Sheridan College

SOURCE: Sheridan Institutional Repository

Student Theses

Honours Bachelor of Computer Science (Mobile Computing)

Fall 12-2-2020

An investigation of Grammar Gender-bias Correction for Google Translate When Translating from English to French

Ahmed Samy Merah
Sheridan College, meraha@sheridancollege.ca

An investigation of Grammar Gender-bias Correction for Google Translate When Translating from English to French

A Thesis

presented to

School of Applied Computing, Faculty of Applied Science and Technology

of

Sheridan College, Institute of Technology and Advanced Learning

by

Ahmed Samy Merah

in partial fulfilment of the requirements

for the degree of

Honours Bachelor of Computer Science (Mobile Computing)

December 2020

An investigation of Grammar Gender-bias Correction for Google Translate When Translating from English to French

by

Ahmed Samy Merah

Submitted to the School of Applied Computing, Faculty of Applied Science and Technology
on December 16, 2020, in partial fulfillment of the
requirements for the degree of
Honours Bachelor of Computer Science (Mobile Computing)

Abstract

This work investigated how to address the Google Translate's gender-bias when translating from English to French. The developed solution is called GT gender-bias corrector that was built based on combining natural language processing and machine learning methods. The natural language processing was used to analyze the original sentences and their translations grammatically identifying parts of speech. The parts of speech analysis facilitated the identification of three patterns that are associated with the gender bias of Google Translate when translating from English to French. The three patterns were labeled simple, intermediate and complex to reflect the structure complexity. Samples of texts that represent the three patterns were generated. The generated texts were used to build a decision-tree-based classifier to automatically detect the pattern to which a text belongs. The GT gender-bias corrector was tested using a survey completed by participants with diverse levels of English and French fluency. The survey analysis showed the success of the corrector in addressing the Google Translate gender-bias for the three patterns identified in this work.

Keywords:

Thesis Supervisor: Dr. El Sayed Mahmoud
Title: Professor, School of Applied Computing

Acknowledgments

This aknowledges Dr. El Sayed Mahmoud's efforts in making this work possible through his guidance.

Contents

1 Introduction **8**

1.1 Problem Context 8

1.2 Terms and Definitions 10

1.3 Problem Statement 11

1.4 Purpose 11

1.5 Motivation 11

1.6 Proposed Work 12

1.7 Thesis Statement 12

1.8 Contributions 13

1.9 Organizations of Thesis 13

2 Literature Review **14**

2.1 Machine Translation (MT) 14

2.1.1 Why high quality Machine Translation is important 14

2.1.2 Statistical Machine Translation (SMT) 17

2.1.3 Rules Based Machine Translation (RBMT) 17

2.1.4 Hybrid Machine Translation (HMT) 17

2.1.5 Neural Machine Translation (NMT) 18

2.2 Translation of Grammatical Gender research 18

2.2.1 Importance of Grammatical Gender 18

2.2.2 Determiners/Articles 19

2.2.3 Nouns 19

2.2.4 Adjectives 19

2.3 Metrics . . . 20
2.3.1 BLEU . . . 20
2.3.2 NIST . . . 20
2.3.3 METEOR . . . 20

3 Methodology **21**
3.1 Proposed System . . . 21
3.1.1 Overview of system . . . 24
3.1.2 Noun swapping . . . 24
3.1.3 Determiner swapping . . . 25
3.1.4 Adjective swapping . . . 25
3.1.5 Machine learning model . . . 25
3.1.6 Example for applying the corrector . . . 26
3.1.7 Determining word endings . . . 27
3.2 Testing Strategy . . . 27
3.2.1 Data . . . 28
3.2.2 Survey . . . 28
3.2.3 Automated tools . . . 32
3.3 Feasibility of the solution . . . 32
3.4 Trials . . . 32
3.5 Language ambiguity . . . 33

4 Findings **34**
4.1 The GT gender bias corrector perfectly address the Google Translate gender bias of simple and Intermediate cases . . . 35
4.2 The GT gender-bias corrector moderately fixes the Google Translate' gender bias of complex cases . . . 40
4.3 Analysis . . . 42

5 Conclusion **43**
5.1 Future work . . . 43

5.1.1 Extending the GT gender bias corrector to other Romance languages . . . 43
5.1.2 Benefits for Google Translate . . . 44
5.1.3 Expanded survey . . . 44
5.1.4 Using more complex machine learning algorithms to improve the performance . . . 44
5.2 Limitations . . . 45
5.2.1 The system can process two sentences only . . . 45
5.2.2 Reverse lemma instead of manually getting word endings . . . 45

Bibliography **46**

List of Figures

1-1 Speakers of different languages present at Google I/O Challenge . . . 9

3-1 General view of the system 22
3-2 Detailed system view 23
3-3 Translation issue example 26
3-4 Group 1: Simple case 29
3-5 Group 2: Intermediate case 30
3-6 Group 3: Complex case 31

4-1 Simple cases responses Google Translate 35
4-2 Simple cases responses GT gender-bias corrector 36
4-3 Intermediate cases responses Google Translate 37
4-4 Intermediate cases responses GT gender-bias corrector 39
4-5 Complex cases responses Google Translate 40
4-6 Complex cases responses GT gender-bias corrector 41

Chapter 1

Introduction

1.1 Problem Context

When French or Romance language speakers with limited English skills need help to correctly communicate their ideas in English, Google Translate seems to be the go-to tool that these speakers flood towards. This has been shown after a 10 year review of Google's Translate success in 2016 where it was found that 92% of Google Translate users were from outside the United States indicating that many users are not native English speakers [1]. This shows that many individuals rely on Google Translate's algorithm for their communication. The main reason for this reliance on Google Translate is its accessibility to users as Google is available worldwide and does not require the translation service to be downloaded or purchased.

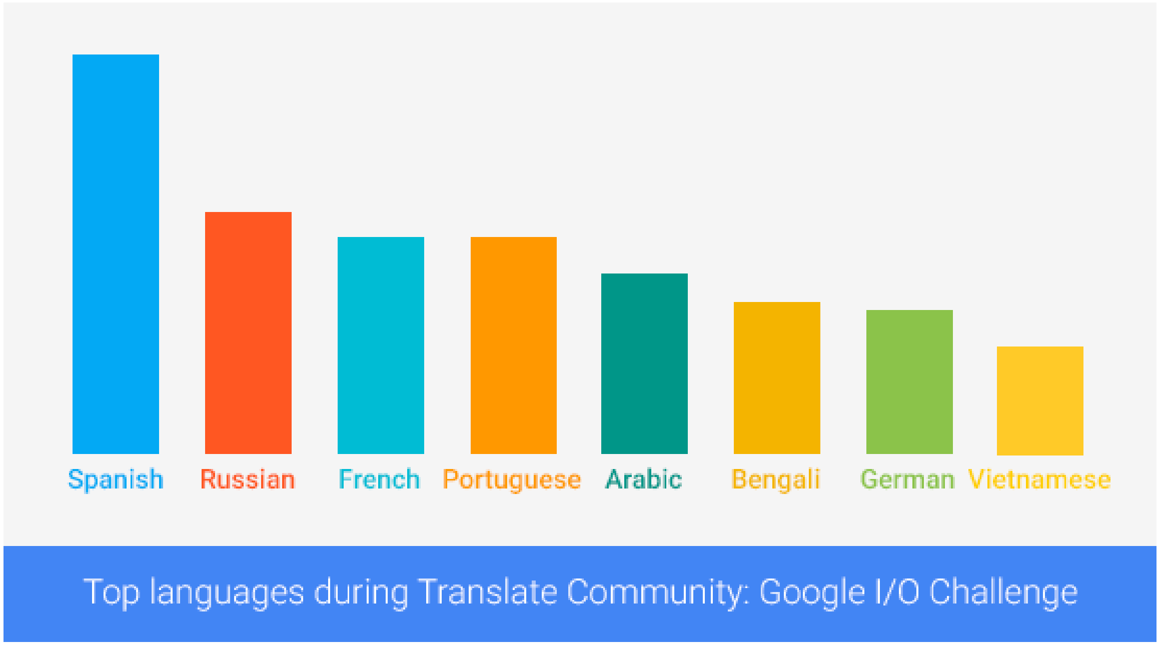


Figure 1-1: Speakers of different languages present at Google I/O Challenge

This graph illustrates the native language of participants at Google I/O Translate Community Challenge. We can see a staggering number of participants speak languages that are part of the Romance languages which is very representative of Google Translate's user base [2]. This large user base often relies on Google Translate to communicate a message to a native or non-native English speaker. The resulting translation can often be distorted or incoherent compared to the meaning of the text in the original language. This incoherence is evident when translating between French and English because French includes a large number of idioms and employs various rules that have no corresponding rules in the English language such as an acute emphasis on the masculine and feminine gender. Idioms, sentiment and grammatical gender are essential to understand the meaning of a French sentence. This work focused on building an add-on system to identify and address translation shortcomings caused by Google Translate's grammatical gender bias.

1.2 Terms and Definitions

1. Google Translate: Software used for translation of many languages developed by Google.
2. Grammatical gender: Class system, composed of two or three classes, whose nouns that have human male and female referents tend to be in separate classes.
3. Romance languages: Languages that originated from Latin such French, Spanish, Italian and Portugese.
4. Natural Language Processing: Field that concerns itself with how machines interpret and use human language.
5. Python: Commonly used programming language.
6. Spacy: Python library for Natural Language Processing.
7. Decision tree: Predictive modeling approach used in Machine learning.
8. Determiner / Article: Word that determines the references of a noun or nouns groups.
9. Noun: Word other than a pronoun that identifies a class of people, places or things.
10. Common noun: Noun used to refer to a class of entities such as a profession (Nurse).
11. Proper noun: Noun used to refer to a single entity such as city names (Toronto).
12. Verb: Words that indicate actions, occurrences and state of being.
13. Adjective: Words that indicate qualities or states of nouns.
14. Accent: Alternative spelling of a letter that can indicate a different sound or a distinction.

1.3 Problem Statement

It is difficult for French speakers with limited English skills and vice versa to communicate without using automatic translation services. However, automatic translation tools often distort the meaning when translating from English to French. This often leads to confusion and false information, especially when it comes to grammatical gender as it affects determiners, articles, nouns and adjectives which can result in a very awkward translation. This work developed an add-on system to Google Translate that addresses the gender issue. The system is called GT gender-bias corrector.

1.4 Purpose

The purpose of this thesis was to develop a new add-on system to identify and address misrepresented grammatical gender in translations from English to French using Google translate. It aims to facilitate translation from English to French so that the meaning and context of the original text is not altered during the translation process. This would help in maintaining the meaning clear to the people involved in communications that rely on such translation. The system will promote communication between English and French speakers which will have positive impacts on trades and social activities.

1.5 Motivation

The commercial and social positive impacts of improving the English-French or other Romance language translation quality is the main motivation of this research. The progress of various Natural Language Processing techniques is another motivation because this improves the feasibility of addressing these translation issues. A third motivation is the ability to extend the system to other Romance languages such as Spanish, Portuguese and Italian. Currently, English is well established as the universal language in business and technology and requires individuals in those fields to be able to have a certain proficiency in the language. By creating a system that can help people supplement their language skills in order to make themselves more clear in business or in colloquial conversations, we can boost productivity and avoid misguided decision making due to low quality translations. As translation blunders can have an enormous impact on a business such as with HSBC's 2009 slogan translation fiasco where millions of dollars needed to be spent on marketing to fix a simple mistake that could have easily been avoided [3]. This research aligns with the current National Research Council of Canada correcting machine translation for the Chinese language family using a software known as YiSi [4].

1.6 Proposed Work

This work consists of the development of an add-on system to Google translate including the design of an experiment to evaluate it. The system consists of an algorithm that combines NLP techniques and a decision tree model. The NLP algorithm analyzes the text to be translated into tokens identifying the parts of speech. The tokens are entered to the decision tree model to detect the text structure complexity. Determining the structure complexity facilitates identifying how to address the gender bias because the detected structure implies the gender of the subject and the objects with their associated parts of speech. The system is a gender-bias corrector that ensures that gender is correctly translated from English to French. The importance of the system is the fact that French makes major distinctions between masculine and feminine which can change the meaning and semantics of linguistic features. This research includes human participants who are shown various translated texts in order to evaluate the ability of the system to address the gender bias.

1.7 Thesis Statement

An add-on system can be developed to maintain the integrity of grammatical gender when translating from English to French using Google Translate. The system combines NLP and a decision tree model to analyze the text and identify the changes needed to address the gender-bias. The text analysis includes identifying the parts of speech to show the structure and the gender of the subjects and objects with their associated parts of speech. The decision tree model detects the structure pattern to which a text belongs which in turn facilitates identifying the required changes to address the gender bias.

1.8 Contributions

This work shows us how to use gender correction for the goal of improving translation quality between English and French. The contribution of this work is:

1. Developing the GT gender-bias corrector.

2. Developing a testing strategy that is extensive enough to determine translation quality.

3. Create guidelines to expand findings to other Romance languages

1.9 Organizations of Thesis

The rest of this thesis includes a literature review, a methodology chapter, findings and a conclusion. The literature review focuses on the prior research made in the field of Machine Translation, Natural Language Processing and their use cases. The literature examines the different paradigms of Machine Translation and the challenges faced when using these methods when it comes to grammatical gender. The methodology describes in details the methods that are applied in this work. This includes the selection of the right data and test scenarios and building the GT gender-bias corrector. The methodology also includes testing performance through the use of human participants with surveys and automated test tools metrics. The findings present the results of the survey and automated testing metrics in determining if the GT gender-bias corrector was successful. The conclusion chapter concludes this work and present the future work and limitations of this research.

Chapter 2

Literature Review

The heavy use and reliance on Google Translate worldwide for communication has been a growing phenomenon in various fields, such as education, business and health-care. Many users report inconsistent context of translated messages compared to the original ones, especially when translating to French and Romance languages [5, 6]. The extensive use of Google Translate has even been observed in emergency health-care contexts where patients have had issues getting their message across and where staff members rely on Google Translate due to non-availability and cost of a human translator[7]. This has prompted many domain experts to investigate the accuracy of translation methods to the English language.

This work investigates the sources of inconsistencies in machine translation methods and identifies a solution to address them. This chapter reviews Machine Translation methods, grammatical gender rules in English and French and evaluation metrics for the translation quality.

2.1 Machine Translation (MT)

2.1.1 Why high quality Machine Translation is important

Machine Translation and its application such as Google Translate have been at the core of general communication in many domains for various purposes such as Higher

Education, Healthcare and Business [8, 7, 9]. The advantages of ameliorated Machine Translation are numerous. A research from the European Parliament in 2016 revealed that communication in common languages whether official or spoken leads to an increase of 44% in trade flows [10]. In order to take advantage of this economic increase that transcends borders and cultural differences, a robust Machine Translation system must be implemented in order to get a high quality translation where grammatical gender is persevered for impactful decision making [10]. Such system would remove the language barrier many businesses have when it comes to expansion in emerging markets as business operations are increasingly becoming global.

The importance of high quality Machine Translation is not only confined to the world of business but is also a concern in education settings [8, 11, 12]. As numerous research papers and articles are written in different languages, it is important to have a reliable Machine Translation system that does not alter the research and the message it is trying to convey. According to Klaus Mundt and Michael Groves, Google Translate has been extensively used by the student body and educators alike to submit work and teach respectively [13]. The major concern of Google Translate use in higher education is transformation of content and use of the tool as a crutch for lack of writing skills [12, 13]. In order for students or educators to properly learn and extract insights from material written in a different language: grammatical gender must be preserved. As current Machine Translation often distorts meaning when it comes to intricate text, this is of utmost importance.

Machine Translation has also been extensively used in healthcare and medicine [7, 14]. As medicine is a very critical subject where miscommunication can have effects of a catastrophically high magnitude, good communication is key to avoid malpractice. According to CRICO Strategies, medical miscommunication was the cause of 30% of medical malpractice cases in the United States between 2009 and 2013 with 24% of those cases leading to death [15]. These miscommunication cases incurred $1.7 billion in costs [15]. Google Translate has proven itself as a semi effective tool for very simple medical translation that medical practitioners should not trust for critical medical communication [14, 16]. Improving Machine Translation in this domain would allow various medical practitioners worldwide to communicate diagnosis with ease to patients in their mother tongue without the worry of malpractice for the healthcare provider. This would also remove the often cumbersome process of having to call translators and deal with scheduling issues especially during emergency situations [7].

Machine Translation also has shortcomings when it comes to colloquialism and common dialects [17]. There is current research being done on Arabic dialect handling using Hybrid Machine Translation which normalizes dialect into standard language as dialects and colloquialism do not follow the same structure as written standard language [18, 17]. As individuals tend to express themselves in colloquial terms for sake of simplicity, it is important that a Machine Translation system can intercept and interpret such terms without giving an erroneous translation in return. Such examples of colloquialism include Verlan which is a common argot in the French language where partial backwards spelling of a word can be used to identify the word itself [19]. A common example of Verlan is the word for "Woman" which in French is "Femme", Verlan uses partial backwards spelling and turns the word into "Meuf" [20]. A Machine Translation system that can adapt to colloquialism and dialects can be very useful in the domain of tourism as the industry often deals with a lot of colloquial terms when performing domain specific tasks such as tours, recommendations and talking to locals.

The purpose of Machine Translation (MT) research is to bridge gaps between various languages for different use cases in a plethora of industries [21]. Many methods have been developed in the field of computational linguistics to produce machine translation [22]. These methods are based on four paradigms: Statistical Machine Translation (SMT), Rules Based Machine Translation (RBMT), Hybrid Machine Translation (HMT) and Neural Machine Translation (NMT) [23, 24, 25, 26]. The category of Hybrid Machine Translation (HMT) between French and English using grammatical gender rules as tools for improved translation will be the focus of this research.

2.1.2 Statistical Machine Translation (SMT)

Statistical Machine Translation is a paradigm of Machine Translation where translation is derived from statistical models that use large text corpuses of different languages [23]. The main benefit of Statistical Machine Translation is the efficiency of the resources used in carrying out the translation as it is not dependant on specific languages [23]. The main drawbacks of the method are the lack of corpus available to create significant statistical models and limited fluency that does not always interpret linguistic rules between languages appropriately [23].

2.1.3 Rules Based Machine Translation (RBMT)

Rules Based Machine Translation is another paradigm of Machine Translation that encompasses other Machine Translation paradigms such as transfer-based, dictionary based and interlingual machine translation [27]. Rules Based Machine Translation uses syntactic and grammatical rules of both languages it intends to translate in order to perform semantic analysis [23]. The advantages of RBMT include: No need for large corpus of texts, each error can be addressed with a new rule, simple to refine as rules are hand written and modular depending on the language [28]. There are certain drawbacks to this method that include: Domain adaptability as new lexical rules have to be written for certain domains which can be often costly and cost of dictionary building [23].

2.1.4 Hybrid Machine Translation (HMT)

Hybrid Machine Translation is a paradigm of Machine Translation that blends various other Machine Translation paradigms, most commonly SMT, RBMT and recently NMT [29]. HMT attempts to use the best techniques of each paradigm in order to get close to a perfect translation [18]. A common way to use HMT is one that Dr.

Hassan Sawaf developed by blending SMT and RBMT into the same process [25]. The most common approach is to use a Rule-Based engine for any preprocessing and applying statistical analysis after in order to reduce resource usage and get a more accurate output [18]. This also helps reduce the labor associated with RBMT as after the rules are in place statistical methods take over and do the grunt work which reduces human error [30].

2.1.5 Neural Machine Translation (NMT)

Neural Machine Translation is the most recent paradigm of Machine Translation that has been developed. NMT uses neural networks to carry out translation tasks [31]. It does not use sub components as the model gets trained on various texts and derives its own conclusions, from this definition NMT is considered an end to end solution [32].

2.2 Translation of Grammatical Gender research

This section presents the relevant research made about grammatical gender and highlights the gap in this research area

2.2.1 Importance of Grammatical Gender

Grammatical gender is a very important component of the French language and fluency in grammatical gender rules is necessary for an authentic translation [33]. As English lacks grammatical gender in its grammatical rules [34], this issue may not seem of critical importance, but it has been shown that mishandling of grammatical gender in translation can compromise the integrity of the sentence [35]. Such examples are of the sentence “I am happy”, which in French depending on the gender can be “Je suis hereux” for masculine and “Je suis heureuse” for feminine [35]. Eva Vanmassenhove from Cornell University has proven that using a gender feature for Machine Translation which uses language pairs improves quality of translation for languages that heavily use grammatical gender which tend to be Romance languages [35]. Gabriel Stanovsky attempted to use gender neutral words in English such as "doctor" and "nurse" in order to evaluate bias in Machine Translation systems in eight languages [36]. The research concluded that the 6 Machine Translation systems tested were prone to bias error in all eight languages [36].

Grammatical gender affects different parts of speech in a sentence such as determiners/articles, nouns and adjectives. French and Romance languages rely on Masculine and Feminine (No neuter) for word endings.

2.2.2 Determiners/Articles

Determiners and articles are words used to indicate if a following noun is singular/-plural and masculine/feminine [37]. Current research is currently focused on the Japanese language's determiners system, Francis Bond built three algorithms that perform text analysis using referential Japanese phrases to determine the appropriate article when translating to Japanese [38].

2.2.3 Nouns

Nouns are words that designate a class of beings, things, places or particular unique parts of the class. In French or Romance languages, each noun has a gender [39]. Some research has been done for Romanian which is an eastern Romance language. Silviu Cucerzan and David Yarowsky from John Hokins University used noun seeding of common nouns that traditionally lean towards one gender in order to model suffixes for translation through context. [40].

2.2.4 Adjectives

Adjectives are words that are used after nouns and verbs to qualify or describe a subject [41]. Gabriel Stanovsky's research investigates biases in certain adjectives such as pretty and handsome which are commonly used for women and men respectively and evaluates bias when used with traditionally masculine and feminine nouns [36]. Such method reduced bias for Spanish, Russian, Ukrainian [36].

2.3 Metrics

Even though translation and natural language is often nuanced and up to interpretation, machine translation still has metrics that need to be attained in order to be considered high quality. The main subjective metric that should always be prioritized is the understanding of the individual on the receiving end of the translation, but due to the fact that we cannot survey every individual the metrics that can be used are BLEU, NIST and METEOR [42].

2.3.1 BLEU

Bilingual Evaluation Understudy Score is a metric to evaluate machine translations from a reference text. It compares the N grams from the reference text with the ones of the hypothesis. From there a score between 0% and 100% is given [43]. 100% indicating a perfect match and 0% indicating no match.

2.3.2 NIST

NIST is another method to evaluate machine translation which is based on BLEU but assigns weight to specific N grams depending on its rarity [44].

2.3.3 METEOR

METEOR is another machine translation metric, it is different from BLEU and NIST due to the fact that it uses stemming and synonymy. It uses precision and recall with a higher emphasis on recall [45].

Chapter 3

Methodology

This chapter presents the methods for building and evaluating the add-on system and its machine learning model in order to address the gender bias of google translate when translating from English to French. The chapter also presents the details of the testing strategy including testing data, test cases, survey structure and automated testing.

3.1 Proposed System

The add-on System identifies issues related to misinterpreted gender and fixes them. The system's algorithm analyzes the original text and its Google translation to identify how to alter the Google translation output in a way that corrects the translation of gender grammar to match the original text. The algorithm is a corrector because it aims to review the grammar related to the gender when translating from English to French and correct it properly using a machine learning model and dependency parsing that classifies the parts of speech and issues that need to be fixed. The algorithm identifies the gender in the French translation based on analyzing the text using parts of speech and a decision tree model that detects the complexity of identifying the gender then makes the appropriate changes in the translated text according to the requirement in order to respect the grammatical gender rules. The algorithm uses tokenization, vectors and dependency parsing through the Spacy library in addition to a customized decision tree model built using Weka to identify the class of changes. Figure 3-1 and Figure 3-2 shows the flowcharts of a general and a specific view of the system respectively.

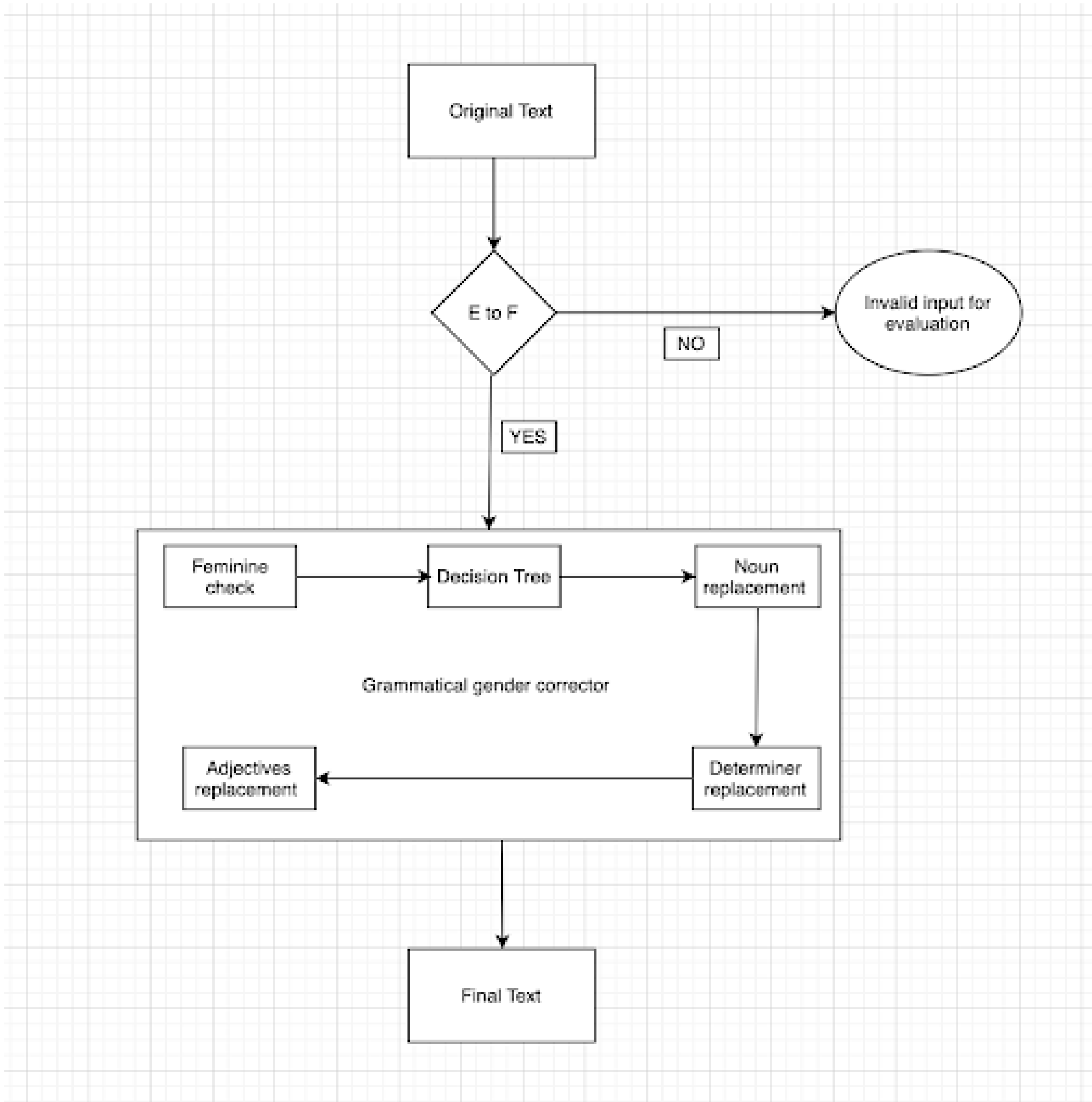


Figure 3-1: General view of the system

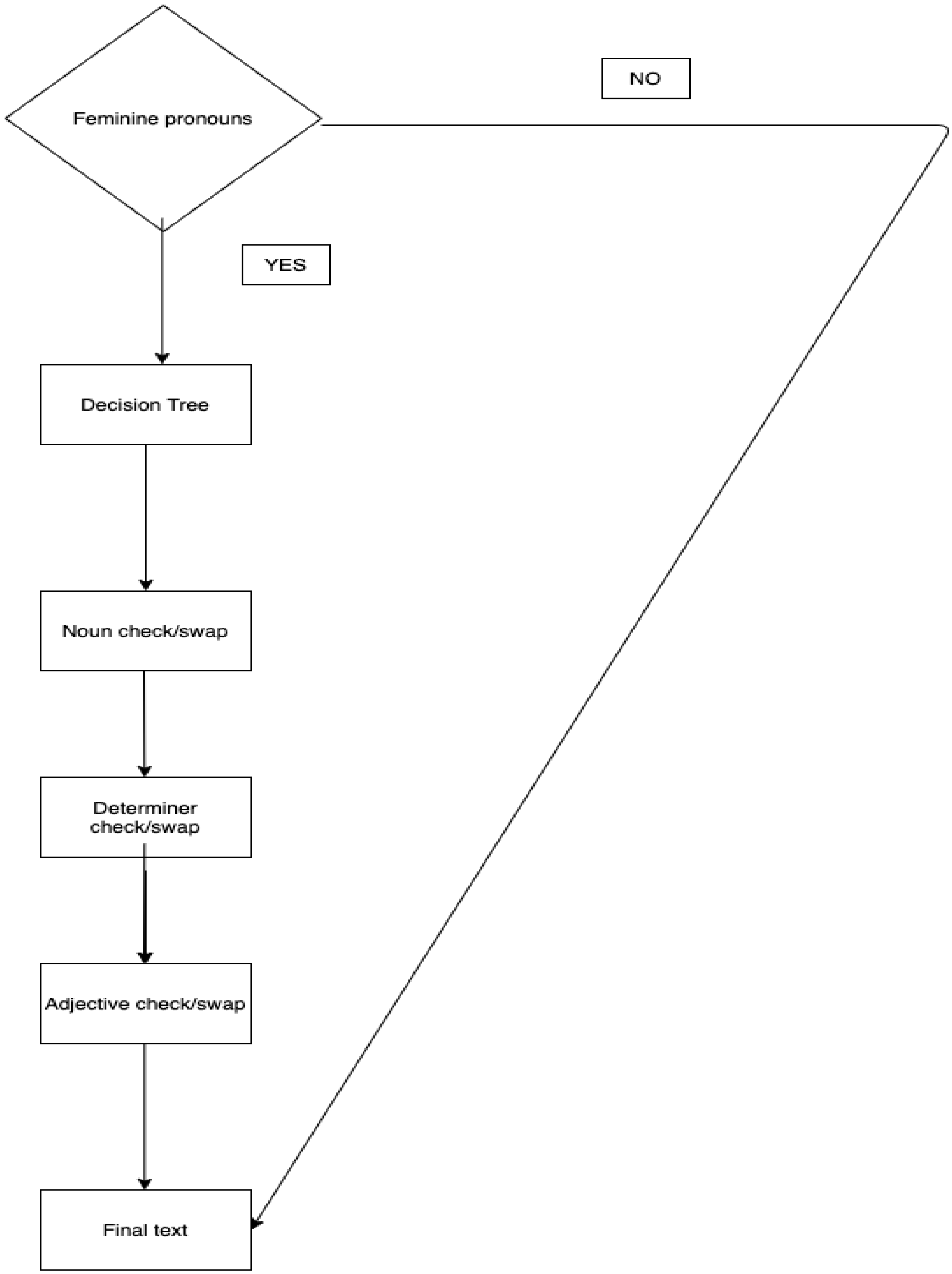


Figure 3-2: Detailed system view

3.1.1 Overview of system

In Figure 3-1, the flow of the system starts with receiving the original text as an input, the system checks the language of the text through the Google Translate API. The system accepts English original text only. The system uses the Google API to translate the text in French then forwards the original text and the french translation to the GT gender-bias corrector for further analysis to address gender related issues in the translation. Figure 3-2 shows a detailed view of the corrector, the first step is to analyze the text and its translation through a decision tree that identifies the changes that need to be made through the dependency parser. If a text has no feminine support words such as “she”, “elle” or “her” there is no feminine subject and therefore no modifications need to be made as Google Translate defaults to masculine translation. When the decision tree gives the result of the changes that need to be made, the parts of speech that are tagged will need to go through modification. If the subject is a masculine noun, the word ending will need to be modified which triggers the need to modify determiners and adjectives associated with that noun per grammar rules, this process is called swapping. The Spacy library uses vectors after swapping the tagged part of speech in order to determine if it is a valid word in the French language using its own model. After finishing all the required swaps, the final output text that respects grammar gender rules is delivered. More details of the components are in the following subsections.

3.1.2 Noun swapping

After text analysis, if there are no feminine support words, no modifications need to be made due to Google Translate’s bias to masculine. If feminine support words are found, nouns affected by the feminine support words are the first words to be modified due to the fact that they affect other groups of words such as determiners and adjectives. Gender of a noun can be accurately determined by the noun ending.

3.1.3 Determiner swapping

If a noun needs to be swapped to the feminine gender, the determiner needs to also change to reflect the noun gender. The only exception is for plural determiners as they remain the same for masculine and feminine (unisex).

3.1.4 Adjective swapping

If a noun is swapped and is followed by an adjective, the adjective has to reflect the noun gender. Just like nouns, gender can be determined by word endings.

3.1.5 Machine learning model

A decision tree needs to be built for this solution, its complexity is associated with the number of nouns and pronouns. A simple translation has only one noun in its first sentence which is by default the subject and what the feminine pronoun in the next sentence refers to. An intermediate case has 2 nouns which are the subject accompanied by an object or complement, the system needs to be able to know which of these nouns is the feminine pronoun referring to which requires training of the model to be able to distinguish the subject with the help of the Spacy dependency parser. Complex translations have at least 2 nouns and more than one pronoun which makes the task more complex in order to identify what parts of speech are affected by which and what needs to be modified. In order to fix a simple case, we only need to change the noun ending and the determiners and adjectives associated with it. For an intermediate case we need to distinguish the subject from the object in order to perform the proper modification. For the complex cases, we need to distinguish which pronouns affect which subject or object in order to translate the proper groups of nouns, determiners and adjectives.

The training is done using a CSV file with attributes that are put through Weka using a J48 decision tree to predict the class of changes needing to be done in a text. The French texts that are part of the model are French text that were put through Google Translate and came out erroneous. Some of the sentences are also English text containing the word "her" that does not have a direct counterpart in French but still needs to be treated for Grammar gender issues. The attributes of the model are as follows.

1. s: Subject
2. c: Object
3. ds: Determiner of the subject
4. dc: Determiner of the object
5. as: Adjective of the subject
6. ac: Adjective of the object
7. ps: Pronoun of the subject
8. pc: Pronoun of the object
9. class: Combination code of what needs to be changed

3.1.6 Example for applying the corrector

Figure 3-3: Translation issue example

As seen in the figure above, the word "boss" defaults to the French masculine "patron" as there is no common English word for "female boss" which in French is "patronne".

This example demonstrates an erroneous translation that needs to be modified in order to respect gender grammar as the noun, determiner and adjective all default to masculine when the subject is feminine when separated by a period.

3.1.7 Determining word endings

Through sanitization of the Google Translate text by replacing Hex values that represent accents in the French alphabet to Unicode characters, we can tokenize sentences and words. We can proceed to load the large French model from Spacy that allows us to get more information from the words through a doc object. This allows us to extract what is known as a tag and a dep, which has all the information about the words, such as word family, gender, number and dependency to other parts of speech. From this we can extract endings of targeted words and pass the necessary changes based on such. We then validate the word through vector checks. Below are common word endings with their gender counterpart.

1. Word endings
2. Masc: ”-on” — Fem: ”-onne”
3. Masc: ”-in” — Fem: ”-ine”
4. Masc: ”-eur” — Fem: ”-euse”
5. Masc: ”-ien” — Fem: ”-ienne”
6. Masc: ”-ou” — Fem: ”-olle”
7. Masc: ”-x” — Fem: ”-se”
8. Masc: ”-c” — Fem: ”-che”

3.2 Testing Strategy

The testing strategy involves evaluating the performance of Google Translate and the add-on translation correction system based on three scenarios. The scenarios include:

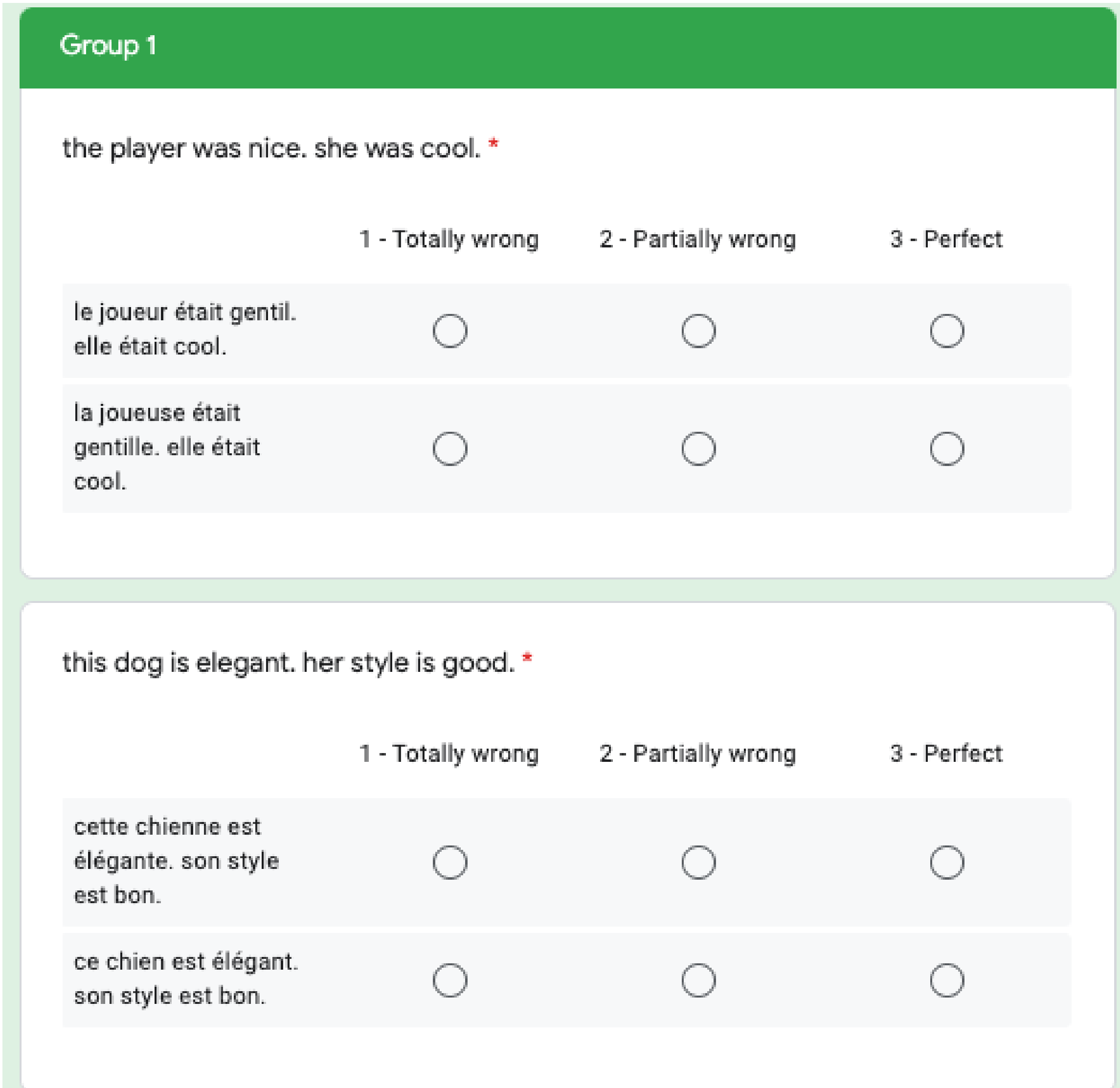
(1) Simple cases, (2) Intermediate cases and (3) Complex cases. This work collects four samples per each scenario to recruit a reasonable base to evaluate the quality of the translations. As English and French are natural languages, we cannot solely rely on automatic analysis and evaluation algorithms. The testing of the solution relies on human participants from diverse backgrounds through surveys. Each reader completes a survey after reading the original sentence in English and rates the Google translate output and the add-on system output. The survey questions collect information that identify and quantify the improvement in translating grammatical gender and the general readability (clarity) of the translated text.

3.2.1 Data

The data that is used for this work is gathered from Larousse and Oxford dictionaries, the two standard dictionaries for French and English respectively. Articles from French newspaper Le Monde were also used in order to generate more cases.

3.2.2 Survey

The survey consists of three sections where each section collects information about the gender bias of one of the three gender-bias patterns identified in this work. Each section includes four diverse questions to test the correctness of our system translation. Each question presents the original sentence and the two translations (without identifying the translator) and asks the participant to select the most correct one (the one with less errors in translating the gender). The questions are shuffled in a certain way so that participants cannot see a pattern. Below are some screenshots showing highlights of the survey.



Group 1

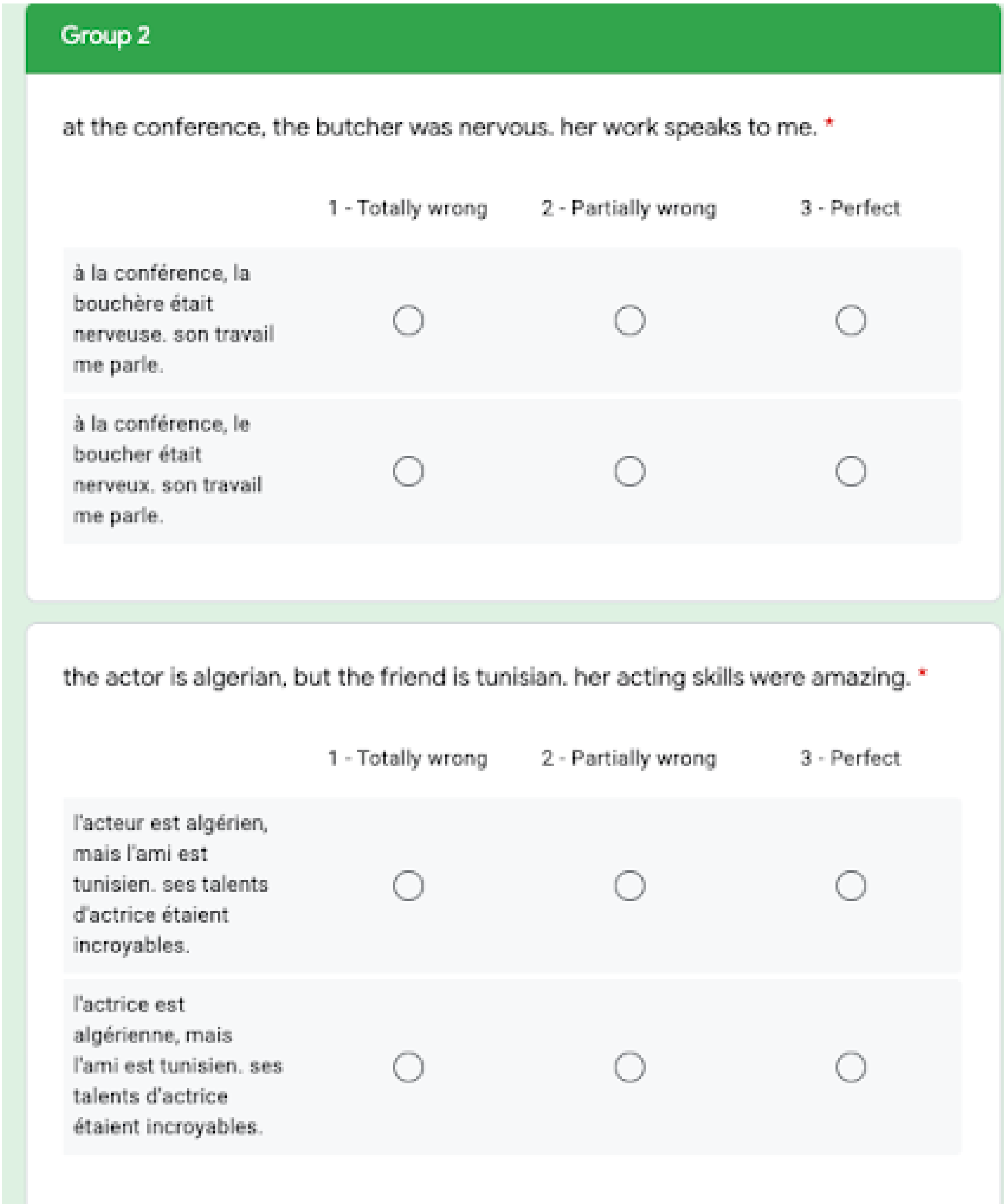
the player was nice. she was cool. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
le joueur était gentil. elle était cool.	○	○	○
la joueuse était gentille. elle était cool.	○	○	○

this dog is elegant. her style is good. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
cette chienne est élégante. son style est bon.	○	○	○
ce chien est élégant. son style est bon.	○	○	○

Figure 3-4: Group 1: Simple case



Group 2

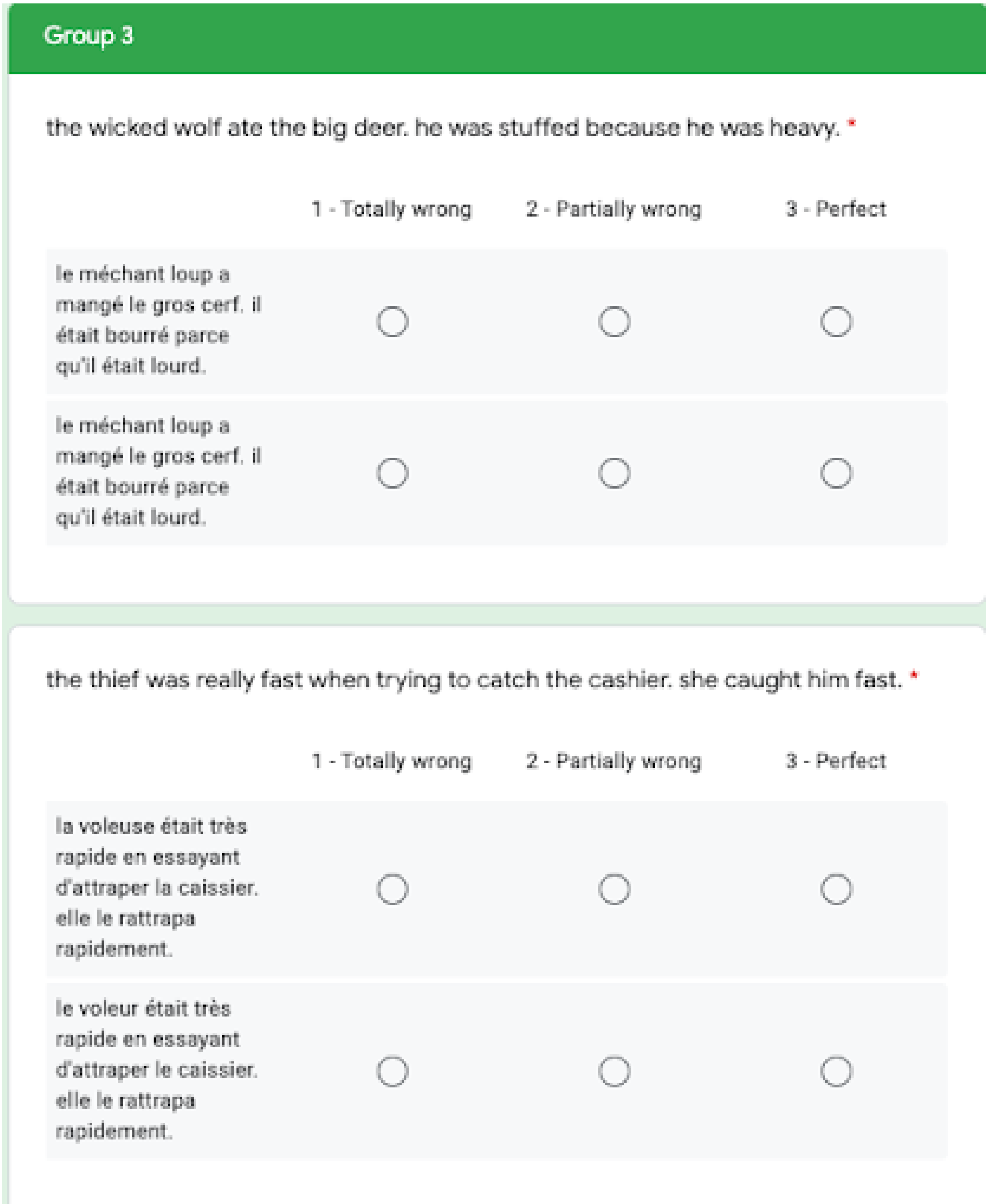
at the conference, the butcher was nervous. her work speaks to me. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
à la conférence, la bouchère était nerveuse. son travail me parle.	○	○	○
à la conférence, le boucher était nerveux. son travail me parle.	○	○	○

the actor is algerian, but the friend is tunisian. her acting skills were amazing. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
l'acteur est algérien, mais l'ami est tunisien. ses talents d'actrice étaient incroyables.	○	○	○
l'actrice est algérienne, mais l'ami est tunisien. ses talents d'actrice étaient incroyables.	○	○	○

Figure 3-5: Group 2: Intermediate case


Group 3

the wicked wolf ate the big deer. he was stuffed because he was heavy. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
le méchant loup a mangé le gros cerf. il était bourré parce qu'il était lourd.	○	○	○
le méchant loup a mangé le gros cerf. il était bourré parce qu'il était lourd.	○	○	○

the thief was really fast when trying to catch the cashier. she caught him fast. *

	1 - Totally wrong	2 - Partially wrong	3 - Perfect
la voleuse était très rapide en essayant d'attraper la caissier. elle le rattrapa rapidement.	○	○	○
le voleur était très rapide en essayant d'attraper le caissier. elle le rattrapa rapidement.	○	○	○

Figure 3-6: Group 3: Complex case

The evaluation of the survey is divided by group, from there the system that has the least wrong responses is considered superior to the other system.

3.2.3 Automated tools

Automated testing and scoring are also applied to the testing strategy of this work. Bilingual Evaluation Understudy (BLEU) is aslo used to score translations the users are given in order to rule out human bias but is not the major decision factor of translation quality, a minimum score of 80 is acceptable. The Machine learning model is evaluated by its ability to classify text properly.

3.3 Feasibility of the solution

There is a large corpus in both languages in order to gather more test cases for grammatical gender due to the variance of word endings and structure. This allows us to build a better machine learning model by providing more examples.

3.4 Trials

The decision of using manual processing for suffix swapping was done after trying to build a Lemma lookup for the Spacy library. A Lemma lookup prototype was possible to create for versions of Spacy below 2.2 but is not compatible with the latest version used in this research. Therefore until a reverse lookup can be built for the newest version of Spacy, manual swapping was the only viable option. Also a decision tree was chosen because of its simplicity and speed in order to improve feasibility in the times given as the use of more complex methods would lead to longer time training and a steeper learning curve.

3.5 Language ambiguity

The system addresses language ambiguity by following standard french rules as described by the normalization institutions such as “l’Académie française” and “l’Office québécois de la langue française” which are the two main boards that regulate standard French. This variety of French is the lingua franca of all Francophones and is considered proper written text.

Chapter 4

Findings

This chapter introduces and discusses the results of the survey conducted to evaluate the output of the GT gender bias corrector developed in this work compared to Google translate. Fifteen unique participants with diverse levels of fluency in English and French have completed the survey presented in the methodology chapter.

The survey consists of twelve texts divided into three groups based on the complexity of text structures. The first group includes text with simple structure that contains a single subject and a single pronoun. The second group includes texts with an intermediate structure that has a single subject, single object and a single pronoun. The third group includes texts with complex structure that has a subject, object and multiple pronouns. The GT gender-bias corrector performed very well on texts with simple cases, and texts with intermediate cases while it performed moderately on the texts with complex structure. In all three groups the GT gender-bias corrector significantly improved Google Translate's output by addressing the gender bias. The participants were asked to rate the sentences using:

1. Totally wrong - Blue

2. Partially wrong - Red

3. Perfect - Orange

The remaining parts of this chapter present and discuss the participants evaluations to the gender bias identified in the three translation groups presented in the survey.

4.1 The GT gender bias corrector perfectly address the Google Translate gender bias of simple and Intermediate cases

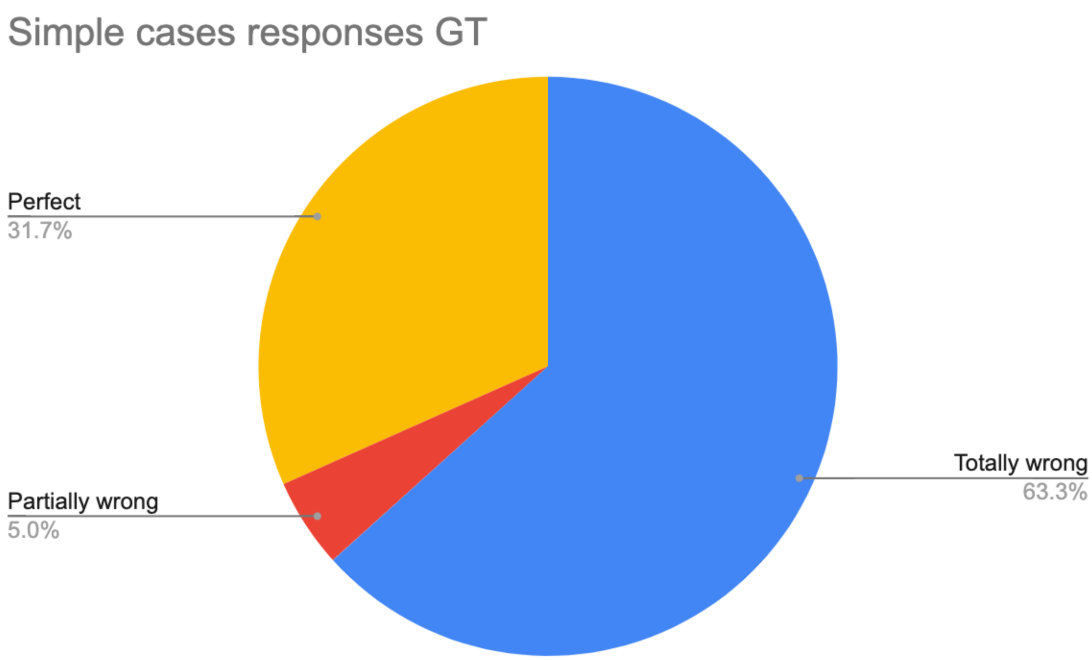


Figure 4-1: Simple cases responses Google Translate

About 63% of respondents found Google Translate's translation of simple cases to be completely wrong when it came to gender grammar. About 31% of respondents found some translations to be perfect with the bulk of these being translations that did not contain feminine pronouns that default to masculine which required no changes.

5% of respondents found them to be partially wrong. According to The Bilingual Evaluation Understudy Score which uses a correct translation as a reference and the Google Translate translation as a hypothesis, on simple cases Google Translate performed like this:

1. 44.18%

2. 54.39%

3. 100.00%

4. 72.27%

This indicates that Google Translate has the bulk of translations correctly in terms of general grammar but still lacks the gender grammar. In the next figure the simple findings of the GT gender-bias corrector findings are illustrated.

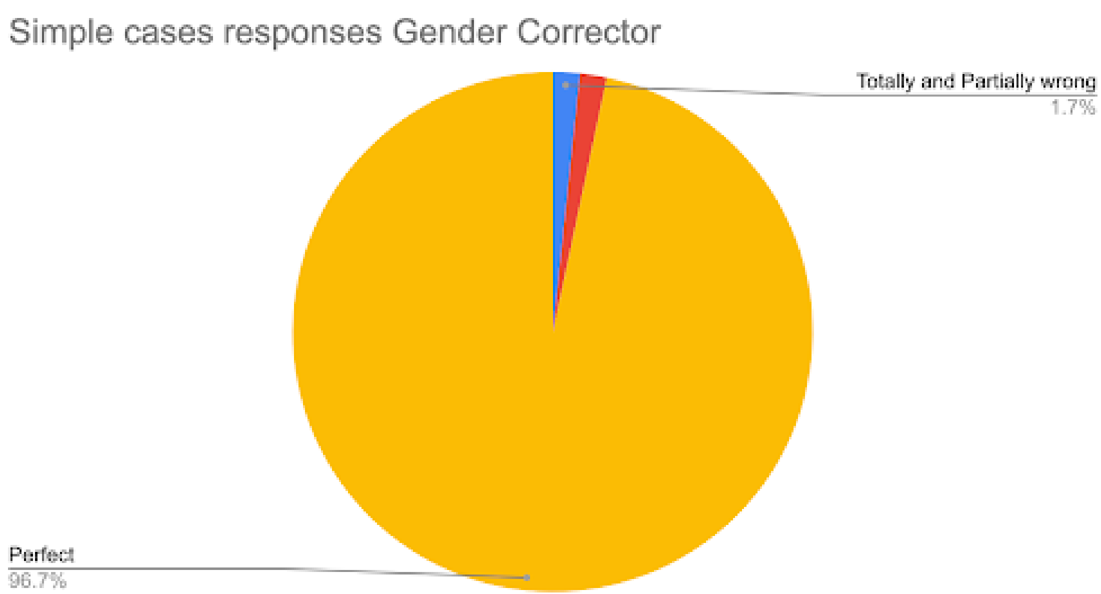


Figure 4-2: Simple cases responses GT gender-bias corrector

Around 97% of respondents believed that the system had a perfect translation of the gender grammar rules in the sentences provided. Included are the translations that kept masculine sentences as is. Only 3.4% believed that the translations were wrong. According to The Bilingual Evaluation Understudy Score, on simple cases the system performed like this:

1. 100.00%
2. 100.00%
3. 100.00%
4. 100.00%

These scores indicate that the system matched exactly with the reference sentence with a score of 100%. Below are the findings for intermediate cases.

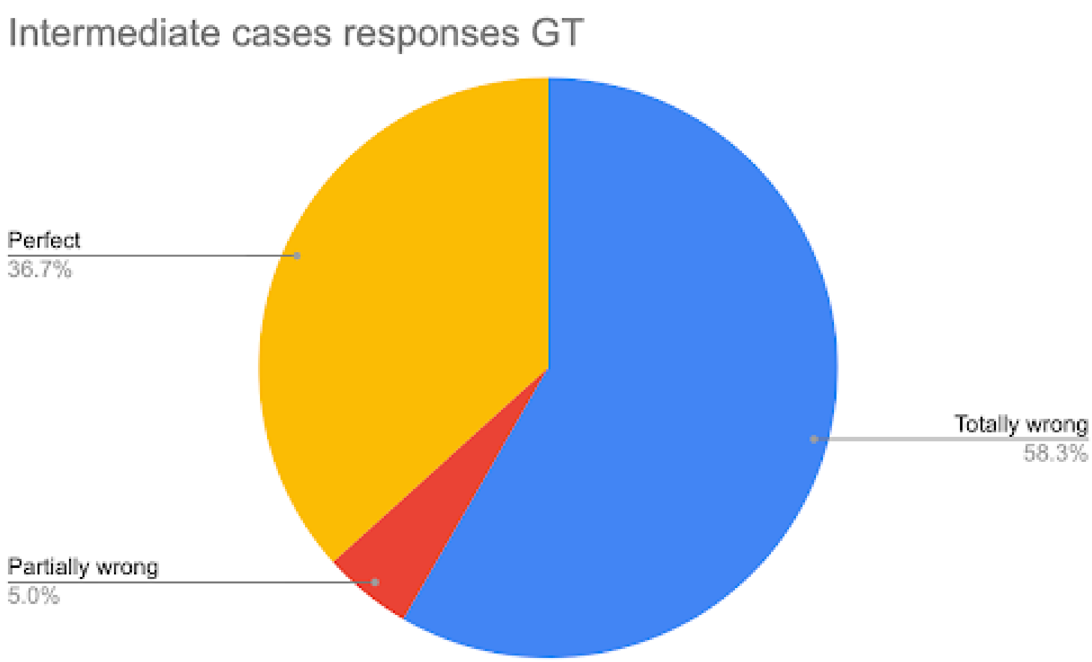


Figure 4-3: Intermediate cases responses Google Translate

In intermediate cases we see a very similar pattern to the simple cases with about 58% of respondents claiming Google Translate's translations were totally wrong with ferminine pronouns and 36% found them to be perfect with the bulk of them being translations that did not need gender grammar changes. BLEU score for the following sentences is as follows:

1. 54.14%

2. 75.4%

3. 58.43%

4. 100.00%

In intermediate cases Google Translate performed slightly better than on simple cases due to BLEU favoring larger texts in its algorithms that Google has mostly right minus the gender grammar. Below are the statistics for the intermediate cases with the GT gender-bias corrector.

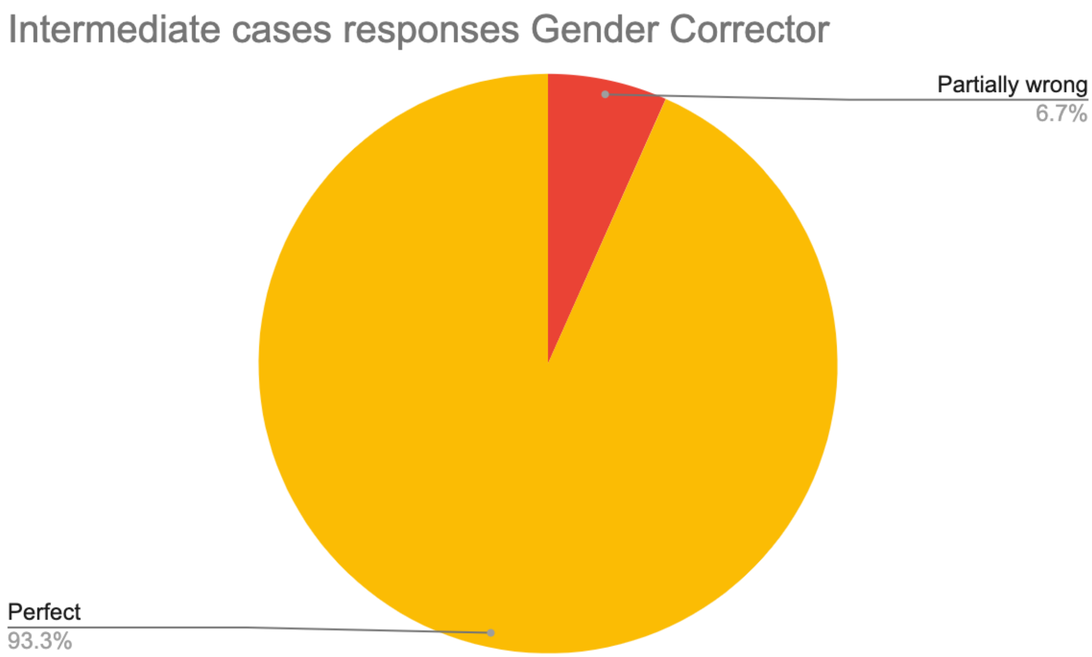


Figure 4-4: Intermediate cases responses GT gender-bias corrector

Similarly to simple cases, the GT gender-bias corrector outperformed stand alone Google Translate with about 93% of respondents claiming the sentences were perfect and only around 6% claiming them to be partially wrong with none saying they were totally wrong. These results include masculine gender sentences that remain unchanged. BLEU scores for them are as follows:

1. 100.00%
2. 100.00%
3. 100.00%
4. 100.00%

These scores indicate that the system matched exactly with the reference sentence with a score of 100%.

4.2 The GT gender-bias corrector moderately fixes the Google Translate' gender bias of complex cases

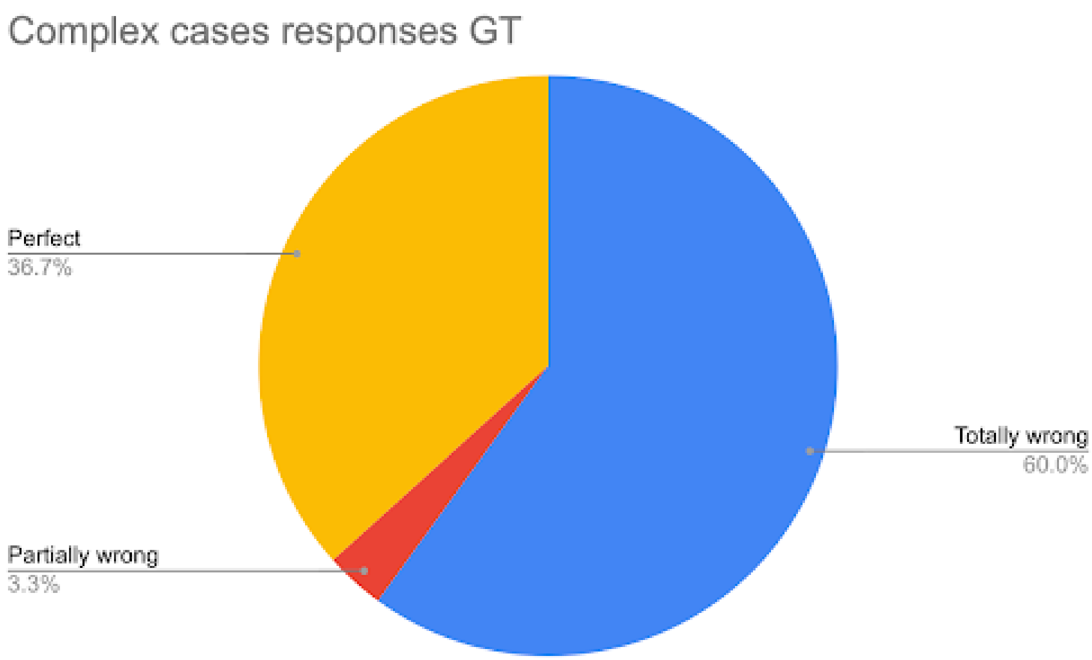


Figure 4-5: Complex cases responses Google Translate

We see a clear pattern here with 60% of respondents claiming Google Translate was totally wrong and about 36% claiming that the translations were perfect with the bulk being masculine gender that needed no modifications. Only about 3% claimed some were partially wrong. Their BLEU scores is as follows:

1. 100.00%
2. 84.09%
3. 85.35%

4. 56.81%

Google Translate performs better in the BLEU score here due to the larger translation of a complex case where Google gets the bulk of the Translation correctly minus the gender rules. Below is the GT gender-bias corrector findings for complex cases.

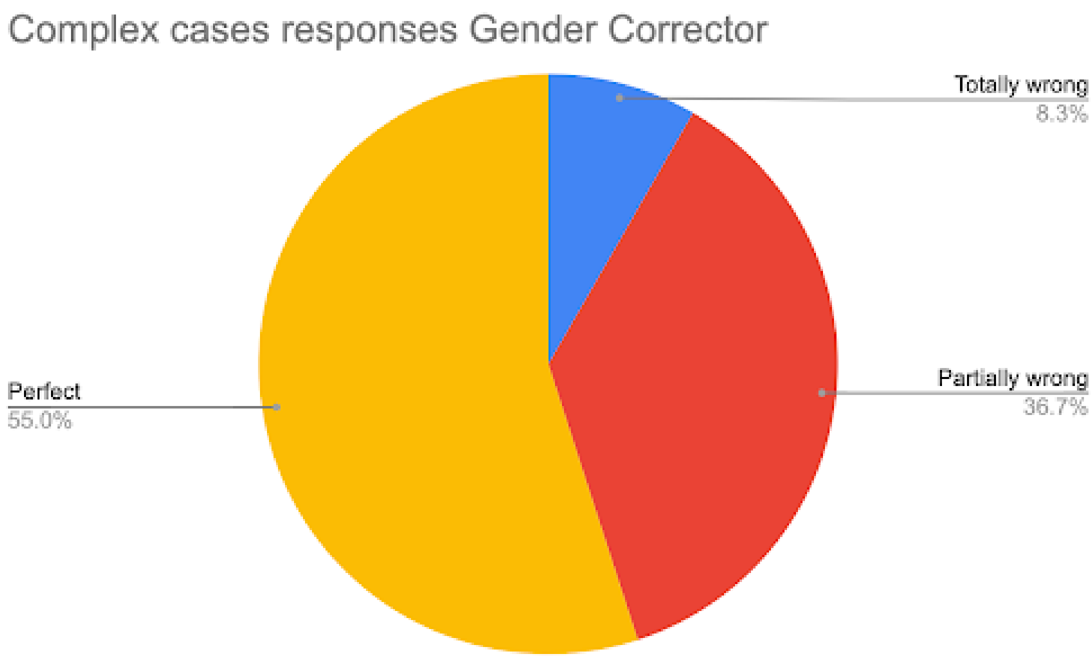


Figure 4-6: Complex cases responses GT gender-bias corrector

Here the GT gender-bias corrector did not perform as well as in the previous cases with more diverse responses which indicates a less precise gender grammar correction but remains more correct than Google Translate. 55% believed the translations were perfect with about 37% believing it was partially wrong and 8% as totally wrong. Their BLEU scores is as follows:

1. 100.00%

2. 80.59%

3. 90.45%

4. 82.55%

Even though Google Translate's BLEU score improved in complex cases, the GT gender-bias corrector still has the upper hand. The reason for lower scores of complex cases is attributed to certain determiners being changed when they were not supposed to. All nouns were properly modified which is the main part of speech that needs to be modified.

4.3 Analysis

In all three categories of translation that were surveyed, the GT gender-bias corrector performed much better than stand alone Google Translate which indicates that Google translate is erroneous and biased towards masculine gender when it comes to multi sentence gender grammar. Such is seen through the survey results, BLEU score and professional translation. The certain disparities in survey responses can be attributed to many factors such as fluency levels as certain candidates are not native speakers and learned French as a second language and are not as fluent with gender rules in French grammar. Another explanation for these disparities is the misunderstanding of the survey instructions and lack of attention to the sentences which is why at least 15 participants were selected.

The reason for the lower accuracy for complex cases might be attribute to certain parts of speech that did not undergo perfect modification which could lead to a sentence not being totally wrong but half wrong hence the higher percentage of partially wrong responses in that category compared to the intermediate and simple cases which would only undergo the modification of only three parts of speech while a complex case could undergo six.

Chapter 5

Conclusion

This research is the first to propose an add-on solution to the Google translate gender-bias problem that negatively impacts a large number of French speakers and speakers of similar languages. Although it cannot process large corpus of texts yet, the developed solution can be fine tuned to process gender of larger corpus by generating more training data to build machine learning models that can process larger text with complex structure. The developed system significantly outperforms Google on all cases tested in this work . This system is the first step towards changing the way we use Google Translate because this is the first tool that communicates with Google translate to address a problem. The need for such gender bias correctors was raised by several stakeholders such as business executives, diplomats and regular language learners. The pressing need of these stakeholders will promote the future of the GT gender-bias corrector.

5.1 Future work

5.1.1 Extending the GT gender bias corrector to other Romance languages

This research focused specifically on French, but all other Romance languages such as Spanish, Portuguese, Italian and Romanian suffer from the exact same issue. The methods and findings of this research can create the framework for fixing gender grammar in these languages as they follow very similar grammar patterns due to their common ancestor, Latin. Delving into techniques to properly translate complex cases and applicability to other languages will be investigated.

5.1.2 Benefits for Google Translate

Google Translate can benefit from the add on system by making users more trustworthy of the system which will promote its use and make it less erroneous in the eyes of its users and stakeholders. This will allow the reduction of mistranslation which promotes trades and social activities.

5.1.3 Expanded survey

An expanded use of the survey can allow us to learn more about the reasons why users rated answers in a certain fashion in order to gain more insights into natural language and machine translation evaluation methods. This allows us to extract how they use the system and their frame of mind when doing so. From there we can also find out why people use Google Translate in order to tailor the system.

5.1.4 Using more complex machine learning algorithms to improve the performance

This work used a J48 decision tree and we achieved 75% accuracy, in the future we should test using more sophisticated models such as CNN (Convolutional Neural Network). The decision tree is a very simple and rudimentary method that was used to prove that the proposed solution had some type of feasibility. Such a solution can be kept for simple and intermediate cases. When it comes to complex cases and beyond, we need a more expandable solution such as Convolutional Neural Networks. With this method we can feed larger texts and tag more parts of speech in order to account for many more scenarios and patterns that the simple attributes we currently have. Neural networks also help us derive deeper insights into what parts of speech need to be processed in our text in order to account for more complex cases. There are more patterns that can be extracted from different texts which incur more attributes that need to be accounted for. The bigger the text, the more parts of speech need to be tagged with the main ones being subjects and objects as they cause other parts of speech to be dependent on them. The number of subjects and objects determines the complexity of a text.

5.2 Limitations

5.2.1 The system can process two sentences only

Some improvements to the system can be done for tagging parts of speech on complex cases and the ability to determine grammar gender changes on large text that contain more sentences. In order to process more sentences we would need more time and available technology for large processing of text and rigorous machine learning model training. We were able to extract patterns on a simple model and made it applicable to many cases, therefore with more examples and better training we can achieve an even more robust system.

5.2.2 Reverse lemma instead of manually getting word endings

Another improvement is the building of a reverse Lemma lookup for the Spacy library which would reduce manual processing for fetching word endings. The Lemmatizer contains all the words related to a specific token in a sentence and therefore from there we can fetch the feminine equivalent of the specific word we are trying to modify without going through rigorous manual processing. This would drastically improve efficiency.

Bibliography

[1] Barak Turovsky. Ten years of google translate. https://www.blog.google/products/translate/ten-years-of-google-translate/, 2016. [Online; Accessed 2020-1-24].

[2] Aaron Babst. Improving google translate during i/o. https://translate.googleblog.com/2015/06/improving-google-translate-during-io.html, 2015. [Online; Accessed 2020-1-24].

[3] Carmen Hiers. Mistranslations cause loss of profits, trust, and reputation for global businesses. https://transformaonline.com/mistranslations-cause-loss-of-profits-trust-and-reputation-for-global-businesses/, 2018. [Online; Accessed 2020-1-25].

[4] National Research Council Canada. Meet yisi, a machine translation teacher who cracks down (nicely) on errors in meaning. https://www.canada.ca/en/national-research-council/news/2019/02/meet-yisi-a-machine-translation-teacher-who-cracks-down-nicely-on-errors-in-meaning.html, 2019. [Online; Accessed 2020-1-25].

[5] Ethan M Balk, Mei Chung, Minghua L Chen, Thomas A Trikalinos, and Lina Kong Win Chang. Assessing the accuracy of google translate to allow data extraction from trials published in non-english languages. 2013.

[6] N Börner, S Sponholz, K König, S Brodkorb, C Bührer, and Charles Christopher Roehr. Google translate is not sufficient to overcome language barriers in neonatal medicine. *Klinische Padiatrie*, 225(7):413–417, 2013.

[7] Sumant Patil and Patrick Davies. Use of google translate in medical communication: evaluation of accuracy. *Bmj*, 349:g7392, 2014.

[8] Alta Van Rensburg, Cobus Snyman, and Susan Lotz. Applying google translate in a higher education environment: Translation products assessed. *Southern African linguistics and applied language studies*, 30(4):511–524, 2012.

[9] William John Hutchins. *The development and use of machine translation systems and computer-based translation tools*. Bahri, 2003.

[10] Michelle Gazzola. Research for cult committee–european strategy for multilingualism: Benefits and costs, 2016.

[11] Gennady Medvedev. Google translate in teaching english. *Journal of Teaching English for Specific and Academic Purposes*, 4(1):181–193, 2016.

[12] Cynthia Ducar and Deborah Houk Schocket. Machine translation and the l2 classroom: Pedagogical solutions for making peace with google translate. *Foreign Language Annals*, 51(4):779–795, 2018.

[13] Klaus Mundt and Michael Groves. A double-edged sword: the merits and the policy implications of google translate in higher education. *European Journal of Higher Education*, 6(4):387–401, 2016.

[14] Fernando Ochoa Leite, Catarina Cochat, Henrique Salgado, Mariana Pinto da Costa, Marta Queirós, Olga Campos, and Paulo Carvalho. Using google translateˆ© in the hospital: A case report. *Technology and Health Care*, 24(6):965–968, 2016.

[15] CRICO Strategies. Malpractice risks in communication failures: 2015 annual benchmarking report. *Boston, MA*, 2015.

[16] Xuewei Chen, Sandra Acosta, and Adam Etheridge Barry. Evaluating the accuracy of google translate for diabetes education material. *JMIR diabetes*, 1(1):e3, 2016.

[17] Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59. Association for Computational Linguistics, 2012.

[18] Hassan Sawaf. Arabic dialect handling in hybrid machine translation. In *Proceedings of the conference of the association for machine translation in the americas (amta), denver, colorado*, 2010.

[19] Jacomine Nortier and Margreet Dorleijn. Multi-ethnolects: Kebabnorsk, perkerdansk, verlan, kanakensprache, straattaal, etc. *Contact languages: A comprehensive guide*, 6:237–270, 2013.

[20] Larissa Sloutsky and Catherine Black. Le verlan, phénomène langagier et social: récapitulatif. *The French Review*, pages 308–324, 2008.

[21] William John Hutchins and Harold L Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

[22] Sandeep Saini and Vineet Sahula. A survey of machine translation techniques and systems for indian languages. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 676–681. IEEE, 2015.

[23] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[24] Remya Rajan, Remya Sivan, Remya Ravindran, and KP Soman. Rule based machine translation from english to malayalam. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 439–441. IEEE, 2009.

[25] Hassan Sawaf, Mohammad Shihadah, and Mudar Yaghi. Hybrid machine translation, October 24 2017. US Patent 9,798,720.

[26] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[27] Sergei Nirenburg. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24, 1989.

[28] Budditha Hettige and Asoka S Karunananda. Computational model of grammar for english to sinhala machine translation. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 26–31. IEEE, 2011.

[29] John Hutchins. Machine translation: A concise history. *Computer aided translation: Theory and practice*, 13(29-70):11, 2007.

[30] Eduard H Hovy. Deepening wisdom or compromised principles?-the hybridization of statistical and symbolic mt systems. *IEEE Expert*, 11(2):16–18, 1996.

[31] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[33] Virginia M Holmes and B Dejean de la Batie. Assignment of grammatical gender by native speakers and foreign learners of french. *Applied Psycholinguistics*, 20(4):479–506, 1999.

[34] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*, 2018.

[35] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 2019.

[36] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

[37] Academemie Francaise. article. https://www.dictionnaire-academie.fr/article/A9A2699, 2020. [Online; Accessed 2020-10-5].

[38] Francis Bond. *Determiners and number in English contrasted with Japanese, as exemplified in machine translation*. University of Queensland, 2001.

[39] Academie Francaise. nom. https://www.dictionnaire-academie.fr/article/A9N0540, 2020. [Online; Accessed 2020-10-5].

[40] Silviu Cucerzan and David Yarowsky. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47, 2003.

[41] Academie Francaise. nom. https://www.dictionnaire-academie.fr/article/A9A0539, 2020. [Online; Accessed 2020-10-5].

[42] Mirjam Sepesy Maucec and Gregor Donaj. Machine translation and the evaluation of its quality. In *Natural Language Processing-New Approaches and Recent Applications*. IntechOpen, 2019.

[43] Jason Brownlee. A gentle introduction to calculating the bleu score for text in python. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/, 2017. [Online; Accessed 2020-11-23].

[44] P. S. Bronsart Gregory A. Sanders Mark A. Przybocki, Kay Peterson. The nist 2008 metrics for machine translation challenge - overview, methodology, metrics, and results. https://www.nist.gov/publications/nist-2008-metrics-machine-translation-challenge-overview-methodology-metrics-and, 2010. [Online; Accessed 2020-11-23].

[45] Alon Lavie Michael Denkowski. Meteor. https://www.cs.cmu.edu/ alavie/METEOR/, 2014. [Online; Accessed 2020-11-23].