HIP Domain Translate: A Linguistically-Grounded HCI Approach to Domain Adaptation

A THESIS PRESENTED
BY
REBECCA L. HAO
TO
THE DEPARTMENT OF COMPUTER SCIENCE AND DEPARTMENT OF LINGUISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
COMPUTER SCIENCE AND LINGUISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH 2020

Thesis advisor: Elena L. Glassman Rebecca L. Hao

HIP Domain Translate: A Linguistically-Grounded HCI Approach to Domain Adaptation

Abstract

Machine translation (MT) has enabled immense communication and shared knowledge throughout the world, yet it still falls short. Specifically, when these systems lack data in a particular domain, their accuracy plummets. Considerable work in data-based and model-based approaches have improved the accuracy of domain-specific MT, but these do not directly address ambiguities and decisions of style that are context-dependent. In this thesis, I present Human-Intelligence Powered (HIP) Domain Translate, a system that leverages post-edits made by users to machine translated texts by generating "fixes" within a domain for users to apply. Though interviews, the collection and analysis of post-edits, and a preliminary six-person user study on the prototype, I demonstrate that this system shows promise as a method to quickly and effectively capture contextual information within a domain that is missing in machine translation outputs.

Contents

Listing of figures

This thesis is dedicated to my parents, from and with whom I find love and curiosity.

Acknowledgments

I would like to thank my senior thesis advisor, Professor Elena Glassman (Computer Science), for her enthusiasm, questions, feedback, guidance, and support throughout this process. It has truly been a pleasure exploring the intersections of HCI and linguistics with Elena and fellow Computer Science and Linguistics joint concentrator Jake Cui, and working through each step of the research process! I'm definitely more excited about HCI and research as a result, and am looking forward to the future! Also, thank you to Professor Kathryn Davidson (Linguistics) for really eye-opening discussions, and your excitement, passion, and patience.

Even though I have dedicated this thesis to them, I wanted to again mention my parents, who have been with me and supported me every step of the way. Their belief in me, excitement about my growth and learning, and willingness to either dive deeply into and work through a topic of discussion together or just give me a big hug, is why I can be reasonably grounded yet also excited to explore.

Finally, thank you to my friends and peers who have given time, love, publicity, expertise, and support to both this project and me – I could not have done this without you all!

If we spoke a different language, we would perceive a somewhat different world.

Ludwig Wittgenstein

1 Introduction

In our increasingly connected world, translation *enables* connection. It allows for increased person-to-person communication – conversations across languages, cultures, and borders. It also enables the sharing of resources and knowledge by making work originally in different languages accessible to a larger audience – expanding our collective knowledge and supporting development.

Translation has been made significantly faster and easier through developments in machine translation (MT) (summary [8]), from the earliest rule-based systems to current systems that leverage neural networks. These approaches work very well when there is a lot of data, but fall short in domain-specific contexts where there may be less data [2]. This thesis is concerned with better understanding where and why these systems fall short in this domain-specific context, what effect this has on users of MT systems, and proposing ways in which we can address these limitations.

1.1 Motivation

The questions pursued and approach taken are largely inspired by an appreciation for the complexity and nuances of language – that though machines are becoming increasingly powerful and able to represent, process, and produce language that is more human-like each day, they lacks common knowledge and the ability to situate itself in context, gathering and making use of knowledge not necessarily present in the words themselves [10].

This limitation is particularly prevalent in the case of domain-specific translation, since there is generally less bilingual data available. This problem is especially significant given a general need for domain-specific translations – for example, in healthcare, there is a pressing need for machine translations in public health due to language barriers contributing to health disparities and time and cost required for language translation [22].

Given the limitations of machine translation, then, it seems relevant and necessary to involve human intelligence more directly. This is where human-computer interaction (HCI) can play a foundational role – concerned with study of how humans and technology interact and designing systems and interfaces by considering and involving the user, perhaps it is possible to involve humans to acquire this missing knowledge. This is not necessarily a novel idea – HCI approaches have been used to harness human intelligence generally, and have even been applied in the case of MT specifically (e.g. [3, 5, 7, 17, 20]). However, as far as I am aware and at the time of writing, there has yet to be an HCI approach that seeks to address and harness domain-specific intelligence in a way that benefits translators more generally across all domains.

1.2 Overview of Contributions

Overall, this thesis attempts to investigate and determine an approach to provide contextual information within a domain that current general machine translation tools lack, and argues for a human-centered approach that leverages human

intelligence (in the form of post-edits) that are understood through linguistic means. Though preliminary in nature, this thesis contributes:

1. Needfinding and discussion (Chapter 3) that reveals real-world problems of domain-specific MT on the users and how these relate to human language knowledge.

2. The collection and analysis of post-edits on translated texts in several domains and language directions (Section 4.3), observing patterns between post-edits and how they may be understood linguistically.

3. A prototype for a system (Section 4.2), Human-Intelligence-Powered (HIP) Domain Translate, that collects and learns "fixes" from these post-edits that can quickly improve a translation within a domain.[1]

4. A six-person user study to test this prototype (Section 4.4), that demonstrates the promise of this interface concept, in providing the MT system with actionable contextual information within a domain that makes refining these translations easier.

Together, these illuminate problems in domain-specific machine translation, and build up an approach to address domain-specific machine translation *generally* that leverages human intelligence through linguistic means.

[1] A demo for this system can be found at `bear.rebeccahao.com`.

Figure 1.2.1: The Vision for HIP Domain Translate

Within a domain, HIP Domain Translate takes edits made to translations and learns linguistically-inspired fixes, and displays these fixes to the user. The user (and subsequent users) can then apply these fixes to make the process of translating within a domain both faster and easier.

2

Background

This project lies on the intersection of a number of seemingly disparate research directions: machine translation (MT) and the problem of domain adaptation, human-computer interaction (HCI) and translation, bilingualism studies, and general findings in theoretical linguistics, in both syntax (sentence structure) and semantics (meaning of language). These interact because though there is considerable work to improve domain-specific MT, these have been primarily through innovations related to the data and models involved [2]. In this thesis, I suggest that an HCI approach that leverages human intelligence may be able to address certain problems in domain adaptation of MT, and that this human intelligence can be taken in and utilized based on our understanding of how language works.

2.1 The Problem of Domain Adaptation for MT

Machine translation refers to the automatic translation of text from one language to another. Given the complexity and flexibility of human language, this is not an easy task. Previously, this was done through rule-based systems [10]. Currently, there are two main approaches to machine translation – statistical machine translation (SMT) and neural machine translation (NMT). SMT leverages pipelines of statistical models whose parameters come from analysis of parallel bilingual text corpora, while NMT uses neural networks to learn an end-to-end model for translation.

Domain adaptation refers to the difficulty that MT has when there is little to no in-domain data. This is both a problem for SMT and NMT and has been addressed with reasonable success, yet still remains a problem and continues to be actively explored [2].

In SMT, a number of techniques have improved performance when there is very little bilingual in-domain data available, a central approach being the interpolation of out-of-domain (with lots of bilingual data) and in-domain (with little bilingual data) language models and biasing towards a specific domain [11]. From there, there have been a number of approaches to further improve performance by leveraging different kinds of data: using in-domain monolingual data, by generating synthetic parallel data by translating monolingual data [1], and using out-of-domain phrase pairs, by weighing them according to their relevance to the target domain and factoring in whether they appear to be related to the domain or general language [6]. Additional approaches look to find more data to use, based on where in-domain models falls short, including mining comparable corpora to find translations of previously unseen words [4].

In NMT, many of the above techniques have also been applied, like the use of monolingual corpora, synthetically generating parallel corpora, and using out-of-domain parallel corpora [2]. Additional approaches, however, involve fine tuning, or training on a rich out-of-domain corpus and then tuning parameters using a resource-poor in-domain corpus [2], or making use of data that is most

similar to in-domain sentence embeddings (using these representations of the sentence that are internal to NMT) [23].

While numerous with unique solutions to domain-specific MT, these improvements for both NMT and SMT seem to be largely data-centric or model-centric, and given the continued work, likely do not fully capture what can be learned and used regarding language.

2.2 HCI Approaches to Machine Translation

A different perspective to consider is that of human-computer interaction (HCI). For machine translation specifically, a number of novel interfaces have been proposed that use mixed-initiative approaches to improve translation, involving human input in different ways and at different points in relation to the machine involvement. Post-editing involves the correction of errors in MT output. There are text-based and graphical post-editing interfaces that learn from edits and reduce effort and time [20]. Predictive Translation Memory (PTM) presents a system for translators to build translations incrementally by considering AI-based machine suggestions that update based on the current state of the translation [7]. Interactive Neural Machine Translation Prediction (INMT) assists translators with on-the-fly hints and suggestions to make the translation process faster and more efficient [17]. Other improvements to MT have come in the form of incremental updating of MT models based on user input [25] and providing explanations for suggestions that an interface gives to improve a translation [3].

There have also been interfaces designed for *specific* domains, though these are less numerous. One example, PHAST, is a collaborative MT system that was designed for "laypeople" in public health (bilingual public health professionals with a deep understanding in public health but not in linguistics or technology) [5]. It is collaborative in that separate individuals can log onto the tool and all contribute post-edits to shared documents.

However, at the time of writing and as far as I am aware, an HCI approach has not yet been used to address the specific problem of domain adaptation for

machine translation more generally. While the first mixed-initiative approaches may implicitly address certain issues with domain-specific MT (i.e. better leveraging of human intelligence), and domain-specific interfaces address problems inherent to translation in a single domain, there does not seem to be a more general-purpose solution for non-professional translators that can bring improvements to in-domain translation across domains, robustly and effectively.

2.3 MT and Linguistics

While machine translation and domain adaptation involve language and have high impact in computer science academia and industry, at least how I understand it, beyond older linguistic rule-based approaches to MT, they exist quite separately and distantly from linguistics. First, machine translation involves writing and modifications of writing. Linguistics, however, is more centrally concerned with spoken and signed language – language that arises *naturally* from people and tends to change over time. From a linguistic perspective, writing is not natural language but an *invention* by humans, with filtering through sometimes prescriptive rules. Also, machine translation involves *translation*, which involves a directed task of moving from one language to another rather than natural use of several languages (bilingualism studies), which again moves us further away from the spontaneous utterances that come naturally from humans without instruction. Yet, in this thesis, I have claimed that linguistics has a *central role* in an HCI approach to domain adaptation – how does it, despite this distance I have described?

2.4 Human Language Errors

I think one way of initially unpacking this is to consider human language errors, and how they may parallel MT errors. If there are parallels between human language errors and MT errors, perhaps there are parallels between other characteristics of human language and what an MT system might be doing, and

in this connection maybe we can understand an MT system's errors through how language works in humans. Given that native speakers rarely make errors, or at least we do not perceive them to in a descriptive (rather than prescriptive) linguistic perspective, we turn then to second language (L2) speakers.

Work by Schwartz and Kroll investigates L2 speakers' word recognition of language-ambiguous words (i.e. cognates and homographs) [18], and observe cross-language lexical competition. Intriguingly, they find that this effect can be influenced by constraining the sentences: both languages are active in low-constraint settings while the effects of this competition are eliminated in high-constraint settings [18]. In some ways, this seems to parallel our discussion of domain and language error – even when there is a certain degree of error, this may be mitigated by constraining the context.

This also suggests a larger interaction between MT and bilingualism – namely, that proficiency in a language may have profound impacts on bilingual processing, but also on translation. This suggests that errors as a result of lower proficiency in a language may be relevant in interpreting MT translations.

Previous work demonstrates that speakers with lower proficiency may have more difficulty when it comes to resolving ambiguity (which requires an understanding of the context) [24]. On the flip side, syntax seems to develop into a shared representation as proficiency in L2 increases [9], which may lead to interference between languages. Rather than being able to extrapolate a trend in this behavior and how it might map to MT translations (because this relationship, if any exists, is not clear), I think this demonstrates the potential of ways one's language may operate internally and the dimensions in which it may have difficulty. While these findings may not directly map to MT specifically, it suggests directions of understanding what MT might be learning, and gives a space of possibilities of what it might not be (or, alternatively, if some kind of bilingual knowledge or data interferes with something else).

Combined with the earlier discussion of lexical ambiguity, these findings suggest that perhaps we can begin by paying attention and understanding the MT through the lenses of ambiguity, the lexicon (specific words), and syntax

(sentence structure).

2.5 Generalizable Linguistic Rules

After establishing a potential parallel and probing the idea that maybe we can understand an MT system through, this takes us to the promise of using generalizable linguistic rules, in order process edits and make improvements across meaning (semantics) and structure (syntax).

There is rich literature in each of these fields in how we humans represent, understand, process, and produce language. Rather than attempt to discuss the vast work in these large areas of linguistics, I will illustrate a specific example of each to suggest potential relevance to the question we explore here: whether certain rules can be learned from edits and those can be made useful in translation.

- **Ambiguity**. Meaning or interpretation of expressions are context-sensitive when their truth relies on the context. How do pronouns get their meaning? Demonstratives (words like "this" or "that")?
- **Lexical ambiguity**. Words can have different interpretations depending on the context. A simple example is that in the sentence "The bat flew away," a ball and bat could be flying together out of a stadium, or a nocturnal mammal could be flapping its wings.
- **Syntax**. Since properties are different in languages, maybe we can determine systematic differences between them. For example, in the case of binding theory (whether certain words can refer to each other based on the sentence structure), pronouns, referential expressions, and anaphors behave differently in different languages.

The first two point toward an importance of context in determining a contextually correct translation, and syntax is an example of a systematic way we could go about processing an MT output.

One way of understanding this discussion of how machine translation, HCI, and linguistics may relate to each other is that I am arguing that a human-centric approach may need some medium for the human intelligence to be understood – namely, in these linguistic rules that we have discovered about ourselves. This is distinct from a general machine learning approach, and is also distinct from needing to "hard code" a number of linguistic rules. Though it remains an open question exactly how we will go about doing this, I investigate whether there is promise in combining these two approaches.

2.6 Learning from Mistakes

Finally, a last relevant topic is existing work in learning from mistakes. Several efforts have furthered our understanding of improving systems involving error correction. For example, similarly to how there is considerable effort in developing translation systems with increasing accuracy, yet translation is inherently ambiguous, as is handwriting – with a novel interface CueTIP, Shilman et al proposed a mixed-initiative system that evolves its results from user corrections, reducing the costs of correction over time, and suggest a number of high-level design principles for mixed-initiative correction interfaces [19].

The argument that attention to errors is useful has also been furthered in MT research. Specifically, error analysis (identifying and classifying MT errors) has been argued to be integral to MT development in that it gives a qualitative view not visible in standard evaluation methods [21]. Post-editing itself does not promote the learning from mistakes, but several interfaces learn from the edits translators make for improvements to the translations and MT models (e.g. [20, 25]).

This demonstrates that learning from mistakes may be promising, and that there are design recommendations that can be considered in relation to this approach.

Putting all of this together, though there have been a number of innovative

interfaces designed for the task of translation, at the time of writing and as far as I am aware, an HCI approach has not yet been used to address the specific problem of domain adaptation. While machine learning and statistical techniques have made progress in addressing the problems of domain adaptation, I am arguing for an HCI approach that makes use of both human intelligence and linguistic analysis that may also begin to address these problems in new ways.

3

Problem Definition

To begin to approach the broader question of users' needs regarding translation and how they relate to translation tools, I performed several needfinding interviews and contextual inquiries. Only by first understanding these needs can we then suggest and demonstrate a solution to them.

3.1 Needfinding

I interviewed and conducted contextual inquiries (combined taking from 30 minutes to 1 hour and 30 minutes) on four individuals (N1-4), each of whom translates semi-regularly but would not consider themselves professional translators. We are interested in this population because we are interested in domain-specific systems that are more general and do not depend on high involvement of domain-experts and professional translators. Rather than seeking

to optimize performance of professional translators specifically, we look to provide a more theoretical suggestion to how to approach domain-specific translation more generally. The individuals varied in occupation and domain-expertise, ranging between a high school student who frequently translates poetry (N3), to educators translating education and development-related texts (N1, N2), to a director in software (N4). Three of the interviews were conducted in Mandarin so quotations from these individuals are translated by me (N1-3), while one was conducted in English (N4). Additional information about the individuals can be found in Appendix B.1.

3.1.1 Interviews

Though these interviews, it was immediately apparent that translation fulfills a number of diverse needs. First, documents may need to be translated for work purposes: for example, N1 has translated American preschool curricula, N2 has translated book chapters, and N4 has translated both user manuals and urban planning documents for grants. However, translation is also useful for learning purposes – N4 likes to read articles and websites in different languages, while N1 also enjoys reading English texts and translating them to understand them more deeply (since Mandarin is his native language). Translation can also be used for basic communication (N4 sometimes translates meeting minutes and Q&A's). Further, translation even can be used for artistic appreciation and enjoyment – N3 translates long poems. These examples are in written form, but translation in a number of modalities is also common – examples include being a real time interpreter (N2) or doing translations of video into subtitles (N2, N3).

It seems like the main goal of translation, then, is to get the semantics and meaning across in a way that is useful to the intended audience. This may vary depending on the audience and need – work materials may require higher quality and attention to detail than glancing over a text for learning purposes. However, how do we make the translations useful to the intended audience? N1 suggests that a translation should not only translate general meaning, but also do so in the

way the target language usually conveys information. For him, his translations to his students are effective because he communicates the ideas "as a Chinese person would express them," so they are easier for his Chinese students to process.

Because of the discussed variance in needs, translation tools like Google and Baidu Translate are used differently as well. For example, N1 actually does not use translation tools when trying to learn on his own from books, while N4 uses the Google Chrome feature that translates an entire web page for you. Some use a computer (N1, N2, N4), and some use mobile phones (N3, while N1 and N2 use phones for OCR to scan text). Depending on the need, these translation tools can be used to translate paragraphs at a time, or sentences, or individual words.

They are useful because the vocabulary is there, the text is already mostly translated, and you "only need to adjust and move things around" (N1). Especially for individuals like N1, who mentioned that he is less confident and slower at speaking and writing in English, lessening cognitive load and improving speed would be greatly beneficial.

Yet despite these pros, translation tools still run into a number of problems: N1 aptly describes that he feels like there are two main challenges, the first being "word choice" and the second "background." Specifically, he and other individuals (N2) felt as though the vocabulary was often not selected correctly, and N4 mentioned how when there's a lot of jargon in a translation she will frequently Google (not Google Translate) the words. The magnitude of this impact is furthered by N1's statement that "I believe that [machine] translation will never reach the level of a human, because it would need to understand cultures, differences in meaning, while preserving style, feeling, and even humor – this is where we need people's creativity." The seeming intractability of this artistry is furthered by how when the "original line is too pretty" in a poem, N3 revisits her translation of that line frequently, five to six times, continuing to look at and revise it.

Intriguingly but perhaps not surprisingly, this conceptual idea of "background" is not just important to translation tools, but to translators themselves. Specifically, language proficiency and an understanding of background is also the

case with people. For example, N2 described when how going from Chinese to English, despite her proficiency in English comprehension, she has difficulty translating because she does not know the connotations and typical usage of a number of English words. This parallel supports the claim that "background" is important to our translations. And similarly to the idea of pragmatics discussed in Section 2.5, there must be ways to represent and unpack "background."

These interviews also highlighted what the translation tool does not know – it is clear that it can learn things about language and translation implicitly, but at least it never explicitly asks for any additional information about the user or the text it is translating (noted by N1).

Overall, these interviews demonstrate that at least how the users understand it, the translation tools do not fit their needs in terms of having the correct vocabulary or style, which they feel could be further informed by context.

3.1.2 Contextual Inquiries

After each interview, we also did a contextual inquiry. Contextual inquires involve asking the participant to perform a task as they would normally, often narrating along the way and open to questioning from the researcher. We frequently ask "why" someone is doing something in order to better understand their motivations and obstacles. Some problems may be difficult to articulate but are visible in practice.

Though less directly relevant to the question at hand, contextual inquires illuminated the magnitude of logistical tasks and difficulty individuals encountered. Doing an OCR from a book passage (N1), or typing out a poem from a phone image (N3). From there, the OCR includes line breaks, so inputting that into Google Translate led to initially extremely inaccurate and strange translations. Additionally, it showed a diversity of solutions in terms of where to keep your running translation, whether in a document (N1, N2) or a phone notes page (N3). I do not aim to address these needs in this thesis, but it is informative to know more about what considerations people looking to translate

texts have beyond the direct interaction with the MT. It is possible that these will influence their interaction and subsequent use of the MT.

Additionally, we again observe a difference between language directions. Namely, that depending on familiarity of the cultural knowledge and context, it also becomes more difficult for human translators. For example, when translating from English to Chinese, N1 would go one sentence by one sentence, trying to understand the meaning of each, while in Chinese it was relatively easy and fast to know what is said so he would look at the paragraph as a whole. In contrast, N4, who is fully fluent in English, when going from Chinese to English, would go paragraph by paragraph with no problems.

3.2 Defining "Domain"

Despite how intuitions that ‘factoring in a domain’ seem reasonable, it is not immediately clear how we should go about defining these "domains." From relevant work, a possibility is to define by specific categories of texts like news commentary or parliamentary speeches [11].

For the purposes of this preliminary exploration of these questions, however, I am curious about our intuitions related to broader domains. From the initial needfinding, it was brought up that the participants felt as though the vocabulary and background required seemed to relate to broader "fields." This may not necessarily hold, but we investigate it here with the intention of refining this definition (which to me, seems like it could be its own research question) at a later date.

Specifically then, in this thesis, we seek to address the MT's shortcomings, which include the problems of insufficient word choice (especially given the context), and the background (including style, structure), We do this by unpacking whether considering a domain or language structure is useful. Perhaps it is possible to still utilize people's creativity, language knowledge, and domain knowledge in an efficient way to quickly address these shortcomings.

4

HIP Domain Translate

In response to the gaps in translation tools and MT, I present **Human-Intelligence-Powered (HIP) Domain Translate**. Beyond a traditional translation interface, it includes "fixes" that can be directly applied to the text that modify the words used, structure, and other details. It takes in users' post-MT-translation edits within a domain to learn these fixes, that both themselves and later users may wish to apply. In this way, users are able to arrive at accurate translations more quickly, and better capture some information that a general translation tool is unable to capture, like specific vocabulary, phrases, or structure that are used in a particular domain or context. Crucially, this choice, while not necessarily as novel of a mixed-initiative task, allows for users who are not necessarily looking to deeply engage with an interface to also benefit from improved translations.

4.1 Scenario: Interacting with Domain-Specific Fixes

Some things are best illustrated by example. Consider two individuals, Eve and Victor, who are both interested in translating texts about education. Eve has a lot of experience translating and considers herself an English-Mandarin bilingual, and is an expert user of HIP Domain Translate, while Victor is newer both the language of Mandarin and the tool.

Eve can enter the tool, and select the domain of "education." With an interface similar to existing translation tools, she can copy or write text that she wishes to be translated into one text box, press "Translate," and receive an MT translation in the second text box. From there, she can edit as she wishes, and when she is satisfied with the translation she can press "Done editing." She sees several quick fixes she can turn on and off on the side and uses them occasionally to replace all instances of a word.

Victor, on the other hand, enters the tool, and also can input the text he wishes to be translated. From there, he explores the quick fixes, toggling them on and off to see what it does to the translation, and keeps a number of them on to immediately improve the quality of the translation in an education context. Feeling satisfied that the semantic meaning is correct and certain contextual information is preserved (i.e. that it makes sense in the education context), he then quickly edits the text for flow and quickly checks up on the grammar.

Critically, users can choose to contribute to the quick fixes in the domain, or they may just use the tool as-is. Very quickly, learning automatically from edits within a domain, the system is able to provide quick fixes that are specific to the domain. In this way, domain-specific information like word choice, style, and means of communication can be accurate to the language at hand.

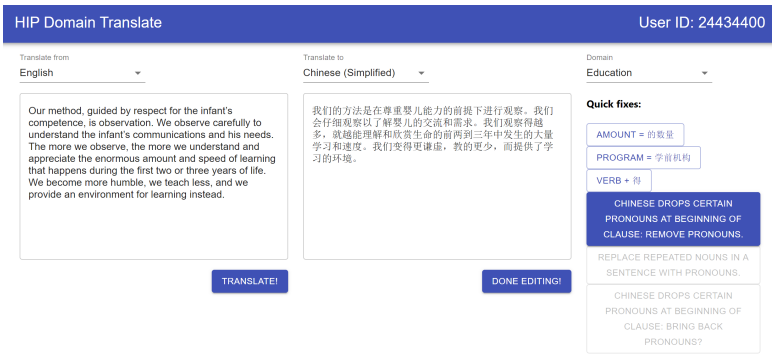

Figure 4.1.1: HIP Domain Translate Interface

4.2 System Implementation

4.2.1 Application

HIP Domain Translate is a full-stack MERN application built using a **R**eact frontend, **E**xpress (**n**odejs) server, and **m**ongoDB database. The React frontend uses Material UI [1], renders the text areas, selections, and buttons, saves cookies of any relevant data input or output, uses the server API to connect to the database, and performs translations on text using Google Translate API. [2] The server is configured to connect to the database, and send critically the (1) original translation from Google Translate and the (2) translation after the user has edited it, so we can compare them (it also sends additional state information for context, but this is less critical). The application is deployed to Google Cloud Platform's App Engine. [3] The code can be found in a Github repository. [4]

[1] https://material-ui.com/
[2] https://cloud.google.com/translate/docs
[3] https://cloud.google.com/appengine
[4] https://github.com/beccahao/hiptranslate

4.2.2 Fixes

Though the vision for the fixes is that they will be learned through linguistic processing (as described in the scenario above), for the purposes of this thesis I use a prototype that involves linguistic rules determined ahead of time, selected after analyzing the data discussed in the next section. Therefore, in this case then, my own analysis is taking the place of the fix-learning the system would in theory do, in an effort to understand whether these fixes are useful within a domain.

Specifically, the fixes that were included were as follows (please see Appendix A.3 for the specific passages that were used).

1. **Lexical replacement.** Translate "amount of" to "的数量" (a more general word) and "Program" to "学前机构" (a word that captures the educational version of a word with many possible meanings).

2. **Verb + 得.** In Chinese, there is a distinction between three /də/'s (Standard Mandarin pronunciation), 的 for noun modification, 得 used for verb modification, 地 for compound modification (adjectives). Some translations used the incorrect /də/, so a rule automatically fixes it after verbs.

3. **Replace repeated nouns.** After the first time a referential expression is mentioned, in English typically you would use pronouns to refer to it based on rules in binding theory (see Background Section 2.5 for more details).

4. **Dropping and adding pronouns.** In Chinese, due to zero anaphora [13], there are less pronouns in the beginning of certain clauses in Chinese than English would. This fix removes certain pronouns when going from English to Chinese, and adds a "<pronoun>" to the text where a pronoun should go when translating from Chinese to English.

I included fixes that varied in domain-specificity to investigate how this impacted engagement with and perceived relevance of the fixes.

We use linguistic reasoning largely because it gives an interpretable framework for understanding what our language generally looks like, and to quickly pinpoint errors with high confidence. This is distinct from the idea of "hard-coding" linguistic rules that was prominent in natural language processing but now has moved towards machine learning. Here, the linguistic rules are a way of understanding our inputs of human intelligence (edits to MT translations).

4.3 Data Collection

In order to better understand what kinds of fixes might be useful, to see if there may be "domain-specific" edits, and to "simulate" the learning process from many edits to arrive at the fixes for the prototype as described in the previous section, I collected preliminary data on edits to translations within a domain. To do this, I sent out an anonymous Google form through various university mailing lists that collected general information about the individual filling out the form (their language proficiency, domain(s) of expertise, occupation), instructions to provide edits to translations, and several optional follow-up questions (requesting experiences, feedback, and any resources they may have used when completing the task).

When providing edits to translations, individuals were asked to either use a generated passage for the language pair (e.g. English to Chinese) and domain (e.g. "Education") or supply their own passage they felt was part of the domain. They were then asked to press "Translate" and a Google Translate-generated translation would appear. After that, they were asked to edit the translation until they felt that the translation sounded natural and was semantically accurate.

4.3.1 Results

Overall, I received 45 translations, from a total of 22 distinct individuals (see Appendix B.2 for more information about them. The form was anonymous, with randomly generated user ID's to identify the individuals). Seven were from English to Spanish, ten were from English to Chinese, 14 were from Spanish to

English, and 14 were from Chinese to English. Since it was up to the participants what texts they submitted edits to translations for, there was a large range of texts: many from the sample paragraphs in the domains of education, psychology, business/economics, literature, social studies, and healthcare/medicine. Though I did not make it so that people could suggest their own domains, I would likely do this in further investigations (and/or do a separate experiment to determine what is a reasonable domain, as discussed in Future Work, Section 5.1). Despite this, an individual (C12) provided a translation for a passage on music, furthering the potential we have begun to observe in understanding the user.

4.3.2 Analysis

There was a great degree of variation across translations, even within the same text. Several notable observations:

1. **Different languages and directions of translation have very different degrees of editing**. Notably, Spanish to English required *very little* modification, if any at all (5/12 identical, non-identical were literature passages 5/12 or required very few simple edits, like changing an article from "a" to "the" or a general, not domain-specific word substitution), while the other language directions required significantly more edits.

2. **Literature in general almost always required restructuring of sentences and word changes**. This seems intuitive, given that there is special attention to style and artistic flourish in literature. I mention this because it provides a contrast with more technical domains, and potentially points to limitations in translation tools that is related to demands of style. How much does style matter within a domain? How much can, and should tools attempt to recreate style? This has begun to be explored by work that shows that translators prefer manual translation of literature over SMT or NMT [14].

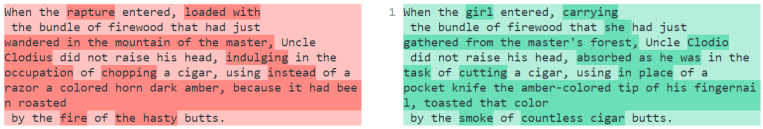

Figure 4.3.1: Edits to a Spanish to English Literature Passage Translation

MT on the left, Post-edited translation on the right

3. **More technical domains seemed to require less edits**. This may be surprising, but it looks like machine translation does very well in technical texts with a decent amount of domain-specific words. This could be explained in that these more technical texts seem to operate okay with the same structure, and can be translated word-by-word.

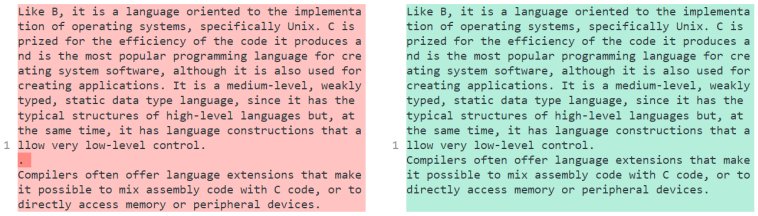

Figure 4.3.2: Edits to a Spanish to English CS Passage Translation

4. **Domains can disambiguate words that may have multiple meanings.** However, the problem seems more to be with words that could have multiple meanings. For example, 经历 /dʒɪŋ.le/ and 体验 /te.jæn/) are both "experience" in English. However, the former refers more to a memory, or that someone has "gotten through" something, while the latter refers more to an interactive experience a person is engaging in, which is more relevant for education (e.g. hands-on activities).

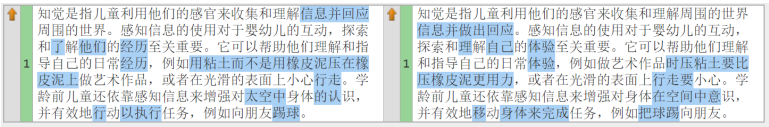

Figure 4.3.3: Edits to a English to Chinese Psychology Passage Translation

"Experience" can take on two different words – in the education context, 体验 (/te.jæn/) makes more sense.

5. **Phrases may be more domain-specific than words**. Though "phrase" is a vague description, frequently, translation tools seem to generate roundabout phrases to convey phrases that succinctly can be described in domain-specific language. For example, that "from the perspective of brain science" can be more aptly "from the scientific perspective", or "pulse of the students" to "the students' heartbeat" (or, as one participant suggested, "the students' vibe").

Overall, this data collection and analysis suggests that considering the domain a translation occurs is valuable, and that there is potential for generalizable rules based on the edits made. Specifically, it suggests that rather than providing domain-specific vocabulary more generally, domains are more useful to disambiguate words that may have multiple meanings. Additionally, domains may have their idiosyncratic ways of conveying certain words (e.g. like how phrases can almost act like proper nouns). If you would like to look around, you can find the diffs of translations at `bit.ly/hipcollect-diffs`.

4.4 User Study

From there, I performed a user study to test the prototype of HIP Domain Translate, in order to investigate the usefulness of the fixes, domain-specific or language-specific.

4.4.1 Participants

I recruited six Mandarin-English bilinguals[5] through various university mailing lists and peers' families. Four participants (P2, P4, P5, P6) were Harvard undergraduate students (within the age range of 18-24 years old), and the other two (P1, P3) were working adults (within the age range of 45-54). Despite how translations and their edits as described in Section 4.3 were recruited through similar channels as this user study, prior to the study, none of the participants had used HIP Domain Translate before. All participants had self-reported English levels of "native or bilingual proficiency" (P2-6) or "full professional proficiency" (P1) and self-reported Mandarin levels of "native or bilingual proficiency" (P1, P3, P4, P5) or "full professional proficiency" (P2, P6).

Participants were split into two groups: a control group (P2, P3, P4) and an experimental group (P1, P5, P6). These groups were balanced for age (two undergraduates and one working adult in each), gender (two females and one male in each), and domain expertise (an individual in each falling under social sciences, natural sciences, and computation). For more information about the specific participants, please reference Appendix B.3.

4.4.2 Methodology

I conducted a 45-minute user study with each participant that looked to understand each participant's language histories, translation experience, and investigate what kind of effect these fixes in the interface had on their translation experience. Specifically, the study was structured as follows:

1. **Initial interview** (10 minutes): First, I asked them about their language history and use, a more refined description of their language proficiency (qualitatively describing skills in speaking, listening and comprehension,

[5]The distinction between Mandarin and Chinese is used because Chinese can refer to the overall language (including written) while Mandarin can refer to the standard dialect of spoken Chinese.

reading, and writing separately), and any experience with translation and translation tools.

2. **Translation tasks** (25 minutes): Next, participants were asked to translate 2-4 different passages depending on the time it took (most translated 3). The order of passages was randomized across participants. All passages were within the "education" domain, and two were in the English to Chinese direction while the other two were the Chinese to English direction. Participants were allowed to use external tools.

3. **Post-task interview** (10 minutes): Finally, participants were asked several follow-up questions, including questions about satisfaction with the translation task generally and the tool itself, what they found challenging and enjoyable, and about their perception and understanding of the MT's accuracy and errors it makes (if it is making errors). Participants in the experimental group were asked about how they think the fixes could fit into a translation workflow, and for their reactions to the fixes themselves.

To elaborate on several of the decisions made above, I modified the initial interview to include more fine-grained detail about language proficiency because from the form that participants filled out to sign up for the study, there were several indications of differences between speaking/comprehension skills and reading/writing skills. Namely, that some participants self-report "native or bilingual proficiency" when they have strong speaking/comprehension skills, but weaker reading/writing skills. This is not unexpected, given the discussed differences (Section 2.3) between natural language (spoken or signed language) and written language (the invention we use in machine translation).

For the translation tasks, the passages were selected from the "education" domain because "education" texts are generally accessible to a wider audience, while still being within a domain with its own patterns and tendencies. Additionally, from the data collection study, the language directions and domain I received the most data from were English to Chinese and Chinese to English

"education" passages (perhaps because of this accessibility to a more general audience), so I had the most information to select useful fixes to prototype.

External tools were allowed in order to simulate the actual environment users would be translating in. Previous work has suggested that using a realistic translation environment will lead to more realistic and useful results [12].

4.4.3 Results

Due to the small number of participants, the quantitative results of the user study are not significant and the qualitative results may not be representative. However, these preliminary quantitative and qualitative results show several improvements in the experimental group over the control group, suggesting that these human-intelligence-powered and linguistically inspired "fixes" are a promising direction in improving machine translation. I will also discuss constructive feedback on the tool and the implications this feedback has on our design.

Quantitative Analysis

Time Taken There was a large amount of variation in the time it took to complete the translations. On average, the control group completed the tasks more quickly. Given the low number of participants and the differences between them, it is difficult to confidently conclude anything from this. The problem is further exacerbated in that there was an inconsistent number of translated passages in each group, due to time constraints (initially aimed to do four, but most participants only had enough time to complete three, while one (P5) completed two). In particular, it seems like language proficiency and interpretation of the instructions had a large effect. For example, because P1 was the only one more comfortable going from English to Chinese (due to more comfort in Chinese) her results are generally the opposite in terms of time required from those who felt more comfortable translating from Chinese to English (everyone else). For differences in interpretation, though participants were encouraged to work as quickly as possible but maintaining accuracy, it was

clear that some valued accuracy and effort while others valued speed. This could be confirmed or denied by comparing time taken and accuracy, but due to the variability and incompleteness (unable to complete all translations due to set up), I have not done this quantitatively (see following section for a qualitative interpretation). Future iterations would likely require refinement – I think the effect of differences in interpretation would be decreased with a larger group of subjects, while differences in language proficiency should be better standardized across experimental groups (and/or only do one direction of translation). There should also be more time in study design to ensure the number of tasks can be completed in the allotted time.

A potential explanation for these observation beyond what has already been discussed may be that users in the experimental group spent additional time exploring and playing with the novel interface elements (the quick fixes). Perhaps this suggests that the study design may need to change, in that exploration time to acquaint oneself with the interface may be important.

	P2	P3	P4	Avg Ctrl	P1	P5	P6	Avg Exp
Passage 1	5:12	2:00	5:04	**4:05**	3:24	10:02	4:28	5:58
Passage 2	11:10	8:20	-	9:45	5:00	-	9:01	**7:00**
Passage 3	6:25	-	4:55	**5:40**	-	8:24	-	8:24
Passage 4	-	4:02	4:00	**4:01**	12:12	-	3:26	7:48
Avg	7:35	4:47	4:39	**5:52**	6:52	9:13	5:38	7:17

Table 4.4.1: Time Taken

Translation Accuracy These values are very preliminary, but to get a sense of approximately how accurate the final translations were, I computed a BLEU score [16] for each translation, using `sentence_bleu` from `nltk.translate`. Translations were compared against a human translation (performed by me with assistance) and the results are reported here. Despite its simplicity (based in n-grams, i.e. word similarity), BLEU has been shown to correlate quite well to human evaluation, but it is disputed whether BLEU is a useful metric, and again,

given the number of subjects in this study and that there is only one reference translation, the interpretations should be taken with low confidence. Additional reference translations would likely improve this, and there are a number of other potentially useful metrics that factor in a more holistic evaluation, like metrics that seek to be more robust by factoring in linguistic features like textual entailment [15].

Interestingly, on average, the group that was the most accurate for each passage was also the fastest. This could be explained by relative proficiency, could be fluke given the methods, and/or it could point to a correlation between accuracy and speed.

	P2	P3	P4	Avg Ctrl	P1	P5	P6	Avg Exp	MT
Passage 1	0.18	0.17	0.25	**0.20**	0.22	0.19	0.14	0.18	0.18
Passage 2	0.19	0.17	-	0.18	0.30	-	0.19	**0.24**	0.14
Passage 3	0.14	-	0.13	**0.13**	-	0.10	-	0.10	0.12
Passage 4	-	0.24	0.28	**0.26**	0.24	-	0.15	0.19	0.24
Avg	0.17	0.19	0.22	**0.19**	0.25	0.14	0.16	0.18	0.17

Table 4.4.2: Translation Accuracy

Satisfaction Generally, it seems like despite it on average taking longer, the experimental group participants both enjoyed themselves more while translating and had a more positive reaction to the tool. Here, a potential confounding factor may again be the relative comfort level of the participants in a particular language. For example, several participants (P2, P5) mentioned frustrations in their own Mandarin ability. However, this is further confounded because several participants (e.g. P2) cited the experience of learning while translating as enjoyable.

Reliance on Other Tools Though not a purely quantitative measure, degree of use of other tools may also be correlated with the difficulty of

	P2	P3	P4	Avg Ctrl	P1	P5	P6	Avg Exp
Process	5	4	4	4.33	5	4.5	5	**4.67**
Tool	4	4	4	4	5	4.5	4	**4.5**

Table 4.4.3: Satisfaction with Translation Task and Tool

translation. This felt relevant especially because Google Translate was used by a number of participants for help with comprehension of Chinese (used either for the English translation, or more commonly, the phonetic alphabet spelling). From the screen recordings of the participants and observation, multiple approaches were taken. Several participants would frequently put single words into Google translate to get their pronunciations (P2) or definitions (P4). Others generally avoided additional tools and mulled over certain words (e.g P6). Perhaps for more generalizable findings, additional study should impose more constraints on resource usage (and/or track it more rigorously). Otherwise, reliance on resources could also have profound impacts on the other metrics in this study (e.g. switching between resources could incur a cognitive cost, or it could make participants both faster and more accurate by quickly providing solutions to their needs).

Qualitative Analysis

An analysis of the participants' responses reveals that the interface and inclusion of fixes is usable and useful, though it may not directly map to increased understanding of the MT's problems.

Tool Usage and Usability All experimental subjects engaged with the fixes, and expressed positive sentiment toward the potential for fixes. In terms of pure usability, participants across groups found the translation interface simple and easy to understand (P1, P3), and participants in the experimental group found the buttons easy to use (P6) and visually appealing (P5). All members of the experimental group expressed that they felt like the fixes were helpful in improving accuracy (P1, P5, P6).

An issue that arose, however, was that for some, without prompting, it was not immediately clear what the fixes actually did, which led to for some, less exploration and usage (P1). P1 expressed that she was confused whether they were settings to be applied during the translation, or fixes that would occur to the text after. There are concrete ways in which we can iterate on this design, however, including instructions, visual cues, and potentially a restructuring of where and how the fixes are displayed.

Fitting into Translation Workflow This approach is also promising because the experimental group responded in diverse ways of how the fixes were helpful and at which points they used (and described they would use) them. This ranged from seeing the fixes as more of a proofreading device (P5) to something that might be useful for someone less proficient in a language to activate immediately to have something more accurate to the domain automatically (P6). Specifically, P5 expressed that mistakes like "Verb + 得" is a common mistake that people often get mixed up and may not immediately notice.

Understanding of MT There did not seem to be a direct relationship between the presence and use of buttons to the understanding of the MT (insightful descriptions below from control subjects P3 and P4). However, given the small number of participants of varying backgrounds this is by no means conclusive and requires further investigation.

P3 highlighted a critical behavior of the MT, noting that the machine is "very accurate," since it is able to really finely detect sentence structure and replace term-by-term. However, "what it is lacking is that two languages don't express things the same way. Matching word by word is *not* a good translation, since language is alive." She also notes how cultural subtleties are not captured. Reminiscent of frustrations discussed in our needfinding (Section 3.1), there is something missing pragmatically in current translations.

Another relevant and specific observation was that the "bridging of phrases is not quite accurate (P4)." Other participants (P2, P5) noted how transition words

did not seem natural.

P6 also touched on the contextual or pragmatic information we have been discussing throughout this thesis. One specific passage includes reference to "Programs" and a "Framework" (capitalized), and P6 suggested that humans are able to implicitly understand that these likely refer to specific programs and a specific framework.

From the data collection and the user study, I have shown broadly that the use of fixes that are learned from data are useful to the user in improving accuracy across language proficiencies, and that the interface was well-received and easy to use. Quantitative measures in the user study are difficult to interpret, and instead point to several considerations for future study. Additionally, the discussion of how users understood the MT strengthens the idea that there is something missing contextually that the MT is unable to capture.

5

Conclusions and Future Work

We began this exploration with the problem of domain adaptation in machine translation. Rather than it merely being a problem with not enough in-domain bilingual data, our needfinding (Chapter 3) demonstrated that this results in nuanced shortcomings of the MT that impact the translator, both in how they approach and understand translations. These shortcomings seem to be largely due to a lack of consideration of context, which results in improper word choice and difficulties in adjusting style and structure. Patterns in edits across domains and differences between domains (from data collection in Section 4.3) suggest that there seems to be potential in leveraging domains to address these issues.

From there, we designed and tested a prototype for a system that collects post-edits from users and learns quick fixes for the user (and other users) to engage with to improve their translations. Though results are difficult to interpret due to the size of the study, the interface was well received – participants

considered it usable and described specific ways to include and use them it a translation workflow.

What makes this exploration and design unique, however, is that it leverages human intelligence to augment a translation system more quickly and robustly, by learning in a linguistically-grounded manner that creates "context" within a domain that the MT was originally lacking. The resulting interface is accessible to people of varying language proficiency and provides user freedom in their translation process.

5.1 Future Work

Though our results are promising, there is plenty of space for refinement and development, both of the questions we are asking and the design of the system itself.

First, I imagine future work in problem definition and the questions we ask:

1. **A more refined definition of "domain."** Given the exploratory nature of this work, it seemed appropriate to keep the definition of "domain" loose and see if there were any observable patterns of edits within broader domains. The level of granularity and precise definition of a domain that can be generalized across most efficiently (and perhaps interpretably) are very much still open questions. Is "academic fields" the most useful definition? Something more granular? Doing a specific task?

2. **Making better use of a rich linguistics literature.** My handling of linguistics rules only begins to scratch the surface of what is out there. What core concepts can we use in our process of learning and leveraging information from post-edits?

3. **Language selection.** The linguistics concepts we end up using may be affected by the languages we study, given the differences between languages. This user study primarily focused on translation between

English and Chinese, and the data collection asked for edits to translations between English, Chinese, and Spanish. While there are notable similarities and differences between these, maybe we want something more generalizable. How different would the learning be across languages?.

4. **Mental models of MT.** Though we have some preliminary findings on how users understand MT and the errors its making, and how it might be affected by visible fixes based on linguistic rules, there is a lot more to explore here. Understanding this is useful because a user's understanding of the MT affects how they use and interpret the model's results. How do users currently understand MT? What is a useful mental model for them, and how do we achieve it?

There is also relevant future work on the interaction, system, and evaluation:

1. **Dynamic fixes that learn from edits:** What happens if we move from a proof-of-concept prototype to a full system? We would first have to consider what kind of learning from edits to fixes we would want to do. Should we leverage machine learning? How will we more precisely define the data, outputs, and other relevant pieces to arrive at generalizable fixes across a domain?

2. **Learning from the current user:** Expanding on the point above, should we immediately learn in real time from the current user? How would we represent the fixes that are generated? Are they different from the fixes learned within the domain across numerous individuals over time?

3. **Rules coming into conflict:** What if activated fixes come into conflict? How should we handle them in a way that is transparent to the user and is still in line with our goal of improving accuracy more generally?

4. **Adjusting rule learning in relation to level or use:** If we continue to pursue a system that learns fixes from edits, how should we weigh the

contributions of each user? Should contributions be opt-in? Should how we consider them change depending on the level of proficiency of the user, and/or if they are aiming for higher fidelity translation vs. gisting?

5. **Reevaluation of the interaction:** Though I used post-editing in order to receive a certain kind of input in order to analyze and produce rules from, and it does seem to capture a degree of "context," this may not be the most efficient or enjoyable solution for translators. Should we adjust the means of interaction in future work? Can we learn from guidelines and previous work on mixed-initiative systems for translation (or otherwise)?

6. **Improved translation accuracy evaluation:** We used a quick and simple accuracy metric for machine translation. What if we evaluated with people instead? Or alternatively, an accuracy metric that considers more than neighbors and word correspondence (for example, trying to better capture meaning using an evaluation based on textual entailment [15]). Because of how there can be many different good translations for a single passage, we might want to explore more holistic accuracy criteria.

7. **Visualization of fixes:** In the prototype, enabling or disabling the fixes only changes the text in the text box. Is it easier to understand, use, and more helpful to include some sort of indication that there is a difference? Alternatively, even if we choose to pursue a different means of interaction in future work, we may wish to consider how to display errors and correction of errors of MT outputs [21].

5.2 Conclusion

While this thesis does not necessarily demonstrate that fixes learned from edits would in fact be useful across a domain, it does point toward prototype versions of these rules being useful to users within a domain – making translation easier, while long-term cultivating a better understanding of MT errors. Despite its

modesty, the hope is that this is a beginning piece of what is a larger effort towards addressing the issue of lack of context and the resulting shortcomings in MT.

I also hope that this thesis serves as a potentially optimistic attempt to consider ways in which HCI, MT, and linguistics may interact. Though extremely different in methodology and perspective, working along their intersections has begun to illuminate to me what diverse solutions and questions may be possible.

To conclude, one way of understanding what we have done here is that we have provided a form of domain-specific context to MT by harnessing collective human intelligence through linguistic means. Provided with this context within the domain, MT gets a little bit smarter, and we, the interconnected species that we are, benefit.

A Materials

A.1 Study Protocols

The scripts and/or instructions used during each study can be found here:

- **Needfinding (en/zh):** bit.ly/hiptranslate-needfindingprotocol
- **Data Collection:** bit.ly/hiptranslate
- **User Study:** bit.ly/hiptranslate-userstudyprotocol

A.2 Application and Code Links

- A running version of HIP Domain Translate can be found at bear.rebeccahao.com (**experimental condition** in user study).
- The application used in the **control condition** in the user study can be found at panda.rebeccahao.com.
- The application used during **data collection** can be found at hiptranslate.rebeccahao.com.

- Code can be found in this Github repository: https://github.com/beccahao/hiptranslate.

A.3 User Study Materials

These were the passages that were given to the participants to translate in the study, and corresponding human translations (used for the BLEU score).

No.	Passages Presented	Human Translation
1	Our method, guided by respect for the infant's competence, is observation. We observe carefully to understand the infant's communications and his needs. The more we observe, the more we understand and appreciate the enormous amount and speed of learning that happens during the first two or three years of life. We become more humble, we teach less, and we provide an environment for learning instead.	我们的方法，是在尊重婴儿能力的指导下的观察。我们仔细观察来理解婴儿的沟通和需要。我们观察得越多，我们越理解并赞赏在生命的头两到三年发生的学习的数量和速度，我们变得更谦和，我们教得少，以提供学习环境来替代。
2	Programs should use the Framework to guide their choices in curriculum and learning materials, to plan daily activities, and to inform intentional teaching practices. Aligning instruction and opportunities for play, exploration, discovery, and problem-solving with the early learning outcomes described in the Framework will promote successful learning in all children. Programs should also use the Framework with families to help them engage in their children's learning. This Framework replaces the 2010 Head Start Child Development and Early Learning Framework.	学前机构应该用这个框架知道课程和学习材料的选择，规划日常活动，告知有目的的教育实践。把游戏，探索，发现和问题解决的指导及机会和框架中描述的学习成果联系起来，可以促进所有儿童的成功学习。学前机构还应该和家庭一起使用这个框架，帮助他们参与孩子的学习。这个框架取代了2010年的启蒙计划的儿童发展和早期学习框架。
3	比如，我让幼儿做"磁铁能吸引什么"的小实验，"用磁铁拉火柴盒"的小实验，先将几个图钉放在火柴盒里，火柴盒就被拉动了，孩子们快活极了。我及时引导幼儿认识到磁铁隔着东西也能吸铁的特性。接着我们又做了让大头针在塑料板上跳舞的小游戏，孩子们边做边玩，非常高兴。	For example, I asked the children to do two experiments: "what can a magnet attract?" and "use a magnet to pull a match box." First, I put several push pins into a match box and the match box moved – the children were very happy. I immediately guided them to understand that magnets would attract objects even through other items. Then, using this new knowledge, we played a game of making pins dance on a plastic board. Children were happily playing while experimenting.
4	做为一个好老师，只有走进学生的心灵才能透视学生的内心世界，摸准学生的脉搏，倾听学生的心声，与学生的情感产生共鸣，从而帮助其排忧解难化解心理压力，排除心理障碍，有效地激发其内在的积极性将教育的要求有机地内化为学生的自觉行动，教育才能有的放矢和行之有效。	To be a good teacher, you should enter your studentsminds, then you will be able to see the students' inner worlds, feel their heartbeats, hear their voices, and resonate with their emotions. Then, you can help them solve problems, reduce psychological pressure, eliminate mental obstacles, take initiative, and understand educational expectations such that they can manage themselves. In these ways, education can be better targeted and more effective.

B

Participant Information

*Starred is language proficiency (**en** is English, **es** is Spanish, **zh** is Mandarin). This is generally self-reported data, but for the user study, is modified after more granular questions, where: (1) is Elementary proficiency, (2) is Limited working proficiency, (3) is Professional working proficiency, (4) is Full professional proficiency, and (5) is Native or bilingual proficiency. These descriptions were used because they are used more generally so they would be familiar and more straightforward to report for the participants.

B.1 Needfinding Participant Information

Code	Age	Gender	en*	zh*	Other Languages	Domain Expertise
N1	55-64	M	3	5	-	Early-Childhood Education
N2	45-54	F	4	5	-	Education
N3	Under 18	F	3	5	-	Literature, Poetry
N4	45-54	M	5	5	Cantonese (5)	GIS Software

B.2 Data Collection Participant Info

Code	Age	Study For	en*	zh/es*	Other Languages	Domain Expertise
C1	18-24	Chinese	5	5	Spanish (1)	Statistics/Economics
C2	18-24	Chinese	5	4		Economics and science
C3		Chinese	5	3		
C4	18-24	Spanish	5	5		Neuroscience
C5	18-24	Spanish	5	2		Computer Science
C6		Chinese	4	2		
C7	18-24	Chinese	5	5		Economics/Statistics
C8	18-24	Spanish	5	3		Government/Education
C9	44-54	Chinese	4	5		Education
C10		Spanish	5	4	Portuguese (3)	
C11	18-24	Spanish	5	2		
C12	18-24	Spanish	5	5		Neuroscience
C13	18-24	Spanish	5	4	Italian (1)	Psychology and Economics
C14	18-24	Spanish	5	5		
C15	18-24	Chinese	5	5	Spanish (1)	Math
C16	18-24	Spanish	5	3	Chinese (2), German (3)	Biology
C17	18-24	Spanish	5	3	Chinese (3)	
C18	18-24	Spanish	5	4	Vietnamese (5)	Math/CS
C19	18-24	Spanish	5	5		Biomedical engineering
C20	18-24	Chinese	5	3		
C21	18-24	Spanish	5	5	Danish (2), French (1)	Government + Politics; Europe
C22	18-24	Spanish	5	2		Neuroscience

B.3 User Study Participant Information

Code	Age	Gender	en*	zh*	Other Languages	Domain Expertise
P1	45-54	F	4	5	-	Education
P2	18-24	F	5	4	Spanish (2)	History of Science
P3	45-54	F	5	5	Cantonese (5)	GIS Software
P4	18-24	M	5	4	Spanish (3), Hokkien (5), Japanese (2)	Biochemistry
P5	18-24	F	5	4	Spanish (1), German (2), Italian (1)	Statistics and Economics
P6	18-24	M	5	4	Spanish (4), Arabic (2)	Social Studies

References

[1] Nicola Bertoldi and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, pages 182–189, 2009.

[2] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.

[3] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206.

[4] Hal Daumé and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, page 407–412, USA, 2011. Association for Computational Linguistics. ISBN 9781932432886.

[5] Kristin Dew, Anne M Turner, Loma Desai, Nathalie Martin, Adrian Laurenzi, and Katrin Kirchhoff. Phast: A collaborative machine translation and post-editing tool for public health. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:492—501, 2015. ISSN 1942-597X.

[6] George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, page 451–459, USA, 2010. Association for Computational Linguistics.

[7] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D. Manning. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, page 177–187, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330695.

[8] Spence Green, Jeffrey Heer, and Christopher D. Manning. Natural language translation at the intersection of ai and hci. *Queue*, 13(6):30–42, June 2015. ISSN 1542-7730.

[9] Robert J. Hartsuiker and Sarah Bernolet. The development of shared syntax in second language learning. *Bilingualism: Language and Cognition*, 20(2):219–234, 2017. doi: 10.1017/S1366728915000164.

[10] Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. ISSN 0036-8075.

[11] Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224–227, 2007.

[12] Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. Assessing post-editing efficiency in a realistic translation environment. In Sharon O'Brien, Michel Simard, and Lucia Specia, editors, *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice (Nice, September 2, 2013)*, pages 83–91, Allschwil, September 2013. European Association for Machine Translation.

[13] Charles Li and Sandra Thompson. Third-person pronouns and zero-anaphora in chinese discourse. *Syntax and Semantics*, 12, 01 1979.

[14] Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262, 2018. ISSN 2211-3711.

[15] Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:*

Volume 1 - Volume 1, ACL '09, page 297–305, USA, 2009. Association for Computational Linguistics. ISBN 9781932432459.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[17] Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. INMT: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108, Hong Kong, China, November 2019. Association for Computational Linguistics.

[18] Ana I. Schwartz and Judith F. Kroll. Bilingual lexical activation in sentence context. *Journal of Memory and Language*, 55(2):197 – 212, 2006. ISSN 0749-596X.

[19] Michael Shilman, Desney S. Tan, and Patrice Simard. Cuetip: A mixed-initiative interface for correcting handwriting errors. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, UIST '06, page 323–332, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933131.

[20] Patrick Simianer, Sariya Karimova, and Stefan Riezler. A post-editing interface for immediate adaptation in statistical machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 16–20, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[21] Sara Stymne. Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, HLT '11, page 56–61, USA, 2011. Association for Computational Linguistics. ISBN 9781932432909.

[22] Anne Turner, Margo Bergman, Megumu Brownstein, Kate Cole, and Katrin Kirchhoff. A comparison of human and machine translation of health promotion materials for public health practice: Time, costs, and quality. *Journal of public health management and practice : JPHMP*, 20, 09 2013.

[23] Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[24] Guowei Zhou, Yao Chen, Yin Feng, and Rong Zhou. Processing of translation-ambiguous words by chinese–english bilinguals in sentence context. *Journal of psycholinguistic research*, 48(5):1133–1161, 2019.

[25] Álvaro Peris and Francisco Casacuberta. Online learning for effort reduction in interactive neural machine translation. *Computer Speech Language*, 58:98 – 126, 2019. ISSN 0885-2308.