Localizing Webpages for Francophone Audiences with Machine Translation

John E. Driscoll

March 20, 2018

Abstract

Localization is the process of translating the text and adapting the visual content of a software product for a specific target region. Given the global nature of the economy, organizations can appeal to non-English speaking consumers by offering localized versions of their software or web application. Currently, the localization process calls for time-consuming and costly translations of text strings by language service providers (LSP) with each additional locale incurring a new cost for LSP's manual translations. The high cost in terms of both time and money makes localizing software intimidating for organizations. This project explores the automation of the translation aspect of the localization process by using machine translation (MT) services. A Python script was developed to scrape HTML data and translate strings to French using Google Translate. The automatically generated translations were scored by human evaluators and the automated metric, BLEU. Participants were surprised at the quality of the translations, which were preserved meaning and were understandable despite some grammatical errors. This research suggests that the translation aspect of localizing software can be at least partially automated then edited for quality and fluency.

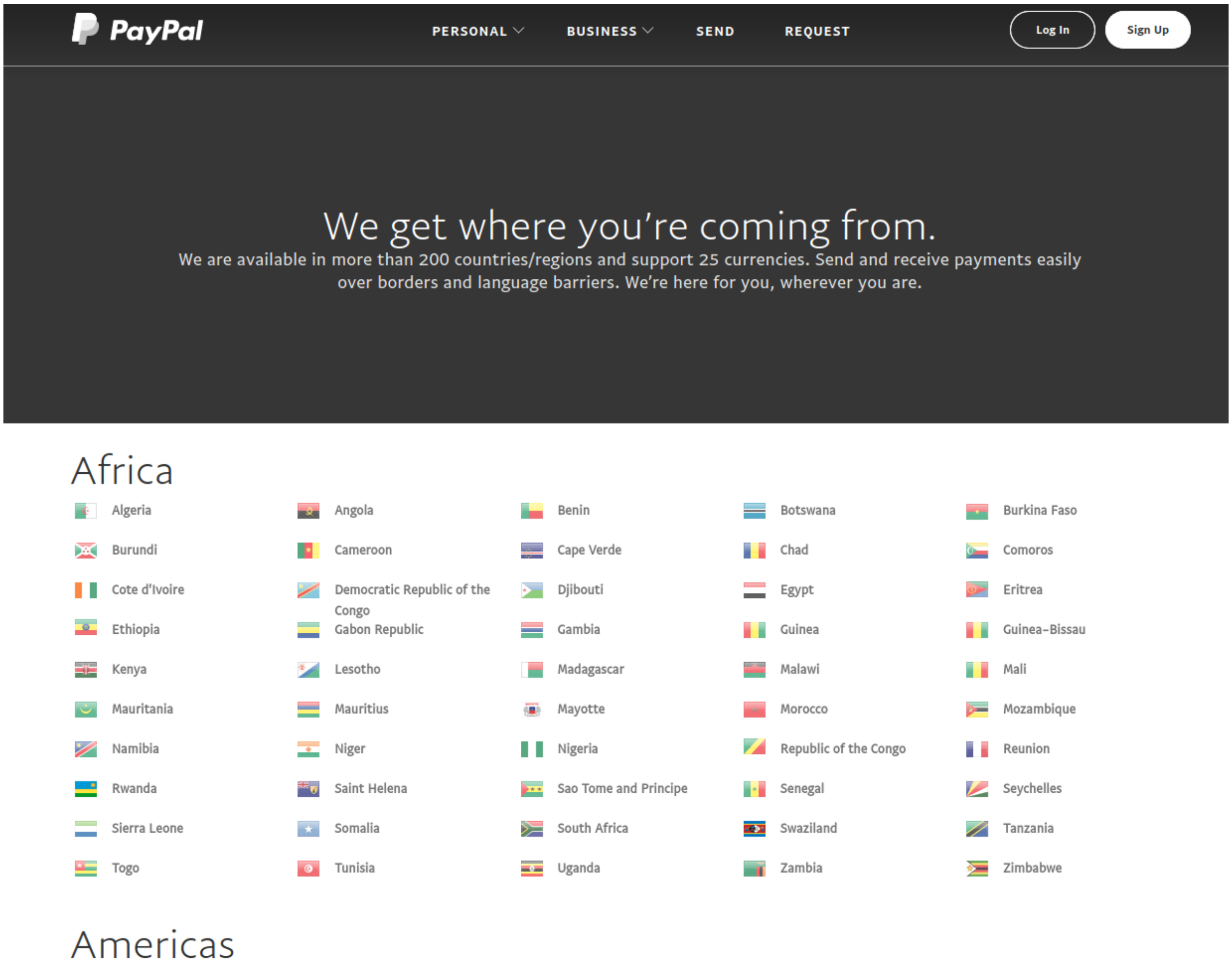
Contents

List of Figures

1 Introduction

In the context of computer science, internationalization and localization are defined in the following manner by the World Wide Web Consortium (W3C): Internationalization is the design and development of a product, application or document content that enables easy localization for target audiences that vary in culture, region, or language. Localization refers to the adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market (a locale). Stated more simply, internationalization refers to the translation of text strings to the language of a locale and the modification of any visual or other non-text content to suit the target locale. Internationalization significantly affects the ease of the product's localization as software that has been internationalized means that developers have already identified the strings and other content in need of change when localizing for any target region or audience. A prime example of fully internationalized web content is pictured in Figure 1, which is a screenshot of Paypal's localized web page options. Paypal offers their localized financial services to over 200 regions, which allows their products to be universally accessible for web users. However, paypal's internationalization and localization efforts do not represent the standard for web accessibilty. Retrofitting a linguistically- and culturally-centered deliverable for a global market is obviously much more difficult and time-consuming than designing a deliverable with the intent of presenting it globally. (Think back to the Y2K effort and trying to ”undo” two-character year fields that were built on the assumption of ”19xx”). So ideally, internationalization occurs as a fundamental step in the design and development process, rather than as an afterthought that can often involve awkward and expensive re-engineering.

Currently, Internationalization in the content management industry deals with the prerequisites of creating content in many languages by identifying the content, which needs to be translated or changed based on the target locale. These prerequisites include technologies and standards related to such things as character encoding, language identification and font selection. Internationalization is a prerequisite of localization, which is the specific adaptation of content to local markets and cultures. Localization typically involves translation and this translation is often outsourced to human translators and other cultural experts for a given market. Unfortunately, the current model for internationalization and localization often leads to a convoluted workflow that inhibits developer efficiency. This convoluted workflow consists of sending strings to human translators and waiting for their manual translations. Once the translated strings are returned and the software can be localized, it also must be checked by editors for quality and fluency in the target langauge. Manual translation is a time-consuming and expensive practice, which leads to exploring the efficiency of machine translation services in place of manual translation. Machine translation is a subfield of computational linguistics that investigates the use of sofware to translate text or speech from one



We get where you're coming from.

We are available in more than 200 countries/regions and support 25 currencies. Send and receive payments easily over borders and language barriers. We're here for you, wherever you are.

Africa

Algeria
Angola
Benin
Botswana
Burkina Faso
Burundi
Cameroon
Cape Verde
Chad
Comoros
Cote d'Ivoire
Democratic Republic of the Congo
Djibouti
Egypt
Eritrea
Ethiopia
Gabon Republic
Gambia
Guinea
Guinea-Bissau
Kenya
Lesotho
Madagascar
Malawi
Mali
Mauritania
Mauritius
Mayotte
Morocco
Mozambique
Namibia
Niger
Nigeria
Republic of the Congo
Reunion
Rwanda
Saint Helena
Sao Tome and Principe
Senegal
Seychelles
Sierra Leone
Somalia
South Africa
Swaziland
Tanzania
Togo
Tunisia
Uganda
Zambia
Zimbabwe

Americas

Figure 1: Paypal's offering of localized web content

language to another. It is possible that the localization process can be partially automated by integrating Machine Translation (MT) into the translation aspect of the process. This project will explore the integration of automated MT techniques with industry-provided APIs. The goal is to increase automation and reduce the amount of human effort necessary to localize software while still generating web pages that are suitably localized for Francophone audiences.

2 Motivation

Automatic localization holds great importance because of the amount of resources that companies spend on the translation process during Internationalization and Localization. Big software companies such as Microsoft and Facebook may be able to effectively budget optimal adaptation of their products to various locales as these companies have significant spending power. However, small software companies and startups that are operating on relatively tight budgets will have trouble with internationalizing their products. This may introduce serious time delays and an increase in costs, which go far beyond just engineering. For instance, big companies typically invest $2 million and 12 to 18 months of their engineering resources in internationalization and delivery of the first foreign language version of their content Wang et al. [11]. Then, fully localizing a software product for one additional language can add up to $100K, whereas Microsoft estimates its costs are $300K or more per product Wang et al. [11]. This can be a tall order for small software companies, but also for big companies to target languages with low strategic value in terms of market share. The fully automated and accurate translation of web applications would be hugely beneficial to developers across the world. For instance Wang et al. [11] said the following, "Another area related to our approach is machine translation. Machine translation has been used in translating web pages in some products. If the technique of machine trannslation is advanced enough to accurately translate the generated web pages in real time, the internationalization and localization of web applications may become unnecessary.

All of this information led me to formulate the following research hypothesis: Can I layer automated MT techniques and integrations with industry-provided software packages to make a translation framework using my own Python middleware that locates need-to-translate strings in HTML webpages and localizes them for Francophone locales while still enabling high-quality web app internationalization and localization?

3 Background and Related Work

The work discussed here focuses primarily on the progressions of Machine Translation systems, as well as the significance of internationalization. With a particular focus on frameworks for improving the efficiency of internationalization and the benefits of Machine Translation in regards to human interaction.

3.1 Machine Translation

Machine Translation of natural human languages began to emerge as a research field in the 1950's as the idea of high-speed high-quality translations of arbitrary texts piqued the interest of both the military and intelligence communities [10]. The first public demonstration of an MT system occurred in 1954 with the Georgetown-IBM experiment, which translated more than 60 Russian sentences to English. Despite the relatively small scale of this demonstration, it was influential in sparking public interest in MT and stimulating large-scale funding for MT research in the following decades [6]. Research and development of MT systems continued through the mid-1960's when a report published by the Automatic Language Process-

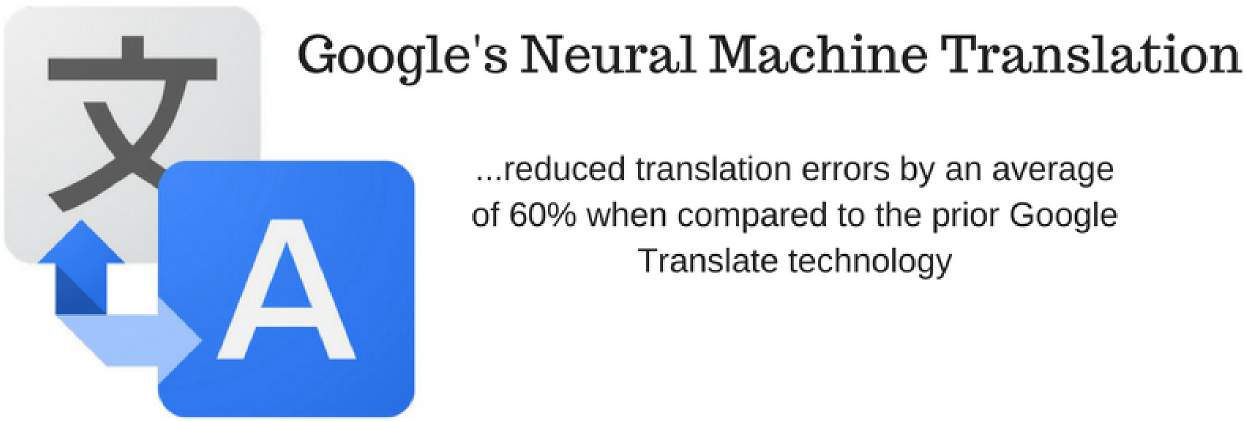

Google's Neural Machine Translation
...reduced translation errors by an average of 60% when compared to the prior Google Translate technology
文
A

Figure 2: Google's NMT service.

ing Advisory Committee (ALPAC), which was formed by the United States government in 1964, bashed the prospects of MT. The report, which was published in 1966, concluded that MT was slower, less accurate, twice as expensive as human translation [6]. ALPAC's report was profoundly influential on researchers in the United States and brought MT development to halt for over a decade.

As computational power improved over time and became less expensive, interest in Machine Translation was revitalized in the early 1980's among researchers in the United States. This decade saw the implentation of statistical models in various areas of computational linguistics including automatic speech recognition, lexicogrpahy and natural language processing [2]. Developers of MT systems emulated this trend and began to rely upon statistical approaches to automated translation. IBM was a pioneer in the statistical approach to Machine Translation and researchers at the Thomas J. Watson Research Center developed their own system during the late 1980's, which is detailed in Brown et al.'s 1990 paper: "A Statistical Approach to Machine Translation." The paper details an approach that translates French sentences to English by relying on statistical models derived from the analysis of bilingual text corpora. IBM's work rekindled interest in the field of Machine Translation and formed the basis of the MT paradigm known as Statistical Machine Translation. Statistical MT encompasses several models for translation including word-based translations, syntax-based translation and phrase-based translation.

The 1990's and 2000's saw the widespread adoption of SMT across the field of Machine Translation. As the availability of large text corpora, such as the Europarl corpus[9], and automated metrics for testing MT systems increased, researchers were able to develop SMT systems without knowledge of the target language for translation. This increase in research and development led to open-source and propietary systems became available to the public and large corporations began to make use of MT systems specifically in the field of software localization [7]. One of the most notable online translation services, Google Translate, launched in 2006 using statistical machine translation, specifically phrase-based translation, to translate input sequences. SMT served as the industry standard for MT systems until the development of Neural Machine Translation (NMT) in 2016. Rather than deriving outputs from statistical models based on bilingual corpora, NMT uses a deep learning approach based on neural networks [12]. At a high level, NMT works in two stages. Firstly, the system takes the entire input sequence and models each word that needs to be translated based on the context of the word (and its possible translations) within the full sentence. Secondly, it translates the word model (not the word itself but the model the neural network has built of it), within the context of the sentence, into the target language. NMT significantly improved the performance of Google Translate as measured by a BLEU score [12].

BLEU is an automated metric for evaluating machine translation output. Presented by IBM researchers in their 2002 paper, "BLEU: a method for automatic evaluation of machine translation," the researchers derived their motivation from the high-cost in terms of both time and money for evaluating machine translation systems with human evaluators. Given the expensive nature of human evaluation approaches, an inexpensive automatic evaluation that is quick and language-independent would be extremely beneficial to the developers of MT systems. To establish the automated metric the researchers pose the following question: How does one measure translation performance? From the viewpoint of the authors, the closer a machine translation is to a professional human translation, the better it is [8]. This notion of gauging translation performance in regards to reference translations is the main idea behind the proposed method, BLEU, which stands for Bilingual Evaluation Understudy. This evaluation system requires two components: a numerical "translation closeness" metric and a corpus of human reference translations. The aforementioned closeness metric is based upon the word error rate metric used by researchers in the field of speech recognition. While outlining the baseline metric for the BLEU method, the researchers note that there are several "perfect" translations for a given source sentence. This variation in what constitutes a potential high-quality translation comes from the subjectivity of a human translator in regards to their word choice and word order. The researchers use a modified n-gram precision metric to capture two aspects of translation quality: adequacy and fluency. Adequacy and fluency are metrics commonly used during human evaluation of translations [1]. BLEU provides a fast method for evaluating the efficacy of MT models, while also correlating with human judgments based on statistical analysis for translation into English from four different languages: French, Spanish, Arabic and Chinese [8].

3.2 Internationalization and Localization

In their, 2013 paper "Teaching Internationalization Internationally," Heines and Kasem present input on why internationalization should be taught. The paper provides a good foundation in regards to the importance of internationalization and the growing role software internationalization and localization will play in our global economy [5]. These researchers advocate for and demonstrate the importance of internationalization. The research team especially emphasizes the point that a true understanding of internalization must come through experiencing first-hand the drawbacks of programs that are not localized to a given locale. American students collaborate via email and online chats with Polish students to receive feedback on their code and its ease of understanding. In this research paper, the American professor had his students develop web pages with internationalization in mind, while the Polish professor had his students modify the web-pages developed by the American students to see how well the design and code of the American students would help when Polish Text was substituted for English text. Given that the research paper chronicles an approach to teaching internationalization to students in two different locales, there were mixed results and feedback based on each students reaction to this unconventional learning process. The major takeaway from this paper was that localization and the cultural changes that developers need to address in order to localize their code is still a very difficult and manual process. Nevertheless software internationalization and localization are crucial to societys growth as we move further into the digital age[5].

Having established the overall importance of internationalization and the challenges it can pose to both developers and users,I will move on the next related work. The 2012 research paper, The Multilingual Web, by D. Filip et al. addresses the tension introduced by my topic through its exploration of the internet and thus web applications as platforms that have users with many different native languages [3]. This research paper is a high-level overview of the Web as a multilingual entity, to which people from across the globe have access. In particular, the paper reports on the MutilingualWeb initiative, which is a collaboration between the W3C Internationalization Activity and the European Commission. The MultilingualWeb initiative was formed as a thematic net- work project, which explores the standards and best practices for supporting the creation, localization and use of multilingual web-based information. For me, the fact that this paper reported on an EC initiative (French) and sought to address best practices in web internationalization and the gaps, which have to be filled in internationalization, was very applicable to my own research. The paper recognizes 6 differing viewpoints in regards to web standardization: Developers, Creators, Localizers, Machines, Users and- Policy Makers. Each of the aforementioned category play a significant role during the life of a web application, whether it is the developer(s) creating the multilingual content or the policy makers dealing with licensing standards for language resources. My main takeaway from this paper is that web internationalization and localization still poses a major problem for those engineers creating the programs as their web applications must go through a series of steps before it is easily accessible to a broad international market [3].

Considering the web and its users as a multilingual entity necessitates easy manners of communication for collaboration on the development side, specifically in the case of teams of developers comprised of both native and non-native English speakers. Two is Better Than One: Improving Multilingual Collaboration by Providing Two Machine Translation Outputs, a 2015 paper by G. Gao et al., introduces the benefits that free MT services (such as Google Translate) can have for cross-language work environments [4]. This Research paper asks the question: Does providing two Machine Translation outputs during multilingual collaboration improve conversational grounding and task management? Many organizations collaborate across national and language boundaries or, at the very least, hire non-native english speakers. Language barriers can make effective teamwork and collaboration difficult as it amplifies the cognitive load already being placed on non-native speakers by their workload. Despite MT's benefits, one of the major drawbacks of machine translation is the meaning of the original message being lost after being processed by MT algorithms and the failure of users to detect these incorrect translations. This paper looked at a low-cost and relatively straightforward solution to make up for the MTs potential drawbacks: provide two MT outputs. In order to study the effectiveness of this solution, the researchers created three different communication conditions: MT with a single translation output for each message, MT with two translation outputs for each message, and English as a common language. In the first two conditions, both native English speaking participants and native Mandarin speaking participants typed and received messages in their native languages. In the third condition, participants typed and received messages in English. The results were too long-winded to include in this annotation, but overall the results suggest that when people speak different native languages, showing two different MT outputs has many benefits and few costs.

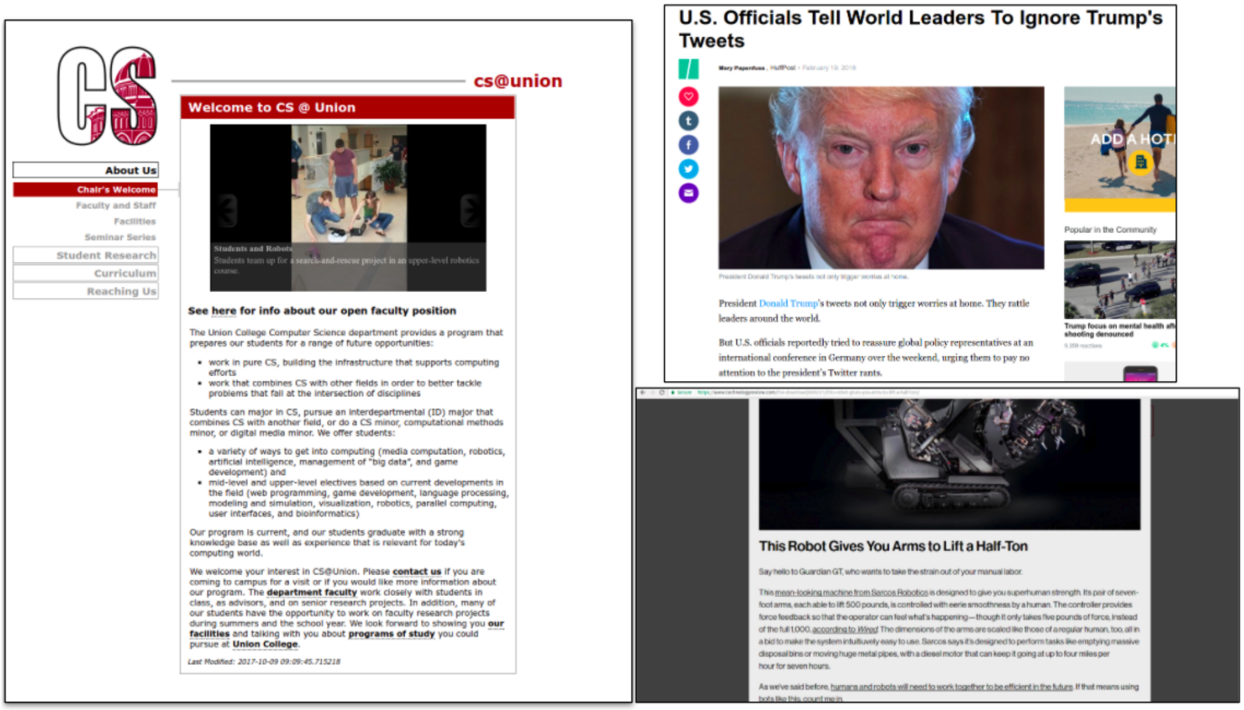
This next paper is the most directly related to my research paper of any related work that has been discussed so far. Expanding on previous research, this paper outlines a framework for automatically locating the strings that must be localized in a web application. Locating Need-to-Translate Constant Strings in Web Applications, a 2012 paper by X. Wang et al., builds off their previous work TranStrL, which is an Eclipse (Java IDE) plugin to automatically locate need to translate strings when internationalizing Java software [11]. This research paper concentrates on the task of identifying constant strings in web applications that need to be translated in order for the application to be effectively internationalized. There are two research questions that this paper explicitly seeks to evaluate 1. How effective is their approach to locating need-to-translate constant strings in web applications? and 2. How effective are their techniques for locating the different types of need-to-translate constant strings in web applications compared with a related approach by the researchers which was designed to locate need-to-translate strings in Java software? The paper notes that developers of web applications need to locate all the hard-coded language-specific elements and externalize them to property files. Furthermore, when locating these language-specific elements, the most time-consuming task is to locate need-to-translate constant strings due to their large number and scattered distribution in the code. However, the arduous task of performing internationalization on web applications is particularly important as once an application is online, users from all over the world can access them. The task of translating and localizing the strings is still heavily reliant on human translation and editing by language experts for specific locales. The results of the research demonstrate that the research teams approach was effective for locating need-to-translate strings due to their programs ability to accurately distinguish between visible and non-visible strings that may go to generated HTML texts. But the most exciting part of the paper came in researchers discussing of related work, in which they mentioned Machine Translation and how MT could be integrated into web and software internationalization to reduce the human workload in internationalization.

4 Data

To test the translation middleware of this project, I have selected 3 web pages of varying structure and content. The following web pages contain content specific to various domains including political content and scientific conten in order to see how MT performs with various domain-specific language. The 3 web pages were the following:

- The Union College CS Department homepage
- A Yahoo News article about President Trump's Tweets
- An Article from The Download, a blog from MIT's Technology Review

These web pages were chosen for experimentation due to their differing content and structures. The HTML data ranges from text-heavy pages such as the Union CS department homepage to pages that contain more visual features and components such as the Yahoo News article about President Trump's tweets. All of the HTML data being used from these three web pages is from one static webpage. Some pages rarely change their content such as the Union CS Department home page, while pages such as Yahoo News change constantly. The web pages are all approximately the same in terms of the length HTML data with the range of the pages being betweenn 250 and 400 lines. Longer web pages were included in the test



Welcome to CS @ Union

Students and Robots
Students team up for a search-and-rescue project in an upper-level robotics course.

See here for info about our open faculty position

The Union College Computer Science department provides a program that prepares our students for a range of future opportunities:

- work in pure CS, building the infrastructure that supports computing efforts
- work that combines CS with other fields in order to better tackle problems that fall at the intersection of disciplines

Students can major in CS, pursue an interdepartmental (ID) major that combines CS with another field, or do a CS minor, computational methods minor, or digital media minor. We offer students:

- a variety of ways to get into computing (media computation, robotics, artificial intelligence, management of "big data", and game development) and
- mid-level and upper-level electives based on current developments in the field (web programming, game development, language processing, modeling and simulation, visualization, robotics, parallel computing, user interfaces, and bioinformatics)

Our program is current, and our students graduate with a strong knowledge base as well as experience that is relevant for today's computing world.

We welcome your interest in CS@Union. Please **contact us** if you are coming to campus for a visit or if you would like more information about our program. The **department faculty** work closely with students in class, as advisors, and on senior research projects. In addition, many of our students have the opportunity to work on faculty research projects during summers and the school year. We look forward to showing you **our facilities** and talking with you about **programs of study** you could pursue at **Union College**.

Last Modified: 2017-10-09 09:09:45.715218

U.S. Officials Tell World Leaders To Ignore Trump's Tweets

Mary Papenfuss , HuffPost · February 19, 2018

President Donald Trump's tweets not only trigger worries at home.

President Donald Trump's tweets not only trigger worries at home. They rattle leaders around the world.

But U.S. officials reportedly tried to reassure global policy representatives at an international conference in Germany over the weekend, urging them to pay no attention to the president's Twitter rants.

This Robot Gives You Arms to Lift a Half-Ton

Say hello to Guardian GT, who wants to take the strain out of your manual labor.

This mean-looking machine from Sarcos Robotics is designed to give you superhuman strength. Its pair of seven-foot arms, each able to lift 500 pounds, is controlled with eerie smoothness by a human. The controller provides force feedback so that the operator can feel what's happening—though it only takes five pounds of force, instead of the full 1,000, according to *Wired*. The dimensions of the arms are scaled like those of a regular human, too, all in a bid to make the system intuitively easy to use. Sarcos says it's designed to perform tasks like emptying massive disposal bins or moving huge metal pipes, with a diesel motor that can keep it going at up to four miles per hour for seven hours.

As we've said before, humans and robots will need to work together to be efficient in the future. If that means using bots like this, count me in.

Figure 3: The three webpages.

data as I wanted the translated output to be a manageable length for participants to work through and provide feedback. The important difference between the pages is their varying content and domain-specific language. The Union CS home page surveys the field of Computer Science at Union College, while the Yahoo News article contains political content. The technology blog from MIT's the Download is concerned with scientific content, however it is less domain-specific than the Union CS home page.

5 Methods and Design

The first step in the development of this framework is understanding two of the most crucial aspects in internationalization and localization: the workflow of a developer and where translation fits into the process of this workflow. Through my research I have gleaned the following as a prototypical workflow in the localization process:

- 1 New product (web application) developed
- 2 Externalize need-to-translate strings
- 3 Send strings for translation
- 4 Translation Queue (comprised of need-to-translate strings sent)

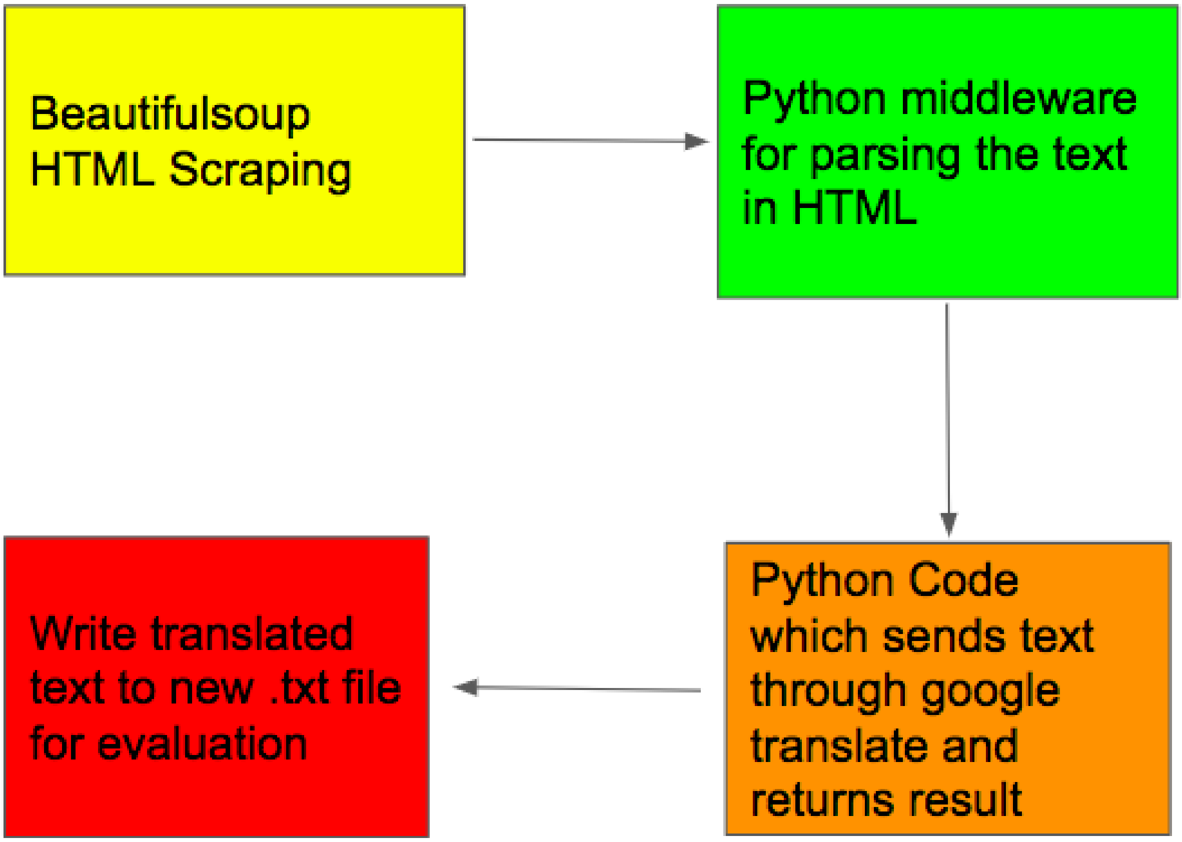
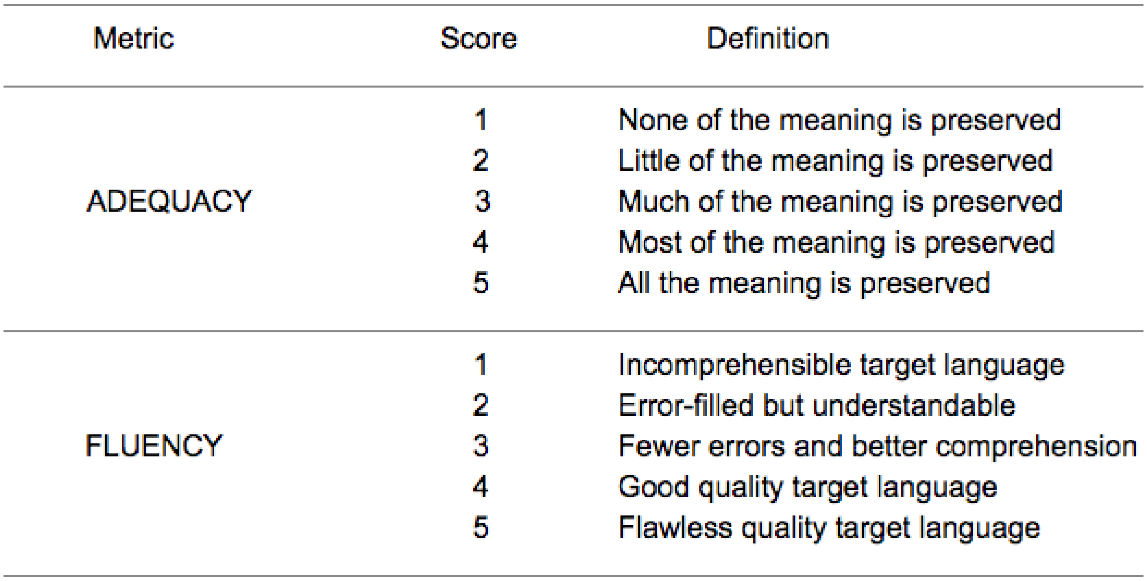

Beautifulsoup HTML Scraping
Python middleware for parsing the text in HTML
Python Code which sends text through google translate and returns result
Write translated text to new .txt file for evaluation

Figure 4: The Translation Pipline

- 5 Initial Translation by LSP (Language Service Providers)
- 6 Translation Review
- 7 Translation Revision
- 8 Final human translation/editing of strings

The list above is representative of a convoluted workflow that costs developers time and money to roll new products out to different locales. Translation plays a pivotal role in the localization of products and one of my goals is to make the role translation plays as unobtrusive as possible on the developer side. This project will focus on automating steps 2-8 in the translation process detailed above. To achieve this I will have the Translation Queue of externalized strings be put through a free MT service, which in this case was Google Translate . The externalization of strings is also an important aspect as this can be a time consuming task without the help of a software such as that developed by X. Wang et al (TranStrL). For dealing with the difficulties of localization, I have created a translation pipeline using open-source MT resource in conjunction with the Python library, BeautifulSoup4. First HTML data is scraped using Urllib and BS4 and converted into a BS4 object. The libraries used were based on a tutorial on Stanford University's website (http://web.stanford.edu/ zlotnick/TextAsData/WebScrapingwithBeautifulSoup.html). Once the HTML data has been converted to a BeautifulSoup object, the strings are parsed from the HTML. The parsing of strings proved particularly difficult as the HTML data varied from web page to web page. Once these need-to-translate are extracted from the HTML, they are then converted to unicode so that they can be translated. The strings must be converted to unicode as they are merely a stream of bytes after BeautifulSoup HTML scraping. The pipeline then uses a python script to send the strings individually through Google Translate or Bing Translate. After returning the translated string, the pipeline replaces the old untranslated string with the newly translated one. Finally, using Python, a new HTML text file is written to act as the translated webpage. In order to evaluate the quality of the automated tranlations, I gathered feedback from French-speakers at Union College and made reference translations for the web pages to generate BLEU scores.

Participants in the experiment gave both quantitative and qualitative feedback on each of the three translated web pages. From a quantitative standpoint, I gathered feedback based on two metrics: adequacy and fluency. The adequacy metric is equivalent to asking: Does the output convey the same meaning as the input sentence? On the other hand, the fluency metric seeks to answer the question: is the output good, fluent french [1]? The french speakers were asked to rate the translations from 1-5 in regards to both adequacy and fluency. On both scales, 1 is the worst possible rating and 5 is the best, however the two



Metric	Score	Definition
ADEQUACY	1	None of the meaning is preserved
	2	Little of the meaning is preserved
	3	Much of the meaning is preserved
	4	Most of the meaning is preserved
	5	All the meaning is preserved
FLUENCY	1	Incomprehensible target language
	2	Error-filled but understandable
	3	Fewer errors and better comprehension
	4	Good quality target language
	5	Flawless quality target language

Figure 5: Tables for Adequacy and Fluency ratings

metrics represent different aspects of what constitutes a good translation. A breakdown of the two scales is provided in Figure 3. From a qualitative standpoint, I gathered feedback based on editing marks made by participants. The translations were shared electronically with the participants and they were able to suggest edits wherever they saw fit to do so. More specifically, I was looking for common trends in the qualitative feedback such as word choice errors, verb tense errors and word order errors.

In order to conduct my experiment, a Google drive was created, which included specific instructions regarding the scoring metrics used, each of the three translated web pages and a Google Form for recording feedback. Prior to evaluating the translations, participants were asked a series of questions in order to gauge their background in French. The documents contained screenshots of the translated web pages along and were broken down into sentence by sentence (or passage by passage) pieces. Each piece of parsed text included the original english text for the translation for the participants to use as a reference. Additionally, I worked with Professor Charles Batson and Union's French language assistant, Amandine Belard, to create actual reference translations for each of the web pages. These reference translations served as gold standards when evaluating the qualitative feedback from the automated translations and allowed BLEU scores to be generated.

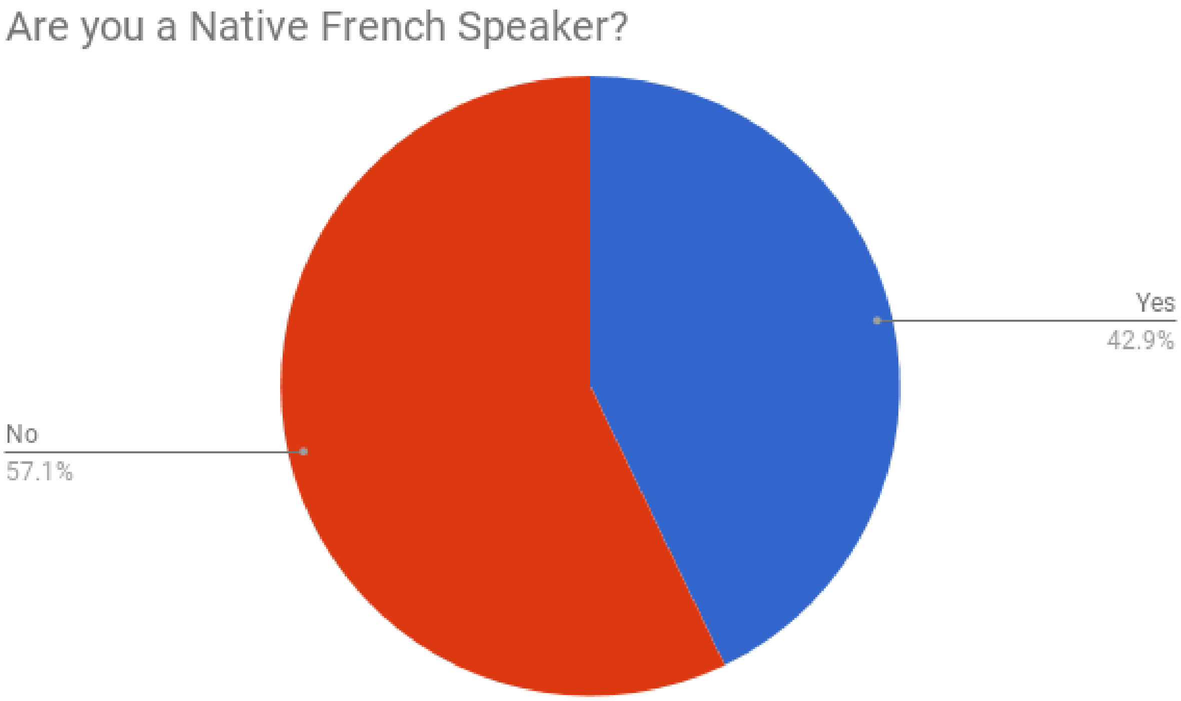


Figure 6: Preliminary Question 1.

6 Results

The results of my feedback have been broken down into three categories: preliminary background questions, quantitative feedback and qualitative feedback.

6.1 Background of Participants

A total of 14 individuals participated in my experiment. Before evaluating the three web page translations, participants were asked three questions regarding their background as French speakers. These questions were asked in order to gauge the fluency of particpants. The questions were:

- Are you a native French speaker?
- Did you grow up around French speakers (e.g. immediate family, relatives etc.
- What is the highest level French course you have taken?

Of the 14 participants, the majority of them were non-native French speakers. On the other hand, the majority of the participants had indeed grown up around French-speakers . The third and final preliminary

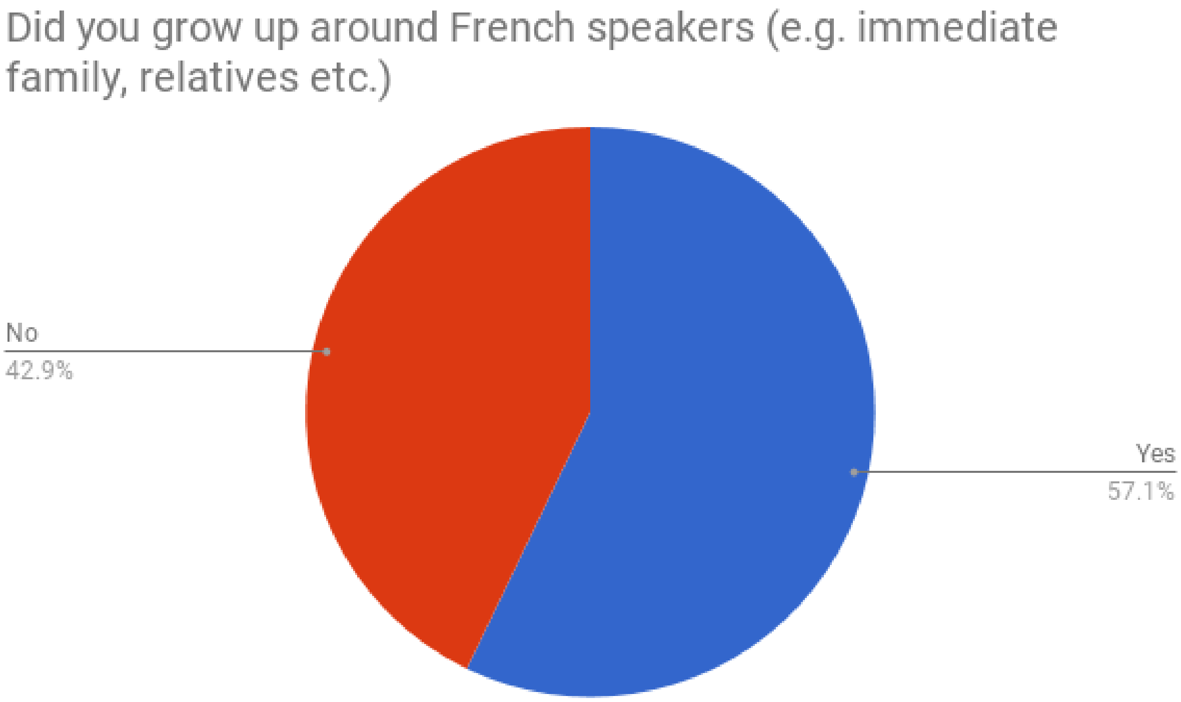


Figure 7: Preliminary Question 2.

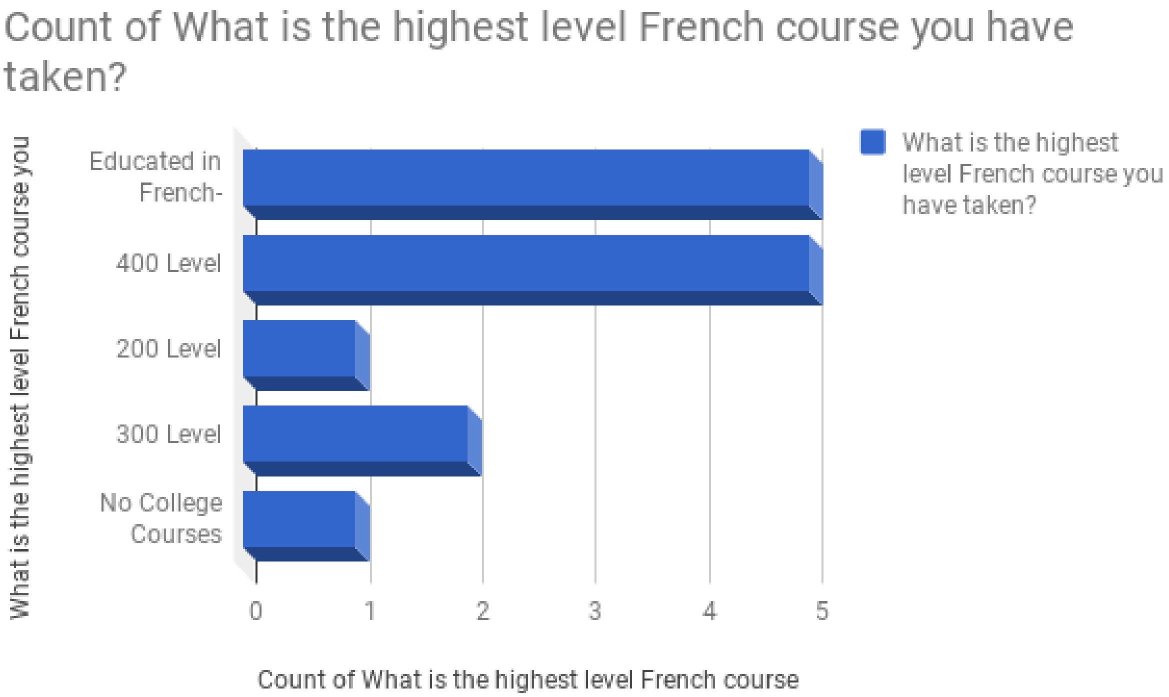


Figure 8: Preliminary Question 3.

question was showed a variance among the participants' educational background in French. Although there were a couple outliers, the majority of participants were either advanced French-speakers (300 or 400 level) or were received their secondary at a French-speaking school. The feedback from these questions affirms that fluent and capable French-speakers are evaluating the translations. All of this data is visualized in Figures 6-8.

6.2 Quantitative Feedback

The quantitative feedback for this experiment was broken down into two metrics, adequacy and fluency, which the participants scored on a 1-5 scale. In terms of adequacy, each of the three translated web pages had an average of at least 4, which indicates that "most of the meaning" was preserved in the automated translations. The Union College CS department home page scored the lowest in terms of adequacy with an average of exactly 4. Several participants noted that the language of this web page was specifically tailored to Computer Science, which resulted in word choice errors in the translation. Specific types of translation errors will be discussed at length in the Qualitative Feedback section. The other two translations averaged only slightly better scores in terms of adequacy with the article about Trump's tweets avegaring just above 4.2 and the technology blog averaging approximately 4.1. This relative consistency in terms of adequacy scores across the three different translations demonstrates that automated translations can provide understandable and meaningful translations to French. Overall, I noticed that several participants were genuinely surprised with the ability of the machine translated outputs to convey much of the meaning from the original text.

The other piece of quantitative feedback for these translations came in the form of fluency scores, which like adequacy were scored on a scale from 1-5. As was previously mentioned, fluency addresses the grammatical correctness of translations in regard to the target language. All of the averaged fluency scores for the translations fell between 3 and 4. A score of 3 indicates "fewer errors and better comprehesion" while a score of 4 represents "good quality target language." The Union CS department home page was also the lowest in terms of average fluency rating at approximately 3.35. As will be discussed in the following section regarding qualitative feedback, I believe the lower score derives from the text being laden with technical terms and phrases specific to the field of Computer Science. The Trump news article and the MIT technology blog were close in terms of their fluency scores with approximately 3.79 and 3.86, respectively. As these web pages were not as domain-specific as the Union CS home page, the automated translations were more fluent. Given the above average scores for both the quantitative metrics and the fact that the scores came from both native and non-native French speakers, I believe that MT systems can be effective as

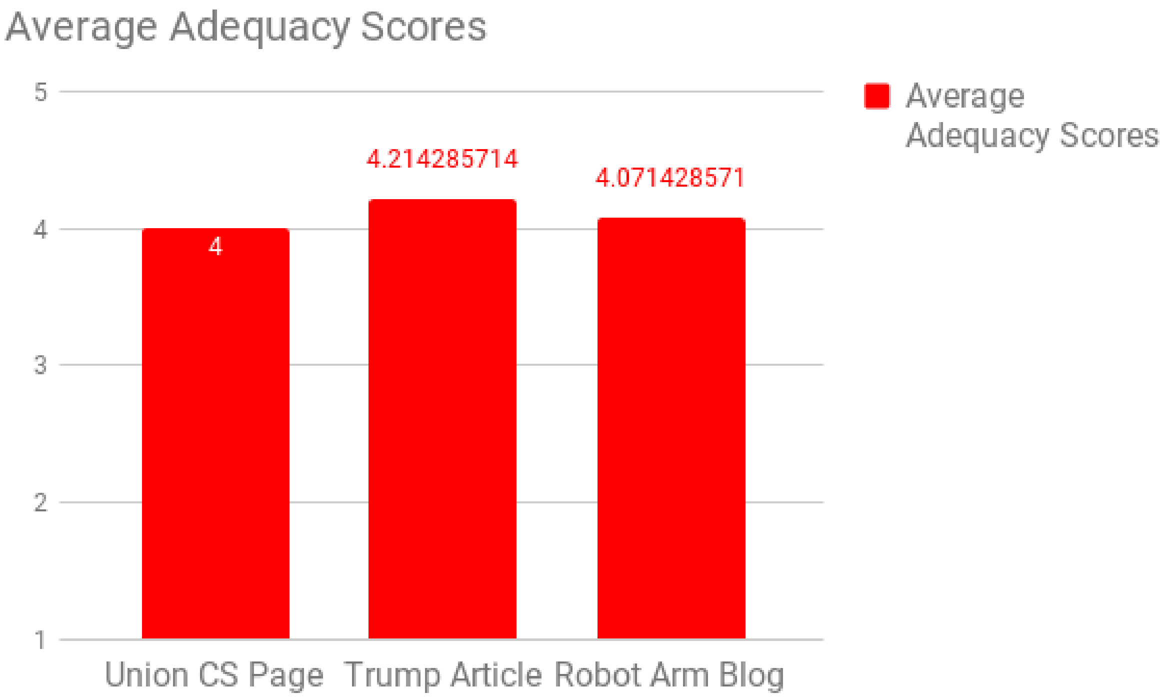


Figure 9: Graphs of averaged adequacy scores for each translation.

preliminary tools for translation when localizing software. The human workload is drastically reduced by only requiring editing from translators rather than starting the translation from scratch.

6.3 Qualitative Feedback

As has been stated, translations and their perceived quality by audiences is highly subjective. However, the suggested edits made by participants for each of the translations showed a common trend among the mistakes made by the MT system. These common mistakes included word choice errors, verb tense errors and word placement errors. This qualitative feedback is important for addressing the specific areas where an MT system does not generate satisfactory output to the target language.

The most common error that participants reported in their suggested edits to the translated texts were word placement errors. Given that this experiment focused on the fidelity of English to French translations, examples of these common errors will not be overly language-specific. An example of word placement error that was noticed by several participants was in the article about Trump's tweets. There was a sentence that started, "Last year," and this was translated to French as "La derniere annee." However derniere, the french adjective for last, has different meanings depending on whether it is before or after the noun it modifies. When derniere comes before a noun as it did in this automatically translated output, it would mean the last

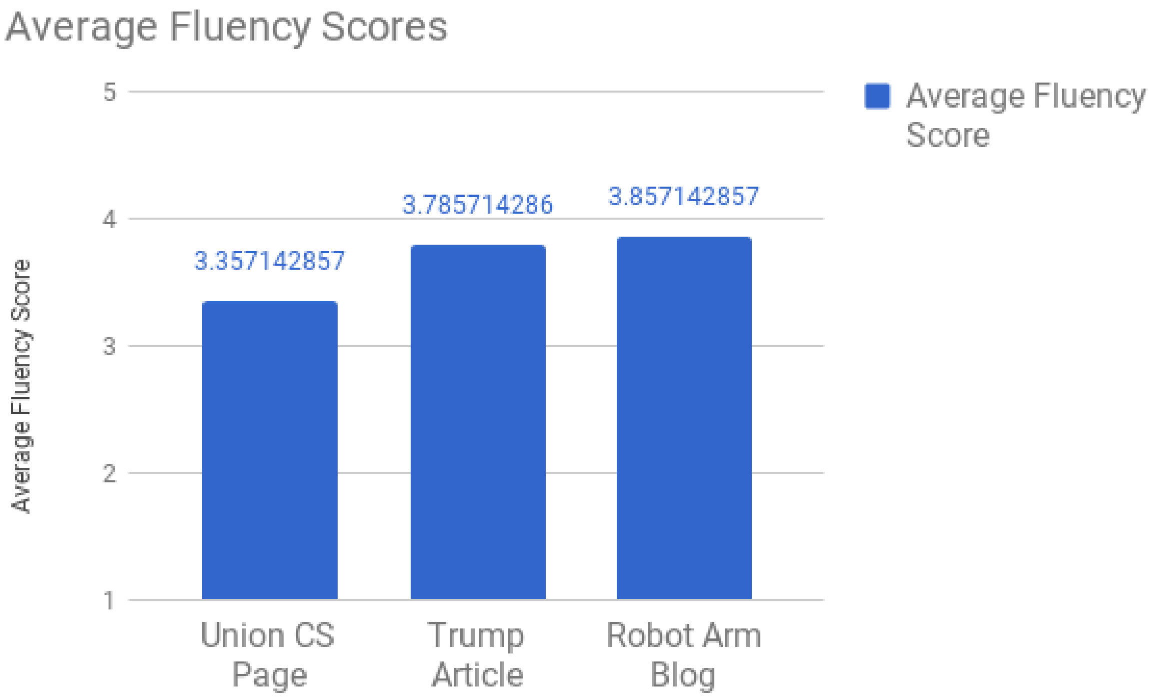


Figure 10: Graphs of averaged fluency scores for each translation.

year as in the final year, which would imply an end of times. The adjective should be before the noun and this text should be translated, “L’annee derniere,” which would mean last year as in the previous year. MT systems struggle with language-specific word placement such as this example.

On the MIT technology blog, there was an example of a verb tense error, which several participants picked up on. The webpage included a dropdown menu that was translated and one of the links in the menu was entitled, “Rewriting Life.” This was translated by the MT system as “Recrire la vie,” which is an infinitive verb form and is not proper french for representing the meaning of the phrase. The MT definition would literally mean “to rewrite life.” Rather this idea is more fluently translated by using a participle verb form. A correct translation of the phrase with the participle would be, “recrivant la vie” as the english input sentence also uses a participle verb. I believe the brevity and lack of context made this particular example difficult for the MT system.

On the Union CS web page, there is a line that ends with phrase, “support computing efforts,” which is translated as “supporte les efforts de calcul.” This is an example of a word choice error as “efforts de calcul” has no real meaning in French as a French-speaker would simply interpret it as “calculation efforts.” Instead a better and more fluent translation would be “support les efforts informatiques,” which makes the phrase specific to the domain of computer science and gives it real meaning in the target language of French. This error and the previous ones may appear minor, but they were noticeable by both native and non-native French-speakers. Thus, they would be most likely be noticeable in a software product or web application and must be addressed. From qualitative feedback such as this, the specific shortcomings of an MT system can be identified.

6.4 Bleu Scores

To calculate BLEU scores, www.letsmt.eu's free interative BLEU score calculator was used. The calculator took in two .txt files, a reference translation and the automated translation. The BLEU scores were all low when compared to the machine translated text in terms of their 4-gram BLEU score (sequences of 4 words). None of the BLEU scores earned over 30 on a scale of 1-100 with 50 being the benchmark for a quality translation that will reduce work for an editor. The Union CS page scored just over 19, the Trump article score just under 24 and the MIT technology blog earned a score of approximately 27. These lackluster BLEU scores is due to the small size of references used, as BLEU scores increase with the size of the test corpus [8].

The results of this automated metric are not shocking as the reference translations differed substantially from the translated texts in terms of word choice and way in which sentences and phrases were structured. There is no one perfect translation for a given input sequence and due to only comparing a small reference text with a short candidate text, the BLEU scores were not able to average out. Producing multiple reference texts for each automatically translated text will be a major aspect of future work in regards to this research. With a larger quantity of reference translations, I am confident that the BLEU scores would be higher.

7 Conclusion

The field of software localization encompasses the adaption of both textual and visual content to a specific target region and language. As this experiment only focused upon automating the role of translation in the localization process, it must be noted that only one aspect of web localization was explored. However, translation accounts for the majority of the human workload in the context of localization and therefore methods for expediting the translation process are important to the development of localization practices.

Although the automated translations were not on par with human translations based on feedback, many participants noted that they were surprised by the quality of the French translations produced by Google Translate. Additionally, all three translated web pages were rated as above average in terms of adequacy and fluency. All fluency scores were at least 4 on a 1-5 scale, which indicates that the automatically generated translations were able to preserve much of the meaning. Based on this feedback, I believe that MT systems could play a role in the localization process by significantly reducing the work for translators and editors. The MT output could provide an initial translation of English text in need of localization for a French-speaking region. This initial translation could then be edited as needed for the target language. As machine translation services continue to improve, MT could eventually be more than just a preliminary tool for the translation aspect of localization.

8 Future Work

Future work on this topic includes expanding all aspects of this experiment from the web pages translated to the amount of participants recruited. The scope of this project was limited by the population of French-speakers at Union College with professor of French and Francophone Studies, Charles Batson, estimating that there were only 20-25 French students, who had the advanced grasp of the language needed to meaningfully participate and provide both quantitative and qualitative feedback. Additionally, I believe that evaluation process could be improved by using MT systems besides Google Translate and providing mulitple translation outputs to participants. Related work has shown significant benefits in regards to providing two or more translation outputs when using MT sysyems. Finally, I believe using a more diverse array of web pages would be beneficial to seeing how MT systems performs in various language domains.

9 Acknowledgements

I would like to thank my advisors Nick Webb and Charles Batson, as well as my parents and siblings for their support and guidance throughout this project. I would also like to thank David Frey for always being a helpful and positive presence in the department.

References

[1] Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. "Adequacy-fluency Metrics: Evaluating MT in the Continuous Space Model Framework". In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 23.3 (Mar. 2015), pp. 472–482. ISSN: 2329-9290. DOI: 10.1109/TASLP.2015.2405751. URL: http://dx.doi.org/10.1109/TASLP.2015.2405751.

[2] Peter F. Brown et al. "A Statistical Approach to Machine Translation". In: *Comput. Linguist.* 16.2 (June 1990), pp. 79–85. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=92858.92860.

[3] David Filip, Dave Lewis, and Felix Sasaki. “The Multilingual Web”. In: *Proceedings of the 21st International Conference on World Wide Web* (Apr. 2012), pp. 251–254.

[4] Ge Gao et al. “Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* (Mar. 2015), pp. 852–863.

[5] Jesse M. Heines and Krzyztof Jassem. “Teaching Internationalization–Internationally”. In: *ITiCSE ’13: Proceedings of the 18th ACM conference on Innovation and technology in computer science education* (July 2013), pp. 93–98.

[6] W J. Hutchins. “MACHINE TRANSLATION: HISTORY AND GENERAL PRINCIPLES”. In: (Jan. 1994).

[7] W J. Hutchins. “The History of Machine Translation in a Nutshell”. In: (Jan. 2005).

[8] Kishore Papinehi et al. “BLEU: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (July 2002), pp. 311–318.

[9] Holger Schwenk. “Building a Statistical Machine Translation System for French Using the Europarl Corpus”. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT ’07. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 189–192. URL: `http://dl.acm.org/citation.cfm?id=1626355.1626380`.

[10] Jonathan Slocum. “A Survey of Machine Translation: Its History, Current Status, and Future Prospects”. In: *Comput. Linguist.* 11.1 (Jan. 1985), pp. 1–17. ISSN: 0891-2017. URL: `http://dl.acm.org/citation.cfm?id=5615.5616`.

[11] Xiaoyin Wang et al. “Locating Need-to-Translate Constant Strings in Web Applications”. In: *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering* (Nov. 2012), pp. 87–96.

[12] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* (2016).