process to take advantage of richer and more elegantly structured linguistic data to improve FE performance.

6.4 CLOSING STATEMENT & FINAL THOUGHTS

To conclude, this thesis has demonstrated that the correct application of the FE process to an MT system can improve its performance. Further development and testing of FE technology and corpora to power it will surely produce even more competitive results which can be used to further refine our FE technology. This in turn should help FE secure itself as an important technology in the MT industry in the future. However in our closing statement, we would like to demonstrate what we have learnt about MT and take the opportunity to share our ideas on where we think the ultimate solution will be found, as this also indicates the direction our research will take after the submission of this thesis.

While we believe FE greatly adds to the success of an MT system, we have not fooled ourselves into believing the process can by itself achieve seamless translation that is on par with a human. FE only deals with text, and text alone is not always enough for an MT system to comprehend the underlying message behind every translation. Regardless of the approach, MT systems that only deal with text eventually confront a semantic barrier. We think Yehoshua Bar-Hillel had a good point with his pen and box argument. MT systems need real word and context based knowledge to be able to rule out translations that are logically flawed and out of context. MT needs to be sensitive to the real world, so it knows that a box cannot fit inside a pen we write with. Text based statistics can resolve this issue to an extent; however we need the MT system to know intuitively by having an understanding of the attributes and functions of real world objects. Considering the Vauquois Triangle shown earlier, no matter what approach is used for MT, the more the process grows in articulation and climbs up the triangle, the more it is likely to fall under the umbrella of AI.

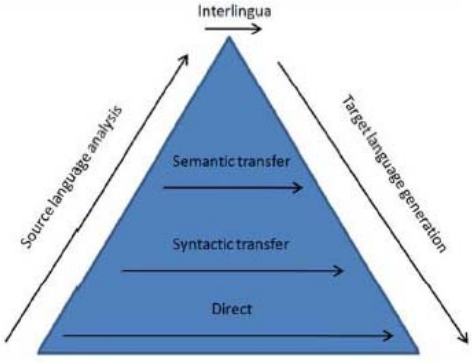


FIG. 84. THE VAUQUOIS TRIANGLE

We believe that the ultimate MT system will have a refined form of translation AI, encouraging the MT system to think for itself as a multi-input system (vision, sound etc) with real world and context based knowledge. The MT system would have profiles of typical attributes and functions that objects have, and then by using information from these profiles, it could solve complex ambiguities.

FIG. 85. OBJECT PROFILING TO OVERCOME SEMANTIC AMBIGUITIES

Let's refer to Fig. 85 and see how word profiling can aid us in translation. Consider the sentence "Open the pen." The MT system may struggle to understand which pen is being talked about with such little information available in the text, and since the definition of a pen is often the one we write with, it is likely the MT system will translate it incorrectly using this assumption. However if the MT system checks over the profiles it has developed for the word pen, it will find that the definition of pen that holds animals in is more likely to be *subjected to the action* of being *opened*, thus clarifying the semantic ambiguity. Let's now consider Bar-Hillel's example, "A box is inside the pen." Considering the profiles again, the typical dimensions of a pen that writes are relatively smaller than the typical dimensions of a box, thus it is likely the box is inside a pen that holds animals.

You may have noticed Fig. 85 resembles what is known as Object Orientated Programming (OOP). Each instance of an object is derived from a class, which is like the blueprint of an object. Classes have defined attributes and functions (which are analogous to adjectives and verbs). For example let's say we have a pen, and it is blue and plastic. This pen object we have is a unique instance of the pen class, which has the attributes of being blue and plastic, and the function of being able to write. Now moving along, what we think should be done here is some reverse engineering. Instead of designing a class, and then instantiating objects, we would like to do it the other way around. We would like to collect a whole lot of unique pen objects from the real world, and analyze them all to define a generic pen class. In other words, build up a profile of what a real world pen is, and if pen has several meanings, then develop several different profiles to accommodate them all. If we have pens such as a marker which is a variation of a pen, we can use inheritance and derive subclasses of the base pen class. In fact much of the theory behind OOP can be applied to natural language. Our example can be observed in Fig. 86 on the following page.

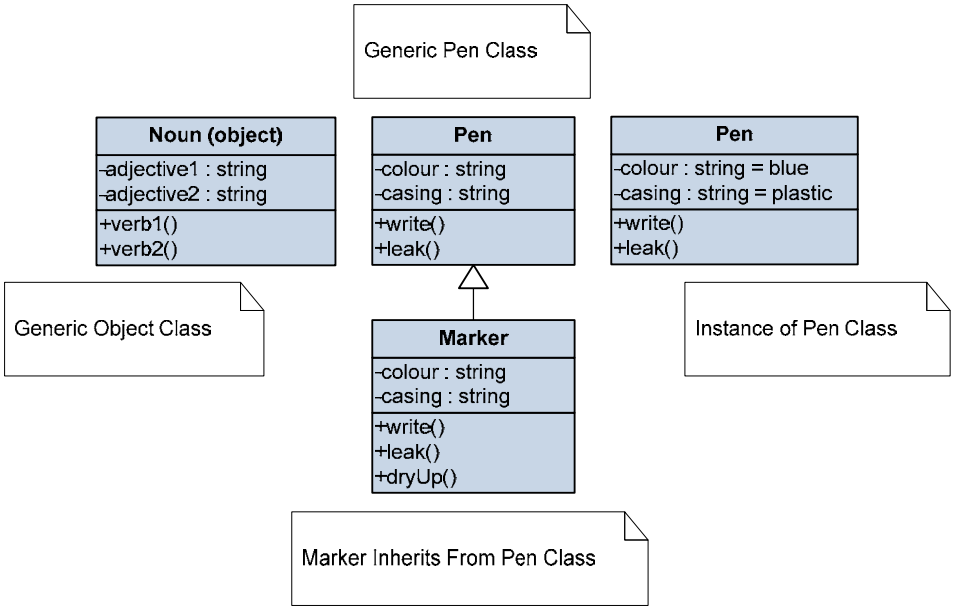


FIG. 86. APPLICATION OF OBJECT ORIENTATED PROGRAMMING THEORY TO NATURAL LANGUAGE

When developing these profiles, text, images, sounds and text describing the smell and feel of an object should be used. Using all this information, we can then put together the typical attributes and functions that each object has. If we structure linguistic information in the style of OOP, we can close the gap between human and machine when it comes to deriving sensible translations. Every time a human hears a word several images, sounds and other knowledge associated with that word appear in their mind, building up a profile of how they would typically interpret that word. From this they can make sensible assumptions and conclusions about the translations they produce. So by forming object profiles, we are giving MT the ability to do this as well and make sensible decisions. Our word profiling implements a semantic framework. Dictionaries of MT systems need to better implement semantic frameworks such as ours, so that a semantic profile of each word can be built up through a variety of multi input linguistic data.

Consider a dictionary that implemented our semantic framework of word profiling. From the images collected about words, we can identify what boxes and pens look like, and how other objects usually interact with them. If our MT system was translating a movie on the fly, the movie could be used as a visual input into our MT system. Images of objects seen in the movie could be compared with those in our semantic dictionary, and help aid with translations when it comes to making semantic decisions that will ensure context based translations. Of course multiple inputs into an MT system would significantly increase the processing and resources required to complete translations; however we can only hope that the future will continue to see improvements in hardware performance. It's possible that visual and other inputs could be ignored and only processed when the MT system actually suffers from some linguistic ambiguities in order to improve the MT system's real time performance.

How can we profile words to build a generic class for each potential object? The internet is a great source, especially because these days there are often websites (i.e. online shopping) that contain specific information about objects. However the challenge in obtaining this information and formulating it into useful profiles is not easy. The web has become more diverse in how information is structured, so any web crawler needs to be quite meticulous and robust in order to adequately collect word profiling information.

Multi input word profiling to create a semantic dictionary is something that would improve the quality of a translation before it reached the FE stage, and FE could also make use of it if it was adapted to do so. After conducting this research, we have realized how the data that feeds an MT system is crucially important to its success. We believe that perhaps structuring the linguistic data available to an MT system is often overlooked, and too much attention is given to the MT system's algorithm. This is why Google is doing so well at the Open MT Evaluations, because they understand how important data is to the translation process.

FE is a fantastic technology we have developed, but the purpose we developed it for was the greater goal of improving the quality of MT. We consider FE to be a worthy process to achieve pure text translation. But in terms of designing the ultimate solution to MT, we understand that a solution that includes extra information about the translation outside of that included in the text is required. As Kaleido K will further endeavour to improve FE technology, the much larger goal of combining it with alternative technologies based on multi input semantic based MT systems will also be pursued. Fig. 87 illustrates a potential model we could use to have a semantic dictionary aid FE in our MT system.

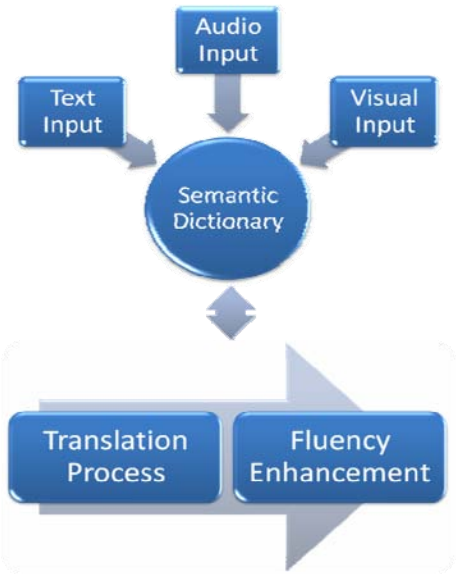


FIG. 87. IMPLEMENTATION OF A DICTIONARY WITH A SEMANTIC FRAMEWORK

MT is a complex problem, and in this thesis we have demonstrated and evaluated how our FE technology is able to tackle it. We have also suggested where we think research efforts should be concentrated in the future and shed light on supporting ideas and methods that also improve the FE process, such as the collecting and structuring of linguistic data. This is now the end of our dissertation, and we hope you as the reader have been educated, enlightened and entertained. You may even wish to follow up on the development of our FE technology, MT systems and linguistic data resources by visiting the Kaleido K website - www.kaleidok.com. Everything you have read in this thesis is still a work in progress. The results we have obtained so far are very promising and the true potential of FE and its supporting components shall be realized in the near future. Look forward to seeing Kaleido K compete in the Open MT Evaluation in upcoming years.

7. APPENDICES

7.1 APPENDIX A

7.1.1 Entrant Details of the NIST 2008 Machine Translation Evaluation[20]

NIST ID	Company	Location
Apptek	Applications Technology Inc.	USA
AUC	The American University in Cairo	Egypt
BASISTECH	Basis Technology	USA
BBN	BBN Technologies	USA
BJUT-MTG	Beijing University of Technology / Machine Translation Group	China
CAS-IA	Chinese Academy of Sciences, Institute of Automation	China
CAS-ICT	Chinese Academy of Sciences, Institute of Computing Technology	China
CAS-IS	Chinese Academy of Sciences, Institute of Software	China
CMU-EBMT	Carnegie Mellon	USA
CMU-SMT	Carnegie Mellon, interACT	USA
CMU-XFER	Carnegie Mellon	USA
Columbia	Columbia University	USA
CUED	University of Cambridge, Dept. of Engineering	UK
Edinburgh	University of Edinburgh	UK
Google	Google	USA
HIT-IR	Harbin Institute of Technology, Information Retrieval Laboratory	China
HKUST	Hong Kong University of Science & Technology	China
IBM	IBM	USA
LIUM	Universite du Maine (Le Mans), Laboratoire d'Informatique	France
MSRA	Microsoft Research Asia	China
NRC	National Research Council	Canada
NTHU	National Tsing Hua University	Taiwan
NTT	NTT Communication Science Laboratories	Japan
QMUL	Queen Mary University of London	UK
SAKHR	Sakhr Software Co.	Egypt
SRI	SRI International	USA
Stanford	Stanford University	USA
UKA	Universitaet Karlsruhe	Germany
UMD	University of Maryland	USA
UPC-LSI	Universitat Politechnica de Catalunya, LSI	Spain
UPC-TALP	Universitat Politechnica de Catalunya, TALP	Spain
XMU-IAI	Xiamen University, Institute of Artificial Intelligence	China
IBM_UMD	IBM / University of Maryland MD	USA
JHU_UMD	Johns Hopkins University / University of Maryland	USA
ISI_LW	USC-ISI / Language Weaver Inc.	USA
MSR_MSRA	Microsoft Research / Microsoft Research Asia	.
MSR_NRC_SRI	Microsoft Research / Microsoft Research Asia / National Research Council Canada / SRI International	.
NICT_ATR	NICT / ATR	Japan
NRC_SYSTRAN	National Research Council Canada / SYSTRAN	.

7.2 APPENDIX B

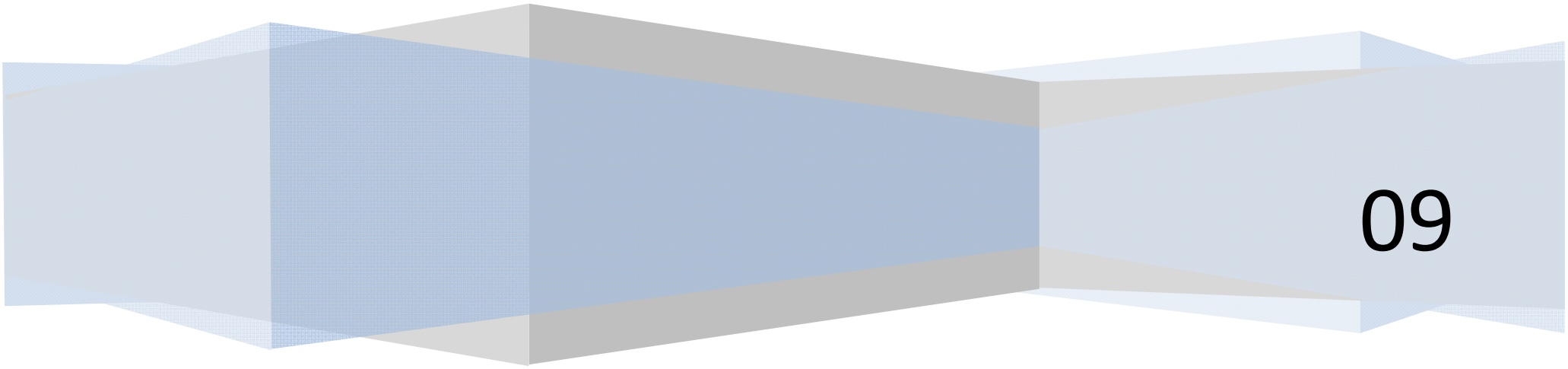
7.2.1 English Market Research Survey

Massey University / Kaleido K

Fluency Enhancement

Applications to Machine Translation

Steve Manion



MARKET RESEARCH SURVEY: PART 1

In the sentences below, find the mistakes and fix them.

- Some sentences do not have any mistakes
- Sometimes, you might need to delete or add to correct the sentence
- A point is given for finding a mistake; a second point is given for fixing the mistake
- The number of mistakes to find and fix is not for you to know
- Lastly, this survey *must* be completed by you, *without* the help of any person or language resource.

1. Would you like some chocolate desert?

2. Often my father and I go to fishing at the lake.

3. I am in my final year of university and it is very difficult.

4. I will graduate this summer and my parents will come to the my ceremoney.

5. Why is your mother and father not able to come your graduation party?

6. After student graduate, they usually sell their books to second hand bookshop.

7. I am not sure how I will find a job next year.

8. I am so exciting about going to Mexico next weekend with john.

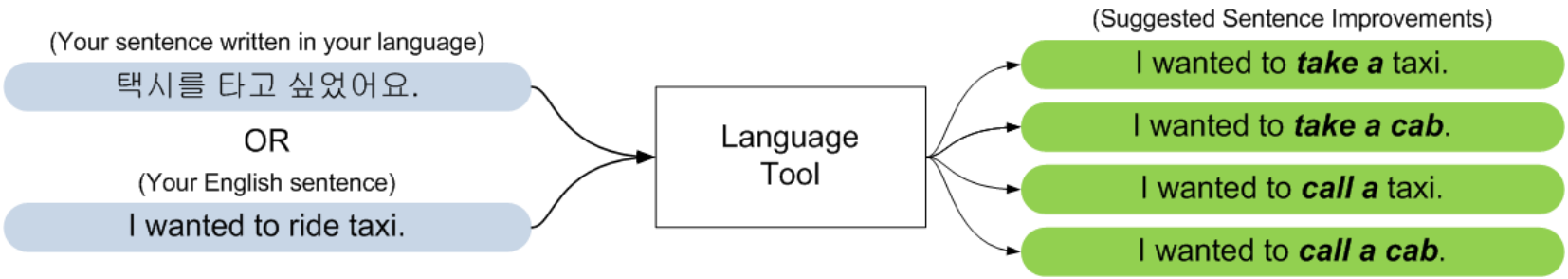
MARKET RESEARCH SURVEY: PART 2

Please read about the tool below. Then answer the questions that follow:

A tool to improve your written English

The tool reads the English sentence you wrote. Then it makes many similar sentences that have the same meaning. The new sentences sound more natural and maybe better express what you want to say. The tool doesn't just fix grammar; it can also completely change the sentence so it sounds like a native speaker said it. This tool can also read your native language, and give you many new English sentences to use in your English writing. The picture below shows how the tool works.

Example:



MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
1	**What do you think of this tool?**	**Poor** **Average** **Good** **Excellent**	
2	**Do you think you would use this tool?**	**No** **Occasionally** **Often**	
3	**Do you think your friends would use this tool? (If your answer is not NO, please comment on who you think would use this tool)**	**No** **Occasionally** **Often**	
4	**If you used the tool, when entering a sentence into it, would you attempt to write the sentence in English, in your own language or both?**	**In English** **In my own language** **Both**	
5	**If you used this tool, how do you want to access it?**	**Download the tool and install it to use offline** **Use it online as a website service**	

MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
6	**If you were to pay for this tool, do you want to pay based on the *length of time used* it or *per sentence improvement*?**	**Length of Time Used** **Per Sentence Improvement**	
7	**If you paid for *length of time used*, how much would you expect to pay for one month's use?**	**$_____USD**	
8	**If you paid *per sentence improvement* and you had 200 sentences improved how much would you expect to pay? (Each sentence could be up to 40 words in length)**	**$_____USD**	
9	**Have you *purchased, seen, or heard* of a tool that restructures and improves your English writing? (Please explain further on this question in the comment box if you can)**	**No** **Purchased** **Seen** **Heard**	
10	**Except English, are there other languages that this tool could help you improve your writing in? (i.e. Chinese, German etc)**	**(Write answer here)**	

MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
11	**Please circle your age group**	**16-23 24-31** **32-39 40+**	
12	**What is your first language?**	**(Write answer here)**	
13	**What is your occupation?**	**(Write answer here)**	
14	**In your everyday life, how often do you need to use English?**	**Never** **Sometimes** **Usually** **Always**	

If you would like to know more about this tool and other Kaleido K language technologies and resources please write your name and email below.* You can also visit our website at **www.kaleidok.com**, join our Facebook group **"Kaleido K Language Community"** or contact Kaleido K directly at **info@kaleidok.com**.

Those who write their name and email go into the draw to win an MP3 player and other prizes!!!

Name: _____

Email: _____

Please indicate what type of emails you would like to receive:

- ☐ Emails related to this Survey / Prize draw
- ☐ Emails related to New Kaleido K Language Technologies and Resources
- ☐ Kaleido K News
- ☐ Kaleido K Special Offers

***Privacy Note:** Please note you will not be spammed or have your email given to any other parties. The emails you will receive will only be of the nature you have selected above.

***** End of Survey *****

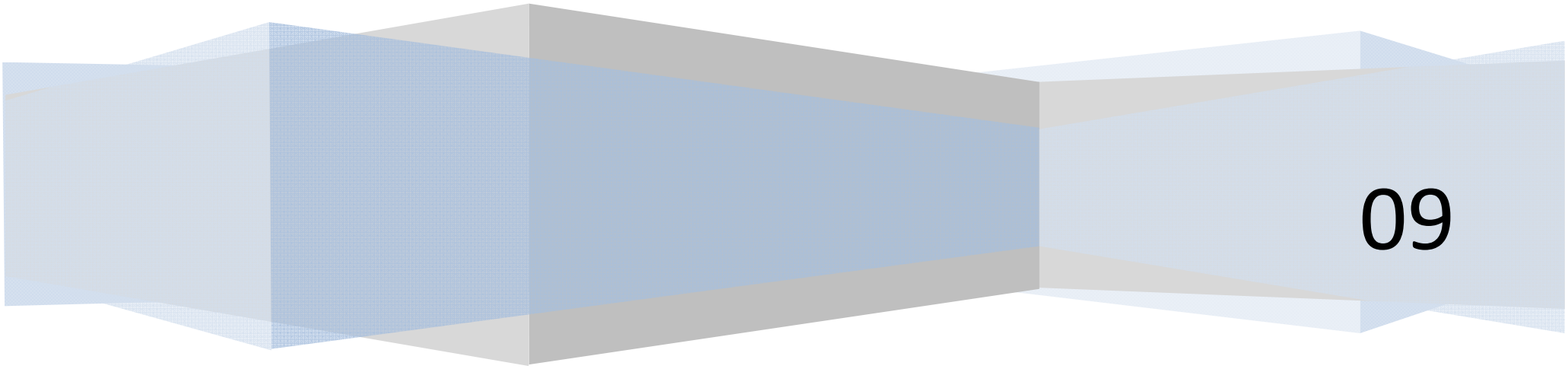
7.2.2 Korean Market Research Survey

Massey University / Kaleido K

Fluency Enhancement

Applications to Machine Translation

Steve Manion



MARKET RESEARCH SURVEY: PART 1

아래의 문장에서, 틀린 부분을 찾아내어 고치시오.

- 몇몇의 문장은 틀린부분이 없을 수도 있습니다.
- 몇개의 문장에서 단어는 문장에 추가 삽입되거나, 지워져야 합니다.
- 각 문장은 2점을 만점으로, 부분점수가 주어집니다.
  (틀린부분을 찾아내었을때는 1점,그 부분을 수정하였을 시에 1점)
- 문장안의 틀린 개수는 주어지지 않습니다.
- 마지막으로, 이 설문조사 중에 인터넷이나 사전등의 사용은 금하고 있으며, 다른사람의 도움없이 혼자서 진행해 주시길 바랍니다. 감사합니다.

1. Would you like some chocolate desert?

2. Often my father and I go to fishing at the lake.

3. I am in my final year of university and it is very difficult.

4. I will graduate this summer and my parents will come to the my ceremoney.

5. Why is your mother and father not able to come your graduation party?

6. After student graduate, they usually sell their books to second hand bookshop.

7. I am not sure how I will find a job next year.

8. I am so exciting about going to Mexico next weekend with john.

MARKET RESEARCH SURVEY: PART 2

아래의 설명을 읽고, 다음장의 질문지에 대답하여 주십시오.

A tool to improve your written English

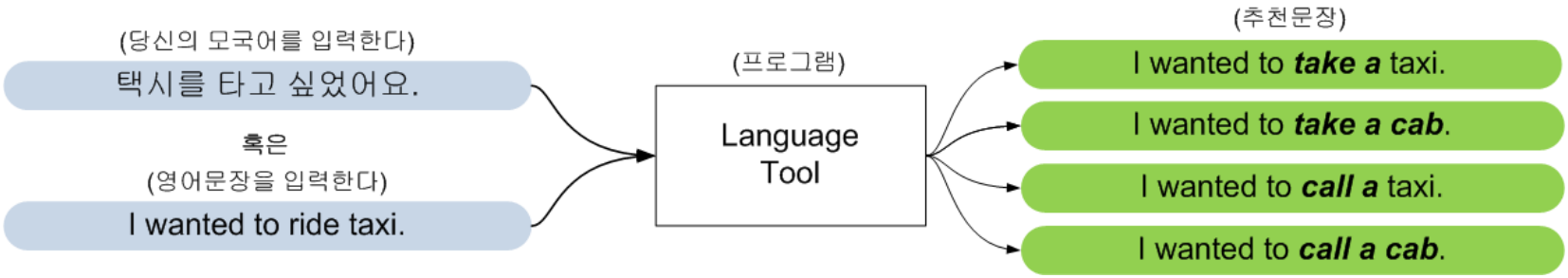
이 프로그램은 당신이 영어로 쓰고자 하는 문장을 읽어낸다.

그리고 나서 그 문장과 비슷하면서도 조금 더 매끄러운 하나의 문장, 또는 두세 개의 비슷한 문장을 생성해낸다. 당신이 이 프로그램에 입력한 문장은 문법적으로만 수정되는 것이 아니라 조금 더 자연스럽고, 일상생활에서 사용하는 문장으로 바뀌어진다.

이 프로그램에서는 당신에게 또 다른 기능을 제공하는데, 그것은 바로 영어에서 영어로의 변환만 가능한 것이 아니라 각국의 언어에서 영어로 번역해주는 기능까지 가지고 있다는 것이다.

아래에 나오는 그림은 주 아이디어를 설명하고 있다.

Example:



MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
1	이 프로그램에 대해서 어떻게 생각하세요?	나쁘다 보통 좋다 대단하다	
2	이 프로그램이 개발된다면 얼마나 사용할 것이라고 생각하나요?	사용하지 않는다 가끔 사용한다 자주 사용한다	
3	당신의 주변사람들도 이 프로그램을 사용할 것이라고 생각하나요? 만약, 아니라고 생각한다면 옆에 이유를 함께 달아주세요)	사용하지 않는다 가끔 사용한다 자주 사용한다	
4	당신이 이 프로그램을 사용하고 있다고 가정하여보자. 문장을 입력하려고 할 때, 모국어와 영어 중 어느 언어를 더 많이 사용할 것 같은가?	영어 모국어 둘다(영어와 모국어)	
5	만약에 당신이 이 프로그램을 사용한다고 할 때, 어떤 방식으로 접근하여 사용하고 싶나?	오프라인에서도 사용할 수 있도록 다운로드 받고싶다 필요할때마다 웹사이트에 들어가 사용하고 싶다	

MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
6	당신이 이 프로그램을 이용하고 돈을 지불하고자 할 때 어떤 방식을 더 선호하십니까?	기간별(ex. 한달.일주일) 문장별(단어개수)	
7	만약 기간별로 프로그램비를 지불해야 한다면, 한달기준 얼마정도가 적당하다고 생각하십니까?	$_____USD	
8	만약 문장별로 프로그램비를 지불해야 한다면, 200 문장을 기준으로 얼마가 적정하다고 생각하십니까? (한문장에 40 단어 이상까지 포용)	$_____USD	
9	전에 이런 프로그램(영어작문을 도와주고 발전시켜주는)을구입하거나 본적이 있나요? (가장 중요한 질문이니 곰곰히 생각하여 주시면 감사하겠습니다.)	아니요 구입한적이있다 본적이있다 들어본적이있다	
10	만약 당신이 작문 실력을 향상시키는 이 프로그램에서 영어 이외에 다른 언어서비스를 제공한다면, 어떤 언어기능을 필요하나요?	(예. 중국어. 독일어 등)	

MARKET RESEARCH SURVEY: PART 2

ID	*Question*	*Selection*	*Comments*
11	당신의 연령이 포함되는 곳에 동그라미 해주세요.	16-23 24-31 32-39 40+	
12	당신의 모국어는 무엇입니까?	(이곳에 답을 써주세요)	
13	당신의 직업은 무엇입니까?	(이곳에 답을 써주세요)	
14	일상생활에서 영어를 얼마나 사용하고 계신가요?	전혀 사용하지 않는다 가끔 사용한다 자주 사용한다 항상 사용한다	

이 프로그램에 대해 더 알고 싶으시거나, 다른 Kaleido K Language 의 기술과 자원이 궁금하시다면 아래에 당신의 성함과 이메일 주소를 남겨 주세요.* 웹사이트(www.kaleidok.com) 를 통해서도 가능합니다. 또한, 국제 친목 사이트인 Facebook 의 클럽**"Kaleido K Language Community"**이나 직접 정보를 받고 싶으신 분은 Kaleido K 의 관계자께 직접 이메일(info@kaleidok.com) 을 보내주시면 감사하겠습니다.

Name: _____

Email: _____

연락처를 기재해주신 분들에 한하여 추첨을 통해 MP3 플레이어 및

다른 경품을 제공해 드리는 행사를 하고 있으니 많은 참여 바랍니다

당신이 받고싶은 종류의 이메일에 체크해주세요.

☐ 이 설문조사와 관련된 이메일

☐ 새로운 Kaleido K Language 의 기술과 자원이 담긴 이메일

☐ Kaleido K 뉴스

☐ Kaleido K 구매/할인 특가 관련 이메일

*지금 이 설문지에 작성하신 이름과 이메일 주소는 관계자 이외에 누구에게도 공개되지 않으며, 당신이 체크 한 것에 관련된 이메일 외 스팸류의 메일은 보내지 않습니다.

***** 설문조사가 끝났습니다. 감사합니다 *****

7.2.3 Raw Data of Market Research Survey

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
1	7	Average	Occas	Occas	Both	Online	Time	2	1	No	No	16-23	Kor	Student	Occas
2	5	Good	Occas	Occas	Both	Online	Time	15	5	No	Chi	16-24	Kor	Student	Often
3	6	Good	Often	Occas	Both	Online	Time	2	2	Purchased	No	16-25	Kor	Student	Often
4	3	Average	Occas	Never	English	Online	Sent	1	2	No	No	16-26	Kor	Student	Occas
5	2	Average	Occas	Occas	Both	Offline	Time	5	5	No	No	16-27	Kor	Student	Occas
6	1	Excellent	Often	Often	Native	Offline	Time	10	3	No	Jap	16-28	Kor	Student	Occas
7	5	Good	Occas	Occas	Native	Online	Time	10	10	No	Jap	16-23	Kor	Student	Occas
8	4	Good	Often	Occas	Native	Offline	Time	20	5	No	Jap	24-31	Kor	Student	Occas
9	4	Good	Occas	Occas	English	Online	Sent	20	2	No	Jap	24-31	Kor	Teacher	Occas
10	12	Excellent	Often	Often	Both	Offline	Time	10	10	No	Chi / Jap	24-31	Kor	PostGrad.	Occas
11	2	Excellent	Often	Often	Native	Offline	Time	25	4	No	Jap	16-23	Kor	Student	Never
12	2	Average	Occas	Often	English	Offline	Time	5	3	Seen	Jap	24-31	Kor	Student	Occas
13	4	Good	Occas	Occas	Native	Online	Time	3	10	No	Spa	24-31	Kor	Student	Occas
14	1	Good	Occas	Occas	Native	Offline	Time	3	5	No	Chi / Jap	24-31	Kor	Student	Occas
15	4	Excellent	Occas	Occas	Native	Offline	Time	10	20	No	Chi	32-39	Kor	Student	Occas
16	14	Excellent	Often	Occas	English	Offline	Time	20	10	No	Fre / Jap / Spa	16-23	Kor	Student	Often
17	5	Good	Occas	Occas	English	Offline	Time	0.5	10	Heard	Jap	24-31	Kor	PostGrad.	Occas
18	2	Good	Occas	Occas	Native	Online	Sent	40	50	No	Chi / Jap	24 - 31	Kor	Trav. Agt.	Occas
19	4	Good	Often	Often	Both	Offline	Time	10	10	Seen	Jap	24-31	Kor	Student	Occas
20	6	Good	Occas	Occas	Native	Online	Time	5	10	Seen	Jap	16-23	Kor	Student	Occas
21	6	Excellent	Occas	Often	Native	Offline	Time	5	10	Seen	No	16-23	Kor	Student	Often
22	4	Good	Occas	Occas	Both	Offline	Time	30	30	No	Jap	24-31	Kor	Student	Occas
23	11	Excellent	Often	Often	Native	Online	Sent	50	0.25	No	Chi	40+	Kor	Librarian	Often

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
24	7	Good	Occas	Occas	Both	Online	Sent	10	1	Heard	No	24-31	Kor	Student	Occas
25	3	Average	Occas	Occas	Native	Offline	Time	10	20	No	Jap	24-31	Kor	Student	Never
26	5	Excellent	Often	Often	Native	Offline	Time	15	5	Seen	Jap	24-31	Kor	Student	Occas
27	1	Good	Often	Often	Both	Online	Time	10	10	No	Jap	24-31	Kor	Student	Occas
28	8	Excellent	Often	Often	Both	Offline	Time	10	10	No	No	24-31	Kor	Student	Occas
29	2	Good	Occas	Occas	Both	Online	Time	5	5	No	Chi	24-31	Kor	Student	Occas
30	4	Good	Often	Often	Native	Offline	Time	10	1	Seen	Jap	24-31	Kor	PostGrad.	Occas
31	1	Good	Occas	Often	Both	Offline	Sent	50	15	Seen	Jap	24-31	Kor	Student	Occas
32	4	Good	Often	Occas	Native	Offline	Time	10	1	Heard	French	24-31	Kor	Student	Occas
33	7	Average	Often	Often	English	Online	Time	10	1	No	French	24-31	Kor	Student	Occas
34	5	Average	Occas	Occas	Native	Offline	Time	30	10	Purchased	No	24-31	Kor	Student	Occas
35	2	Excellent	Often	Often	Both	Online	Time	200	100	No	Chi	24-31	Kor	Student	Often
36	4	Good	Often	Often	Both	Offline	Time	2	2	No	Jap	16-23	Kor	Student	Occas
37	8	Good	Often	Often	Native	Online	Time	100	10	Heard	Chi	16-23	Kor	Soldier	Occas
38	6	Average	Often	Often	Native	Offline	Time	10	10	Purchased	Chi	16-23	Kor	Student	Often
39	17	Good	Occas	Occas	Native	Online	Sent	100	15	Seen	Chi	24-31	Kor	Student	Occas
40	13	Average	Occas	Never	English	Online	Time	15	1	No	(A) Chi / Grk / Lat	16-23	Kor	Student	Often
41	6	Excellent	Often	Often	Both	Online	Time	10	20	Seen	French	16-23	Kor	Student	Occas
42	4	Poor	Never	Occas	English	Online	Sent	10	5	Heard	German	24-31	Kor	Student	Occas
43	3	Good	Occas	Occas	Both	Offline	Time	5	5	No	French	16-23	Kor	Student	Occas
44	6	Good	Often	Often	English	Online	Time	6.5	0.5	No	No	16-23	Kor	Student	Occas
45	6	Excellent	Often	Occas	English	Online	Sent	80	25	No	German	16-23	Kor	Student	Occas
46	13	Average	Occas	Occas	English	Online	Sent	50	5	Purchased	Jap	16-23	Kor	Student	Often
47	10	Good	Often	Occas	Both	Online	Sent	50	20	No	No	16-23	Kor	Student	Often
48	7	Average	Occas	Occas	Both	Offline	Time	30	30	No	Spa	16-23	Kor	Student	Occas
49	10	Excellent	Often	Often	Both	Offline	Time	10	10	Seen	Jap	16-23	Kor	Student	Occas

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
50	0	Excellent	Often	Often	English	Offline	Time	4	4	No	Jap	24-31	Kor	Student	Never
51	10	Good	Often	Often	English	Offline	Sent	35	25	No	Jap	24-31	Kor	Student	Never
52	2	Excellent	Occas	Occas	Native	Offline	Time	10	10	Seen	French	24-31	Kor	Fur. Des.	Never
53	4	Excellent	Occas	Occas	English	Offline	Time	0	0	No	No	16-23	Kor	Student	Occas
54	0	Good	Occas	Occas	Native	Offline	Time	10	2	No	Jap	24-31	Kor	Fr. Lance	Occas
55	0	Excellent	Occas	Often	Native	Offline	Time	4	16	No	Chi / Jap	24-31	Kor	Student	Never
56	3	Excellent	Often	Occas	Native	Offline	Time	11	11	No	Chi	24-31	Kor	Student	Occas
57	7	Good	Occas	Occas	English	Offline	Time	10	10	No	Chi	24-31	Mand	Student	Often
58	6	Good	Often	Often	English	Offline	Time	10	200	Purchased	Fre	24-32	Arab	Student	Always
59	6	Average	Occas	Often	English	Offline	Time	14	50	Purchased	Arab	24-33	Arab	Student	Always
60	3	Good	Occas	Occas	English	Offline	Time	14	10	Heard	Spa	24-34	Mand	Student	Often
61	6	Good	Occas	Occas	English	Online	Time	5	10	No	No	32-39	Thai	Lecturer	Occas
62	11	Good	Occas	Occas	Native	Online	Time	5	5	Seen	Jap	24-32	Kor	Student	Always
63	13	Excellent	Occas	Occas	English	Offline	Time	10	10	No	No	16-23	Mand	Student	Often
64	6	Good	Occas	Occas	Both	Offline	Time	15	20	No	Chi	16-23	Cant	Student	Always
65	2	Excellent	Occas	Occas	Native	Offline	Time	10	10	No	Chi	16-23	Kor	Student	Occas
66	1	Good	Often	Occas	English	Offline	Time	10	10	No	Jap	16-23	Kor	Student	Never
67	0	Good	Often	Often	Both	Offline	Sent	10	3	No	Chi	24-31	Kor	Student	Occas
68	10	Excellent	Occas	Occas	Native	Online	Sent	0.5	0.5	No	Chi	16-23	Kor	Student	Occas
69	14	Good	Occas	Often	Native	Online	Sent	10	10	No	Jap	16-23	Kor	Student	Occas
70	3	Good	Occas	Occas	Both	Offline	Time	30	30	Seen	No	16-23	Kor	Student	Occas
71	6	Good	Occas	Occas	English	Offline	Time	20	10	No	Chi / Jap	16-23	Kor	Student	Occas
72	2	Good	Occas	Occas	Native	Online	Time	5	5	Heard	Chi	16-23	Kor	Student	Often
73	2	Excellent	Often	Often	Native	Offline	Time	20	5	No	Chi	16-23	Kor	Student	Often
74	6	Good	Occas	Occas	Native	Offline	Time	4	2	Heard	Chi	24-31	Kor	Student	Occas
75	14	Good	Often	Often	English	Online	Time	30	50	Seen	Ger / Jap	16-23	Kor	Student	Occas

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
76	2	Good	Occas	Occas	Both	Offline	Sent	50	20	No	Jap	40+	Kor	Teacher	Never
77	5	Excellent	Often	Occas	Native	Offline	Time	20	25	No	Spa	16-23	Kor	Student	Occas
78	0	Excellent	Occas	Occas	Native	Offline	Time	10	10	No	Jap	40+	Kor	Farmer	Never
79	0	Good	Occas	Occas	Both	Offline	Time	3	3	No	Chi	24-31	Kor	Student	Often
80	9	Excellent	Often	Often	English	Offline	Time	10	10	No	Fre / Jap	24-31	Kor	Student	Often
81	6	Good	Often	Often	Native	Offline	Time	20	20	Seen	French	24-31	Kor	Student	Often
82	8	Good	Occas	Occas	Native	Offline	Sent	10	5	Seen	Ita	24-31	Kor	Student	Occas
83	7	Good	Occas	Occas	Native	Offline	Time	10	5	No	No	24-31	Kor	Student	Occas
84	1	Average	Occas	Occas	English	Offline	Time	20	10	No	Chi / Jap	16-23	Kor	Student	Never
85	2	Good	Often	Often	Native	Offline	Time	10	10	Seen	Chi / Jap	16-23	Kor	Student	Occas
86	4	Average	Occas	Occas	Both	Online	Time	320	40	Purchased	Jap	16-23	Kor	Student	Always
87	16	Excellent	Often	Often	English	Offline	Time	5	5	No	Jap	24-31	Jap	Banker	Always
88	8	Average	Often	Often	Native	Offline	Time	3	2	No	No	16-23	Kor	Student	Often
89	4	Good	Occas	Occas	Native	Online	Sent	3	5	No	Chi / Jap	16-23	Kor	Student	Occas
90	2	Good	Often	Occas	Native	Online	Time	10	3.5	No	Chi / Jap	24-31	Kor	Student	Occas
91	5	Excellent	Often	Often	Both	Offline	Time	20	0.5	No	Jap	24-31	Kor	Student	Occas
92	8	Good	Often	Occas	English	Offline	Time	3	2	No	Chi / Jap	16-23	Kor	Student	Occas
93	4	Good	Occas	Never	English	Online	Sent	0	0	Heard	Chi / Jap	16-23	Kor	Student	Never
94	5	Average	Occas	Never	Both	Offline	Time	0	0	Heard	Chi / Jap	24-31	Kor	Student	Occas
95	0	Good	Occas	Occas	Native	Offline	Sent	1	1	No	Chi	16-23	Kor	Student	Occas
96	0	Good	Occas	Occas	Native	Offline	Time	1	1	No	No	24-31	Kor	Student	Often
97	0	Average	Occas	Occas	Both	Online	Time	100	100	No	No	24-31	Kor	Student	Occas
98	2	Good	Often	Often	Native	Offline	Time	10	5	Seen	Jap	24-31	Kor	Student	Often
99	3	Good	Occas	Occas	Native	Offline	Sent	1	1	Heard	Jap	16-23	Kor	Student	Never
100	4	Excellent	Often	Often	Native	Offline	Time	100	10	Heard	Jap	24-31	Kor	Student	Occas
101	0	Good	Often	Often	Native	Online	Time	5	1	Seen	Jap	24-31	Kor	Student	Occas

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
102	3	Average	Occas	Occas	Native	Offline	Time	10	3	No	Fre / Jap	24-31	Kor	Student	Occas
103	6	Good	Occas	Occas	Native	Offline	Time	2	1	Seen	Chi / Jap	24-31	Kor	Student	Occas
104	2	Good	Often	Occas	Both	Offline	Time	7	7	Seen	Jap	24-31	Kor	Student	Occas
105	2	Average	Occas	Occas	English	Offline	Time	5	2	No	Chi / Fre / Jap	24-31	Kor	Student	Occas
106	0	Average	Occas	Often	English	Offline	Time	15	5	Seen	Chi	24-31	Kor	Student	Never
107	0	Average	Occas	Occas	Native	Offline	Time	2	0	No	French	16-23	Kor	Student	Occas
108	8	Good	Often	Often	Both	Offline	Time	15	5	Seen	Chi / Jap	24-31	Kor	Student	Occas
109	10	Good	Often	Often	Both	Online	Time	30	50	Seen	Jap	24-31	Kor	Student	Never
110	3	Excellent	Often	Occas	Native	Offline	Sent	2	2	Heard	German	16-23	Kor	Student	Occas
111	0	Good	Occas	Occas	Both	Online	Time	10	10	Seen	French	16-23	Kor	Student	Occas
112	12	Excellent	Often	Often	Both	Offline	Sent	100	100	No	Jap	24-31	Kor	Student	Occas
113	10	Excellent	Often	Often	Native	Online	Time	50	10	No	No	24-31	Kor	Student	Occas
114	6	Good	Often	Often	Both	Offline	Time	50	10	Seen	No	24-31	Kor	Student	Occas
115	10	Good	Occas	Occas	Native	Online	Time	20	10	Seen	Jap	24-31	Kor	Student	Occas
116	3	Good	Often	Often	Native	Offline	Time	5	5	Seen	Jap	16-23	Kor	Student	Occas
117	0	Good	Often	Often	Both	Offline	Time	10	2	Seen	Jap	24-31	Kor	Student	Often
118	1	Excellent	Often	Often	Both	Online	Time	3	1	No	Jap	16-23	Kor	Student	Occas
119	3	Average	Never	Never	Native	Offline	Time	10	5	Seen	Chi	16-23	Kor	Student	Occas
120	6	Average	Occas	Occas	Native	Offline	Time	10	5	No	Jap	16-23	Kor	Student	Occas
121	1	Good	Often	Occas	Both	Online	Time	0	0	Seen	Jap	16-23	Kor	Student	Occas
122	7	Average	Occas	Occas	Native	Offline	Time	2	1	Heard	No	24-31	Kor	Student	Occas
123	3	Average	Occas	Occas	English	Online	Time	5	1	Heard	No	16-23	Kor	Student	Occas
124	6	Average	Never	Occas	Native	Online	Time	10	10	Heard	Fre / Jap	16-23	Kor	Student	Occas
125	5	Excellent	Occas	Occas	English	Offline	Time	5	1	Seen	No	16-23	Kor	Student	Occas
126	8	Good	Often	Often	Native	Online	Sent	1	10	Seen	Chi	24-31	Mand	Student	Often
127	11	Average	Occas	Occas	English	Offline	Time	0	0	No	French	24-31	Kor	Student	Occas

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
128	9	Good	Occas	Occas	English	Offline	Time	10	3	No	Fre / Ita	24-31	Kor	PostGrad.	Occas
129	6	Good	Often	Often	Both	Offline	Sent	2	1.5	No	Chi	24-31	Kor	Student	Occas
130	4	Average	Occas	Occas	Native	Offline	Time	10	3	Seen	Jap	24-31	Kor	Student	Occas
131	2	Good	Occas	Often	Both	Offline	Time	10	10	No	Chi	24-31	Kor	Student	Occas
132	2	Excellent	Occas	Occas	English	Offline	Time	25	25	No	Chi	24-31	Kor	Student	Occas
133	6	Excellent	Often	Often	Native	Offline	Time	100	5	No	Chi / Jap	24-31	Kor	PostGrad.	Occas
134	5	Good	Occas	Occas	English	Offline	Time	100	10	No	Jap	24-31	Kor	Student	Occas
135	8	Good	Occas	Occas	Native	Online	Time	5	5	Heard	Jap	16-23	Kor	Student	Occas
136	6	Good	Occas	Occas	Native	Offline	Sent	5	5	Seen	No	24-31	Kor	PostGrad.	Occas
137	4	Average	Occas	Occas	Both	Offline	Time	5	5	Seen	Chi	24-31	Kor	Student	Always
138	3	Good	Often	Often	Native	Offline	Time	50	15	No	No	16-23	Kor	Student	Occas
139	10	Average	Occas	Occas	Both	Offline	Time	30	20	No	Spa	24-31	Kor	Student	Occas
140	8	Good	Occas	Occas	Both	Offline	Time	10	10	Heard	No	24-31	Kor	Student	Never
141	3	Good	Often	Occas	Both	Offline	Time	15	3	Seen	Chi / Spa	16-23	Kor	Student	Often
142	2	Good	Often	Occas	Native	Offline	Time	5	10	No	Jap	16-23	Kor	Student	Occas
143	6	Excellent	Occas	Often	Native	Offline	Time	1	0.5	No	Jap	24-31	Kor	No Job	Never
144	4	Good	Occas	Occas	Both	Offline	Time	5	5	Heard	Jap	24-31	Kor	Student	Occas
145	4	Average	Occas	Occas	Native	Online	Sent	300	100	No	Chi	24-31	Kor	Student	Occas
146	0	Good	Often	Often	Both	Online	Time	5	1	Heard	French	16-23	Kor	Student	Often
147	2	Good	Often	Often	Native	Offline	Time	100	5	Seen	Jap	24-31	Kor	Student	Occas
148	5	Good	Often	Often	Native	Online	Time	20	20	No	Chi / Fre / Jap	24-31	Kor	Student	Occas
149	7	Good	Often	Often	Both	Online	Sent	10	20	No	Chi	40+	Kor	Teacher	Occas
150	4	Good	Occas	Occas	Native	Online	Time	10	20	Heard	Jap	40+	Kor	Teacher	Occas
151	0	Average	Occas	Never	Native	Online	Time	10	5	No	Jap	24-31	Kor	Student	Never
152	8	Average	Occas	Occas	Native	Online	Time	10	10	No	Chi / Jap	24-31	Kor	Student	Occas
153	8	Good	Occas	Occas	Native	Offline	Sent	10	5	No	Chi / Jap	24-31	Kor	Student	Occas

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
154	0	Average	Never	Never	Native	Online	Time	50	50	No	Jap	24-31	Kor	Student	Occas
155	5	Good	Occas	Occas	English	Offline	Time	5	3	No	German	16-23	Kor	Student	Occas
156	4	Good	Occas	Occas	English	Offline	Sent	10	10	No	Chi	16-23	Kor	Student	Occas
157	2	Excellent	Often	Often	Both	Online	Time	10	25	No	No	16-23	Kor	Student	Never
158	6	Good	Often	Occas	Native	Online	Time	50	30	No	Chi	16-23	Kor	Student	Often
159	19	Good	Occas	Often	English	Offline	Time	25	17.5	No	Spa	16-23	Kor	Student	Often
160	3	Excellent	Often	Often	Native	Offline	Time	25	15	No	Chi / Jap	16-23	Kor	Student	Often
161	7	Good	Occas	Occas	Native	Online	Time	100	20	Seen	Chi	16-23	Kor	Student	Occas
162	0	Excellent	Often	Occas	Native	Online	Time	100	200	Purchased	No	24-31	Kor	Student	Occas
163	1	Good	Often	Often	Native	Offline	Time	1	1	No	No	16-23	Kor	Student	Often
164	8	Good	Often	Often	Native	Offline	Time	5	10	Heard	Jap	24-31	Kor	Student	Occas
165	0	Poor	Never	Never	Native	Online	Sent	50	13	No	German	24-31	Kor	Student	Never
166	0	Average	Occas	Never	English	Offline	Time	5	3	Heard	Chi	24-31	Kor	Student	Often
167	8	Average	Often	Often	Both	Offline	Time	2.5	1.5	Purchased	French	24-31	Kor	Student	Always
168	5	Good	Occas	Occas	English	Offline	Time	10	4	Seen	No	24-31	Kor	Student	Occas
169	3	Average	Often	Occas	English	Offline	Time	7	4	Seen	Jap	24-32	Kor	Student	Often
170	2	Good	Occas	Occas	Native	Offline	Time	8	4	Seen	Chi	24-33	Kor	Student	Never
171	8	Good	Occas	Often	English	Online	Sent	100	20	No	Chi / Jap / Spa	24-31	Kor	Student	Occas
172	9	Average	Never	Occas	Both	Offline	Time	8	5	Heard	No	16-23	Kor	Student	Often
173	5	Average	Occas	Occas	English	Offline	Time	20	10	No	Jap	24-34	Kor	Student	Occas
174	6	Good	Often	Occas	English	Online	Time	25	15	No	No	24-35	Kor	Student	Often
175	3	Good	Occas	Occas	English	Online	Time	270	50	No	No	24-31	Kor	Student	Occas
176	8	Good	Often	Often	English	Online	Time	10	5	Heard	Chi / Jap	16-23	Kor	Student	Occas
177	8	Excellent	Occas	Often	English	Online	Time	0	0	Purchased	No	24-31	Kor	Student	Often
178	10	Excellent	Often	Occas	Native	Online	Time	50	10	Seen	No	24-31	Kor	Student	Never
179	0	Good	Occas	Occas	English	Online	Time	10	2	No	Chi / Jap	24-31	Kor	Student	Often

ID	LVL	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
180	3	Good	Occas	Occas	English	Offline	Time	0	0	No	No	24-31	Kor	Student	Occas
181	8	Excellent	Occas	Often	English	Offline	Time	25	8	Heard	No	16-23	Kor	Student	Occas
182	8	Good	Occas	Occas	Native	Offline	Time	20	20	No	Jap	16-23	Kor	Student	Never
183	8	Excellent	Often	Often	Native	Online	Sent	20	8	No	No	16-23	Kor	Student	Occas
184	8	Excellent	Occas	Often	Both	Offline	Time	20	10	Heard	Chi / Jap / Spa	24-31	Kor	Student	Occas
185	9	Good	Often	Often	English	Online	Time	20	20	Seen	Chi	24-31	Kor	Student	Often
186	0	Good	Often	Occas	English	Offline	Time	5	5	No	Chi / Fre	24-31	Kor	Student	Occas
187	4	Good	Occas	Occas	English	Offline	Time	0	0	No	Chi	24-31	Kor	Student	Occas
188	5	Average	Occas	Occas	Both	Offline	Sent	10	20	Heard	Chi	24-31	Kor	PostGrad.	Often
189	9	Good	Never	Occas	Native	Offline	Time	1	1	No	Jap	24-31	Kor	Student	Occas
190	10	Good	Occas	Occas	Both	Offline	Time	10	4	No	Spa	24-31	Kor	Student	Occas
191	7	Good	Often	Often	Both	Offline	Time	10	10	Seen	Chi	24-31	Kor	Student	Occas
192	10	Good	Occas	Often	Native	Online	Sent	5	5	No	No	34-40	Kor	Sch. Dir.	Often
193	6	Good	Often	Occas	English	Offline	Time	10	5	No	Fre / Spa	26-33	Kor	Secretary	Often
194	18	Excellent	Often	Often	English	Offline	Time	2.5	2.5	No	Fre / Jap	16-23	Eng	Admin.	Always
195	18	Excellent	Occas	Occas	English	Online	Time	5	5	No	No	24-31	Eng	Teacher	Always
196	18	Good	Occas	Often	English	Online	Time	5	4	No	Chi / Jap	16-23	Eng	Rsrh Asst.	Always
197	12	Excellent	Occas	Occas	English	Online	Time	40	40	No	Fre / Spa	24-31	Eng	Teacher	Always
198	18	Good	Occas	Occas	English	Online	Time	5	2	No	French	16-23	Eng	Teacher	Always
199	20	Excellent	Often	Occas	Both	Online	Time	5	15	No	No	24-31	Eng	Teacher	Always
200	20	Good	Occas	Often	Both	Online	Time	30	50	No	Chi	24-31	Eng	Teacher	Always

7.3 APPENDIX C

7.3.1 Abbreviations

BLEU – Bilingual Evaluation Understudy

CBMT – Corpus Based Machine Translation

CR – Casting Radius

EBMT – Example Based Machine Translation

FE – Fluency Enhancement

ILM – Ideal Language Model

LDC – Linguistic Data Consortium

LSS – Link Structure Score

MEMT – Example Based Machine Translation

MLB – Map Level Base

MIT – Massachusetts Institute of Technology

MT – Machine Translation

NIST – National Institute of Standards and Technology

NOS – N-Gram Order Span

OOP – Object Orientated Programming

PR – Prioritization Regime

RBMT – Rule Based Machine Translation

RLM – Realistic Language Model

SHMT – Sequential Hybrid Machine Translation

SIR – Structural Integrity Ratio

SMT – Statistical Machine Translation

WFM – Word Frequency Map

WFML – Word Frequency Map Level

7.3.2 Technical Terms

Corpus/Corpora – A large collection of writings of a specific kind or on a specific subject

N-Gram – a sequence of words, with each token representing a gram. The amount of grams is the equivalent to how many tokens are in the word sequence, for example unigram, bigram etc

Semantic – of, pertaining to, or arising from the different meanings of words or other symbols

Syntactic – of, pertaining to, or arising from the grammatical rules of language

Web Crawler – software that searches the web and caches the data of web pages

7.3.3 References

1. As Easy As Pie. *The Phrase Finder.* [Online] The Phrase Finder, 2009. [Cited: March 7th, 2009.] http://www.phrases.org.uk/meanings/as-easy-as-pie.html.

2. **Ward, Andrew.** Victor, the Wild Boy of Aveyron. *Feral Children.* [Online] Feral Children. [Cited: February 26th, 2009.] http://www.feralchildren.com/en/showchild.php?ch=victor.

3. **University, Bristol.** Language and Gesture. *Bristol University News.* [Online] Bristol University, September 12th, 2003. [Cited: November 9th, 2008.] http://www.bris.ac.uk/news/2003/256.

4. **Linguistic Data Consortium.** LDC Catalog. *Linguistic Data Consortium.* [Online] September 19th, 2006. [Cited: March 5th, 2008.] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

5. *Translating from under-resourced languages: Comparing direct transfer against pivot translation.* **Babych, Bogdan, Hartley, Anthony and Sharoff, Serge.** Copenhagen : University of Leeds, 2007. MT Summit XI. pp. 29-35.

6. **Carbonell, Jamie G. and Brown, Ralf D.** Generalized Example-Based Machine Translation. *School of Computer Science, Carnegie Mellon.* [Online] November 29th, 2004. [Cited: January 22nd, 2009.] http://www.cs.cmu.edu/~ralf/ebmt/general.html.

7. **Knight, Kevin and Koehn, Philipp.** What's New In Statistical Machine Translation - Tutorial from Information Sciences Institute at the University of Southern California. *Statistical Machine Translation.* [Online] 2003. [Cited: June 28th, 2008.] http://people.csail.mit.edu/koehn/publications/tutorial2003.pdf.

8. *Rule-Based Translation With Statistical Phrase-Based Post-Editing.* **Simard, Michel, et al.** Prague : National Research Council Canada, 2007. Association for Computational Linguistics 2007. pp. 203-206.

9. *Multi-Engine Machine Translation with an Open-Source Decoder for Statistical Machine Translation.* **Chen, Yu, et al.** Prague : Saarland University / Deutsche Forschungszentrum für Künstliche Intelligenz (German Research Center for Artificial Intelligence) / Cologne University, 2007. Association for Computational Linguistics. pp. 193-196.

10. **Papineni, Kishore, et al.** *BLEU: A Method for Automatic Evaluation of Machine Translation.* New York : IBM Research Division, 2001.

11. **Zhang, Ying (Joy) and Vogel, Stephan.** Measuring Confidence Intervals for MT Evaluation Metrics. *Carnegie Mellon University.* [Online] October 2004. [Cited: July 28th, 2009.] projectile.sv.cmu.edu/research/public/talks/mt_workshop2004.ppt.

12. *Evaluation of Machine Translation and its Evaluation.* **Turian, Joseph P., Shen, Luke and Melamed, I. Dan.** New Orleans : Proceedings of the MT Summit IX, 2003.

13. **Peters, Susanne.** Systran. *Universität Basel.* [Online] November 9th, 2001. [Cited: March 3rd, 2008.] http://pages.unibas.ch/LIlab/staff/tenhacken/Applied-CL/3_Systran/3_Systran.html#history.

14. **Google.** Google Translate FAQ. *Google.* [Online] [Cited: March 6th, 2008.] http://www.google.com/help/faq_translation.html.

15. **Hutchins, John.** Machine Translation: A Concise History. [book auth.] Chan Sin Wai. *Computer aided translation: Theory and practice.* Hong Kong : Chinese University of Hong Kong, 2007.

16. *Interactive Translation of Japanese to Korean.* **Manion, Steve Lawrence, Punchihewa, Amal and De Silva, Liyanage.** Colombo, Sri Lanka : International Conference on Information and Automation, 2006.

17. **Gerber, Laurie.** Machine Translation: Ingredients for Productive and Stable MT Deployments. *Translation Directory.* [Online] January 2009. [Cited: February 12th, 2009.] http://www.translationdirectory.com/articles/article1945.php.

18. **M, Robert.** Treebank Tokenization. *University of Pennsylvania.* [Online] October 24th, 1997. [Cited: March 5th, 2009.] http://www.cis.upenn.edu/~treebank/tokenization.html.

19. **Banko, Michele and Brill, Eric.** *Scaling to Very Very Large Corpora for Natural Language Disambiguation.* Redmond : Microsoft Research, 2001.

20. **Technology, National Institute of Standards and.** NIST Open MT08 OFFICIAL Results. *National Institute of Standards and Technology.* [Online] June 6th, 2008. [Cited: June 26th, 2008.] http://www.nist.gov/speech/tests/mt/2008/doc/mt08_official_results_v0.html.

21. **IBM Press.** IBM Archives: 701 Translator. *IBM Website.* [Online] January 8th, 1954. [Cited: February 16th, 2008.] http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html.

22. **Meaningful Machines.** Meaningful Machines Technologies. *Meaningful Machines.* [Online] [Cited: February 26th, 2008.] http://www.meaningful.com/technologies.htm.

23. **Norvig, Peter.** Google Developers Day US: Peter Norvig Seminar. *Youtube.* [Online] June 5th, 2007. [Cited: March 8th, 2008.] http://www.youtube.com/watch?v=nU8DcBF-qo4.

24. *Analysis of the New Zealand English and Maori Online Translator.* **Laws, Mark, Kilgour, Richard and Watts, Michael.** Otago : Otago University, 2000.

25. *Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System?* **Zhang, Ying, Vogel, Stephan and Waibel, Alex.** Lisbon (Portugal) : Proceedings of LREC, 2004.

26. *Spectral Clustering for Example Based Machine Translation.* **Gangadharaiah, Rashmi, Brown, Ralf and Carbonell, Jaime.** New York : Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 41–44, 2006.

27. *Hybrid Machine Translation Applied to Media Montoring.* **Sawaf, Hassan, Gaskill, Braddock and Veronis, Michael.** McLean : AppTek Inc., 2008.

28. *Retrospect and prospect in computer-based translation.* **Hutchins, John.** Singapore : Proceedings of MT Summit VII, 1999.

29. *MIT helps create image-recognition software.* **Chandler, David.** 27, Boston : TechTalk - Serving The MIT Community, 2008, Vol. 52.

30. *Statistical Machine Translation: Foundations and Recent Advances.* **Och, Franz Josef.** Phuket : MT Summit X, 2005. MT Summit. pp. 1-72.