UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departament de Sistemes Informàtics i Computació

Some Contributions to Interactive Machine Translation and to the Applications of Machine Translation for Historical Documents

Ph.D. Dissertation

Miguel Domingo

Supervised by Prof. Francisco Casacuberta

November 2021

PRHLT

GOBIERNO DE ESPAÑA

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

GENERALITAT VALENCIANA

Unión Europea
FEDER

Invertimos en su futuro

The research leading to this thesis has been partially funded by *Ministerio de Economía y Competitividad (MINECO)* under projects *SmartWays* (grant agreement RTC-2014-1466-4), *CoMUN-HaT* (grant agreement TIN2015-70924-C2-1-R) and *MISMISFAKEnHATE* (grant agreement PGC2018-096212-B-C31); *Generalitat Valenciana* under projects *ALMAMATER* (grant agreement PROMETEOII/2014/030) and *DeepPattern* (grant agreement PROMETEO/2019/121); the *European Union* through *Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER)* from Comunitat Valenciana (2014–2020) under project *Sistemas de frabricación inteligentes para la indústria 4.0* (grant agreement IDIFEDER/2018/025); and the *PRHLT research center* under the research line *Machine Learning Applications.* We gratefully acknowledge the support of *NVIDIA Corporation* with the donation of a GPU used through this thesis. Additionally, I would like to thank Andrés Trapiello and *Ediciones Destino* for granting us permission to use their book in our researches.

Document prepared and typeset in LaTeX.

Acknowledgments

It feels strange to be writing these words. A mixed feeling of happiness and nostalgia. It feels like it was yesterday when I attended to my first university lecture, while the day before I was entering middle school. Where did the time go?

As a child, I used to dream about becoming a scientist: a white coat, a lab, astonishing inventions, lots of incomprehensible formulas... Yes, I was heavily influenced by literature and TV—after all, mad scientist are the best characters, right? However, what really thrilled me was the concept of research: solving problems that no one had solved before, while creating *cool* things for someone's benefice.

Life went on, and I somehow managed to combine most of the things I grew up loving: science, literature, languages and history. This dissertation is, thus, the climax of that dream. I can now go back to child me and proudly tell him: attaboy, you did great!

All this would not have been possible without the love and support of so many people. First I would like to thank my advisor, Paco, for giving me the opportunity to develop my thesis at the PRHLT. Without his support and guidance I would not be writing this today.

To my labmates and the rest of my PRHLT colleges. Especially to my MT seniors—Mara and Álvaro—and Lorenzo, who is also writing the last words of his dissertation right now. It's been a pleasure sharing so many moments with you all.

To my family, for all their love and unconditional support. Especially to my mum, who has always encouraged me to pursue my dreams and gives me strength to go forwards.

Finally, I would like to thank my *inconclusive* friends for bearing with my time constraints and for patiently waiting for me to end this phase of my life.

Valencia, November 15th 2021.

Abstract

Historical documents are an important part of our cultural heritage. However, due to the language barrier inherent in human language and the linguistic properties of these documents, their accessibility is mostly limited to scholars. On the one hand, human language evolves with the passage of time. On the other hand, spelling conventions were not created until recently and, thus, orthography changes depending on the time period and author. For these reasons, the work of scholars is needed for non-experts to gain a basic understanding of a given document.

In this thesis, we tackle two tasks related with the processing of historical documents. The first task is *language modernization* which, in order to make historical documents more accessible to non-experts, aims to rewrite a document using the modern version of the document's original language. The second task is *spelling normalization.* The aforementioned linguistic properties of historical documents suppose an additional challenge for the effective natural language processing of these documents. Thus, this task aims to adapt a document's spelling to modern standards in order to achieve an orthography consistency.

We affront both task from a machine translation perspective, considering a document's original language as the source language, and its modern/normalized counterpart as the target language. We propose several approaches based on statistical and neural machine translation, and carry out a wide experimentation that shows the potential of our contributions—with the statistical approaches yielding equal or better results than the neural approaches in most of the cases. For the language modernization task, this experimentation includes a human evaluation conducted with the help of scholars and a user study that verifies that our proposals are able to help non-experts to gain a basic understanding of a historical document without the intervention of a scholar.

As with any machine translation problem, our applications are not error-free. Thus, to obtain perfect modernizations/normalizations, a scholar needs to supervise and correct the errors. This is a common procedure in the translation industry. The interactive machine translation framework aims to reduce the effort needed for obtaining high quality translations by embedding the human agent and the translation system into a cooperative correction process. However, most interactive protocols follow a left-to-right strategy. In this thesis, we developed a new interactive protocol that breaks this left-to-right barrier. We evaluated this new protocol in a machine translation environment, obtaining large reductions of the human effort. Finally, since this interactive framework is of general application to any translation problem, we applied it—our new protocol together with one of the classic left-to-right protocols—to language modernization and spelling normalization. As with machine translation, the interactive framework diminished the effort required for correcting the outputs of an automatic system.

Resumen

The following abstract has been automatically translated from its English version using Google Translate *and manually post-edited. The post-editing process can be seen at Appendix C.*

Los documentos históricos son una parte importante de nuestra herencia cultural. Sin embargo, debido a la barrera idiomática inherente en el lenguaje humano y a las propiedades lingüísticas de estos documentos, su accesibilidad está principalmente restringida a los académicos. Por un lado, el lenguaje humano evoluciona con el paso del tiempo. Por otro lado, las convenciones ortográficas no se crearon hasta hace poco y, por tanto, la ortografía cambia según el período temporal y el autor. Por estas razones, el trabajo de los académicos es necesario para que los no expertos puedan obtener una comprensión básica de un documento determinado.

En esta tesis abordamos dos tareas relacionadas con el procesamiento de documentos históricos. La primera tarea es la *modernización del lenguaje* que, a fin de hacer que los documentos históricos estén más accesibles para los no expertos, tiene como objetivo reescribir un documento utilizando la versión moderna del idioma original del documento. La segunda tarea es la *normalización ortográfica.* Las propiedades lingüísticas de los documentos históricos mencionadas con anterioridad suponen un desafío adicional para la aplicación efectiva del procesado del lenguaje natural en estos documentos. Por lo tanto, esta tarea tiene como objetivo adaptar la ortografía de un documento a los estándares modernos a fin de lograr una consistencia ortográfica.

Ambas tareas las afrontamos desde una perspectiva de traducción automática, considerando el idioma original de un documento como el idioma fuente, y su homólogo moderno/normalizado como el idioma objetivo. Proponemos varios enfoques basados en la traducción automática estadística y neuronal, y llevamos a cabo una amplia experimentación que ratifica el potencial de nuestras contribuciones –en donde los enfoques estadísticos arrojan resultados iguales o mejores que los enfoques neuronales para la mayoría de los casos–. En el caso de la tarea de modernización del lenguaje, esta experimentación incluye una evaluación humana realizada con la ayuda de académicos y un estudio con usuarios que verifica que nuestras propuestas pueden ayudar a los no expertos a obtener una comprensión básica de un documento histórico sin la intervención de un académico.

Como ocurre con cualquier problema de traducción automática, nuestras aplicaciones no están libres de errores. Por lo tanto, para obtener modernizaciones/normalizaciones perfectas, un académico debe supervisar y corregir los errores. Este es un procedimiento común en la industria de la traducción. La metodología de traducción automática interactiva tiene como objetivo reducir el esfuerzo necesario para obtener traducciones de alta calidad uniendo al agente humano y al sistema de traducción en un proceso de corrección cooperativo. Sin embargo, la mayoría de los protocolos interactivos siguen una estrategia de izquierda a derecha. En esta tesis desarrollamos un nuevo protocolo interactivo que rompe con esta barrera de izquierda a derecha. Hemos evaluado este nuevo protocolo en un entorno de traducción automática, obteniendo grandes reducciones del esfuerzo humano. Finalmente, dado que este marco interactivo es de aplicación general a cualquier problema de traducción, lo hemos aplicado –nuestro nuevo protocolo junto con uno de los protocolos clásicos de izquierda a derecha– a la modernización del lenguaje y a la normalización ortográfica. Al igual que en traducción automática, el marco interactivo logra disminuir el esfuerzo requerido para corregir los resultados de un sistema automático.

Resum

The following abstract has been automatically translated from its Spanish version using SisHiTra[1]*, a translation tool developed at the* PRHLT Research Center[2]*, and manually post-edited. The post-editing process can be seen at Appendix C.*

Els documents històrics són una part important de la nostra herència cultural. No obstant això, degut a la barrera idiomàtica inherent en el llenguatge humà i a les propietats lingüístiques d'aquests documents, la seua accessibilitat està principalment restringida als acadèmics. D'una banda, el llenguatge humà evoluciona amb el pas del temps. D'altra banda, les convencions ortogràfiques no es van crear fins fa poc i, per tant, l'ortografia canvia segons el període temporal i l'autor. Per aquestes raons, el treball dels acadèmics és necessari perquè els no experts puguen obtindre una comprensió bàsica d'un document determinat.

En aquesta tesi abordem dues tasques relacionades amb el processament de documents històrics. La primera tasca és la *modernització del llenguatge* que, a fi de fer que els documents històrics estiguen més accessibles per als no experts, té per objectiu reescriure un document utilitzant la versió moderna de l'idioma original del document. La segona tasca és la *normalització ortogràfica.* Les propietats lingüístiques dels documents històrics mencionades amb anterioritat suposen un desafiament addicional per a l'aplicació efectiva del processat del llenguatge natural en aquests documents. Per tant, aquesta tasca té per objectiu adaptar l'ortografia d'un document als estàndards moderns a fi d'aconseguir una consistència ortogràfica.

[1] `http://demosmt.prhlt.upv.es/sishitra/`.

[2] `https://www.prhlt.upv.es`.

Dues tasques les afrontem des d'una perspectiva de traducció automàtica, considerant l'idioma original d'un document com a l'idioma font, i el seu homòleg modern/normalitzat com a l'idioma objectiu. Proposem diversos enfocaments basats en la traducció automàtica estadística i neuronal, i portem a terme una àmplia experimentació que ratifica el potencial de les nostres contribucions –on els enfocaments estadístics obtenen resultats iguals o millors que els enfocaments neuronals per a la majoria dels casos–. En el cas de la tasca de modernització del llenguatge, aquesta experimentació inclou una avaluació humana realitzada amb l'ajuda d'acadèmics i un estudi amb usuaris que verifica que les nostres propostes poden ajudar als no experts a obtindre una comprensió bàsica d'un document històric sense la intervenció d'un acadèmic.

Com ocurreix amb qualsevol problema de traducció automàtica, les nostres aplicacions no estan lliures d'errades. Per tant, per obtindre modernitzacions/normalitzacions perfectes, un acadèmic ha de supervisar i corregir les errades. Aquest és un procediment comú en la indústria de la traducció. La metodologia de traducció automàtica interactiva té per objectiu reduir l'esforç necessari per obtindre traduccions d'alta qualitat unint a l'agent humà i al sistema de traducció en un procés de correcció cooperatiu. Tot i això, la majoria dels protocols interactius segueixen una estratègia d'esquerra a dreta. En aquesta tesi desenvolupem un nou protocol interactiu que trenca amb aquesta barrera d'esquerra a dreta. Hem avaluat aquest nou protocol en un entorn de traducció automàtica, obtenint grans reduccions de l'esforç humà. Finalment, atès que aquest marc interactiu és d'aplicació general a qualsevol problema de traducció, l'hem aplicat –el nostre nou protocol junt amb un dels protocols clàssics d'esquerra a dreta– a la modernització del llenguatge i a la normalitzaciò ortogràfica. De la mateixa manera que en traducció automàtica, el marc interactiu aconsegueix disminuir l'esforç requerit per corregir els resultats d'un sistema automàtic.

Preface

Due to their importance as part of our cultural heritage, many natural language processing (NLP) researches are focused on historical documents. However, the linguistic properties of these documents create additional challenges for these researches. Machine translation (MT) focus on generating automatic translations from a source language into a target language. In this thesis, we focus on two tasks related to tackling some of these linguistic challenges for the processing of historical documents: language modernization—which revolves around the linguist evolution of the document's language throughout the years—and spelling normalization—which tries to account for the orthography inconsistencies resulted from the lack of spelling conventions.

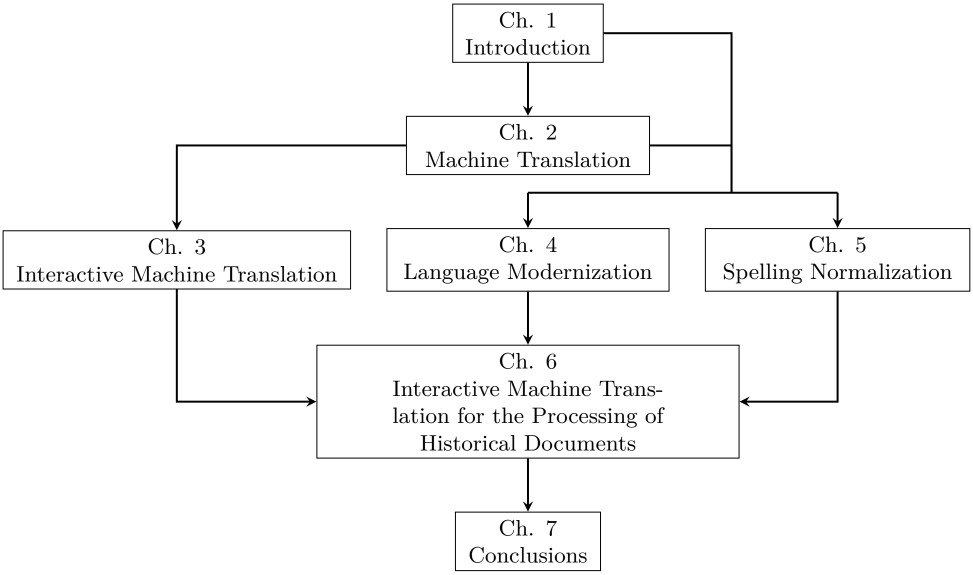
We reformulate both tasks as translations problems in which we want to translate a document from their original language to either its modern or spelling-normalized counterpart. Then, we propose several MT-based approaches to tackle each of these tasks.

However, the MT problem is still far from solved. Thus, its automatic translations need to be reviewed and corrected in order to obtain high quality translations. The interactive machine translation (IMT) field proposes an interactive framework in which machine and human work together to generate those translations. We propose a new IMT protocol that reduces the human effort needed for that collaboration and applied this field into the aforementioned tasks in order to help scholars create better modernizations/normalizations.

The scientific goals of this thesis are divided into two main groups:

1. **Machine translation applications to historical documents**. We propose and study several machine translation applications to two tasks related with the processing of historical documents: language modernization and spelling normalization.
2. **Interactive machine translation**. We develop a new interactive protocol and apply this new protocol and the classical prefix-based protocol to the processing of historical documents.

This dissertation is structured in 7 chapters that relate as follows:



The content of each chapter is:

Chapter 1 introduces machine translation and the two task related with the processing of historical documents which are part of the scientific goals of this thesis: language modernization and spelling normalization.

Chapter 2 defines the machine translation framework used in this thesis.

Chapter 3 reviews the classic prefix-based interactive-predictive paradigm and proposes a new protocol that tries to overcome some prefix-based limitations. This new protocol is implemented and evaluated.

Chapter 4 describes the work we conducted on the language modernization task. It includes an automatic and a human evaluation, and a user study that assess our proposals.

Chapter 5 describes the work we conducted on the spelling normalization task, and the experiments we conducted in order to assess our proposals.

Chapter 6 applies the protocol developed on Chapter 3 and the classic prefix-based protocol to the processing of historical documents. An automatic evaluation assesses the advantages of this methodology. Finally, an online demonstrator of the prefix-based protocol is developed.

Chapter 7 draws the main conclusions of this thesis and reviews the scientific contributions and publications derived from it as well as future research lines.

These chapters are complemented by the following appendixes:

Appendix A studies the impact of the number of byte-pair encoding merge operations in the language modernization task (on Chapter 4).

Appendix B contains the questionnaire from the user study conducted for the language modernization task (on Chapter 4).

Appendix C showcases the post-editing process conducted for generating the Spanish and Valencian versions of the abstract.

Contents

Acronyms

ART approximate randomization testing

BLEU bilingual evaluation understudy

BPE byte pair encoding

CBMT character-based machine translation

CBNMT character-based neural machine translation

CBSMT character-based statistical machine translation

CER character error rate

CM confidence measures

FDA feature decay algorithm

HMM hidden Markov alignment models

HTR handwritten text recognition

IMT interactive machine translation

INMT interactive neural machine translation

KSR key stroke rate

LSTM long short-term memory

MAR mouse action rate

MERT minimum error rate training algorithm

MT machine translation

NLP natural language processing

NMT neural machine translation

RBMT rule-based machine translation

ReLU rectified linear unit

RNN recurrent neural network

SMT statistical machine translation

TER translation error rate

WSR word stroke rate

XML eXtensible Markup Language

Chapter 1

Introduction

Pero mudo y absorto y de rodillas
como se adora a Dios ante su altar,
como yo te he querido…, desengáñate,
así… ¡no te querrán!

(***Rima LIII***. Gustavo Adolfo Bécquer.)

But mute and absorbed and on my knees
as God is worshiped before his altar,
as I have loved you …, deceive yourself,
like this … they won't love you!

(***Rhyme LIII***. Google Translate.)

Contents

Language is one of the most important attributes of humankind. It gives them the capacity to communicate between them, which has allowed the development of societies and the advancement of science. However, language diversity and its evolution with the passage of time creates great challenges for communication.

Natural language processing (NLP) studies human language with the aim of developing systems that are able to understand and generate natural language at the human level. One of the most challenging tasks in this field is machine translation (MT), which aims to reduce the challenges in communication by automatically translating a sentence from a natural language to another. This problem has the interest of the translation industry, which has MT integrated into their production workflow. However, automatic translations need to be reviewed and corrected by human translator in order to achieve high quality standards.

Other challenging NLP tasks revolve around the processing of historical data. With the aim of their preservation and their dissemination, historical manuscripts are transcribed into digital documents. This allows for many applications which are focused on facilitating the study of the document's data. For example, finding one or more words on a set of documents. However, the nature of language create additional difficulties for these applications.

In this thesis, we work on improving the interactive machine translation (IMT) framework in order to reduce the human effort for achieving high quality translations. Then, we apply the MT field to two tasks related with the processing of historical documents and the language-related challenges.

1.1 Machine translation

Problem-free global communication is an old human dream. It can be tracked back to the 17^{th} century (Hutchings, 2004) with the idea of creating a universal language. However, most ideas relied too much on philosophical concepts (Descartes, 1970) and this language was never found. These ideas proliferated during 18^{th} and 19^{th} century. It wasn't until 20^{th} century that the firsts machines developed for automatically translating languages appeared. In 1933, Petr Trojanskij patented two electromechanical devices for their use as translation dictionaries. Seeing with retrospective, this idea was similar to the encoder–decoder framework, which is the current state of the art in MT. However, this idea remained somewhat isolated at the USSR.

At the end of World War II and the beginning of the Cold War, the expertise in breaking enemy codes derived in the idea of using computers for translating from one language into another. However, an excessive optimism about MT—which was though to be solved in three to five years[1]—resulted on a severe cut of funds and on the research being severally halted for a decade.

At the 1970s and 1980s, MT research was focused on the so-called rule-based machine translation (RBMT), which relied on linguistic information for creating a set of translation rules. However, the increase in computational capacity lead way to a new family of methods: corpus-based MT. They relied on statistical methods (Brown et al., 1993) and their capabilities and potential were rapidly acknowledged by the scientific community. For over 20 years, statistical machine translation (SMT) became the state of the art in MT. However, these systems seemed to have reached a performance plateau until the arrival of neural machine translation (NMT) in recent years (Sennrich, 2016). This novel approach quickly became the new state of the art, opening new research directions, questions and challenges.

1.1.1 Taxonomy

MT is frequently classified (e.g., Koehn, 2010) into two main paradigms: rule-based approaches, which relies on the creation of sets of rules by human experts that are able to extract the meaning of a source language and represented in a target language; and corpus-based, which automatically infer the translation of a sort text by extracting information from parallel training data.

Rule-based systems

RBMT was one of the first MT approaches. Their systems are based on linguistic information that is extracted from an analysis of the languages involved within the translation process. Depending on the type of analysis and level of extraction, they can be classified into the following groups (shown in order of depth):

1. **Direct systems**: These systems translate word by word, replacing words from one language into another. They rely on translation dictionaries and, occasionally, on a morphological analysis of the source text. Direct translation was one of the earliest attempts to MT (Kay, 1973).

2. **Transfer systems**: These systems analyze and generate an abstract representation of the source text. Then, this representation is converted into an

[1] `http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html`.

abstract representation of the target text. Finally, this last representation is transformed into the target language.

3. **Interlingual systems**: These systems aim to construct a universal representation of all languages—called interlingua (Vauquois, 1971)—by performing a deep analysis of the languages. They perform translations by analyzing the source language and converting it into an abstract, universal representation provided by the interlingua. This representation is, then, converted to the target language, generating the correspondent translation. Under this framework, the interlingua is a language independent construction which can, therefore, bridge any language pair (Peris, 2019).

RBMT systems have the advantage of being easier to control since their rules are created by humans. Thus, more rules can be added to address potentials flaws in the systems. Moreover, they are very useful for low-resource scenarios. However, the design of the rules is a hard and costly process that requires a human expert in both languages.

Corpus-based systems

Corpus-based systems infer translations by training their MT systems from a collection of bilingual parallel data. Furthermore, they do not require the knowledge of the language involved. Thus, they can be applied to any language combination provided that there are training data available. Their main disadvantage, however, is that data collection are not always available and can be costly to produce. These methods follow two main approaches:

Example-based MT: These systems are able to obtain translations by analogy of the source sentence with respect to a bilingual data collection (Nagao, 1984). Translations are performed by searching for similar sentences in the data collection—which is available to the system at runtime—and recombining them to construct the final translation.

Statistical machine translation: These systems rely on a statistical framework in which it is assumed that a translation probability between source and target sentences can be computed. Therefore, the translation process consists in obtaining the string in the target language with the highest translation probability. The statistical framework relies on a probability distribution which is approximated by a mathematical model whose parameters are estimated from a collection of data. Depending on the probability distribution, several SMT systems have been proposed:

- **Word and phrase-based SMT**: For years, this approach has been the state of the art in MT. This approach is one of the central pieces of this thesis. A description of it can be found at Section 2.1.
- **Neural machine translation**: In this approach, the probability distribution is exclusively modeled with a neural network. It is the current state of the art in MT, and it is also a central piece of this thesis. A description of NMT can be found at Section 2.2.
- **Other models**: other models, such as stochastic finite-state transducers (Casacuberta and Vidal, 2007) or stochastic context-free grammars (Wu, 1997) have been explored under the SMT formalization.

1.2 Historical documents processing

Historical documents have an outstanding cultural value in subjects as diverse as literature, botanic, mathematics, medicine or religion. They are a unique public asset, forming the collective and evolving memory of our societies (Romero et al., 2019). For this reason, there is an increasing need of converting these documents to digital form, leaving place to many tasks that revolve around the processing of historical documents. For example, generating automatic transcriptions (Toselli et al., 2010, 2017), creating search queries capable of finding all occurrences of one or more words (Rogers and Willett, 1991; Ernst-Gerlach and Fuhr, 2006), generating word frequency lists (e.g., Baron et al., 2009) or creating NLP tools which provide automatic annotations to identify and extract linguistic structures such as relative clauses (Hundt et al., 2011) or verb phrases (Fiebranz et al., 2011; Pettersson et al., 2013).

However, a common problem in most tasks is that the language characteristics of historical documents create additional difficulties and limits their accessibility mostly to scholars. In this thesis, we tackle two tasks related with theses linguistic challenges: language modernization and spelling normalization.

1.2.1 Language modernization

Despite being an important part of our cultural heritage, historical documents are mostly accessible to scholars. This is due to the nature of human language–which evolves with the passage of time—and the linguistic properties of these documents: the lack of spelling conventions causes that their orthography changes depending on the time period and author. This increases the difficulty of comprehending historical documents. Thus, for their preservation, in order to make them reachable to a broader audience a scholar is typically in charge of producing a comprehensive contents document which allows non-experts to locate and gain a basic understanding of a given document (e.g., Monk, 2018).

Example 1.1: Example of modernizing the language of a historical document. The original text is a fragment from *Hamlet* by *William Shakespeare.* The modernized version was obtained from Crowther (2003).

Original	**Modernized**
To be, or not to be? That is the question	The question is: is it better to be alive or dead?
Whether 'tis nobler in the mind to suffer	Is it nobler to put up
The slings and arrows of outrageous fortune,	with all the nasty things that luck throws your way,
Or to take arms against a sea of troubles,	or to fight against all those troubles
And, by opposing, end them?	by simply putting an end to them once and for all?
To die, to sleep—No more—	Dying, sleeping—that's all dying is—
and by a sleep to say we end	a sleep that ends
The heartache and the thousand	all the heartache and shocks that
natural shocks That flesh is heir to—	life on earth gives us—
'tis a consummation devoutly to be wished!	that's an achievement to wish for.
To die, to sleep.	To die, to sleep—
To sleep, perchance to dream	to sleep, maybe to dream.

Language modernization aims to tackle this language barrier by generating a new version of a historical document, written in the modern version of the document's original language. Example 1.1 shows an example of modernizing a document. In this case, part of the language structures and rhymes have been lost. However, the modern version is easier to read and comprehend by a broader audience. This problem is also present in poetry translation since the entwinement between sound and word and sense cannot be truly replicated in a different language (Ilonka, 2018). However, translating a poem from one language into another is a way of sharing cultural practices and ideologies across languages (Rajvanshi, 2015).

Language modernization can be a controversial topic since it implies an alteration of the original document (e.g., the manual modernization of *El Quijote* rose a controversy in Spain (Flood, 2015)). However, it is manually applied to classic literature in order to make understandable to contemporary readers works that had been relegated to scholars due to the hardness of their comprehension (Rodríguez Marcos, 2015).

Finally, while the language richness present in historical documents is also part of our cultural heritage, the goal of language modernization is limited to make historical documents accessible to a general audience.

1.2.2 Spelling normalization

Human language is constantly evolving over time. Additionally, spelling conventions were not created until recently. Thus, orthography changes depending on the author and time period. Sometimes, this variety is astonishing. Laing (1993) pointed out that, for instance, the data in *LALME*[2] indicate 45 different forms recorded for the pronoun *it*, 64 for the pronoun *she* and more than 500 for the preposition *through*. This linguistic variations are present in historical documents and have always been a concern for scholars in humanities (Bollmann, 2018).

Since they are an important part of our cultural heritage, the interest in effective NLP for historical documents is on the rise (Bollmann, 2018). However, the aforementioned linguistic problems suppose an additional challenge. Spelling normalization aims to solve these problems. Its goal is to achieve an orthography consistency by adapting the document's spelling to modern standards. Example 1.2 shows an example.

Example 1.2: Example of adapting a document's spelling to modern standards. Characters that need to be adapted are denoted in **red**. Its modern versions are denoted in **purple**. Example extracted from F. Jehle (2001).

Original	**Normalized**
Y al desarmarle, como **e**l se imagina**u**a	Y al desarmarle, como **é**l se imagina**b**a
que aquellas tra**y**das y lle**u**adas que le	que aquellas tra**í**das y lle**v**adas que le
desarma**u**an eran algunas principales	desarma**b**an eran algunas principales
señoras y damas de aquel castillo,	señoras y damas de aquel castillo,
les di**x**o con mucho dona**y**re:	les di**j**o con mucho dona**i**re:
“Nunca fuera ca**u**allero	“Nunca fuera ca**b**allero
de damas ta**m**bien ser**u**ido,	de damas ta**n** bien ser**v**ido,
como fuera don Qui**x**ote	como fuera don Qui**j**ote
quando de su aldea vino:	cuando de su aldea vino:
don**z**ellas cura**u**an d**e**l,	don**c**ellas cura**b**an d**e é**l,
princesas del su ro**z**ino.”	princesas del su ro**c**ino.”

[2] Linguistic Atlas of Late Medieval English.

Spelling normalization could be seen as a subtask of language modernization since, in a sense, updating the spelling to modern standard is part of modernizing the document's language. However, both tasks have a different goal in mind: Language modernization seeks to make historical texts easier to comprehend, and their target audience are non-experts. Spelling normalization limits itself to standardize orthography and its target audience are scholars. Moreover, unlike language modernization, spelling normalization is a monotone problem: there are no word reorders between the original sentence and its target version. Furthermore, most changes happen at a character level. Thus, the spelling normalization problem can be tackled using MT techniques that are not feasible for the language modernization problem.

Chapter 2

Machine Translation

Y si caigo, ¿qué es la vida?
Por perdida ya la di,
cuando el yugo del esclavo,
como un bravo, sacudí.

(***La canción del pirata***. José de Espronceda.)

And if I fall what is life?
For loss I already gave it,
when the yoke of the slave,
like a bravo, shook.

(***The pirate song***. Google Translate.)

Contents

In this chapter, we present and describe the machine translation (MT) framework that shall be used throughout this thesis. We start by formalizing the statistical framework of statistical machine translation (SMT) and, then, its neural approach: neural machine translation (NMT).

2.1 Statistical machine translation

SMT deploys a statistical framework into the MT problem (see Section 1.1). This framework relies on a probability distribution in which the parameters of its mathematical model are estimated from the parallel training data. Given a source sentence x_1^J of length J and a target sentence y_1^I of length I, SMT assumes that a translation probability $Pr(y_1^I \mid x_1^J)$ can be computed. Thus, for each source sentence x_1^J, its goal is to find the target sentence with the highest probability ($\hat{y}_1^{\hat{I}}$) (Brown et al., 1993):

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^I} Pr(y_1^I \mid x_1^J) \tag{2.1}$$

Which can be leveraged applying Bayes' theorem (Bayes, 1763):

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^I} Pr(y_1^I) \cdot Pr(x_1^J \mid y_1^I) \tag{2.2}$$

This last equation is frequently referred as SMT's *fundamental equation.* The term $Pr(y_1^I)$ represents the language model, which measures the well-formedness of the target language sentence; and the term $Pr(x_1^J \mid y_1^I)$ represents the translation model, which captures the correlation between the source and target sentences.

To model this translation process, word alignments were introduced (Brown et al., 1993). Given a source word x_j, this word is aligned to a set of target word positions $\mathbf{a}_j = i_1, \ldots, i_l$ following a generative perspective. Therefore, this alignment implies that the source word x_j generates the target words $y_{i_1}, \ldots, y_{i_l}$. Since alignments cannot be observed during the training process, to model it a hidden variable a_1^J is required, yielding:

$$Pr(x_1^J \mid y_1^I) = \sum_{a_1^J \in \mathscr{A}(x_1^J, y_1^I)} Pr(x_1^J, a_1^J \mid y_1^I) \tag{2.3}$$

where $\mathscr{A}$ represents all the possible alignments between x_1^J and y_1^I.

Through the years, many alignments models have been proposed, starting with the five original models (frequently known as IBM models 1 to 5) from Brown et al. (1993). However, an important breakthrough was achieved with the arrival of the so-called *log-linear model*: a weighted log-linear combination of feature functions which are estimated independently (Och and Ney, 2002; Koehn, 2010):

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^I} \left\{ \sum_{n=1}^{N} \lambda_n \cdot log(f_n(y_1^I, x_1^J)) \right\} \tag{2.4}$$

The most common features include: a (target) language model, bidirectional translation models and a reordering model; among others (Koehn, 2010).

Some of the most popular instantiations of log-linear models include phrase-based models (Zens et al., 2002), hierarchical models (Chiang, 2007) and neural models (see Section 2.2). From this point forward, when we talk about SMT we will be referring to phrase-based SMT.

Thus, the three main challenges of SMT are:

1. **Model definition**: development of models which are able to approximate the probability distribution $Pr(y_1^I \mid x_1^J)$.
2. **Parameter estimation**: After defining the model, its parameters need to be estimated from the training data. This data usually consists in a collection of parallel sentence-aligned documents of translated sentences.
3. **Search problem**: After estimating the parameters, translations can be computed by searching for the target language string with the highest probability. This is also known as decoding.

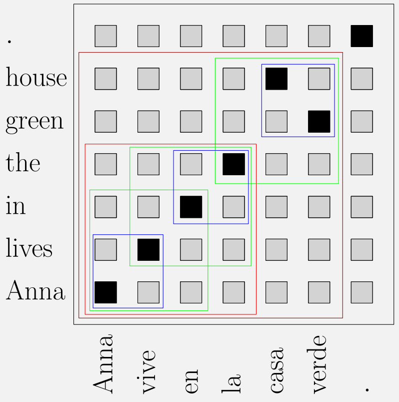
2.1.1 Phrased-based statistical machine translation

Phrased-based models are the most popular of the SMT's log-linear model, constituting and alternative to overcome the limitations of the word based models (Koehn, 2010). They are based in the concept of segmenting the sentence pairs into word sequences known as *phrases* so that the number of source and target phrases are the same (k) and so that a given source phrase is only aligned with a single target phrase (and vice versa).

Thus, translating a source sentence x_1^J into its target equivalent y_1^I consists in the following steps:

1. Divide x_1^J into K source phrases $(\tilde{x}_1, \dots, \tilde{x}_k, \dots, \tilde{x}_K)$.
2. Translate each source phrase into a target phrases $(\tilde{y}_1, \dots, \tilde{y}_k, \dots, \tilde{y}_K)$.
3. Reorder the target phrases to complete the translation y_1^I.

Example 2.1: Example of how phrases are extracted from a word alignment matrix within a phrase-based model for the source sentence *Ana vive en la casa verde*. Example by Chinea-Rios (2019).



(a) Word matrix.

Source phrase	**Target phrase**
Ana	Anna
vive	lives
en	in
la	the
casa	house
verde	green
.	.
Ana vive	Anna lives
vive en	lives in
en la	in the
la casa verde	the green house

(b) Phrases extracted.

Phrase tables are another important step when learning phrase-based models. These tables contain all the $(\tilde{x}_k, \tilde{y}_k)$ observed during training and includes all the values of each feature function assigned to that phrase pair. Given a sentence pair (x_1^J, y_1^I), $(\tilde{x}_k, \tilde{y}_k)$ is a phrase pair $(x_{j_1}^{j_2}, y_{i_1}^{i_2})$ by a symmetrized alignment if the set of target positions linked to source positions in $j_1, \dots, j_2$ by the alignment is included in $i_1, \dots, i_2$ and vice versa.

Among the different heuristic techniques studied for producing phrase-based models, the most commonly used is based on the relative frequencies of the phrase pairs that are extracted from word alignment matrices (Och, 2002). Like world alignment models, phrase-based models assume that an alignment variable captures the relationship between source and target phrases. Example 2.1 shows an example of the word aligned sentence pair and the bilingual phrases extracted from a given sentence. **(a)** showcases the alignment matrix. Black squares represent word alignments while extracted phrases are highlighted with a rectangle comprising one or more squares. **(b)** lists the phrases that can be extracted from the matrix.

The features $h_m(\cdot,\cdot)$ included in phrase-based models are:

1. The inverse translation probability:

$$Pr(\tilde{x}_k \mid \tilde{y}_k) = \frac{\text{count}(\tilde{y}_k, \tilde{x}_k)}{\text{count}(\tilde{y}_k)} \tag{2.5}$$

   where count$(\tilde{y}_k, \tilde{x}_k)$ is the number of times that the phrases $\tilde{x}_k$ and $\tilde{y}_k$ were extracted together throughout the whole training corpus; and count$(\tilde{y}_k)$ is the total count for phrase $\tilde{x}_k$.

2. Direct translation probability, which is similar to the inverse translation probability but computed in the reverse translation direction:

$$Pr(\tilde{y}_k \mid \tilde{x}_k) = \frac{\text{count}(\tilde{y}_k, \tilde{x}_k)}{\text{count}(\tilde{x}_k)} \tag{2.6}$$

3. Direct and inverse lexical translation probabilities. These features were defined by Zens et al. (2002). They attempt to account for the lexical soundness of each phrase pair.

4. The phrase penalty which, like the word penalty feature, it implements a constant cost during decoding. This penalty is accumulated per phrase.

Model tuning

Eq. (2.4) SMT's log-linear model. In this model, the log-linear weight (λ) adjust the importance of each single model within the specific task. The inherent idea is that good values for a certain task might not be the appropriate values for other tasks (e.g., a translation model trained on a domain in which sentences tend to be very long might not perform well in a domain with shorter sentences. Thus, we will need to adjust the weights conveniently to reflect this fact). This process is frequently known as *tuning*.

Among the many methods for optimizing the log-linear weights, the most popular algorithm is minimum error rate training algorithm (MERT) (Och, 2003). Given a parallel tuning set (commonly known as *development* set) $\{(x^{(a)}, y^{(a)})\}_{a=1}^{A}$ composed of A sentences, an initial weight-vector λ is chosen and an n-best list from the decoder is obtained. Then, the iterative process starts. Initially, the starting point is the weight-vector λ. But, on the following iterations, the starting point is the best weight-vector from the previous iteration. After each iteration, the decoder is run again to obtain new n-best lists that are merged with the existing ones. Additionally, MERT uses a number of additional random points in vector space to avoid poor local optima. The iterative process stops when there are either no changes in the weight-vector or no new translations in the n-best list.

The goal of MERT is to minimize the error count $E(r_1^I, y_1^I)$ by scoring translation hypothesis against a set of reference translations $\{r^{(a)}\}_{a=1}^{A}$. Assuming that $E(r_1^A, y_1^A) = \sum_{a=1}^{A} E(r_a, y_a)$ (i.e., error count is additively decomposable by sentence), this results in the following optimization problem:

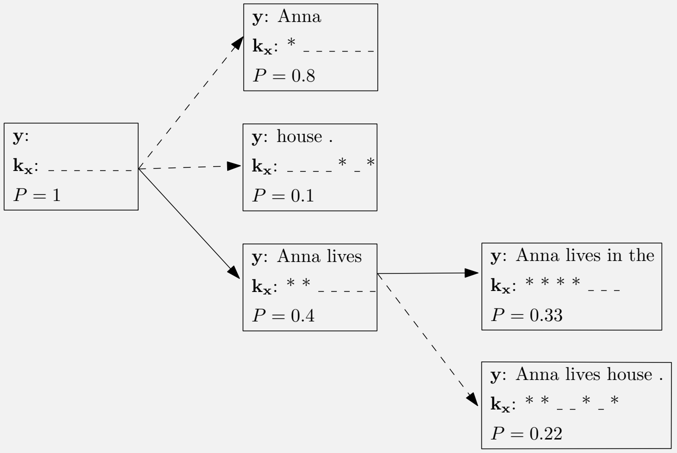
$$\hat{\lambda} = \arg\min_{\lambda} \left\{ \sum_{a=1}^{A} E(r_a, \hat{y}(x_a; \lambda)) \right\} \tag{2.7}$$

MERT has two critical drawbacks: it relies on having a fair amount of data for tuning, and it only relies on the data from the development set. Thus, these problems can lead to over-fitting to the specific characteristics of the development set.

Decoding

Given a source sentence x_1^J, the goal of the decoding process is to find its best translation hypothesis $\hat{y_1^I}$. In general, it is a hard problem since there is an exponential number of possible translations. Therefore, an exhaustive search of all possible translations—including scoring them and selecting the best one—is computationally very expensive. In fact, decoding is an NP-hard problem (Knight, 1999). To overcome this, different heuristic have been proposed to obtain a translation which is very close to $\hat{y_1^I}$. Some examples of these methods are the multi-stack depth-first decoding algorithm (Ortiz-Martínez, 2011); greedy strategies (Germann et al., 2001; Wang, 1998); and the search algorithm by Tillmann and Ney (2003), which is an adaptation of the classic algorithm for speech recognition (Jelinek, 1997).

Example 2.2: Example of the decoding procedure for the source sentence *Ana vive en la casa verde.*. $\mathbf{k_x}$ which words of the source sentence x_1^J have been translated at that point. _ indicates that the word x_i has not been translated yet and * indicates that it has already been translated.



Example 2.2 illustrates the procedure of translating the sentence *Ana vive en la casa verde.* following the phrases extracted in Example 2.1. Initially, the empty hypothesis is expanded into several partial hypotheses using different phrases. This leads to different coverage vectors k_x, which indicates which words of the source sentence x_1^J have already been translated. The hypothesis probability is computed as a product. Therefore, translating more source words leads to a lower probability mass being assigned to that specific hypothesis. Hypothesis expansion is done by expanding first those hypotheses with the highest probabilities. Thus, the algorithm prefers to expand hypotheses with fewer translated words. To avoid this, the algorithm uses the coverage vector. This allows that only those hypotheses with the same amount of translated words can compete among each other.

2.2 Neural machine translation

NMT adapts the SMT statistical framework (see Section 2.1) for its modeling with a neural model. Applying the chain rule of the probability over Eq. (2.1), this expression can be factorized into:

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^I} \prod_{i=1}^{I} Pr(y_i \mid y_1^{i-1}, x_1^J) \tag{2.8}$$

where x_1^J is a source sentence of length J and y_1^I is a target sentence of length I.

A neural model with parameters Θ can directly model this probability. Taking logarithms for the sake of numerical stability, we obtain:

$$\hat{y}_1^{\hat{I}} \approx \arg\max_{I, y_1^I} \sum_{i=1}^{I} \log p(y_i \mid y_1^{i-1}, x_1^J; \Theta) \tag{2.9}$$

Θ is usually estimated from a parallel corpora $\{(x^{(s)}, y^{(s)})\}_{s=1}^S$ composed of S sentences. The training objective usually consists in finding the set of parameters $\hat{\Theta}$ that minimizes the minus log-likelihood on the training set:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{s=1}^{S} \sum_{i=1}^{I^{(s)}} \log p(y_i \mid y_1^{i-1^{(s)}}, x^{(s)}; \Theta) \tag{2.10}$$

where $I^{(s)}$ is the length of the s-th target sentence and $y_1^{i-1^{(s)}}$ denotes the s-th target sentence up to the position $i-1$.

NMT tackles the SMT problem by means of a neural network whose components are jointly estimated. More precisely, is addresses the three main challenges posed by SMT (see Section 2.1) as follows:

1. **Model definition**: NMT's models are large neural networks that direct approximate the probability distribution $Pr(y_1^I \mid x_1^J)$. More information about their architectures is detailed at Section 2.2.1.
2. **Parameter estimation**: Model's parameters are typically estimated by means of gradient descend, under a maximum likelihood approach. This

represents a key difference with respect to SMT: NMT's parameters are jointly estimated while SMT's parameters are trained independently and combined by means of the log-linear model.

3. **Search problem**: most NMT systems use beam search (Lowerre, 1976) for finding the best translation.

2.2.1 Architectures

In this thesis, we are going to center on two different NMT architectures: recurrent neural network (RNN) (Jordan, 1990; Elman, 1990; Hochreiter and Schmidhuber, 1997) encoder-decoder with attention and Transformer (Vaswani et al., 2017). Note that, however, different architectures have been proposed thorough the years. For example, ConvNets (Kalchbrenner and Blunsom, 2013); convolutional sequence-to-sequence with attention mechanisms (Gehring et al., 2017); or multidimensional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks (Bahar et al., 2018).

RNN encoder-decoder with attention

Fig. 2.1 illustrates the architecture of the RNN encoder-decoder with attention.

Encoder The system's input is a sequence of tokens $x_1^J = x_1, \dots, x_J$. Each element belongs to a finite vocabulary χ and is codified with a unique index from 1 to $|\chi|$. Each word x_j is, then, projected into a continuous space by means of an embedding matrix:

$$\mathbf{x}_j = \mathbf{E}_s(x_j) \tag{2.11}$$

where $\mathbf{E}_s \in \mathbb{R}^{e \times |\chi|}$ is the embedding matrix of the source language; being e the embedding size and $E_s(x_j)$ the row of the embedding matrix corresponding to the element x_j.

A bidirectional LSTM[1] RNN encoder (f_e) processes the sequence $\mathbf{x}_1, \dots, \mathbf{x}_J$. The concatenation of their hidden states is often used as the combination function of forward and backward layers. Thus, we obtain a sequence of states—known as *annotations*—from the sequence of the embeddings that model the dependencies across the sequence. They are computed as:

[1] In this thesis we are only going to use LSTM. However, other type of cells are possible.

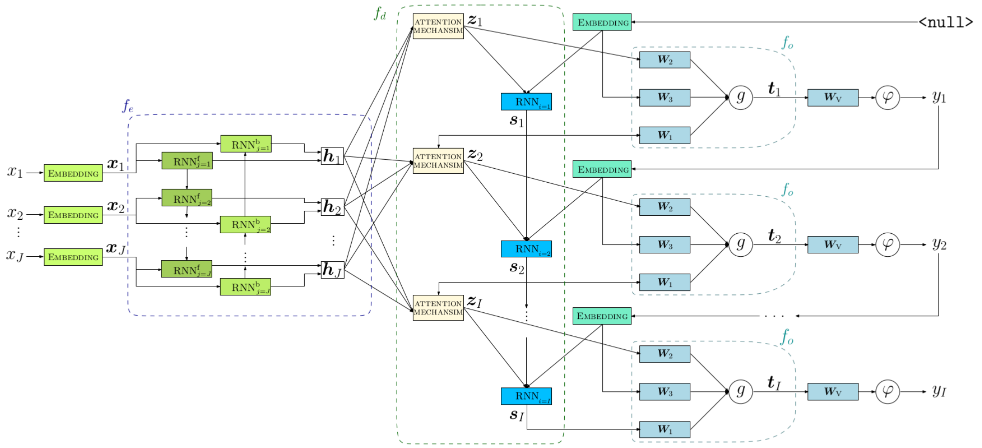


Figure 2.1: RNN encoder–decoder with attention. $x_1, \dots, x_J$ is the sequence of source words, projected into the continuous space by means of an embedding matrix. This sequence of embeddings are processed by an encoder f_e: a bidirectional RNN, yielding a sequence of annotations $(\mathbf{h}_1^J)$. This sequence is the input of the decoder RNN f_d, consisting in decoder RNN with an attention mechanism, followed by a deep output function (f_o) and a fully-connected output layer $(\mathbf{W}_V)$. Finally, the softmax function (φ) is used for obtaining the probabilities of the target words. Figure extracted from Peris et al. (2017).

$$\mathbf{h}_1^J = f_e(x_1^J) \tag{2.12}$$

where $\mathbf{h}_1^J$ is a sequence of J annotations in which each element $h_j \in \mathbf{h}_1^J; 1 \leq j \leq J$ can be seen as a representation of size k of the elements around the position j of the source sequence. This RNN encoder can be composed of several stacked layers (Wu et al., 2016). If the encoder is a deep network, $\mathbf{h}_1^J$ is made of the hidden states from the top layer in the stack.

Decoder The decoder models the translation probability factorized in Eq. (2.8). It is composed of an RNN with attention mechanism followed by a deep output layer. The decoder provides the RNN of autoregressive capabilities by performing the recurrence over the sequence of previously generated tokens, which are introduced to the decoder via their embedding following Eq. (2.9). As in the encoder, we made use of an LSTM architecture.

The attention mechanism bridges the sequence of annotations computed by the encoder with the hidden state of the decoder. At each decoding step i, the attention mechanism computes a context vector $\mathbf{z}_i$ as follows:

$$\mathbf{z}_i = \varphi(a(\mathbf{s}_{i-1}, \mathbf{h}_1^J))\mathbf{h}_1^J \tag{2.13}$$

where φ is the softmax function and $\varphi(a(\mathbf{s}_{i-1}, \mathbf{h}_1^J))$ computes the attention weights of the annotations at the i-th decoding step. Thus, at this time step, the decoder computes a hidden state $\mathbf{s}_i$ considering the context vector ($\mathbf{z}_i$), the word embedding of the previous generated token ($\mathbf{E}_t(y_{i-1})$) and the previous hidden state ($\mathbf{s}_{i-1}$) following:

$$\mathbf{s}_i = f_d(\mathbf{E}_t(y_{i-1}), \mathbf{s}_{i-1}, \mathbf{z}_i) \tag{2.14}$$

where f_d is the recurrent function with attention; $\mathbf{s}_i \in \mathbb{R}^q$ is the hidden state of the decoder RNN, of dimension q; $\mathbf{E}_t \in \mathbb{R}^{d \times |\mathcal{Y}|}$ is the embedding matrix of the target language; $\mathcal{Y}$ is the finite target vocabulary; and d is the dimension of the target word embedding.

Usually, the first state of the decoder is initialized according to the function f_i:

$$\mathbf{s}_0 = f_i(\mathbf{h}_1^J) \tag{2.15}$$

This function is frequently defined as a number of fully-connected layers with an average representation of the annotations (Sennrich et al., 2017) or the last state of the backward encoder (Bahdanau et al., 2015) as input. The decoder's output state ($\mathbf{s}_i$) is combined with the context vector ($\mathbf{z}_i$) computed by the attention mechanism and the embedding of the previously generated word ($\mathbf{E}_t(y_{i-1})$) in a deep output layer (Pascanu et al., 2013), which applies the function f_o to obtain an l-sized intermediate representation:

$$f_o(\mathbf{s}_i, \mathbf{z}_i, \mathbf{E}_t(y_{i-1})) = g(\mathbf{W}_1\mathbf{s}_i, \mathbf{W}_2\mathbf{z}_i, \mathbf{W}_3\mathbf{E}_t(y_{i-1}) + \mathbf{b}) \tag{2.16}$$

where the non-linear function g is usually a tanh; $\mathbf{W}_1 \in \mathbb{R}^{l \times q}$, $\mathbf{W}_2 \in \mathbb{R}^{l \times k}$, $\mathbf{W}_3 \in \mathbb{R}^{l \times d}$ and $\mathbf{b} \in \mathbb{R}^l$ are trainable weights.

Using a vocabulary-sized fully-connected layer, this intermediate representation is projected to the space of the target vocabulary. Then, we apply the softmax function to obtain a probability distribution over the target vocabulary $\mathbf{p}_i$:

$$\mathbf{p}_i = \varphi(\mathbf{W}_V\mathbf{t}_i + \mathbf{b}_V) \tag{2.17}$$

where $\mathbf{W}_V \in \mathbb{R}^{|\mathscr{Y}| \times l}$ and $\mathbf{b}_V \in \mathbb{R}^{|\mathscr{Y}|}$ are the weights to learn and $\mathbf{p}_i$ is the probability distribution defined by Eq. (2.9). At a time step i, the probability of the token y is given by its corresponding position in $\mathbf{p}_i$:

$$p(y_i = y \mid y_1^{i-1}, x_1^J; \Theta) = \bar{\mathbf{y}}^\top \mathbf{p}_i \tag{2.18}$$

where $\bar{\mathbf{y}} \in [0,1]^{|\mathscr{Y}|}$ is the one-hot codification of the token y.

Transformer

This model is based on the application of attention mechanisms. Thus, it is able to compute different representations of the source and target sequences. Moreover, since the model does not have any recurrences, the training can be parallelized to a greater extent than RNN models. Furthermore, it is capable to model larger context (Agrawal et al., 2018). The Transformer model has gained popularity in the recent years, becoming the new state of the art in MT (Barrault et al., 2020). However, it suffers from a weakness: it is extremely sensitive to hyperparameters, making it hard to find a working configuration. Additionally, they require large amounts of training data for yielding a good performance, which is a problem when working with historical data due to its scarceness (Bollmann and Søgaard, 2016).

Similarly to RNN models, Transformer also follows an encoder–decoder approach: a representation of the source sequence is computed by the encoder and, then, the decoder generates the translated sentence from this representation. Fig. 2.2 illustrates the model's architecture.

Positional information Like RNN, Transformer's inputs and outputs are the sequence of elements of source and target sentences, shifted one time-step to the right following the teacher forcing training scheme.

The elements from the discrete vocabulary spaces are projected into a continuous space via embedding matrices. Then, Eq. (2.12) is applied to the input sequence x_1^J, obtaining a sequence of J embeddings of dimension d_m: $\mathbf{x}_1, \ldots, \mathbf{x}_J$. Additionally, due to the recurrence being dropped from the model, positional information needs to be injected into the sequence representation. This can be done via positional encodings (Gehring et al., 2017). These encodings are a sequence of vectors $\mathbf{e}_1, \ldots, \mathbf{e}_J$ which provide positional information to the sequence. Each vector is a m-dimensional vector constructed using sinusoidal signals ac-

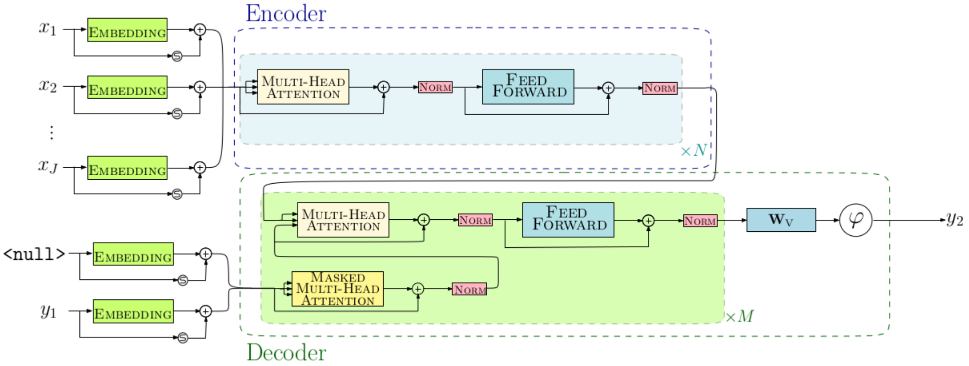


Figure 2.2: The Transformer model. The input of the system is a sequence of words $x_1, \ldots, x_J$, projected into the continuous space via an embedding matrix. These embeddings are augmented with positional information to have a notion of sequentiality. The encoder is a stack of N layers. Each layer features a multi-head attention mechanism followed by a feed-forward layer. The decoder is another stack of M layers. Previous words are encoded similarly as input words, but using a masked multi-head attention mechanism. Next, input and output representations are combined through another multi-head attention mechanism and feed-forward layers. The representation of the last decoder layer is projected to the target language vocabulary space. Finally, a softmax function computes the probabilities in this space. Figure extracted from Peris et al. (2017).

cording to its position within the sequence. Thus, each element from $\mathbf{e}_j$ is defined as:

$$e_{j,k} = \begin{cases} \sin(\frac{j}{1000^{\frac{2k}{d_m}}}) \text{ if } k \text{ is even} \\ \\ \cos(\frac{j}{1000^{\frac{2k}{d_m}}}) \text{ if } k \text{ is odd} \end{cases}, \text{ for } 0 \leq k \leq d_m \text{ and } 1 \leq j \leq J \qquad (2.19)$$

To obtain a sequence of position-aware embeddings, these positional information is added to the regular embeddings:

$$\bar{\mathbf{x}}_1, \ldots \bar{\mathbf{x}}_J = \mathbf{x}_1 + \mathbf{e}_1, \ldots, \mathbf{x}_J + \mathbf{e}_J \qquad (2.20)$$

Multi-head attention Multi-head attention is an extension of the regular attention mechanism, which allows learning representations of different sub-spaces at different positions. Since it is devised for non-recurrent architectures, there are no dependencies between the different query vectors. Thus, the operations can be parallelized by stacking T' query vectors as a sequence of queries $q_1^{T'}$; where each $q_{t'} \in \mathbb{R}$; for $1 \leq t' \leq T'$.

The multi-head attention model computes this attention in parallel, over H different, learned projections (of size d_m) of the queries, keys and values. These projections are computed as follows:

$$\begin{aligned} \bar{\mathbf{q}}_1^{T'} &= \mathbf{W}_q \mathbf{q}_{t'};\ \text{for } 1 \leq t' \leq T' \\ \bar{\mathbf{k}}_1^{T} &= \mathbf{W}_k \mathbf{k}_{t};\ \text{for } 1 \leq t \leq T \\ \bar{\mathbf{v}}_1^{T} &= \mathbf{W}_v \mathbf{v}_{t};\ \text{for } 1 \leq t \leq T \end{aligned} \tag{2.21}$$

where each $\mathbf{W}_q \in \mathbb{R}^{d_m \times q}$, $\mathbf{W}_k \in \mathbb{R}^{d_m \times k}$ and $\mathbf{W}_v \in \mathbb{R}^{d_m \times v}$ are the trainable matrices.

Once the queries, keys and values have been projected, the multi-head attention model applies an attention mechanism in parallel all the elements of these sequences:

$$\mathbf{H}_h = \varphi(a'(\mathbf{q}_1^{T'}, \mathbf{k}_1^{T}))\mathbf{v}_1^{T} \tag{2.22}$$

where $a'(\mathbf{q}_1^{T'}, \mathbf{k}_1^{T})$ denotes an attention function that computes a sequence of scores for each one of the elements of the sequence $\mathbf{q}_1^{T'}$ in parallel and normalized applying the softmax function (φ). $\mathbf{H}_h \in \mathbb{R}^{T' \times H d_m}$ for $1 \leq h \leq H$ are called heads and the typical attention function used for multi-head attention is the scaled dot product (Vaswani et al., 2017).

The heads are concatenated into a matrix ($[\mathbf{H_1}; \ldots; \mathbf{H_H}] \in \mathbb{R}^{T' \times H d_m}$) and projected into an output space of o dimensions by means of a trainable matrix $\mathbf{W}_o \in \mathbb{R}^{H d_m \times o}$. This way, the multi-head attention function is defined as:

$$\gamma(\mathbf{q}_1^{T'}, \mathbf{k}_1^{T}, \mathbf{v}_1^{T}) = [\mathbf{H_1}; \ldots; \mathbf{H_H}]\mathbf{W}_o \tag{2.23}$$

Encoder The encoder is a stack of N layers—the structure of which is a multi-head attention mechanism—followed by a feed-forward network. All sublayers have residual connections, whose results are normalized via a normalization layer after applying dropout. For the sake of simplicity, the layer normalization, stack of layers and dropout operations will be omitted in the notation. Thus, the input of this layer is a sequence of J inputs of d_m dimensions $(\mathbf{h}_1^J)$. Therefore, the encoder is defined as:

$$\mathbf{h}_1^J = f_F(\mathbf{h}_1^J + \gamma(\mathbf{h}_1^J, \mathbf{h}_1^J, \mathbf{h}_1^J)) + \mathbf{h}_1^J + \gamma(\mathbf{h}_1^J, \mathbf{h}_1^J, \mathbf{h}_1^J) \tag{2.24}$$

where γ is the multi-head attention defined in Eq. (2.23) and f_F is a 2-layered feed-forward network, with a rectified linear unit (ReLU) and linear activations:

$$f_F(\mathbf{h}_1^J) = \text{ReLU}(\mathbf{h}_j\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \forall \mathbf{h}_j \in \mathbf{h}_1^J \tag{2.25}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_n \times d_F}$, $\mathbf{b}_1 \in \mathbb{R}^{d_f}$, $\mathbf{W}_2 \in \mathbb{R}^{d_F \times d_m}$ and $\mathbf{b}_2 \in \mathbb{R}^{d_m}$ are the parameters to learn.

The inputs to the first layer of the encoder stack $(\mathbf{h}_1^J)$ are position-aware embeddings computed by Eq. (2.20), which are also regularized via dropout. Note that the attention is applied over the same sequence by the encoder. Thus, it is a self-attentional system, which computes representations at an intra-sequence level.

Decoder The decoder is another stack of M layers which can be separated in two different parts: The first part applies self-attention—similarly to the encoder—to encode the sequence of shifted outputs. The second part performs inter-sequences and generates the target sequence, bridging together the representation of both self-attention modules.

As previously described, positional information is injected to the sequence of shifted output embeddings, producing a sequence $\bar{\mathbf{a}}_1^I = \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_I$ of I embeddings of the shifted output. Then, similarly as the encoder, it applies a self-attention mechanism (γ_1) to this sequence. Note that, to prevent the decoder to look into future elements of the sequence, this attention is masked. The self-attended representation $(\mathbf{a}_1^I)$ is computed as:

$$\mathbf{a}_1^I = \mathbf{a}_1^I + \gamma_1(\mathbf{a}_1^I, \mathbf{a}_1^I, \mathbf{a}_1^I) \tag{2.26}$$

where m is the m-th decoding layer (for $0 \leq m \leq M$) and $\mathbf{a}_i \in \mathbb{R}^{d_m}$.

For bridging together the representations obtained from the input and the shifted output sequences, the second part of the decoder must compute an inter-sequence attention. To that effect, an inter-sequential multi-head attention mechanism (γ_2) is applied. The keys and values from this mechanism come from the source sequence ($\mathbf{h}_1^J$). The queries come from the target sequence ($\mathbf{a}_1^I$). Like the encoder, the multi-head attention mechanism is followed by a feed-forward network that computes an output sequence of I elements ($\mathbf{z}_1^I$) at each decoder layer (for $0 \leq m \leq M$). This output is the input of the following layer in the decoder stack:

$$\mathbf{z}_1^I = f_F(\mathbf{a}_1^I + \gamma_2(\mathbf{a}_1^I, \mathbf{h}_1^J, \mathbf{h}_1^J)) + \mathbf{a}_1^I + \gamma_2(\mathbf{a}_1^I, \mathbf{h}_1^J, \mathbf{h}_1^J) \tag{2.27}$$

Then, following the stack of layers of the decoder, the same fully-connected layer with a softmax activation than for RNN (Eqs. (2.17) and (2.18)) are applied to the decoder's outputs ($\mathbf{z}_1^I$):

$$\mathbf{p}_i = \varphi(\mathbf{W}_V \mathbf{z}_i + \mathbf{b}_v) \tag{2.28}$$

where $\mathbf{W}_V \in \mathbb{R}^{|\mathcal{Y}| \times d_m}$ and $\mathbf{b}_v \in \mathbb{R}^{|\mathcal{Y}|}$ are the weights to learn.

2.2.2 Subword NMT

The size of the vocabulary is a limiting problem in NMT: while MT is an open-vocabulary task, NMT models require finite vocabularies. Source and target vocabulary's elements are mapped into unique indexes and projected to the continuous space via embedding matrices. Thus, the size of these matrices are proportional to the vocabulary size. Furthermore, a normalization through the full target vocabulary needs to be computed for the output layer. Thus, models with a very large vocabulary are impractical to use. Moreover, while using the system, new unknown words can appear and the system should be able to deal with them.

To mitigate this problem, Sennrich et al. (2016b) proposed to use a compression algorithm to encode words as sequences of smaller units and use this sequences of subwords as translations units—instead of using sequences of words. More precisely, they proposed the use of the byte pair encoding (BPE) (Gage, 1994) algorithm which iteratively merges the most common pair of bytes into a single, unused byte.

Applied to word segmentation, BPE starts by splitting the data into characters, adding a special symbol that denotes the end of a word. Then, it iteratively merges the two most common consecutive symbols into a new, unused symbol. This process is repeated until the stopping criterion—typically, a set number of merges—is reached. The size of the final vocabulary will be equal to the number of merge operations conducted plus the size of the initial vocabulary (after splitting the data into characters). Additionally, since the special end-of-word symbol is maintained during the whole process, reverting the encoding can be trivially done with a replace operation. Thus, this technique is able to represent words with different granularity: the rarest words will be represented as sequences of subwords while the most common words will be maintained at a word level or a closer representation.

BPE has become one of the standard methods for building NMT systems (Wu et al., 2016). It is usually applied jointly to source and target corpora (provided that both languages share the same alphabet). This prevents segmentation inconsistencies and ensures that the same words (e.g., proper nouns) are always segmented in the same way.

2.2.3 Synthetic data generation

Since there is not always enough parallel training data available, a common procedure in MT is to generate synthetic data from a monolingual corpus. In this section, we present the most frequent method used by the MT community—especially in resource-poor scenarios (Poncelas et al., 2018)—which is the one that will be used in this thesis. Additionally, we present a data selection technique in order to filter the monolingual data and select only the segments that will benefit more our system.

Backtranslation

Backtranslation (Sennrich et al., 2016a) is a technique for generating a synthetic data from a monolingual corpus in the target language. Given an MT system to translate from the target language into the source language, the idea is to translate the target monolingual data with this system to create a synthetic source. Then, the resulting translations become the source part of the synthetic parallel data and the monolingual corpus will become the target part.

Data selection

Data selection aims to increase the training data by selecting from corpora from other domains only those sentences that are similar to the text to translate. This can be leveraged for creating better synthetic data: instead of generating the synthetic corpus from the complete monolingual corpus (which often consists in all the monolingual data available), we first apply a data selection technique in order to filter it to only the sentences most similar to the source sentences from the test.

In this thesis, we made use of one of the most common data selection techniques: feature decay algorithm (FDA) (Biçici and Yuret, 2015). Given an in-domain document (which could be part of the training data or, if no data is available, the document to translate) and the out-of-domain data (which could be either parallel corpora or monolingual data), this technique aims to maximize the coverage of a set of features extracted from the in-domain document.

Chapter 3
Interactive Machine Translation

aa KAMI-SAMA onegai
futari dake no Dream Time kudasai
o-ki ni iri no usa-chan daite kon'ya mo OYASUMI

(***Fuwa Fuwa Time***. Ho-kago Tea Time.)

Oh, Kanae, please
Please only have two of them.
You and Usage are good for you, okay, good night

(***Fluffy fluffy***. Google Translate.)

Contents

Due to their inability to produce error-free translations, human translators need to revise and correct machine translation (MT) hypothesis in a process known as *post-editing*. Interactive machine translation (IMT) proposes a collaborative framework in which human and machine work together to produce the final high-quality translations. In one of the first approaches, users corrected the leftmost wrong word from the system's hypothesis. With this correction, they indicated the system that all the words that preceded the correction were also right. The system reacted to their feedback by generating a new suffix that completed the validated prefix to form a new hypothesis. This procedure was repeated until users were satisfied with the system's hypothesis. This approach is known as prefix-based IMT and will be described with more detail in Section 3.2.

In the rest of the chapter, we review the state of the art in IMT. Then, we describe one of the classical left-to-right protocols. After that, we propose a new protocol that breaks this left-to-right limitation, and we extend it with an active prediction module that helps users to make corrections. Following that, we conduct a series of experiments in order to assess the quality of our proposals. Then, we show and comment the results of the evaluation. Finally, we qualitatively analyze the strengths and weakness of our protocol and reach some conclusions.

3.1 State of the art

The IMT paradigm was introduced during the *TransType* project (Foster et al., 1997) and was further developed during *TransType2* (Barrachina et al., 2009) and *CasMaCat* (Alabau et al., 2013). Sanchis-Trilles et al. (2008) extended the user's protocol by profiting from the use of the mouse for validating a prefix and suggesting a new suffix each time a user clicked on a position to type a word. González-Rubio et al. (2010) improved the prefix generation step by making use of confidence measures to assist users in the prefix validation. Koehn et al. (2014) proposed a character-based approach for the suffix generation: for each character of the word correction that users type, the system's provide a new suffix. Torregrosa et al. (2014) made a similar proposal, but their system offered several suffix suggestions for the user to select. Alabau et al. (2011) and Alabau et al. (2014) integrated handwriting and speech recognition by introducing multimodal interaction into the IMT environment. Nepveu et al. (2004); Ortiz-Martínez (2016) applied online learning techniques to their systems in order to profit from the user's feedback to improve their systems. Azadi and Khadivi (2015) improved the search strategy for the suffix generation. Marie and Max (2015) introduced a touch-based interaction to iteratively improve translation quality. A new framework in which, at each iteration, a user corrected the most critical error from the translation hypothesis was presented (Cheng et al., 2016).

With the rise of neural machine translation (NMT), the interactive framework was deployed into the neural systems (Knowles and Koehn, 2016; Wuebker et al., 2016; Peris et al., 2017). Online learning techniques were added (Peris and Casacuberta, 2019). Kreutzer and Riezler (2019) proposed the use of self-regulation strategies that learn which type of feedback to query from a human teacher. The use of confidence measures (CM) was integrated into interactive neural machine translation (INMT) (Knowles and Koehn, 2018; Navarro and Casacuberta, 2021a). Santy et al. (2019) presented a demonstrator of an INMT system that assists human translators with on-the-fly hints and suggestions. A user study comparing IMT and INMT was conducted with the help of professional translators (Daems and Macken, 2019). Knowles et al. (2019) conduced a user study with professional translators to evaluate the productivity of prefix-based INMT. Reinforcement and imitation learning was applied (Lam et al., 2019). Syntax-aware INMT was proposed (Gupta et al., 2020; Zhao et al., 2020). Lin et al. (2021) proposed to use word-level autocompletors. Finally, bandit feedback was applied (Navarro and Casacuberta, 2021b).

3.2 Prefix-based IMT

In this protocol, the system proposes an initial translation y_1^I of length I. Then, the user reviews it and corrects the leftmost wrong word y_i. Inherently, this correction validates all the words that precede this correction, forming a validated prefix $\tilde{y}_1^i$, that includes the corrected word $\tilde{y}_i$. Immediately, the system reacts to this user feedback ($f = \tilde{y}_1^i$), generating a suffix $\hat{y}_{i+1}^{\hat{I}}$ that completes $\tilde{y}_1^i$ to obtain a new translation of x_1^J : $\hat{y}_i^I = \tilde{y}_1^i \, \hat{y}_{i+1}^{\hat{I}}$. This process is repeated until the user accepts the system's complete suggestion. Fig. 3.1 illustrates this protocol.

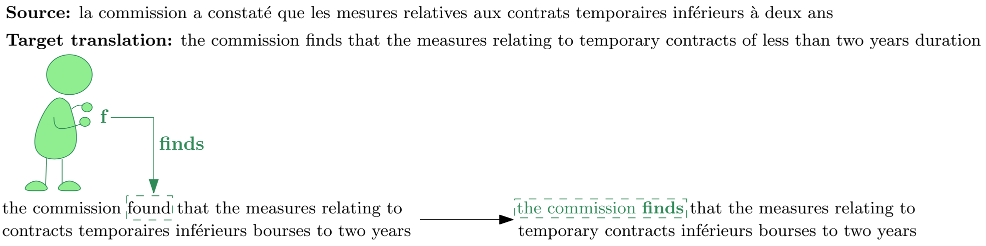
The suffix generation was formalized by Barrachina et al. (2009) as follows:

$$\hat{y}_{i+1}^{\hat{I}} = \arg\max_{I, y_{i+1}^I} Pr(y_{i+1}^I \mid x_1^J, f = \tilde{y}_1^i) \tag{3.1}$$

which can be rewritten as:

$$\hat{y}_{i+1}^{\hat{I}} = \arg\max_{I, y_{i+1}^I} Pr(\tilde{y}_1^i \, y_{i+1}^I \mid x_1^J) \tag{3.2}$$

This equation is very similar to Eq. (2.1): at each iteration, the process consists in a regular search in the translations space but constrained by the prefix $\tilde{y}_1^i$.



Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

the commission found that the measures relating to contracts temporaires inférieurs bourses to two years

the commission **finds** that the measures relating to temporary contracts inférieurs bourses to two years

Figure 3.1: Illustration of the prefix-based IMT protocol. The user corrects the leftmost wrong word (*found*), validating the prefix *the commission finds*. Then, the system generates a new hypothesis coherent with the user's feedback.

3.2.1 Implementation

Our implementation of the prefix-based IMT protocol was done following the procedure described by Barrachina et al. (2009) of exploring a word graph and generating the best suffix for a given prefix. For each sentence to translate, we generated a word graph. Then, considering the word graph as a weighted finite-state automaton, we parsed the validated prefix over the correspondent word graph—from the initial state to any other intermediate state—to find the best path that accounted for the prefix. Finally, we obtained the corresponding suffix searching for the best path from the intermediate state to the final state.

Our implementation of prefix-based IMT is, therefore, consistent with Barrachina et al. (2009), considering that we generate word graphs with *Moses* (Koehn et al., 2007) while they used finite state translators.

3.3 Segment-based IMT

This protocol extends the human–computer collaboration from prefix-based IMT. Now, at each iteration, users can validate segments (sequences of words), combine consecutive segments—deleting all the words between them (if any)—to create a larger one or correct a word. Fig. 3.2 illustrates this methodology.

Like with the prefix-based protocol, the process starts with the system suggesting an initial translation. Then, users review it and validate those sequences of words which they consider to be correct. Following that, they can delete words between

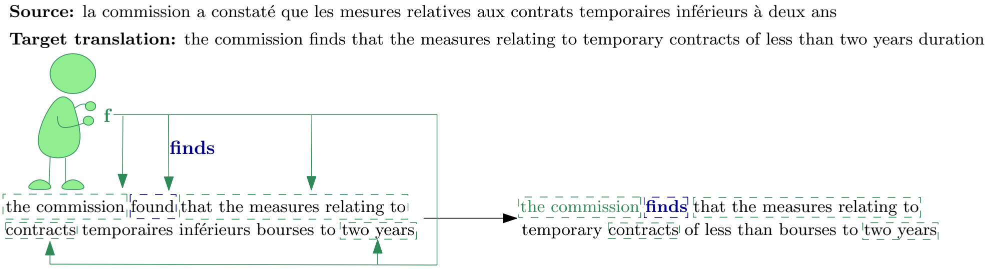

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans
Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration
f
finds
the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years
the commission finds that the measures relating to
temporary contracts of less than bourses to two years

Figure 3.2: Illustration of the segment-based IMT protocol. The user validates the segments the commission , that the measures relating to , contracts and two years and corrects the word *found*. Then, the system generates a new hypothesis coherent with the user's feedback.

validated segments to create a larger segment. Finally, they correct a word. Example 3.1 exemplifies the possible user actions.

Example 3.1: Example of the possible user actions in segment-based IMT. The process starts with a user validating the correct word sequences (If you have been exposed , you should , go and your doctor for tests). Then, they delete some words (~~consult~~) to create a bigger segment (If you have been exposed , you should go). Finally, they make a word correction (**to** which is added between two validated segments).

Reference: If you have been exposed , you should go to your doctor for tests

Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: If you have been exposed , you should consult go your doctor for tests

Words deletion: If you have been exposed , you should ~~consult~~ go your doctor for tests

Word correction: If you have been exposed , you should go **to** your doctor for tests

These three actions constitute the user's feedback, which has the form $\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$; where $\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$ is the sequence of N correct segments validated by the user in an interaction. Each segment is defined as a sequence of one or more target words. Therefore, each action taken by the user modifies the feedback differently. Thus, a user can:

1. Validate a new segment, inserting a new segment $\tilde{\mathbf{f}}_i$ in $\tilde{\mathbf{f}}_1^N$.

2. Delete words between two segments, merging two consecutive segments $\tilde{\mathbf{f}}_i$, $\tilde{\mathbf{f}}_{i+1}$ into a new one.

3. Introduce a word correction. This is introduced as a new one-word validated segment, $\tilde{\mathbf{f}}_i$, which is inserted in $\tilde{\mathbf{f}}_1^N$.

The first two actions are optional: at a given iteration, users might not validate new segments or delete words. The last action is mandatory: the word correction triggers the system to react to the user's feedback, starting a new iteration of the process.

The system's reaction to the user's feedback results in a sequence of new translation segments $\widehat{\mathbf{h}}_0^{N+1} = \widehat{\mathbf{h}}_0, \dots, \widehat{\mathbf{h}}_{N+1}$. That means, an $\widehat{\mathbf{h}}_i$ for each pair of validated segments $\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i+1}$, being $1 \leq i \leq N$; plus one more at the beginning of the hypothesis, $\widehat{\mathbf{h}}_0$; and another at the end of the hypothesis, $\widehat{\mathbf{h}}_{N+1}$. The new translation of x_1^J is obtained by alternating validated and non-validated segments: $\hat{y}_1^I = \widehat{\mathbf{h}}_0, \tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_N, \widehat{\mathbf{h}}_{N+1}$. The goal is to obtain the best sequence of translation segments, given the user's feedback and the source sentence:

$$\widehat{\mathbf{h}}_0^{N+1} = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0^{N+1} \mid x_1^J, \tilde{\mathbf{f}}_1^N) \tag{3.3}$$

which can be rewritten as:

$$\widehat{\mathbf{h}}_0^{N+1} = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0, \tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_N, \mathbf{h}_{N+1} \mid x_1^J) \tag{3.4}$$

This last equation is very similar to the classical prefix-based IMT equation (Eq. (3.2)). Now, the search is performed in the space of possible substrings of the translations of x_1^J, constrained by the sequence of segments $\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_N$, instead of being limited to the space of suffixes constrained by $\tilde{y}_1^i$, as in Eq. (3.2).

3.3.1 Implementation

Moses (Koehn et al., 2007) decoder has an eXtensible Markup Language (XML) scheme which allows us to specify the desired translation of parts of a sentence. Taking advantage of this scheme, we are able to validate segments of a translation hypothesis without altering the models. More precisely, we use the *exclusive* mode of this scheme, which only takes into account the given translation of a part of a sentence, ignoring any phrases from the phrase table that overlap with that span. With this, we can constrain the search process to follow Eq. (3.4). Example 3.2 shows an example of a sentence in XML markup language. More details about the XML scheme are presented at Section 3.3.1.

Example 3.2: Example of a sentence in XML markup language in which we are able to specify the desired translation for some parts of the sentence: *Si vous avez été exposé , vous devriez* must be translated as *If you have been exposed , you should* and *votre médecin pour des tests* as *your doctor for tests.*

```
<x translation=‶The commission" >la commission</x>
<x translation=‶finds" >a constaté</x> <x translation=‶that the measures
relating to contracts" >que les mesures relatives aux contrats</x>
temporaires inférieurs à <x translation=‶two years" >deux ans</x>
```

We built a prototype that manages the interaction between user and the statistical machine translation (SMT) system. It takes into account the user's feedback, generates a new translation with *Moses* and suggests the new hypothesis to the user. This has an average response time of 90 ms[1]—which, according to Nielsen (1993), is below *"the limit for having a user feel that the system is reacting instantaneously"*.

In segment-based IMT, the user's feedback comes from three different actions: validating segments, correcting words and merging segments. However, the first two actions affect the generation of the new XML markup sentence in the same way. Therefore, we only apply two different operations to the XML:

Segment validation: for each segment validated by the user, we align the words of that target segment with their correspondent source words (phrase alignments) and generate an XML tag to indicate the desired translation of those source words.

Word correction: Each time a user corrects a word, the new word is aligned with its correspondent source words using hidden Markov alignment models (HMM) (Vogel et al., 1996) to compute the alignment probability between the new word and the non-validated source words. Then, we generate an XML tag to indicate that the translation of those source words is the val-

[1] Tested on a machine with an Intel i5 CPU at 3.1 GHz.

idated word. These alignments are computed with *mgiza* (Gao and Vogel, 2008).

This software is open-sourced and publicly available[2].

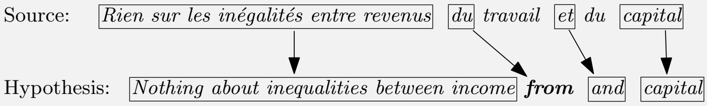
XML generation

To construct the XML sequence, we needed to associate the validated target segments with their corresponding source words. To that effect, we defined a source segment as a sequence of source words which are associated to a validated segment. In this section, we describe and discuss some problems that arose during the implementation of the segment-based protocol and the design decisions we took for overcoming these problems.

Segment reorders

Since users validate segments taking into account their order of appearance in the hypothesis, this order must be maintained in future iterations. To that effect, we made use of the *wall* reordering constraint. This feature ensures that all words left to a wall are translated before considering the rest of the sentence. Thus, the reordering model is no longer able to reorder words located in different sides of a wall. Example 3.3 illustrates an example of an XML instance using walls.

Example 3.3: Example of the usage of the XML scheme's *wall* feature to avoid that the reordering model alters the order of the validated segments.)

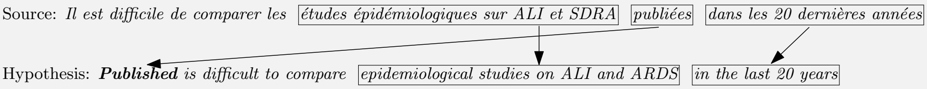

Source:
Rien sur les inégalités entre revenus
du
travail
et
du
capital
Hypothesis:
Nothing about inequalities between income
from
and
capital

XML: <x translation=‶Nothing about the inequalities between income" >
Rien sur les inégalités entre revenus</x><wall/>
<x translation=‶from" >du</x><wall/> travail <x translation=‶and" >
et</x><wall/> du <x translation=‶capital" >capital</x><wall/>
Translation: *Nothing about inequalities between income* *from* *work* *and* *capital*

[2] https://github.com/midobal/sb-imt.

Another problem related with the order of the segments arises when source and target segments have a different order. This could cause a wrong reordering of the target segments in future translation hypotheses. Our user model assumes that validated segments will not be reordered. Thus, we must ensure that the segment's order is not altered along the process. To that effect, after generating the new translation hypothesis with *Moses*, we reorder the validated segments to match the ordering provided by the user. Example 3.4 illustrates how to generate a translation using this solution.

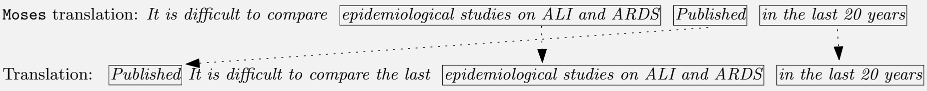
Example 3.4: Example of a sentence in XML markup language in which source and target segments are ordered differently. The user has validated the segments **Published** , epidemiological studies on ALI and ARDS and in the last 20 years). However, due to the difference in order between source and target, these segments have a different order in the new hypothesis (epidemiological studies on ALI and ARDS , **Published** and in the last 20 years). As a solution, after creating the XML and generating the translation with *Moses*, we reorder the validated segments from translation to match the order indicated by the user. Arrows represent alignments between source and target validated segments. Dashed arrows represent the change in position of target validated segments.



Source: *Il est difficile de comparer les* *études épidémiologiques sur ALI et SDRA* *publiées* *dans les 20 dernières années*

Hypothesis: ***Published*** *is difficult to compare* *epidemiological studies on ALI and ARDS* *in the last 20 years*

XML: `Il est difficile de comparer les <x translation=‟epidemiological studies on ALI and ARDS" >études épidémiologiques sur ALI et SDRA</x> <x translation=‟published" >publiées</x> <x translation=‟in the last 20 years" >dans les 20 dernières années</x>`

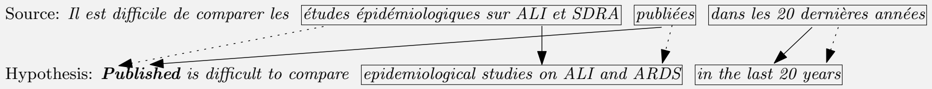


`Moses` translation: *It is difficult to compare* *epidemiological studies on ALI and ARDS* *Published* *in the last 20 years*

Translation: *Published* *It is difficult to compare the last* *epidemiological studies on ALI and ARDS* *in the last 20 years*

We need to take into account that, since the translation is constructed following the order of the source segments, this solution affects the language model. The XML scheme is strongly affected by the order in which the translation is constructed. Thus, the translation generated by *Moses* needs to be reordered to be coherent with the user's feedback. An alternative to this solution is to alter the way in which the XML is build: the first target segment is assigned as a translation of the first source segment; the second target segment as the translation of the second source segment; etc. Example 3.5 shows an example of this solution.

Example 3.5: Alternative to the solution to Example 3.4 for a sentence in XML markup language in which source and target are ordered differently. Dashed arrows represent how these segments have been aligned in the XML. To ensure that the validated segments respect the order indicated by the users in future translations, we modify the XML construction. Now, instead of their corresponding translation, the first source segment (*études épidémiologiques sur ALI et SDRA*) is assigned the translation of the first target segment (*Published*) and the second source segment (*publiées*) is assigned the translation of the second target segment (*epidemiological studies on ALI and ARDS in the last 20 years*). Arrows represent alignments between source and target validated segments.


Source: *Il est difficile de comparer les* *études épidémiologiques sur ALI et SDRA* *publiées* *dans les 20 dernières années*
Hypothesis: ***Published*** *is difficult to compare* *epidemiological studies on ALI and ARDS* *in the last 20 years*

XML: `Il est difficile de comparer les <x translation=‟Published">études épidémiologiques sur ALI et SDRA</x><wall/> <x translation=‟epidemiological studies on ALI and ARDS">publiées</x><wall/> <x translation=‟in the last 20 years">dans les dernières 20 années</x><wall/>`

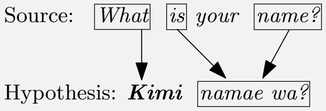
Translation: *It is difficult to compare the* *Published* *epidemiological studies on ALI and ARDS* *in the last 20 years*

This strategy has the benefit of not affecting the language model. However, the translation assigned to a given source segment might not be the real one. We tested both approaches and came to the conclusion than penalizing the language model is more severe than affecting the translation and reordering models. Thus, we followed this second strategy.

Non-consecutive corresponding sources

Due to word reordering between languages, a validated segment might be aligned with more than one source segment. Therefore, if we assigned to each source segment their corresponding translation, we could end up altering the word order in the target segment. To avoid this, we decided to assign the complete validated segment to the leftmost source segment and an empty translation to the rest of the source segments. Example 3.6 shows an example in which this situation happens.

Example 3.6: Example of generating an XML sequence when a validated segment has been originated by multiple source segments. The validated segment *namae wa?* has been generated by the source segments *is* and *name?* . If we assign to each source segment their corresponding translation (option 1) then the next translation hypothesis is wrong (the segment validated by the user has become a mixture of two new segments in a different order). However, if we assign the whole validated segment as the translation of the leftmost source segment (option 2), the new translation hypothesis is coherent with the user's feedback. Note that, for the sake of clarity, we have used a romanized version of Japanese for this example.



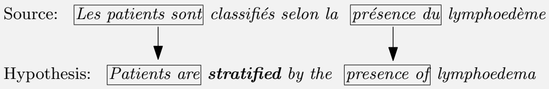
Option 1:
XML: <x translation="Kimi" >What</x> <x translation="wa" >is</x> your <x translation="namae?" >name?</x>
Translation: *Kimi* *wa* *no* *namae?*

Option 2:
XML: <x translation="Kimi" >What</x> <x translation="namae wa?" >is</x> your <x translation="" >name?</x>
Translation: *Kimi* *no* *namae wa?*

Words without corresponding source segment

Each time a user corrects a word, we need to find its corresponding source segment to generate the XML (see Section 3.3.1). However, there are cases in which we are unable to find it. For example, if the new word is an out-of-vocabulary or if its alignment probability with the sources is very low. When that happens, we are unable to generate an XML instance that accounts for the word correction. To cope with this problem, we create a new artificial source at the end of the sentence. Then, we generate the XML considering this artificial source as the corresponding source of the word corrected by the user. Example 3.7 shows an example in which this situation appears.

Example 3.7: Example of a sentence in XML markup language in which we were unable to find the corresponding source segment of the user word correction (***stratified***), due to the low probability of its alignment with its original source word (*classifiés*). As a solution, we artificially added a new source (.) at the end of the sentence and assigned the word correction as its translation.



XML: `<x translation=‟Patients are" >Les patients sont</x> classifiés selon la <x translation=‟presence of" >présence du</x> lymphoedème <x translation=‟stratified" >.</x>`

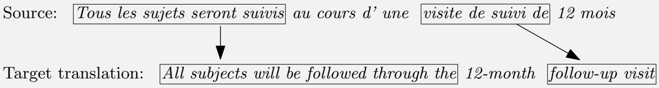
Translation: *Patients are* *classifiés* *stratified* *by the* *presence of* *lymphoedema*

Spurious words

Sometimes, there are source words which do not have a direct correspondence with the user's desired translation. We called them *spurious words*, and they represent a challenge for our system. Since the XML is generated by assigning the validated targets to their correspondent sources, spurious words never get into the XML. Thus, they end up generating undesired translations. Moreover, in the cases in which we fail to align the word correction with its correspondent source segment (see Section 3.3.1), these source segments became spurious words: they never get a translation associated to them and, thus, *Moses* keeps translating them.

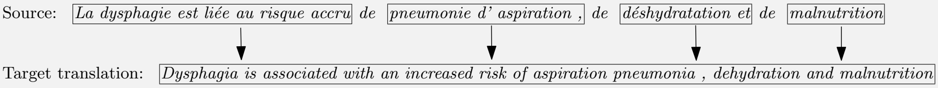
As a result of these untreated sources, users need to either merge more segments or to increase the number of times they input an end-of-translation stroke. Thus, the user effort increases. This problem represents a major challenge within our proposal which we aim to address in future works. Examples 3.8 and 3.9 reflect this problem.

Example 3.8: Example of the increase in the number of mouse actions due to spurious source words. The source words *au cours d' une* do not have a target translation. However, *Moses* translates them as *course of a*. Additionally, at some point of the session, the user's correction *12-month* failed identifying their correspondent source segment. Therefore, *Moses* is generating an undesired translation for them (*12 months*). For this reason, prior to validating the translation, the users have to perform two additional merge operations.


Source: Tous les sujets seront suivis au cours d' une visite de suivi de 12 mois
Target translation: All subjects will be followed through the 12-month follow-up visit

Hypothesis: *All subjects will be followed through the* *course of a* *12-month* *12 months* *follow-up visit*
User feedback: *All subjects will be followed through the 12-month follow-up visit*

Example 3.9: Example of the increase in the number of word strokes due to spurious source words. The source words *de*, *de* and *de* do not have a target translation. However, *Moses* translates them as *of of of*. Prior to validating the translation, the user must type the special end-of-translation stroke (#) to indicate to the system that the validated parts of the hypothesis conform their desired translation.


Source: La dysphagie est liée au risque accru de pneumonie d' aspiration , de déshydratation et de malnutrition
Target translation: Dysphagia is associated with an increased risk of aspiration pneumonia , dehydration and malnutrition

Hypothesis: *Dysphagia is associated with* *an* *increased risk of aspiration pneumonia , dehydration* *and malnutrition* *of of of*
User feedback: *Dysphagia is associated with* *an* *increased risk of aspiration pneumonia , dehydration* *and malnutrition* #

3.4 Segment-based IMT with active prediction

With the aim of improving the word correction step of the IMT process, we propose a variant of our segment-based protocol that include an active prediction module that suggest to the users which word they should correct first. This module is based on CM and assumes that correcting first the word with the least confidence shall lead to the largest improvement in future iterations. Thus, the system suggests the non-validated word from the hypothesis with the least confidence value.

Following prior works that applied CM in IMT (Ueffing and Ney, 2005; González-Rubio et al., 2010), we implemented a word-level CM based on IBM Model 1 (Brown et al., 1993)—similar to the one described by Ueffing and Ney (2005)—and a word-level CM based on HMM. Given that time constraints are crucial in IMT, these implementations result suitable due to their speed. Given a source sentence $x_1^J = x_1, \dots, x_J$ and its translation hypothesis $y_1^I = y_1, \dots, y_I$, the confidence value of a word y_i ($c(y_i)$) is given by:

$$c(y_i) = \max_{1 \leq j \leq J} p(y_i \mid x_j) \tag{3.5}$$

where x_j is a source word at position j; J is the length of the source sentence; and $p(y_i \mid x_j)$ is the lexicon probability given by either the IBM Model 1 or the HMM.

Finally, we implemented a random baseline in which the word to correct is randomly selected from the non-validated segments.

3.5 Experimental framework

In this section, we describe the framework of the experiments conducted in order to assess our proposals. We start by presenting the corpora, continue by describing how we built our systems and end by commenting the automatic evaluation metrics and how we simulated users working on the different interactive scenarios.

3.5.1 Corpora

Following prior IMT works (Tomás and Casacuberta, 2006; Barrachina et al., 2009), we tested our proposal with five different corpora:

EMEA[3] (Tiedemann, 2009b): a collection of medical documents from the *European Medical Agency*.

EU[4] (Barrachina et al., 2009): a collection of documents from the *Bulletin of the European Union*.

TED[5] (Federico et al., 2011): a collection of public speeches from a variety of topics.

Xerox (Barrachina et al., 2009): a collection of *Xerox*'s printer manuals.

Europarl (Koehn, 2005): a collection of proceedings from the European Parliament. We used WMT[6]'s *news-test2013* for development and *news-test2015* as test.

All datasets were kept true-cased, except for the Chinese–English language pair from TED, since Chinese has no case information. All datasets were tokenized using the standard tool provided by the *Moses* (Koehn et al., 2007) toolkit. Chinese sentences were split into words using the Stanford word segmenter (Tseng et al., 2005). Table 3.1 shows the main features of the corpora.

3.5.2 Systems

All systems were trained with *Moses* (Koehn et al., 2007), following the standard procedure: we estimated a 5-gram language model—smoothed with the improved KneserNey method—using *SRILM* (Stolcke, 2002), and optimized the weights of the log-linear model with minimum error rate training algorithm (MERT) (Och, 2003).

[3] `http://www.statmt.org/wmt14/medical-task/`.

[4] `https://doi.org/10.5281/zenodo.5653096`.

[5] `https://wit3.fbk.eu/mt.php?release=2013-01`.

[6] `http://www.statmt.org/wmt15/translation-task.html`.

Table 3.1: Corpora statistics. K denotes thousands and M millions. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. **Fr** denotes French; **En**, English; **De**, German; **Es**, Spanish and **Zh**, Chinese.

		EMEA		**EU**		**TED**	
		Fr–En	**De–En**	**Es–En**	**Fr–En**	**Zh–En**	**Es–En**
Train	$|S|$	1.1M	1.1M	214.5K	982.7K	107.0K	160.2K
	$|T|$	14.3/17.0M	13.3/14.5M	6.0/5.4M	20.7/18.9M	1.9/2.1M	3.0/3.2M
	$|V|$	71.0/80.0K	128.0/71.0K	84.0/70.0K	161.4/150.4K	55.0/41.7K	89.0/61.7K
Val.	$|S|$	500	500	400	400	934	887
	$|T|$	12.0/10.0K	10.0/10.0K	12.0/10.0K	11.5/10.1K	21.5/20.1K	19.1/20.1K
	$|V|$	2.9/2.7K	3.2/2.8K	3.0/2.7K	2.9/2.6K	3.8/3.2K	4.1/3.4K
Test	$|S|$	1000	1000	800	800	1664	1570
	$|T|$	27.0/21.0K	21.0/21.0K	23.0/20.0K	22.5/20.0K	33.2/31.9K	30.7/32.0K
	$|V|$	4.5/4.5K	5.7/4.5K	4.7/4.2K	4.5/4.0K	4.5/3.7K	5.1/3.9K

		Xerox		**Europarl**	
		Es–En	**Fr–En**	**Fr–En**	**De–En**
Train	$|S|$	55.7K	51.8K	2.0M	1.9M
	$|T|$	0.8/0.7M	0.5/0.6M	60.5/54.5M	49.8/52.3M
	$|V|$	16.8/14.0K	24.8/13.7K	160.0/131.2K	394.6/129.1K
Val.	$|S|$	1012	964	3000	3000
	$|T|$	16.0/14.4K	10.7/10.9K	73.7/64.8K	63.4/64.8K
	$|V|$	1.8/1.6K	1.7/1.5K	11.5/9.7K	12.7/9.7K
Test	$|S|$	1125	984	1500	2169
	$|T|$	10.1/8.4K	11.9/12.5K	29.9/27.2K	44.1/46.8K
	$|V|$	2.0/1.9K	2.2/1.8K	6.3/5.6K	10.0/8.1K

3.5.3 Metrics

The quality of our interactive protocol is assessed according to the following metrics:

Word Stroke Ratio (WSR) (Tomás and Casacuberta, 2006): Measures the number of words edited by the user, normalized by the number of words in the final translation. In this work, we assume that the edition of a word has a constant cost (one word stroke), independently of its length. This metric is computed as part of our user simulation (see Section 3.5.4).

Mouse Action Ratio (MAR) (Barrachina et al., 2009): Measures the number of mouse actions made by the user, normalized by the number of characters in the final translation. In the prefix-based protocol, the user makes a mouse action each time they need to edit a word (to position the prompt), plus an additional action per sentence to validate the final translation. The segment-based protocol expands those mouse actions. Now, the user makes two actions each time they validate a segment (clicking at the start and at the end of the segment), and two more each time they delete some words located between segments[7] (same procedure as selecting segments but using the right button of the mouse). In this work, we assumed that the cost of a mouse action is more similar to the cost of typing a character than to the cost of typing a word. Therefore, we normalized the mouse actions with respect to characters. This metric is computed as part of our user simulation (see Section 3.5.4).

Additionally, to evaluate the quality of the initial translations and the difficulty of each task, we used the following well-known metrics:

BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002): Computes the geometric average of the modified n-gram precision, multiplied by a brevity factor that penalizes short sentences. In order to ensure consistent BLEU scores, we used *sacreBLEU* (Post, 2018) for computing this metric.

Translation Error Rate (TER) (Snover et al., 2006): Computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation. We computed this metric using the official *tercom* software[8].

Additionally, we applied approximate randomization testing (ART) (Riezler and Maxwell, 2005)[9]—with $10,000$ repetitions and using a p-value of 0.05—to determine whether two systems presented statistically significance.

3.5.4 User simulation

Due to the time and economic costs of conducting frequent human evaluations during system deployment, we carried out an automatic evaluation with simulated users. These users had as goal to generate the translations from the reference. Different user simulations are implemented accordingly to the protocol to evaluate.

The main drawback with this evaluation is that we are not considering the cognitive efforts of the interactive protocols. For prefix-based IMT, the user reviews the sentence to select the prefix. However, changes in the suffix from one iteration to the next one have also a cognitive impact in the user—especially when

[7] One mouse action is enough for selecting or deleting a one-word segment: the user would simply click on the word.

[8] `https://www.cs.umd.edu/~snover/tercom/`.

[9] We used the following software for doing the computations: `https://github.com/midobal/mt-scripts/tree/master/art`.

correct segments disappear on consecutive hypothesis. This impact is removed in the segment-based protocol. However, the cognitive effort of selecting the correct segments to validate enters into play.

Therefore, in this work we are ignoring the cognitive efforts for both protocol. Moreover, we are simulating the best case scenarios in which users are able to select all correct segments from a translation hypothesis. In a future work, we should study the actual cognitive effort of each protocol as well as the performance of the segment-based IMT when validating only a subset of all the correct segments.

Prefix-based simulation

At each iteration, the user compares, from left to right, the system's hypothesis with the reference. Once it detects a different word, the user corrects it, validating a new prefix in the process. The cost of this correction is one mouse action and one word stroke. Then, reacting to the user feedback, the system generates a new suffix that completes the prefix to conform a new translation hypothesis. This process is repeated until the hypothesis and the reference are the same.

Segment-based simulation

For this simulation, we are assuming that validated word segments must be in the same order as in the reference (i.e., the desired translation). Thus, segments that need to be reordered are not validated. Furthermore, validated segments must maintain the same order in successive iterations. We are aware that more complex user models could contemplate the possibility of reordering validated segments. However, we left this as a future line of work. Additionally, for the sake of simplicity and without loss of generality, we assume that the user always corrects the leftmost wrong word.

At each iteration, the user compares the system's hypothesis with the reference, computing the longest common subsequence (Apostolico and Guerra, 1987) between them. With this, we obtain the common word segments. Then, the user validate them and increase the number of mouse actions—one action for each one-word segment, two actions for each multi-word segment. After that, the user checks, from left to right, if any pair of consecutive validated segments should be merged into a single segment (i.e., they appear one after the other in the reference but are separated by some words in the hypothesis). If they can then, for each pair of validated segments to merge, the user deletes the words between them—increasing mouse actions in one when deleting one word, and in two when deleting more than one word. Finally, to account for the word correction, the user compares the system's hypothesis against the reference, correcting the left-most different word (which increases in one the mouse actions and word strokes). Finally, the system generates the XML instance and obtains a new translation hypothesis with the help of *Moses*. This process is repeated until the hypothesis matches the reference. Example 3.10 exemplifies this simulation.

Example 3.10: Follow up to Example 3.1 to exemplify how user actions are simulated. In the **segment validation**, we compute the longest common subsequence between hypothesis and reference, obtaining the segments: *If you have been exposed , you should*, *go* and *your doctor for tests*). After that, in the **word deletion**, since the first two validated segments appear together in the reference, we delete the word between them (*consult*) to create a bigger validated segment (*If you have been exposed , you should go*). Finally, in the **word correction**, we look for the leftmost reference word not included in a validated segment (to) and add it to the target in its correspondent position. Validating or deleting words have a cost of one mouse action for one-word segments, and two mouse actions for multiple-word segments. A word correction has a cost of one mouse action and one word stroke.

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: 2 + 1 + 2 = 5

Words deletion: $\boxed{\text{If you have been exposed , you should } \cancel{\text{consult}} \text{ go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: 1

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
Mouse actions: 1
Word strokes: 1

Total mouse actions: 7
Total word strokes: 1

Segment-based with active prediction simulation

The user simulation for segment-based active prediction is identical to the simulation of the vanilla segment-based protocol (see Section 3.5.4). The only difference is that, instead of always correcting first the leftmost wrong word, we correct the word indicated by the active prediction module. This is simulated by computing the confidence of each non-validated target word which is either next to a segment (the word right before or after the segment) or it is the first or last word from the hypothesis. This limitation is necessary for the simulation in order to know the user's correction. The active prediction module selects the word with the least confidence and indicates the user that it should be the next word to correct.

3.6 Evaluation

We now present the results of the experimental session. We first present the results obtained by the main approaches. Then, we present the results of the active prediction system.

3.6.1 Main approaches

Table 3.2 compares the user effort results of the segment-based against the prefix-based approach. Prefix-based results were obtained following Barrachina et al. (2009) and are similar to those reported in the literature (Tomás and Casacuberta, 2006; Barrachina et al., 2009)—taking into account that we are generating the word graphs using *Moses* version 3. The quality of the initial translation is shown as indicative of the difficulty of each task.

The segment-based approach improves significantly, in comparison to the prefix-based, the effort required for typing corrections (yielding diminishes of up to 47 points of word stroke rate (WSR)). However, this reduction comes with an increase in the number of mouse actions (from 5 up to 25 points of mouse action rate (MAR)), which is always smaller than the effort reduction.

In the case of *EMEA*, the segment-based approach obtains a reduction of 17 to 40 points of WSR, at the expenses of increasing the MAR in 10 points. Since the initial translation quality of the French–English tasks is higher than the German–English tasks, this last pair of languages obtains the highest effort reduction.

Something similar happens with *EU*. In this case, the initial translation quality is higher for all language pairs. Therefore, the effort reduction is smaller. Nonethe-

Table 3.2: Results of the segment-based IMT approach in comparison with the prefix-based approach. All values are reported as percentages. Differences between each approach are statistically significant in all cases.

				Prefix-based		Segment-based	
Corpus	**Language**	**BLEU** [↑]	**TER** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
EMEA	Fr–En	30.5	48.6	57.8	12.4	33.6	21.6
	En–Fr	29.8	52.6	58.4	12.5	41.7	21.7
	De–En	23.4	57.6	70.9	14.1	31.0	24.4
	En–De	15.7	64.8	74.9	12.0	35.6	23.1
EU	Es–En	47.3	40.8	45.6	10.2	30.5	16.0
	En–Es	47.9	41.1	44.6	9.7	31.9	14.8
	Fr–En	52.1	36.2	37.3	7.5	26.3	14.4
	En–Fr	51.3	38.6	38.8	7.3	29.4	12.8
TED	Zh–En	11.7	76.2	83.1	22.4	36.1	35.8
	En–Zh	8.7	83.3	86.3	55.7	60.0	80.0
	Es–En	36.5	42.7	51.1	12.9	31.7	22.9
	En–Es	31.3	47.7	53.2	12.3	36.7	22.8
Xerox	Es–En	52.2	31.8	35.8	10.5	20.0	20.4
	En–Es	60.8	27.3	28.3	7.9	21.9	14.3
	De–En	32.2	54.6	62.7	15.1	29.2	26.9
	En–De	24.1	64.5	68.3	12.6	32.7	23.6
Europarl	Fr–En	26.5	51.4	58.7	13.9	30.2	30.3
	En–Fr	26.5	55.6	61.4	13.5	31.5	28.4
	De–En	19.2	61.1	73.3	17.7	34.4	30.8
	En–De	15.3	68.4	75.0	15.0	33.1	25.9

less, the segment-based approach obtains a typing reduction of 10 to 15 points of WSR, with an increase in the number of mouse actions of 5 to 6 points of MAR.

TED achieves the highest effort reduction, since it contains the language pair with the lowest initial quality translation (Chinese–English, with 8.7/11.7 points of bilingual evaluation understudy (BLEU) (Papineni et al., 2002) and 83.3/76.2 points of translation error rate (TER) (Snover et al., 2006)). In this case, the effort reduction consists in an improvement of 26 to 47 points of WSR, at the expenses of an increase of 13 to 25 points of MAR. It is worth mentioning the high value of the mouse effort when translating to Chinese, which is most likely due to this language containing very few characters per word. The Spanish–English tasks, containing a higher initial translation quality, are more similar to the previous task. They obtain a 20 points reduction of the typing effort, with an increase of 10 points of the mouse effort.

Table 3.3: Results of the segment-based approach with an active prediction system that suggests the order in which words are corrected. In the regular segment-based protocol, the word corrected is the leftmost wrong word. *IBM*$_1$ implements world-level CM based on IBM model 1. *HMM* implement world-level CM based on HMM. *Random* is a baseline in which the word to correct is selected randomly. All values are reported as percentages. Differences between each method are not statistically significant between one another.

				Segment-based with active prediction					
		Segment-based		**IBM$_1$**		**HMM**		**Random**	
Corpus	**Language**	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
EMEA	Fr–En	33.6	21.6	35.1	23.4	35.5	22.9	35.7	22.8
	En–Fr	41.7	21.7	41.2	23.3	41.8	22.5	41.9	22.0
	De–En	31.0	24.4	30.3	24.3	30.7	24.6	30.0	24.1
	En–De	35.6	23.1	35.0	22.6	35.2	22.6	34.7	22.6
EU	Es–En	30.5	16.0	30.7	17.6	31.2	17.2	31.0	17.0
	En–Es	31.9	14.8	31.2	16.7	31.6	16.0	31.7	15.8
	Fr–En	26.3	14.4	26.9	15.7	27.2	15.5	27.2	15.4
	En–Fr	29.4	12.8	29.4	13.8	29.6	13.7	29.6	13.5
TED	Zh–En	36.1	35.8	35.8	35.4	35.9	35.4	34.9	35.0
	En–Zh	60.0	80.0	60.3	85.5	60.9	83.3	60.9	81.8
	Es–En	31.7	22.9	32.0	24.7	32.3	24.4	32.2	24.2
	En–Es	36.7	22.8	36.6	24.7	37.1	24.0	37.1	23.7
Xerox	Es–En	20.0	20.4	20.1	20.4	20.1	20.5	19.9	20.1
	En–Es	21.9	14.3	22.3	15.2	22.6	14.9	22.6	14.7
	De–En	29.2	26.9	29.3	26.7	29.2	26.6	29.0	26.5
	En–De	32.7	23.6	32.1	22.6	32.3	22.5	32.0	22.7
Europarl	Fr–En	30.2	30.3	29.8	29.7	29.8	29.7	29.4	29.6
	En–Fr	31.5	28.4	30.9	27.7	31.1	27.6	30.4	27.5
	De–En	34.4	30.8	34.3	30.7	34.5	30.7	33.6	30.2
	En–De	33.1	25.9	32.6	25.4	32.6	25.4	32.1	25.3

Xerox has similar results to the previous corpora. The Spanish–English tasks contain a higher initial translation quality, and so the effort reduction is lower (7 to 15 points of WSR at the expenses of an increase of 6 to 10 points of MAR). The German–English tasks, having a lower translation quality, have 33 to 36 points of reduction of the typing effort, and an increase of 11 points of the mouse effort.

Finally, both language pairs from *Europarl* behave similarly. They obtain a typing effort reduction of 28 to 35 points of WSR, with an increase in the number of mouse actions of 11 to 16 points of MAR.

3.6.2 Segment-based IMT with active prediction

The experiments in which we added an active prediction module that suggests which word should be corrected first (see Section 3.4) have been unsuccessful. Correcting first the non-validated word with the lowest confidence value has failed at improving the translation quality of the next hypothesis, resulting in the same amount of user effort (both in terms of word corrections and mouse actions).

Table 3.3 shows the results comparing the regular segment-based approach (which always corrects the leftmost wrong word first) with the world-level CM approaches based on IBM model 1 and HMM, and the random baseline. All strategies obtained similar results, which leads to the conclusion that the order in which corrections are made does not affect the overall user effort.

Example 3.11: Prefix-based IMT session for translating a French sentence into English. The process starts (at *IT-0*) with the system suggesting an initial translation. Then, at iteration 1, the user corrects the leftmost wrong word (**go**). With this action, the user is inherently validating the prefix If you have been exposed , you should . The system takes this user feedback into account and suggests a new hypothesis. Similarly, at iteration 2, the user corrects the leftmost wrong word (**to**). The session ends when the user accepts the last translation suggested by the system.

source $(\mathbf{x})$: Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation $(\tilde{\mathbf{y}})$: If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	If you have been exposed , you should **go** your doctor for tests
	MT	If you have been exposed , you should go consult your doctor for tests
IT-2	**User**	If you have been exposed , you should go **to** your doctor for tests
	MT	If you have been exposed , you should go to consult your doctor for tests
IT-3	**User**	If you have been exposed , you should go to **your** your doctor for tests
	MT	If you have been exposed , you should go to your doctor for tests
END	**User**	If you have been exposed , you should go to your doctor for tests

3.7 Qualitative analysis

Example 3.11 showcases an IMT session for translating a sentence from French to English following the prefix-based protocol. A total of three iterations (in which the user corrects the leftmost wrong word) are needed to achieve the user's desired translation. In comparison, using the prefix-based approach (Example 3.12), this translation can be achieved in only two iterations in which the user validates two segments, deletes a word for creating a bigger segment and makes two words corrections.

To better understand the experimental results, we display some examples which reflect the system's weaknesses.

Example 3.13 showcases an example in which the apparition of some spurious words—source words which do not have a direct correspondence with the words from the desired translation (see Section 3.3.1)—produce a cumbersome behavior. The session starts with the system proposing an initial translation. The user, then, validates some word segments and makes a correction. This correction (***Early-onset***), however, is an out-of-vocabulary. For this reason, the system is unable to associate it with its correspondent source segment (*apparition précoce* and, thus, it keeps offering a translation for them in following iterations. Therefore, at the next iteration, the user has to do a merge operation (uniting the first validated segment with the start of the sentence) to delete those undesired translated words. Furthermore, this problem persists during the following iterations and the user has to keep merging more validated segments to cope with the problem. Additionally, the correction made at iteration two (***occurring***) produces also an error, which increases the problem further. This, together with the spurious words contained in the source sentence (*L'*, *de*, *la* and *septicémie*), results in the user having to make ten extra mouse actions (two per each pair of segments merged) to cope with the problem.

Having translations of spurious words and words for which the user has already typed a translation is a fairly common problem. However, while it is present in more than half of the cases, it typically consists in a few words at some point of the session and does not have a cumbersome effect.

Example 3.12: Segment-based IMT session for translating a French sentence into English. The process starts (at *IT-0*) with the system suggesting an initial translation. Then, at iteration 1, the user validates the correct word segments ($\boxed{\text{if you have been exposed , you should}}$, and $\boxed{\text{your doctor for tests}}$) and types a word correction (**go**). With this information, the system suggests a new hypothesis. At iteration 2, the user deletes a word (*consult*) to create a larger segment ($\boxed{\text{if you have been exposed , you should go}}$) and types a new word correction (**to**). The session ends when the user accepts the last translation suggested by the system.

source ($\mathbf{x}$): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation ($\hat{\mathbf{y}}$): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
IT-2	**User**	$\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{to}}$ $\boxed{\text{your doctor for tests}}$
END	**User**	If you have been exposed , you should go to your doctor for tests

Finally, Example 3.14 depicts a case in which the system has an undesired behavior. The combination of an out-of-vocabulary word (*gens*), a significant word reordering (the first and second halves of the source sentence are reordered in the target sentence) and 4 spurious words (*un*; *,*; *la* and *l'*) makes the system unable to generate good translations. In fact, the initial hypothesis only contains two correct word segments of one word each. Therefore, the user had to type more word corrections and merge more validated segments. In this case, however, the increase in merge operations are mostly caused by the system failing in reordering the translation. The untranslated part of the first half of the source sentence keeps getting translated after the first validated segments. Thus, the user had to merge segments to delete those undesired translated words.

Manually post-editing the initial hypothesis would have taken 10 word strokes plus 11 mouse actions. The segment-based protocol has resulted in 8 word strokes plus 33 mouse actions. Nonetheless, this is an infrequent example of the system's behavior.

Example 3.13: Example of a segment-based IMT session in which the apparition of spurious words results in a cumbersome behavior. Words in *italic* represent undesired translations produced by the system.

source (x): L' apparition précoce de la septicémie néonatale est définie comme une septicémie qui se produit dans les 7 premiers jours de vie

target translation ($\hat{\mathbf{y}}$): Early-onset neonatal sepsis is defined as occurring within the first 7 days of life

IT-0	**MT**	The onset early neonatal sepsis is defined as sepsis which occurs within 7 days of life
IT-1	**User**	**Early-onset** onset early $\boxed{\text{neonatal sepsis is defined as}}$ sepsis which occurs $\boxed{\text{within}}$ $\boxed{\text{7 days of life}}$
	MT	*The onset of the early* $\boxed{\text{Early-onset}}$ $\boxed{\text{neonatal sepsis is defined as}}$ sepsis which occurs $\boxed{\text{within}}$ $\boxed{\text{7 days of life}}$
IT-2	**User**	$\boxed{\text{Early-onset}}$ $\boxed{\text{neonatal sepsis is defined as}}$ **occurring** which occurs $\boxed{\text{within}}$ $\boxed{\text{7 days of life}}$
	MT	$\boxed{\text{Early-onset}}$ *early development of* $\boxed{\text{neonatal sepsis is defined as}}$ *sepsis which occurs* $\boxed{\text{occurring}}$ $\boxed{\text{within}}$ $\boxed{\text{7 days of life}}$
IT-3	**User**	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within}}$ **the** $\boxed{\text{7 days of life}}$
	MT	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within}}$ *early development* $\boxed{\text{the}}$ *sepsis which product* $\boxed{\text{7 days of life}}$
IT-4	**User**	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within the}}$ **first** *which product* $\boxed{\text{7 days of life}}$
	MT	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within the}}$ *early onset sepsis which product* **first** $\boxed{\text{7 days of life}}$
IT-5	**User**	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within the first}}$ $\boxed{\text{7 days of life}}$
	MT	$\boxed{\text{Early-onset neonatal sepsis is defined as occurring}}$ $\boxed{\text{within the first}}$ $\boxed{\text{7 days of life}}$
END	**User**	Early-onset neonatal sepsis is defined as occurring within the first 7 days of life

Example 3.14: Example of a segment-based IMT session in which the system has an undesired behavior. Words in *italic* represent undesired translations produced by the system.

source (x): À un certain moment de leur vie , la plupart des gens vont souffrir de l' acné
target translation (ŷ): Most people will suffer from acne at some point in their life

IT-0	**MT**	To a certain time of life , most gens will experience acne
IT-1	**User**	**Most** a certain time of life , most gens will experience acne
	MT	To a certain time of life , the Most of gens will experience acne
IT-2	**User**	Most **people** gens will experience acne
	MT	Most To a certain time of life , *the of* people will experience acne
IT-3	**User**	Most people will **suffer** acne
	MT	Most people will To a certain time of life , *the of* suffer acne
IT-4	**User**	Most people will suffer **from** acne
	MT	Most people will suffer To a certain time from their life *the* acne
IT-5	**User**	Most people will suffer from acne **at**
	MT	Most people will suffer from acne *To a* certain at their life *the*
IT-6	**User**	Most people will suffer from acne at **some** their life *the*
	MT	Most people will suffer from acne at *To a* some their life *, the*
IT-7	**User**	Most people will suffer from acne at some **point** their life *, the*
	MT	Most people will suffer from acne at some *To a* point *the European* their life
IT-8	**User**	Most people will suffer from acne at some point **in** *European* their life
	MT	Most people will suffer from acne at some point *To a* in *,* their life
IT-9	**User**	Most people will suffer from acne at some point in their life
	MT	Most people will suffer from acne at some point in their life
END	**User**	Most people will suffer from acne at some point in their life

3.8 Interactive neural machine translation

Peris et al. (2017) developed a prefix and a segment-based protocols for INMT. While I helped in the design of the segment-based protocol, these contributions are part of Álvaro Peris' Ph.D. thesis (Peris, 2019). Therefore, they will not be studied in this dissertation. Nonetheless, they will be used in Chapter 6 for applying the interactive framework to the processing of historical documents (see Section 1.2).

3.8.1 Prefix-based INMT

This protocol is the neural equivalent of the prefix-based IMT protocol (see Section 3.2). It was formalized by Peris et al. (2017) as follows:

$$p(\hat{y}_{i'} \mid \hat{y}_1^{i'-1}, x_1^J, f = \tilde{y}_1^i; \Theta) = \begin{cases} \delta(\hat{y}_{i'}, \tilde{y}_{i'}), & \text{if } i' \leq i \\ \bar{\mathbf{y}}_{i'}^\top \mathbf{p}_{i'} & \text{otherwise} \end{cases} \tag{3.6}$$

where x_1^J is the source sentence; $\tilde{y}_1^i$ is the validated prefix together with the corrected word; Θ are the models parameters; $\bar{\mathbf{y}}_{i'}^\top$ is the one hot codification of the word i'; $\mathbf{p}_{i'}$ contains the probability distribution produced by the model at time-step i; and $\delta(\cdot, \cdot)$ is the Kronecker delta:

$$\delta(\hat{y}_{i'}, \tilde{y}_{i'}) = \begin{cases} 1, & \text{if } \hat{y}_{i'} \equiv \tilde{y}_{i'} \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

This is equivalent to a forced decoding strategy (as in Eq. (2.18)) and can be seen as generating the most probable suffix given a validated prefix, which fits into the statistical framework deployed by Barrachina et al. (2009).

3.8.2 Segment-based INMT

This protocol is the neural equivalent of the segment-based IMT protocol (see Section 3.3). It was formalized by Peris et al. (2017) as follows:

$$p(y_{i_n+i'} \mid y_1^{i_n+i'-1}, x_1^J, f_1^N; \Theta) = \mathbf{y}_{i_n+i'}^\top \mathbf{p}_{i_n+i'}, \quad 1 \leq i' \leq \hat{l}_n \tag{3.8}$$

where $f_1^N = f_1, \ldots, f_N$ is the feedback signal and $f_1, \ldots, f_N$ are a sequence of non-overlapping segments validated by the user; each alternative hypothesis y (partially) has the form $y = \ldots, f_n, h_n, f_{n+i}, \ldots$; g_n is the non-validated segment; and l_n is the size of this non-validated segment and is computed as follows:

$$\hat{l}_n = \underset{0 \leq l_n \leq L}{\arg\max} \frac{1}{l_N + 1} \sum_{i'=i_n+1}^{i_n+l_n+1} \log p(y_{i'} \mid y_1^{i'-1}, x_1^J; \Theta) \tag{3.9}$$

3.8.3 Comparison of INMT against IMT

Table 3.4 showcases some results comparing INMT versus IMT. The reported INMT results are from Peris (2019)—we selected the experiments which used the exact same datasets than us. Their systems were trained using *NMT-Keras* (Peris and Casacuberta, 2018). long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) systems featured a single-layered bidirectional LSTM encoder, using concatenation as fusion operator. The decoder function was another single-layered recurrent neural network (RNN) (Jordan, 1990; Elman, 1990; Hochreiter and Schmidhuber, 1997) with conditional LSTM units, with an additive attention mechanism. All layer dimensions were set to 512 and layer normalization—with decay $\lambda = 10^{-4}$ and dropout $p = 0.1$—was applied to all non-recurrent connections. Regarding their Transformer systems, word embeddings and d_m were set to 512. Each multi-head attention layer had 8 heads, with parallel projections of size 64. Hidden and output dimensions of the feed-forward layers were 2048 and 512, respectively. A dropout $p = 0.1$ was set for all layers. Embeddings were scaled by a factor of $\sqrt{d_m}$. Finally, the encoder and decoder stacked 6 or 4 layers, depending on the size of the training data.

Additionally, since we are only reporting their results, and we do not have access to their systems, we did not conduct any study of the statistical significance between systems.

With an exemption in which differences are most likely not statistically significant, INMT with an RNN architecture yielded the best results for the prefix-based approach. INMT with a Transformer architecture yielded better results than IMT—except for a case in which yielded worse results according to WSR and the same results according to MAR—but worse results than the RNN architecture.

The best results for the segment-based approach were yielded by IMT, with a few exceptions which may or may not be statistically different. This improvements in terms of WSR, however, come with an increase in terms of MAR, which was to be expected since our segment-based protocol contemplates more user actions than its neural counterpart (INMT does not contemplate the ability to merge segments). Nonetheless, we need to remember that we are assuming that mouse actions have a smaller effort than word corrections. Therefore, SMT models seem to deal better with the segment generation than NMT.

Finally, we evaluated the initial translation quality of all systems, observing that the difference between them were not significant enough to affect the comparison. Most likely, this is influenced by the size of the corpora being fairly small (see Table 3.1).

Table 3.4: Comparison of INMT versus IMT. $INMT_{\text{RNN}}$ stands for INMT using an RNN architecture and $INMT_{\text{Trans.}}$ stands for INMT using a Transformer architecture. INMT results are from Peris (2019).

		Prefix-based						**Segment-based**					
		INMT$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**		**INMT**$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**	
		WSR [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
TED	Zh–En	54.9	14.2	60.1	14.3	83.1	22.4	51.2	21.2	49.2	20.4	36.1	35.8
	En–Zh	68.1	28.9	66.7	29.6	86.3	55.7	58.4	64.2	56.6	62.5	60.0	80.0
Xerox	Es–En	30.7	7.2	37.4	8.3	35.8	10.5	29.1	12.5	35.5	13.2	20.0	20.4
	En–Es	28.4	7.3	32.1	8.0	28.3	7.9	22.7	7.5	30.2	12.7	21.9	14.3
	De–En	38.4	9.4	42.2	10.0	62.7	15.1	35.1	13.3	39.9	14.1	29.2	26.9
	En–De	55.1	10.8	56.5	11.2	68.3	12.6	50.9	14.9	54.7	16.0	32.7	23.6

		Translation quality					
		INMT$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**	
		BLEU [↑]	**TER** [↓]	**BLEU** [↑]	**TER** [↓]	**BLEU** [↑]	**TER** [↓]
TED	Zh–En	13.7	75.7	11.5	76.7	11.7	76.2
	En–Zh	9.3	76.7	8.2	77.6	8.7	83.3
Xerox	Es–En	59.0	28.6	53.9	32.1	52.2	31.8
	En–Es	63.5	27.5	60.5	28.3	60.8	27.3
	De–En	36.2	51.1	31.3	54.9	32.2	54.6
	En–De	25.4	63.0	23.2	64.3	24.1	64.5

3.9 Conclusions

In this chapter, we have developed a new IMT protocol that allows the user to validate the correct parts of a translation hypothesis, breaking the left-to-right constrains present in most protocols. We implemented this protocol using a feature of the *Moses* toolkit, and tested it and compared it with the prefix-based protocol in a simulated environment. Results show that the segment-based approach succeeds in overcoming the prefix-based limitation of only correcting the prefix, resulting in a reduction of the user effort. This effort improvement results in a substantial decrease of the typing effort, at the expenses of an increase in the number of mouse actions.

This increase is mostly due to the system's main weaknesses: failing to find the corresponding sources of the user word corrections and source words which do not have a direct correspondence with the words from the desired translation. In both cases, those sources generate undesired translations, resulting in the user having to merge more segments to cope with this problem.

The segment-based methodology successfully takes advantage of the correct parts of a translation hypothesis. This is reflected in the results of the tasks which had the lowest initial translation quality. With one exception, these tasks have been the ones to have the greatest improvement of the user effort. This exception has been the English–Chinese task which, unable to take advantage from them, has needed a greater number of user corrections.

We have also tested an active prediction protocol to assist the user in the correction step of the process. In this protocol, the system informed the user about which word should be corrected first to improve the quality of the next hypothesis. We implemented this protocol using different approaches that relied on the use of CM. In all cases, results did not present statistical differences between each approach. Thus, we concluded that changing the order in which words are corrected had no effect in the overall user effort. Most likely, since the XML scheme takes profit from the word correction only to generate those phrases located near that word, the only effect that altering the order in which the user makes corrections has is to change which parts of the sentence are corrected first.

As future work, we need to improve the way in which the system finds the corresponding source words of a user correction, and how the system deals with source words without a direct correspondence with the goal translation. Additionally, we want to develop new protocols to assist the user in the segment validation step of the process. Furthermore, our user model only validated segments which were ordered in the same way as in the desired translation. In future works, we want to explore additional approaches such as allowing the user to reorder segments. Finally, we assumed that making a mouse action is less of an effort than typing a word and, thus, that the increase in the mouse effort pays off with respect to the significant reduction of the typing effort. However, we should test our proposal with real users to obtain actual measures of the effort reduction.

3.10 Publications

Some of our contributions to the IMT field were accepted for publication at international conferences and journals:

- Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. "Interactive neural machine translation". *Computer Speech & Language*, 45:201–220, 2017. JCR Q2.

  Contributions: *I helped in the design of the segment-based INMT protocol.*

- Miguel Domingo and Álvaro Peris and Francisco Casacuberta. "Segment-based interactive-predictive machine translation". *Machine Translation Journal*, 31:163–185, 2017.

Contributions: *extension and in-depth study of the segment-based protocol and use of CM.*

- Miguel Domingo, Álvaro Peris, and Francisco Casacuberta. “Interactive-predictive translation based on multiple word-segments”. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 282–291, 2016. CORE B. **Best paper award**.

  Contributions: *segment-based protocol.*

Chapter 4
Language Modernization

¿De qué me sirve tener
Aptitud para mi oficio,
Si no tengo el ejercicio
Que la hace desenvolver?

(***La Mano Derecha y la Izquierda***. Miguel Agustín Príncipe.)

What is the use of having
Aptitude for my trade,
If I don't have the exercise
What makes it unwrap?

(***The Right Hand and the Left***. Google Translate.)

Contents

Language modernization aims at generating newer versions of historical documents, written in the modern version of their original language. In this chapter, we present our contributions to this field (see Section 1.2.1). We start by summarizing the state of the art in language modernization. Then, we describe the different approaches we took to tackle this problem. After that, we present the experimental framework followed to assess our proposals and the results of the evaluation conducted. Later, we showcase the results of the user study we conducted to assess whether modernization successfully decreases the difficulty of comprehending historical documents. Finally, we qualitatively analyze the overall results and draw some conclusions.

4.1 State of the art

Language modernization has been manually applied to literature for centuries. One of the most well-known examples is *The Bible*, which has been adapted and translated for generations in order to preserve and transmit its contents (Given, 2015). Classic literature is also frequently modernized in order to bring it closer to a contemporary audience (e.g., *No Fear Shakespeare*[1]; *Odres Nuevos*[2]; *El Quijote* (Trapiello, 2015)). However, on the literature we find that, while normalizing orthography to account for the lack of a spelling convention has been extensively research for years (see Section 1.2.2), automatic modernization of historical documents is a young research field.

One of the first related works was a shared task for translating historical text to contemporary language (Tjong Kim Sang et al., 2017). The task was focused on normalizing the document's spelling. However, they also approached language modernization using a set of rules. After that, to the best of our knowledge, the rest of the works are the ones presented in this thesis. The exception are Sen et al. (2019), whose proposal consisted in—using a neural machine translation (NMT) approach—augmenting the training data by extracting pairs of phrases and adding them as new training sentences. Additionally, Peng et al. (2021) proposed a method for generating a modernized summary of a historical document.

[1] https://www.sparknotes.com/shakespeare/.

[2] https://www.castalia.es/libros?tipo=coleccion&letra=O&nombre=49&other_page=1.

4.2 Approaches

In this section, we present our different approaches to the language modernization problem. All approaches tackle modernization as a machine translation (MT) task. They consider the original language of a historical document as the source language, and its modern version as the target language.

4.2.1 Statistical machine translation

This approach is based on statistical machine translation (SMT) (see Section 2.1). Given a parallel training corpora in which, for each sentence of a given historical document, its corresponding modernized version its available, a phrased-based SMT system is trained. The resulting system will be the modernization system that shall be used for modernizing the language of new documents.

4.2.2 Neural machine translation

This approached is based on NMT (see Section 2.2). Like with the SMT approach (see Section 4.2.1), the language modernization system is obtained by training an NMT system from a parallel set of training data. This approach has two different variants depending on the neural architecture used for training the NMT systems: recurrent neural network (RNN) (Jordan, 1990; Elman, 1990; Hochreiter and Schmidhuber, 1997) with long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) (which, for simplicity's sake we will refer from now on simply as the *LSTM architecture*) or Transformer.

4.2.3 NMT enriched with modern documents

This approach is an extension of the NMT approach (see Section 4.2.2). One of the frequent problems in historical natural language processing (NLP) researches is the scarce availability of suitable training data (Bollmann and Søgaard, 2016). This is specially troublesome to NMT systems, which need large quantities of training data. For this reason, in this approach we aim at enriching the neural systems by taking profit from modern documents.

Our approach is based on the standard method for creating parallel synthetic corpus (see Section 2.2.3), which are regularly used when building state-of-the-art NMT systems—especially in resource-poor scenarios (Poncelas et al., 2018). This method takes a monolingual corpus in the target language and an MT system which has been trained to translate from the target language to the source language. Then, the synthetic data is generated by translating the monolingual corpus with the MT system. The resulting data becomes the source of the synthetic corpus, and the monolingual data becomes the target.

Using this method, we propose to collect modern documents from the same language as the original document to create the monolingual corpus. Additionally, since the closer to the document's domain the data is the more we can profit from it, we use feature decay algorithm (FDA) (Biçici and Yuret, 2015) to filter the modern documents and generate a smaller and more relevant subset. Finally, using this monolingual data we:

1. Train an inverse SMT modernization system—since SMT is less affected by the problem of scarce availability of training data—using the modernized version of the training dataset as source, and the original version as target.

2. Use this system to translate the modern monolingual data, obtaining a new version of the documents which, hopefully, is able to capture the same linguistic characteristics that the original documents have. This new version, together with the original modern document, conform the synthetic parallel data.

3. Train an NMT modernization system using the synthetic corpus.

4. Fine-tune the system by training a few more steps using the original training data.

4.3 Experimental framework

In this section, we describe the framework of the experiments conducted in order to assess our proposals. We start by presenting the corpora, continue by describing how we built our systems and end by commenting the automatic evaluation metrics.

4.3.1 Corpora

We now present the corpora used throughout our experimental sessions. Table 4.1 contains the corpora statistics.

Table 4.1: Corpora statistics. |*S*| stands for number of sentences, |*T*| for number of tokens and |*V*| for size of the vocabulary. *Modern documents* refer to the monolingual data used to create the synthetic data. M denotes millions and K thousands.

		Dutch Bible		**El Quijote**		**OE-ME**	
		Original	**Modernized**	**Original**	**Modernized**	**Original**	**Modernized**
Train	|*S*|	35.2K		10K		2716	
	|*T*|	870.4K	862.4K	283.3K	283.2K	64.3K	69.6K
	|*V*|	53.8K	42.8K	31.7K	31.3K	13.3K	8.6K
Validation	|*S*|	2000		2000		500	
	|*T*|	56.4K	54.8K	53.2K	53.2K	12.2K	13.3K
	|*V*|	9.1K	7.8K	10.7K	10.6K	4.2K	3.2K
Test	|*S*|	5000		2000		500	
	|*T*|	145.8K	140.8K	41.8K	42.0K	11.9K	12.9K
	|*V*|	10.5K	9.0K	8.9K	9.0K	4.1K	3.2K
Modern documents	|*S*|	3.0M		2.0M		6.0M	
	|*T*|	76.1M	74.1M	22.3M	22.2M	67.5M	71.6M
	|*V*|	1.7M	1.7M	210.1K	211.7K	290.2K	287.4K

Dutch Bible

This corpus consists in a collection of different versions of the Dutch Bible (i.e., Dutch translation of the Bible written in different centuries). Among others, it contains a version from 1637—which we consider as the original version—and another from 1888—which we consider as the modern version (using 19th century Dutch as if it were modern Dutch). This corpus was generated as part of a shared task[3] on automatic linguist annotation (Tjong Kim Sang et al., 2017).

As *modern documents* for enriching the neural systems (see Section 4.2.3), we collected all the 19th century Dutch books available at the *Digitale Bibliotheek voor de Nederlandse letteren*[4] and generated a monolingual corpus, which will be used for generating synthetic training data (see Section 2.2.3).

El Quijote

This corpus consists of the well-known 17th century Spanish novel by Miguel de Cervantes and a recent modern translation. We built this corpus using a version (F. Jehle, 2001) of the original 17th century Spanish novel by Miguel de Cervantes, and a 21st century version modernized by Andrés Trapiello[5] (Trapiello, 2015).

[3] `https://ifarm.nl/clin2017st/`.

[4] `http://dbnl.nl/`

[5] We were granted permission to use their work in our research. However, we are not allowed to make the resulting dataset publicly available.

The first step was to split each document into sentences. Since the 17th century version was faithful to the original manuscript (in which each document line is formed by a very few words), we replaced line breaks by spaces to create a single sentence, and removed empty lines. For consistency, we did the same to the 21st century version. After that, we split each document into sentences by adding line breaks to relevant punctuation (i.e., dots, quotation marks, admiration marks, etc). Then, to ensure consistency, we checked special symbols (e.g., quotation marks) and made sure that the same character was used in both versions. Finally, in order to create a parallel corpus, we aligned both documents using *Hunalign* (Varga et al., 2005) and conducted a manual revision to correct the alignment errors.

As *modern documents* for enriching the neural systems (see Section 4.2.3), we collected the Spanish monolingual data from *OpenSubtitles* (Lison and Tiedemann, 2016)—a collection of movie subtitles in different languages.

OE-ME

This corpus consists in the original 11th century English text *The Homilies of the Anglo-Saxon Church* and a 19th century version—which we consider as *modern English.* It was generated and kindly given to us by Sen et al. (2019).

Similarly to *El Quijote*, we collected the English monolingual data from *OpenSubtitles* (Lison and Tiedemann, 2016)—a collection of movie subtitles in different languages— for the *modern documents* used for enriching the neural systems (see Section 4.2.3).

4.3.2 Systems

SMT systems were trained with *Moses* (Koehn et al., 2007), following the standard procedure: we estimated a 5-gram language model—smoothed with the improved KneserNey method—using *SRILM* (Stolcke, 2002), and optimized the weights of the log-linear model with minimum error rate training algorithm (MERT) (Och, 2003). SMT systems were used both for the SMT modernization approach (see Section 4.2.1) and for generating synthetic data[6] (see Section 4.2.3).

We built NMT systems using *OpenNMT-py* (Klein et al., 2017). For the LSTM architecture (Hochreiter and Schmidhuber, 1997), we used long short-term memory units (Gers et al., 2000), with all model dimensions set to 512. We trained the

[6] We applied FDA using the official software: `https://github.com/bicici/FDA`.

system using Adam (Kingma and Ba, 2014) with a fixed learning rate of 0.0002 and a batch size of 60. We applied label smoothing of 0.1 (Szegedy et al., 2015). At inference time, we used beam search with a beam size of 6. In order to reduce vocabulary, we applied joint byte pair encoding (BPE) (Gage, 1994) to all corpora, using 10, 000 merge operations. NMT systems were trained using synthetic data and, then, they were fine-tuned with the training data.

For the Transformer architecture (Vaswani et al., 2017), we used 6 layers; Transformer, with all dimensions set to 512 except for the hidden Transformer feed-forward (which was set to 2048); 8 heads of Transformer self-attention; 2 batches of words in a sequence to run the generator on in parallel; a dropout of 0.1; Adam (Kingma and Ba, 2014), using an Adam beta2 of 0.998, a learning rate of 2 and Noam learning rate decay with 8000 warm up steps; label smoothing of 0.1 (Szegedy et al., 2015); beam search with a beam size of 6; and joint BPE applied to all corpora, using 10, 000 merge operations[7].

4.3.3 Metrics

Since we are approaching modernization from an MT perspective, we saw fitting to adopt some of the most well-known evaluation metrics from MT. This decision was later supported both by the scholar's evaluation (see Section 4.4.2)and the user study (see Section 4.4.3). In both cases, results correlated with the ones from the automatic metrics. Thus, in order to assess our proposal, we made use of:

Translation Error Rate (TER) (Snover et al., 2006): This metric computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation. We computed this metric using the official *tercom* software[8].

BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002): This metric computes the geometric average of the modified n-gram precision, multiplied by a brevity factor. In order to ensure consistent bilingual evaluation understudy (BLEU) (Papineni et al., 2002) scores, we used `sacreBLEU` (Post, 2018) for computing this metrics.

Additionally, we made use of approximate randomization testing (ART) (Riezler and Maxwell, 2005)[9]—with 10, 000 repetitions and using a p-value of 0.05—to determine whether two systems presented statistically significance.

[7] More information regarding the number of merge operations can be found at Appendix A.

[8] `https://www.cs.umd.edu/~snover/tercom/`.

[9] We used the following software for doing the computations: `https://github.com/midobal/mt-scripts/tree/master/art`.

4.4 Evaluation

We now present the evaluation conducted in order to assess our proposals. Using the automatic metrics previously described, we conducted an initial evaluation that included all approaches and their variants. After that, we present two human evaluations for the main approaches: one involving scholars and another involving non-experts users.

4.4.1 Automatic metrics

Table 4.2 presents the results of the automatic evaluation. All approaches significantly improved the modernization quality. The enriched NMT with an LSTM architecture approach yielded the best results in all cases, with the SMT approach performing quite similarly. In fact, results from both approaches were not significantly different except for the *Bible* experiment, in which case the neural approach performed slightly better (0.5 points with respect to translation error rate (TER) (Snover et al., 2006) and around 4 points with respect to BLEU).

Table 4.2: Experimental results. *Baseline* system corresponds to evaluating the quality of the original document with respect to the modernized version. All results are significantly different between all approaches except those denoted with†. [↓] indicates that the lowest the value the highest the quality. [↑] indicates that the highest the value the highest the quality. Best results are denoted in **bold**.

Approach	**Dutch Bible**		**El Quijote**		**OE-ME**	
	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]
Baseline	57.9	12.9	44.2	36.3	91.0	2.8
SMT	11.5	77.5	**30.7**†	**58.3**†	**39.6**†	**39.6**†
NMT_{LSTM}	13.8	79.6	55.1	39.8	82.7	12.8
$\text{NMT}_{\text{Transformer}}$	**11.1**†	**81.7**†	38.4	49.3	54.7	27.3
Enriched NMT_{LSTM}	**11.1**†	**80.6**†	**31.9**†	**57.3**†	**44.3**†	**35.9**†
Enriched $\text{NMT}_{\text{Transformer}}$	18.2	70.6	36.7	51.0	47.2	31.0

With respect to the other approaches, it is worth mentioning how the enriched approaches significantly improve the modernization quality of their counterparts. Except for the Transformer-based approach for the *Bible* task, which performs significantly worse. Considering how well its counterpart performed—together with the enriched LSTM approach, they yielded the best results for this task—this is most likely due to the synthetic data making the system learn a broader domain.

4.4.2 Scholars

4 Scholars specialized in classic Spanish literature helped us perform this evaluation. For this reason, we chose *El Quijote* corpus (see Section 4.3.1). The evaluation was conducted over 100 sentences, which we chose randomly—making sure that the modernized versions were different to the original sentences. We showed each sentence together with its modernization—50 sentences modernized with the SMT approach and another 50 with the NMT approach—and asked the scholars to give a rating according to the quality of the following aspects:

- **Fluency**: how fluid does the modernized sentence sound?
- **Lexical meaning**: how correct is the lexicon of the modernized sentence?
- **Syntax**: how correct is the syntactic construction of the modernized sentence?
- **Semantic**: is the meaning of the original sentence preserved in the modernized sentence?
  - ***1***: the meaning is lost.
  - ***2***: a great part of the meaning is lost.
  - ***3***: half the meaning is lost.
  - ***4***: part of the meaning is lost.
  - ***5***: the meaning remains.
- **Modernization**: how appropriate is the modernization?

Example 4.1 showcases an example of a question. To avoid any bias, we shuffled the sentences and did not give any detail to the evaluators about how modernizations had been produced. Note that the evaluation was conducted using only our two best approaches: SMT and enriched NMT with an LSTM architecture. Table 4.3 shows the results of the evaluation.

Example 4.1: Example of a question from the human evaluation.

In a scale from 1 to 5 (being 1 the lowest score and 5 the highest) indicate the modernization quality according to the following features:

Original sentence: Admiraronse los hombres assi de al figura como de las razones de don Quixote, sin entender la mitad de lo que en ellas decir queria.

Modernized sentence: Se admiraron los hombres tanto de la figura como de las palabras de don Quijote, sin entender la mitad de lo que en ellas decir quería.

	1	2	3	4	5
Fluency	○	○	○	○	○
Lexical meaning	○	○	○	○	○
Syntax	○	○	○	○	○
Semantic	○	○	○	○	○
Modernization	○	○	○	○	○

Table 4.3: Results of the scholars' evaluation. Values correspond to the average score for all sentences of each approach. 1 Is the lowest score and 5 is the highest.

Scholar	**SMT approach**				
	Fluency	**Lexical meaning**	**Syntax**	**Semantic**	**Modernization**
Scholar$_1$	5.0	4.3	4.3	4.6	3.9
Scholar$_2$	2.1	1.9	2.0	2.1	2.0
Scholar$_3$	3.2	3.1	2.9	2.9	3.1
Scholar$_4$	4.5	3.9	4.6	4.3	4.0
Average	3.7	3.3	3.4	3.5	3.2
	Enriched NMT$_{\text{LSTM}}$ approach				
	Fluency	**Lexical meaning**	**Syntax**	**Semantic**	**Modernization**
Scholar$_1$	4.8	4.0	4.0	4.1	4.0
Scholar$_2$	2.0	1.9	1.9	1.9	1.9
Scholar$_3$	3.3	3.2	2.9	3.0	3.1
Scholar$_4$	3.8	3.5	3.7	3.7	3.5
Average	3.7	3.3	3.4	3.5	3.2

While the automatic evaluation (see Section 4.4.1) did not show any significant differences between the SMT and NMT approaches, the human evaluators slightly preferred SMT over NMT. Scores vary considerably depending on the evaluator—scholar$_1$ and scholar$_4$ gave higher scores than scholar$_2$ and scholar$_3$. However, all evaluators agreed that fluency is the strongest point of both approaches. In general, scores are above the average, which seems to correlate with the automatic evaluation.

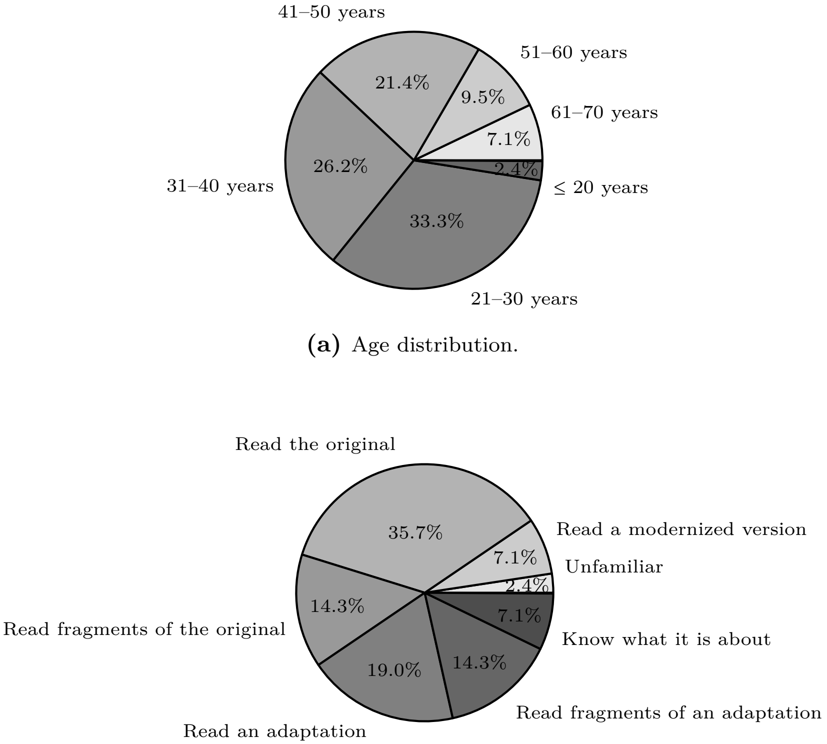
When we asked evaluators about their opinion, they commented that the main problems were related with punctuation and diacritical marks. They also mentioned that, sometimes, part of the sentence was lost in the modernization—a known issue related with NMT (Wu et al., 2016). Additionally, scholar$_1$ commented that, overall, the quality of the modernization was acceptable. However, scholar$_2$ commented that if they had to correct the mistakes, they would have preferred to do the modernization from scratch. In Chapter 6, we will introduce an interactive machine translation (IMT) methodology that will allow scholars to perform modernization in a more productive way.

4.4.3 Non-experts

With the help of 42 participants, we conducted a user study in order to assess whether language modernization is able to make historical document accessible to a broader audience by decreasing the difficulty of comprehending them. Since the participants were Spanish, we conducted the study using *El Quijote* (see Section 4.3.1). Considering that *El Quijote* is well-known in Spain, we asked participants about their familiarity with it. Fig. 4.1 shows some information about the user's age and their familiarity with this literary piece.

Most participants were between 20 and 50 years old, but there were also older and younger people. With one exception, all participants were familiar with *El Quijote* to some extent. In fact, 35.7% of them had read the original version of the novel.

The study consisted in several questions in which we showed two sentences to the user—the original sentence and its modernized version (generated using one of the two systems under study)—and asked them to select which sentence was easier for them to read and comprehend, if both of them had the same difficulty, or if they thought that both sentence did not have the same meaning. The selected sentences were the same used in the scholars' evaluation (see Section 4.4.2). Like in the scholar's evaluation, the study was conducted using only our two best approaches: SMT and enriched NMT with an LSTM architecture (for the sake of simplicity we shall refer to it simply as the *NMT approach* for the rest of this section). In order to avoid any bias, the order in which sentences appeared (i.e., the original sentence and its modernized version) was randomized, as well as the use of the different approaches. Example 4.2 shows an example of a question. More information regarding the questionnaire can be found at Appendix B.



(a) Age distribution.

(b) Familiarity with *El Quijote*.

Figure 4.1: Information about study participants.

Table 4.4 presents the results of the study. Despite the users' familiarity with *El Quijote*, modernization succeeded in making the document (in this case, the individual sentences) easier to comprehend. No matter the modernization approach, users selected the modernized version in the majority of the cases. In most of the remaining cases, users did not find any significant difference with respect to the original sentence.

When comparing both approaches, we observe that the SMT approach yielded better results: Users selected 61.4% of their modernized versions, while they only selected a 50.9% of the sentences modernized by the NMT approach. Additionally, the SMT approach only introduced errors in 7.8% of the cases—the NMT introduced them in 20.3% of the cases—and its modernized versions were harder to comprehend only in 3.2% of the cases—versus a 6.4% of the cases for the NMT approach. Therefore, despite neither the automatic nor the scholar's evaluation yielded significant differences between both approaches, the user study showed that the SMT approach produced versions easier to read and comprehend more successfully than the NMT approach.

Example 4.2: Example of a question.

Select the sentence which is easier for you to read and comprehend:

○ Riose don Quixote, y pidio que quitassen otro lienço, debaxo del qual se descubrio la imagen del patron de las Españas a cauallo, la espada ensangrentada, atropellando moros y pisando cabeças, y, en viendola, dixo don Quixote:

○ Se rió don Quijote, y pidió que quitasen otro lienzo, debajo del cual se descubrió la imagen del patrón de las Españas a caballo, la espada ensangrentada, atropellando moros y pisando cabezas y viéndola, dijo don Quijote:

○ Indifferent.

○ Both sentences do not have the same meaning.

Table 4.4: Results of the user study. Values correspond to the percentage of cases—for each modernization system—in which the users selected that option. *Original* means that users understood better the original version. *Modernized* means that users understood better the modernized version. *Indifferent* means that users did not find any significant differences between the original and modernized versions. *Not equal* means that users felt that the meaning between both version differed.

	Original	**Modernized**	**Indifferent**	**Not equal**
SMT	3.2	61.4	27.6	7.8
NMT	6.4	50.9	22.3	20.3

4.5 Qualitative analysis

In this section, we analyze some examples that showcase the strengths and weaknesses of our approaches, as indicated by the automatic evaluation and by the participant's perception during the user study.

4.5.1 Initial evaluation

Example 4.3 showcases a comparison of how our best approaches behave for the *Dutch Bible* corpus. The SMT approach is able to achieve an error-free modernization while both NMT approaches only make one mistake each: the word *Allerhoogste* which the LSTM-based approach replaces by *sterren* and the Transformer approach replaces by *sterke*. Note that both suggested words and the original word (*Stercke*) share the prefix *ster* (with capital *s* in the case of the original world). Most likely, this difference in modernizations is caused by the subword algorithm: maintaining the same lower-cased version of the original prefix and suggesting different suffix for each system.

Example 4.3: An example comparing how our best approaches behave for the *Dutch Bible* corpus. Mistakes in the modernization are denoted in red.

Original: Want de Allerhooghste gedenckt uwer, ende de Stercke en heeft uwer niet vergeten inde versoeckinge.
Modernized (reference): Want de Allerhoogste gedenkt uwer, en de Almachtige heeft uwer niet vergeten in de verzoeking.

SMT: Want de Allerhoogste gedenkt uwer, en de Almachtige heeft uwer niet vergeten in de verzoeking.
Enriched NMT$_{\text{LSTM}}$: Want de Allerhoogste gedenkt uwer, en de sterren heeft uwer niet vergeten in de verzoeking.
Enriched NMT$_{\text{Transformer}}$: Want de Allerhoogste gedenkt uwer, en de sterke heeft uwer niet vergeten in de verzoeking.

Example 4.4 contains another comparison for the *Dutch Bible* corpus. In this case, the SMT approach has made a small mistake: while *'t gene* has been correctly modernized as *hetgeen*, the symbol ´ remained in the hypothesis. The LSTM-based approach made a casing mistake with the word *rechter* and another mistake with the word *indertijd*, which divided into two different words: *der tijd*. The Transformer-based approach made a similar mistake with the word *indertijd*, which became *in de tijd*. Once more, these similarities are due to an error caused by the subwords algorithm.

Example 4.4: Another example comparing how our best approaches behave for the *Dutch Bible* corpus. Mistakes in the modernization are denoted in red.

Original: Ende also hy een rechtveerdich Rechter is, so heeft hy van u lieden inder tijdt genomen, 't gene hy gegeven hadde.
Modernized (reference): En alzo hij een rechtvaardig rechter is, zo heeft hij van ulieden indertijd genomen, hetgeen hij gegeven had.

SMT: En alzo hij een rechtvaardig rechter is, zo heeft hij van ulieden indertijd genomen, ' hetgeen hij gegeven had.
Enriched NMT$_{\text{LSTM}}$: En alzo hij een rechtvaardig Rechter is, zo heeft hij van ulieden der tijd genomen, hetgeen hij gegeven had.
Enriched NMT$_{\text{Transformer}}$: En alzo hij een rechtvaardig rechter is, zo heeft hij van ulieden in de tijd genomen, hetgeen hij gegeven had.

Example 4.5 showcases a comparison of how our best approaches behave for the *OE-ME* corpus. In this case, both the SMT and the LSTM-based approaches behave similarly: they make the same two mistakes modernizing two words, and the words *should have* are missing. The difference is that the SMT approach has left the original word *abræce* unmodernized while the LSTM approach has removed this word from its modernization. Additionally, the LSTM approach has two more words missing: *A* and *it*. The Transformer-based approach is only missing these two last words. However, most of its modernization is wrong.

Finally, Example 4.6 contains another comparison for the *OE-ME* corpus. The SMT approach makes a small reordering mistake—which does not affect to the readability/comprehension which is the main goal of language modernization. It also fails to modernize the word *Ælmihtiges*, which leaves in its original form, and makes two word mistakes (*will* and *dwell*). The LSTM approach also makes these two words mistakes. However, it correctly modernizes the rest of the sentence, except for a missing comma after the word *Michael*—which, again, does not affect to the readability/comprehension. Finally, the Transformer approach behaves similarly to the LSTM, but making more mistakes (one of which is related to the casing of a word).

Example 4.5: An example comparing how our best approaches behave for the *OE-ME* corpus. Mistakes in the modernization are denoted in red. ␣ indicates that a word is missing in the modernization.

Original: Mare wundor wæs, þæt hé of deaðe arás, þonne he cucu of ðære rode abræce.
Modernized (reference): A greater miracle it was, that he arose from death, than that he living should have broken from the cross.

SMT: A greater miracle it was, that he arose from death, when ␣ he quick ␣ ␣ abræce from the cross.
Enriched NMT$_{\text{LSTM}}$: ␣ Greater miracle ␣ was, that he arose from death, when ␣ he quick ␣ ␣ ␣ from the cross.
Enriched NMT$_{\text{Transformer}}$: ␣ Greater miracle ␣ was, that noble of death arose when he had been quick raised from the cross of darkness.

Example 4.6: Another example comparing how our best approaches behave for the *OE-ME* corpus. Mistakes in the modernization are denoted in red. ␣ indicates that a word is missing in the modernization.

Original: Ic eom Michahel se heah-engel Godes Ælmihtiges, and ic symle on his gesihðe wunige.
Modernized (reference): I am Michael, the archangel of God Almighty, and I continue ever in his sight.

SMT: I am the archangel Michael Ælmihtiges of God, and I will ever dwell in his sight.
Enriched NMT$_{\text{LSTM}}$: I am Michael the archangel of God Almighty, and I will ever dwell in his sight.
Enriched NMT$_{\text{Transformer}}$: I am Michael the archangel of God the Almighty, and I will always dwell in His sight.

4.5.2 User perception

Now, we proceed to show some behavioral examples of our modernization approaches and how they were perceived by the participants of the user study. Remember that these modernizations are from *El Quijote* and were generated using the SMT and enriched NMT with LSTM approaches (see Section 4.4.3).

Example 4.7: Example of a successfully modernized sentence.

Original version: Si sois seruidos, respondio don Quixote, holgaria de verlas, pues imagines que con tanto recato se lleuan, sin duda deuen de ser buenas.
Modernized version: Si sois a bien —respondió don Quijote, alegraría de verlas, pues imágenes que con tanto recato se llevan, sin duda deben de ser buenas.

Example 4.7 showcases an example in which, as indicated per the users, the modernization helped in comprehending the meaning of the sentence. It is worth noting, however, that the modernization contains a small grammatical mistake (it should tell ***me** alegraría de verlas*) and a small orthographic mistake (an em dash is missing after *don Quijote*). Nonetheless, it is worth remembering that the goal of modernization is limited to decreasing the comprehension difficulty. On Chapter 6 we shall present techniques for helping scholars create error-free modernizations.

Example 4.8: Example of modernization in which the modernized version is similar to the original version.

Original version: Huuolo de conceder don Quixote, y assi lo hizo.
Modernized version: Huéolo de conceder don Quijote, y así lo hizo.

Example 4.8 shows an example in which both versions are pretty similar. Only three words have been modified during the modernization—and one of them (*huéolo*) is a mistake introduced by the use of BPE (see Section 2.2.2). Despite this, there are people who found the modernized version easier to read; a great majority that found no difference between them; and a few people that either preferred the original version or considered that they did not have the same meaning.

Example 4.9: Example of modernization in which users preferred the original version over the modernized one.

Original version: Ofreciosele el gallardo pastor, pidiole que se viniesse con el a sus tiendas;
Modernized version: Se le rosó el gallardo pastor, pile dio que se viniese con él a sus tiendas;

Example 4.9 contains an example in which the original sentence is easier to understand than its modernized version. Despite that the users considered that both versions have the same meaning, the modernized one is harder to comprehend since the first half of the sentence does not make much sense. In fact, if we have a look at the human evaluation[10], scholars considered that the modernized version was more or less fluent, but it had poor lexical meaning, syntax and semantic.

Example 4.10: Example of modernization in which part of the content is missing.

Original version: Simplicissimo eres, Sancho, respondio don Quixote,
Modernized version: Simplicissimo eres, Sancho,

Example 4.10 shows a clear example of an unsuccessful modernization: the modernized version is missing part of the content and, on top of that, the one that is present is a copy of the original content and has not been modernized.

Example 4.11: Example of a modernization consisting in modernizing the orthography.

Original version: Este cauallero fue vno de los mejores andantes que tuuo la milicia diuina;
Modernized version: Este caballero fue uno de los mejores andantes que tuvo la milicia divina;

Example 4.11 presents a case in which the modernization consisted only in modernizing the orthography. Despite this, all but a few participants indicated that the modernized version was easier to read and to comprehend than the original version.

Example 4.12: Example of a modernization that alters the meaning of the original sentence.

Original version: Este si que es cauallero y de las esquadras de Christo;
Modernized version: Este sí que es caballero y de las escuadras de aquí;

Example 4.12 showcases an example in which a mistake in one word (*aquí* instead of *Cristo*) alters the meaning of the original sentence. Most participants indicated

[10] Remember that we used the same data for both evaluations.

that the sentences did not have the same meaning. However, it is worth noting that a few of them indicated that the modernized version was easier to read and to comprehend. Most likely, they did not realize the mistake in the last word and focused their attention in the rest of the sentence.

Example 4.13: Example of a modernization in which users do not have a clear preference.

Original version: ¡Santiago, y cierra España! ¿Está por ventura España abierta, y de modo, que es menester cerrarla, o qué ceremonia es esta?
Modernized version: ¡Santiago, y cierra España! ¿—Está por ventura España abierta, y de modo, que es menester cerrarla, o qué ceremonia es esta?

Finally, Example 4.13 shows an example in which the modernization did not seem to help much. 57% Of the users selected the option *indifferent*, 31% preferred the original version, 10% selected the modernized version and one person selected that both version did not have the same meaning.

4.6 Conclusions

In this chapter, we have presented our contributions to the language modernization task. In order to tackle the language barrier inherent in historical documents and make them easier to understand and accessible to a broader audience, we proposed several modernization approaches based on MT.

We evaluated our proposal both automatically and with the help of 4 scholars specialized in Spanish classical literature. Both evaluations showed that our approaches succeeded in making historical documents easier to comprehend by a more general audience.

Finally, we conducted a user study to corroborate these results. The study was conducted using the main SMT and NMT approaches. 42 Volunteers, of different age and background, participated in this study. Results showed that modernization successfully decreased the comprehension difficulty. In most of the cases, users chose the modernized version as the easiest to read and comprehend. However, there is still room for improvement. Sometimes, the modernization introduced errors that made users feel that the meaning had been changed. Other times, users did not find any significant difference between the original version and its modernization. When comparing the SMT and NMT approaches, the

NMT approach made a bigger number of errors and the user chose its modernized versions as the best option fewer times than with the SMT approach.

Despite that results showed that modernization had successfully improved the understanding of historical documents, we have to take into consideration that language-related losses may appear during the process (e.g., Example 1.1 shows an example in which part of the language structures and rhymes disappear). Nonetheless, the goal of modernization is limited to bringing understanding of historical documents to a general audience.

As a future work, we would like to tackle the main problems pointed out during the scholar evaluation and the user study. Mainly, punctuation, diacritical marks, the introduction of non-existent words and loosing parts of the given sentence. We would also like to conduct a new evaluation involving more scholars and more languages and datasets, and a new user study for different languages and datasets.

4.7 Publications

Some of our contributions to the language modernization task were accepted for publication at international conferences and journals:

- Miguel Domingo and Francisco Casacuberta. "Modernizing historical documents: A user study". *Pattern Recognition Letters*, 133:151–157, 2020. JCR Q2.

  Contributions: *human evaluation and user study.*

- Miguel Domingo and Francisco Casacuberta. "A machine translation approach for modernizing historical documents using back translation". In *Proceedings of the International Workshop on Spoken Language Translation*, pages 39–47, 2018.

  Contributions: *NMT and enriched NMT approaches.*

- Miguel Domingo, Mara Chinea-Rios, and Francisco Casacuberta. Historical documents modernization. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 295–306, 2017. CORE B.

  Contributions: *SMT approach.*

Chapter 5
Spelling Normalization

Here we go! go! hashiri-tsudzukeru
dare ni mo tomerare wa shinai
mirai no jibun e to
Give a reason for life todoketai

(***Give a reason***. Hayashibara Megumi.)

Here we go! go! Keep running
No one can stop you
To myself in the future
Give a reason for life

(***Give a reason***. Google Translate.)

Contents

Spelling normalization aims to account for the lack of orthography conventions in historical documents by adapting the document's spelling to modern standards. In this chapter, we present our contributions to this field (see Section 1.2.2). We start by summarizing the state of the art in spelling normalization. Then, we describe the different approaches we took to tackle this problem. After that, we present the experimental framework followed to assess our proposals, and the results of the evaluation conducted. Finally, we qualitatively analyze the overall results and draw some conclusions.

5.1 State of the art

Some approaches to spelling normalization include creating an interactive tool that includes spell checking techniques to assist the user in detecting spelling variations (Baron and Rayson, 2008). A combination of a weighted finite-state transducer, combined with a modern lexicon, a phonological transcriber and a set of rules (Porta et al., 2013). A combination of a list of historical words, a list of modern words and character-based statistical machine translation (SMT) (Scherrer and Erjavec, 2013). A multitask learning approach using a deep bi-long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) applied at a character level (Bollmann and Søgaard, 2016). The application of a token/segment-level character-based SMT approach to normalize historical and user-created words (Ljubešic et al., 2016). The use of rule-based machine translation (RBMT), character-based machine translation (CBMT) and character-based neural machine translation (CBNMT) (Korchagina, 2017). Domingo and Casacuberta (2018) evaluated word-based and character-based MT approaches, finding character-based to be more suitable for this task and that SMT systems outperformed neural machine translation (NMT) systems. Tang et al. (2018), however, compared many neural architectures and reported that the NMT models are much better than SMT models in terms of character error rate (CER). Finally, Hämäläinen et al. (2018) evaluated SMT, NMT, an edit-distance approach, and a rule-based finite state transducer, and advocated for a combination of these approaches to make use of their individual strengths.

Additionally, researchers working on handwritten text recognition (HTR) generate what they call *modern transcripts.* These transcripts are newer versions of the documents in which not only the spelling has been updated to match modern standards but also abbreviated words have been expanded, and some writing mistakes have been fixed. Therefore, part of it is similar to spelling normalization but taking into account that it is done at the same time as the document is being transcribed—taking as input the manuscript's image. On the contrary, spelling normalization is applied to a document which has already been transcribed.

5.2 Approaches

In this section, we present our different approaches to the spelling normalization problem.

5.2.1 Statistical dictionary

This approach is based on a statistical dictionary and is intended to be an additional baseline. We computed *IBM's model 1* (Och and Ney, 2003) to obtain word alignments from source and target of the training set. Then, for each source word, we selected as its translation the target word which had the highest alignment probability with that source word. Finally, at translation time, we translated each source word with the translation that appeared in the dictionary. If a given word did not appear in the dictionary, then we left it untranslated.

5.2.2 Initial approaches

As our initial approach to spelling normalization, we decided to use two simple approaches based on machine translation (MT).

Statistical machine translation

This approach is based on phrased-based SMT (see Section 2.1). Considering the document's language as the source language and its normalized version of that language as the target language, we propose to use SMT to adapt the document's spelling to modern standards.

Neural machine translation

This approach is based on NMT (see Section 2.2). Like the previous approach, considering the document's language as the source language and its normalized version of that language as the target language, we propose to use NMT to adapt the document's spelling to modern standards. There are two different variants for this approach depending on the architecture used of the neural systems: recurrent neural network (RNN) (Jordan, 1990; Elman, 1990; Hochreiter and Schmidhuber,

1997) with LSTM (which, for simplicity's sake we will refer from now on simply as the *LSTM architecture*) or Transformer.

5.2.3 Character-based MT

CBMT takes the vocabulary problem to the limit (see Section 2.2.2) by working at a character level—dropping significantly the chances of getting an out-of-vocabulary word. While it was already being research for SMT (Tiedemann, 2009a; Nakov and Tiedemann, 2012), its interest has increased with the rise of NMT. Some approaches to CBNMT consist in using hierarchical NMT (Ling et al., 2015), a character level decoder (Chung et al., 2016), a character level encoder (Costa-Jussà and Fonollosa, 2016) or—for alphabets in which words are composed by fewer characters—by constructing an NMT system that takes advantage of that alphabet (Costa-Jussà et al., 2017).

Since in spelling normalization changes frequently occur at a character level, it seemed fitting to use a character-based strategy. From all the different CBMT techniques from the literature, we decided to use the simplest approach: splitting words into characters and consider each character as a token. The reason for selecting this approach was because spelling normalization is a much simpler problem than MT and most CBMT techniques seemed far too complex (including the need of bigger training datasets) for this task. Nonetheless, we tried additional techniques for the CBNMT approaches.

CBSMT

This approach is based on character-based statistical machine translation (CBSMT). Considering the document's language as the source language and its normalized version as the target language, this approach follows a CBSMT strategy: the document's words are split into characters and, then, conventional SMT is applied.

CBNMT

This approach is based on CBNMT. More precisely, we make use of the following CBNMT strategies:

- **CBNMT**: This technique uses the simplest character level strategy. Words from both the source and the target are split into characters.

- **SubChar**: This technique combines a sub-word level (see Section 2.2.2) and a character level strategies. Source words are split into sub-words and target words into characters.

- **CharSub**: This technique combines a character level and a sub-word level strategy. Source words are split into characters and target words into sub-words.

For each CBNMT technique, we propose a different normalization approach. Considering the document's language as the source language and its normalized version as the target language, each approach follows one of the aforementioned CBNMT strategies. Source and target words are split into either characters or sub-words (depending on the technique) and, then, conventional NMT is applied to train the normalization system. Moreover, each approach has two different variants depending on the neural architecture used for training the NMT systems (see Section 2.2): LSTM or Transformer.

5.2.4 CBNMT enriched with modern documents

This approach is an extension of the CBNMT approach (see Section 5.2.3). One of the frequent problems in historical natural language processing (NLP) researches is the scarce availability of suitable training data (Bollmann and Søgaard, 2016). This is specially troublesome for NMT systems, which need large quantities of training data. For this reason, in this approach we aim at enriching the neural systems by taking profit from modern documents.

Our approach is based on the standard method for creating parallel synthetic corpus (see Section 2.2.3), which are regularly used when building state-of-the-art NMT systems—especially in resource-poor scenarios (Poncelas et al., 2018). This method takes a monolingual corpus in the target language and an MT system which has been trained to translate from the target language to the source language. Then, the synthetic data is generated by translating the monolingual corpus with the MT system. The resulting data becomes the source of the synthetic corpus, and the monolingual data becomes the target.

Using this method, we propose to collect modern documents from the same language as the original document to create the monolingual corpus. Then, using this monolingual data we:

1. Train a CBSMT system—since SMT is less affected by the problem of scarce availability of training data— using the normalized version of the training dataset as source, and the original version as target.

2. Use this system to translate the modern documents, obtaining a new version of the documents which, hopefully, is able to capture the same orthography inconsistencies that the original documents have. This new version, together with the original modern document, conform a synthetic parallel data which can be used as additional training data.

3. Combine the synthetic data with the training dataset, replicating several times the training dataset in order to match the size of the synthetic data and avoid overfitting (Chatterjee et al., 2017). We chose this methodology over training with the synthetic data and fine-tuning with the training dataset (like we did in the previous task; see Section 4.2.3) due to this task having a certain similarity with automatic post-editing.

4. Use the resulting dataset to train the enriched CBNMT system.

5.3 Experimental framework

In this section, we describe the framework of the experiments conducted in order to assess our proposals. We start by presenting the corpora, continue by describing how we built our systems and end by commenting the automatic evaluation metrics.

5.3.1 Corpora

In order to assess our proposals, we made use of the following corpora:

Entremeses y Comedias[1] (F. Jehle, 2001): A 17th century Spanish collection of comedies by Miguel de Cervantes. It is composed of 16 plays, 8 of which have a very short length. Each line corresponds to the same line from its original manuscript.

Quijote[2] (F. Jehle, 2001): The 17th century Spanish two-volumes novel by Miguel de Cervantes. Each line corresponds to the same line from its original manuscript.

Bohorič (Ljubešić et al., 2016): a collection of 18th century Slovene texts written in the old Bohorič alphabet.

[1] https://users.pfw.edu/jehle/wcce.htm.

[2] https://users.pfw.edu/jehle/wcdq.htm.

Gaj (Ljubešić et al., 2016): a collection of 19th century Slovene texts written in the Gaj alphabet.

As reflected in Table 5.1, the size of the corpora is small. Thus, the need of generating synthetic training data (see Section 2.2.3). As *modern documents*, we selected half a million sentences from OpenSubtitles (Lison and Tiedemann, 2016), a collection of movie subtitles in different languages. We selected the same Spanish sentences for *Entremeses y Comedias* and *Quijote*, and the same Slovene sentences for *Bohoric* and *Gaj*.

Table 5.1: Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens, $|V|$ for size of the vocabulary and $|W|$ for the number of words whose spelling does not match modern standards. M denotes millions and K thousand.

		Entremeses y Comedias	**Quijote**	**Bohorič**	**Gaj**
Train	$\|S\|$	35.6K	48.0K	3.6K	13.0K
	$\|T\|$	250.0/244.0K	436.0/428.0K	61.2/61.0K	198.2/197.6K
	$\|V\|$	19.0/18.0K	24.4/23.3K	14.3/10.9K	34.5/30.7K
	$\|W\|$	52.4K	97.5K	33.0K	32.7K
Development	$\|S\|$	2.0K	2.0K	447	1.6K
	$\|T\|$	13.7/13.6K	19.0/18.0K	7.1/7.1K	25.7/25.6K
	$\|V\|$	3.0/3.0K	3.2/3.2K	2.9/2.5K	8.2/7.7K
	$\|W\|$	1.9K	4.5K	3.8K	4.5K
Test	$\|S\|$	2.0K	2.0K	448	1.6K
	$\|T\|$	15.0/13.3K	18.0/18.0K	7.3/7.3K	26.3/26.2K
	$\|V\|$	2.7/2.6K	3.2/3.2K	3.0/2.6K	8.4/8.0K
	$\|W\|$	3.3K	3.8K	3.8K	4.8K
Modern documents	$\|S\|$	500.0K	500.0K	500.0K	500.0K
	$\|T\|$	3.5M	3.5M	3.0M	3.0M
	$\|V\|$	67.3K	67.3K	84.7K	84.7K

5.3.2 Systems

SMT systems were trained with *Moses* (Koehn et al., 2007), following the standard procedure: we estimated a 5-gram language model—smoothed with the improved KneserNey method—using *SRILM* (Stolcke, 2002), and optimized the weights of the log-linear model with minimum error rate training algorithm (MERT) (Och, 2003). SMT systems were used both for the SMT modernization approach (see Section 4.2.1) and for generating synthetic data (see Section 4.2.3).

We built NMT systems using *OpenNMT-py* (Klein et al., 2017). For the LSTM architecture (Hochreiter and Schmidhuber, 1997), we used long short-term memory units (Gers et al., 2000), with all model dimensions set to 512. We trained the system using Adam (Kingma and Ba, 2014) with a fixed learning rate of 0.0002 and a batch size of 60. We applied label smoothing of 0.1 (Szegedy et al., 2015). At inference time, we used beam search with a beam size of 6. In order to reduce vocabulary, we applied joint byte pair encoding (BPE) (Gage, 1994) to all corpora, using 10,000 merge operations. NMT systems were trained using synthetic data and, then, were fine-tuned with the training data.

For the Transformer architecture (Vaswani et al., 2017), we used 6 layers; Transformer, with all dimensions set to 512 except for the hidden Transformer feed-forward (which was set to 2048); 8 heads of Transformer self-attention; 2 batches of words in a sequence to run the generator on in parallel; a dropout of 0.1; Adam (Kingma and Ba, 2014), using an Adam beta2 of 0.998, a learning rate of 2 and Noam learning rate decay with 8000 warm up steps; label smoothing of 0.1 (Szegedy et al., 2015); beam search with a beam size of 6; and joint BPE applied to all corpora, using 10,000 merge operations.

Finally, we implemented the statistical dictionary using *mgiza* (Gao and Vogel, 2008) for computing *IBM's model 1* (Och and Ney, 2003).

5.3.3 Metrics

In order to assess our proposal, we made use of the following well-known metrics:

Character Error Rate (CER): number of character edit operations (insertion, substitution and deletion), normalized by the number of characters in the final translation. We computed this metric using a custom script[3].

Translation Error Rate (TER) (Snover et al., 2006): number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation. We computed this metric using the official *tercom* software[4].

BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002): geometric average of the modified n-gram precision, multiplied by a brevity factor. In order to ensure consistent bilingual evaluation understudy (BLEU) (Papineni et al., 2002) scores, we used `sacreBLEU` (Post, 2018) for computing this metrics.

We selected CER for being wildly used in the literature and, since we are approaching spelling normalization from an MT point of view, we selected TER

[3] `https://github.com/midobal/mt-scripts/tree/master/wer`.

[4] `https://www.cs.umd.edu/~snover/tercom/`.

and BLEU (which are two of the most frequent metrics for evaluating MT systems).

Additionally, we applied approximate randomization testing (ART) (Riezler and Maxwell, 2005)—with $10,000$ repetitions and using a p-value of 0.05—to determine whether two systems presented statistically significant differences.

5.4 Evaluation

We decided to conduct the evaluation of our proposals in two different steps: we first evaluate our main approaches and, then, we assess the quality of the different CBNMT variants. Finally, we compared our best approaches against a similar task done in HTR.

5.4.1 Main approaches

Table 5.2 presents the experimental results of our main proposals. In all cases, the CBSMT approach yielded the best results. When measuring in terms of CER, the CBNMT and enriched CBNMT approaches using a Transformer architecture yielded results as good as the ones of the CBSMT approach (there were no statistical difference between them) for *Entremeses y Comedias* and *Quijote*, and between the enriched CBNMT approach with an LSTM architecture and the CBSMT approach for *Quijote*. Similarly, translation error rate (TER) (Snover et al., 2006) and BLEU show no statistical differences between the CBSMT and the CBNMT with Transformer for *Quijote*.

In general, the worst results are yielded by the neural approaches. Except for *Quijote*, these approaches do not improve any of the baselines. Most likely, this is due to how small the training data is (see Table 5.1), which affects significantly the neural models. Furthermore, in the case of *Bohorič* and *Gaj*, any of the neural approaches is able to improve the baselines. Most likely, this is also related with the small size of the training corpora, and to the nature of the Slovene language—specially in the case of *Bohorič*, whose documents were written while the Slovene language was having a big restructuring.

Finally, in all cases, the character-based approaches significantly improved the results yielded by their word/sub-word counterparts according to all metrics. Moreover, with an exception, profiting from modern documents improved their results even more. The exception was with the CBNMT Transformer approach for *Quijote* and *Gaj*. In those two cases, this approach did not present statistical

Table 5.2: Experimental results. Baseline system corresponds to considering the original document as the document to which the spelling has been normalized to match modern standards. SD is the statistical dictionary. *Trans.* stands for Transformer. All results are significantly different between all systems except those denoted with † and ‡ (respectively). Best results are denoted in **bold**.

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
Baseline	8.1	28.0	47.0	7.9	19.5	59.4	21.7	49.0	18.0	3.5	12.3	72.6
SD	7.8	18.9	66.8	3.9	5.5	89.3	16.2	20.7	56.1	7.6	8.8	79.8
SMT	6.7	8.0	82.1	$5.3^\ddagger$	4.5	91.1	9.0	15.1	63.0	2.8	5.2	82.6
NMT$_{\text{LSTM}}$	18.0	15.2	72.2	10.2	8.1	84.4	41.4	33.9	36.7	36.0	28.3	50.4
NMT$_{\text{Trans.}}$	27.5	43.9	34.3	$5.5^\ddagger$	18.5	60.6	43.2	66.4	12.6	12.0	18.4	68.8
CBSMT	$\mathbf{1.3}^\dagger$	**4.4**	**91.7**	$\mathbf{2.5}^\dagger$	$\mathbf{3.0}^\dagger$	$\mathbf{94.4}^\dagger$	**2.4**	**8.7**	**80.4**	**1.4**	**5.1**	**88.3**
CBNMT$_{\text{LSTM}}$	$1.7^\ddagger$	12.0	82.7	2.7	$4.3^\ddagger$	$93.3^\ddagger$	29.4	39.5	48.7	31.5	36.9	53.1
En. CBNMT$_{\text{LSTM}}$	$1.7^\ddagger$	13.3	79.4	$\mathbf{2.2}^\dagger$	$4.0^\ddagger$	$93.2^\ddagger$	28.6	38.3	49.5	30.5	35.4	54.9
CBNMT$_{\text{Trans.}}$	$\mathbf{1.4}^\dagger$	6.1	88.0	$\mathbf{1.9}^\dagger$	$\mathbf{3.3}^\dagger$	$\mathbf{93.9}^\dagger$	$26.2^\dagger$	$30.6^\dagger$	$60.0^\dagger$	$29.9^\dagger$	$32.1^\dagger$	$60.0^\dagger$
En. CBNMT$_{\text{Trans.}}$	$\mathbf{1.1}^\dagger$	5.1	89.7	$\mathbf{2.4}^\dagger$	5.1	89.7	$25.7^\dagger$	$29.8^\dagger$	$60.8^\dagger$	$30.0^\dagger$	$32.0^\dagger$	$60.2^\dagger$

differences. Our best guess is that, considering the particularities of the task, the modern documents were not helpful for those two cases.

5.4.2 Additional CBNMT approaches

Table 5.3 presents the experimental results of using additional CBNMT approaches. Like with the standard CBNMT approaches, while they yielded significant improvements for *Entremeses y Comedias* and *Quijote*—with a few exceptions—almost no approach was able to improve the baselines for *Bohorič* and *Gaj*. The only approach able to significantly improve the baselines was the enriched *CharSub*. The aforementioned exceptions for *Entremeses y Comedias* were the *SubChar* approaches and the *CharSub* approaches with a Transformer architecture. Those approaches were not able to improve any baselines for *Entremeses y Comedias*. Additionally, all the new approaches were successfully able to profit from modern documents to improve results.

Overall, the best results were achieved using the Transformer architecture over enriched strategies. However, depending on the corpus, a different approach yielded the best results. Moreover, in all cases, no approach was able to improve the results yielded by the CBSMT strategy.

Table 5.3: Experimental results. Baseline system corresponds to considering the original document as the document to which the spelling has been normalized to match modern standards. *Trans.* stands for Transformer. All results are significantly different between all systems except those denoted with †, ‡ and * (respectively). Best results are denoted in **bold**.

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
CBNMT$_{\text{LSTM}}$	1.7†	12.0	82.7	2.7‡	4.3†	93.3‡	29.4†	39.5†	48.7	31.5†	36.9	53.1
SubChar$_{\text{LSTM}}$	23.3	32.8	54.1	**2.2**†	**3.7**‡	93.8‡	36.7	47.7	39.4	32.7	37.3†	52.4†
CharSub$_{\text{LSTM}}$	5.8	18.2	75.2	3.7	5.8	89.8	67.9	83.8	5.3	37.2	48.1	36.3
En. CBNMT$_{\text{LSTM}}$	1.7†	13.3	79.4†	**2.2**†	4.0†	93.2‡	28.6‡	38.3	49.5	30.5†	**35.4**‡	**54.9**‡
En. SubChar$_{\text{LSTM}}$	37.8	35.8	59.3	**2.3**†	**3.3**‡	**94.9**†	29.5†	**36.9**	**51.5**	31.5	**35.9**‡	**54.3**‡
En. CharSub$_{\text{LSTM}}$	3.8	15.2	78.9†	**2.3**†	4.1†	93.0‡	**27.5***	39.6†	47.2	**29.4**	37.2†	52.3†
CBNMT$_{\text{Trans.}}$	**1.4**‡	6.1	88.0	**1.9**†	**3.3**‡	93.9‡	26.2	30.6‡	60.0†	29.9	32.1*	60.0*
SubChar$_{\text{Trans.}}$	21.2	33.1	64.8	2.6‡	**3.7**‡	93.5‡	28.6‡	33.4	55.2	30.9†	32.7*	59.2*
CharSub$_{\text{Trans.}}$	12.2	42.4	72.1	3.2	4.8	91.4	59.1	68.8	14.9	9.1	11.6	79.1
En. CBNMT$_{\text{Trans.}}$	**1.1**‡	**5.1**	**89.7**	**2.4**†	5.1	89.7	25.7	29.8‡	60.8†	30.0†	32.0*	60.2*
En. SubChar$_{\text{Trans.}}$	43.2	56.5	66.4	**2.4**†	**3.2**‡	**94.4**†	27.3*	31.8	57.8	30.6†	32.6*	59.1*
En. CharSub$_{\text{Trans.}}$	11.9	41.8	72.5	**2.4**†	**3.5**‡	93.9‡	**8.8**	**11.5**	**79.3**	**6.5**	**7.2**	**87.2**

5.4.3 Generation of modern transcripts

Researchers in the HTR field perform a task similar to spelling normalization. It is known as generating *modern transcripts* (Romero et al., 2019), which consist in transcripts of historical documents in which:

- Archaic words forms are replaced with modern spellings.
- Abbreviated words are expanded.
- Writing mistakes are fixed.
- Punctuation marks are modified or added according to present rules.

Given an image that contains a historical manuscript they train an optical model that recognizes the image's content and generates a modern transcript. Due to their similarity with spelling normalization—which is frequently applied to literal transcripts (also known in the literature as *diplomatic transcripts*) of historical manuscripts—we decided to compare how well our approaches performed for normalizing a diplomatic transcript versus generating a modern transcript.

Romero et al. (2019) kindly proportioned us the dataset used in their work and the transcripts they generated. Therefore, we trained our best approaches with this data and compare the results with the ones they achieved. However, due to the differences between both tasks—the original transcripts contain unique abbreviations which need to be expanded as part of the process—our methodology for generating synthetic data for enriching the neural models is not suitable enough. Therefore, we decided to use only the vanilla CBNMT approaches and leave the enriched versions for a future work.

Table 5.4: Experimental results of using our spelling normalization approaches for generating modern transcripts. Baseline system corresponds to considering the diplomatic transcript's hypotheses as the modern transcript. *HTR* is Romero et al. (2019) approach. Results are the average of the four-blocks cross-validation. All results are significantly different between all systems except those denoted with †. Best results are denoted in **bold**.

Approach	**CER** [↓]	**TER** [↓]	**BLEU** [↑]
Baseline	23.6	45.6	30.7
HTR	**18.2**†	**33.2**†	**51.0**†
CBSMT	**18.5**†	**33.5**†	**50.6**†
CBNMT$_{\text{LSTM}}$	20.5	37.4	45.2
CBNMT$_{\text{Transformer}}$	19.0	34.5	49.0

Table 5.4 presents the experimental results. The CBSMT approach yielded the same results as the ones achieved by Romero et al. (2019), while the CBNMT approaches improved the baseline but performed slightly worse than the other approaches. Therefore, this seems to indicate that performing spelling normalization techniques over a diplomatic transcript achieves similar results to generating a modern transcript. Nonetheless, we have to take into account that we applied our spelling normalization approaches without taking into account the particularities of this task: expanding abbreviated words, correcting mistakes, etc. Additionally, we trained our systems using error-free transcripts (from the ground truth). However, in this case we are normalizing the diplomatic transcript generated by Romero et al. (2019), which contains errors introduced by their system. Since our system is not prepared to deal with those errors, they are hurting its performance. To better asses this statement, we used the ground truth's diplomatic transcript to simulate how our system would have performed normalizing an error-free transcript (see Table 5.5).

Results show how normalizing an error-free diplomatic transcript achieves much better results (up to three times better, according to CER and TER). Thus, it verifies that our performance is being hurt by the transcription errors of the HTR

Table 5.5: Experimental results of using our best spelling normalization approach for generating modern transcripts from an error-free diplomatic transcript.

Approach	CER [↓]	TER [↓]	BLEU [↑]
CBSMT	5.9	10.5	66.5

system. In a future work, we shall address how to deal with these errors as well as the additional particularities of this task.

Finally, while our system has not been able to improve the results by Romero et al. (2019), it has a similar performance. Therefore, while their system has the advantage of generating modern transcripts from a manuscript's image, our system is more suitable in cases in which the document has already been transcribed and only needs to be normalized.

5.5 Qualitative analysis

In this section, we showcase some examples to analyze the behavior of our best approaches.

5.5.1 Main approaches

Example 5.1 showcases an example from *Entremeses y Comedias*. In this case, due to the high quality of all systems, all approaches generated the same normalization, updating two out of the three characters which needed to be normalized.

Example 5.1: Example of normalizing a sentence from *Entremeses y Comedias* with our main approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	¡O mal logrado moço! Salid fuera;
Normalized:	¡Oh mal logrado mozo! Salid fuera;
CBSMT:	¡Oh mal logrado mozo! Salí fuera;
Enriched CBNMT$_{\text{LSTM}}$:	¡Oh mal logrado mozo! Salí fuera;
Enriched CBNMT$_{\text{Trans.}}$:	¡Oh mal logrado mozo! Salí fuera;

Example 5.2 presents an example from *Quijote*. The CBSMT approach generates an error-free normalization while the CBNMT approaches successfully normalize three out of the four characters that needed to be updated.

Example 5.2: Example of normalizing a sentence from *Quijote* with our main approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	“Para esso se yo vn buen remedio”, dixo el del Bosque;
Normalized:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;
CBSMT:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;
En. CBNMT$_{\text{LSTM}}$:	“Para es␣o se yo un buen remedio”, dijo el del Bosque;
En. CBNMT$_{\text{Trans.}}$:	“Para es␣o se yo un buen remedio”, dijo el del Bosque;

Example 5.3 contains an example from *Bohorič*. Once more, the CBSMT approach generated an error-free normalization while the CBNMT approaches made two different mistakes each.

Example 5.3: Example of normalizing a sentence from *Bohorič* with our main approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	vadljajo ali lófajo, de bi svédili, kdo jim je kriv te nefrezhe.
Normalized:	vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč␣e.
CBSMT:	vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč␣e.
	nesreč␣e.
En. CBNMT$_{\text{LSTM}}$:	vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.
	nesreč ␣e.
En. CBNMT$_{\text{Trans.}}$:	vadeljajo ali losajo, da bi ␣zvedeli, kdo jim je kriv te nesreč ␣e.

Finally, Example 5.4 showcases an example from *Gaj*, which was the hardest of the tasks. In this case, all approaches failed to normalize the same three characters.

Example 5.4: Example of normalizing a sentence from *Gaj* with our main approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	mislili so povsod, de nihče iz zlate vasí beračevati ne more.
Normalized:	mislili so povsod, da nihče iz zlate vasi berači␣␣ti ne more.
CBSMT:	mislili so povsod, da nihče iz zlate vasi bračevati ne more.
En. CBNMT$_{\text{LSTM}}$:	mislili so povsod, da nihče iz zlate vasi beračevati ne more.
En. CBNMT$_{\text{Trans.}}$:	mislili so povsod, da nihče iz zlate vasi beračevati ne more.

5.5.2 Additional CBNMT approaches

Now, we proceed to study the behavior of each CBNMT normalization approach when normalizing a sentence from each dataset.

Example 5.5: Example of normalizing a sentence from *Entremeses y Comedias* with the CBNMT approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	¡O mal logrado moço! Salid fuera;
Normalized:	¡Oh mal logrado mozo! Salid fuera;
Enriched CBNMT$_{\text{LSTM}}$:	¡Oh mal logrado mozo! Salí fuera;
Enriched CBNMT$_{\text{Trans.}}$:	¡Oh mal logrado mozo! Salí fuera;
Enriched SubChar$_{\text{LSTM}}$:	¡Oh mal logrado mozo! ␣␣␣␣␣ ␣␣␣␣␣␣
Enriched SubChar$_{\text{Trans.}}$:	¡O␣ mal logrado mozo! ␣␣␣␣␣ ␣␣␣␣␣␣
Enriched CharSub$_{\text{LSTM}}$:	¡Oh mal logrado mozo! Salí fuera;
Enriched CharSub$_{\text{Trans.}}$:	¡Oh mal logrado mozo! Salí fuera;

Example 5.6: Example of normalizing a sentence from *Quijote* with the CBNMT approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	“Para esso se yo vn buen remedio”, dixo el del Bosque;
Normalized:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;
En. CBNMT$_{\text{LSTM}}$:	“Para es␣o se yo un buen remedio”, dijo el del Bosque;
En. CBNMT$_{\text{Trans.}}$:	“Para es␣o se yo un buen remedio”, dijo el del Bosque;
En. SubChar$_{\text{LSTM}}$:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;
En. SubChar$_{\text{Trans.}}$:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;
En. CharSub$_{\text{LSTM}}$:	“Para es␣o se yo un buen remedio”, dijo el del Bosque;
En. CharSub$_{\text{Trans.}}$:	“Para es␣o sé yo un buen remedio”, dijo el del Bosque;

Example 5.5 shows an example from *Entremeses y Comedias*. In this case, all the CBNMT and CharSub approaches behave equally: they successfully normalized the two characters that needed to be updated but introduced one new mistake (the same one in all cases). The SubChar approach with an LSTM architecture was also able to normalize those characters, but the last part of the sentence is missing. Finally, the SubChar approach with a Transformer architecture behaved similarly, but failing to normalize one of the two characters.

Example 5.6 contains an example from *Quijote*, in which the normalization consists in updating four characters. The SubChar approaches and the CharSub with a Transformer architecture successfully generated an error-free normalization, while the other approaches made one mistake (the same one in all cases).

Example 5.7: Example of normalizing a sentence from *Bohorič* with the CBNMT approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original:	vadljajo ali lófajo, de bi svédili, kdo jim je kriv te nefrezhe.
Normalized:	vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč␣e.
En. CBNMT$_{\text{LSTM}}$:	vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.
En. CBNMT$_{\text{Trans.}}$:	vadeljajo ali losajo, da bi ␣zvedeli, kdo jim je kriv te nesreč ␣e.
En. SubChar$_{\text{LSTM}}$:	vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.
En. SubChar$_{\text{Trans.}}$:	vadlijo ali losajo, da bi ␣zvedili, kdo jim je neskrite v nes rene.
En. CharSub$_{\text{LSTM}}$:	valjo ali jokajo, da bi ␣zvedili, kdo jim je kri te nesreč␣e.
En. CharSub$_{\text{Trans.}}$:	vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč ␣e.

The normalization from *Bohorič* (see Example 5.7) consists in updating ten different characters. In this case, each approach behave differently, making from 1 (in the case of CharSub with Transformer) up to 10 mistakes (in the case of SubChar wit Transformer, which succeeded normalizing some characters but introduced new errors in the process).

Example 5.8: Example of normalizing a sentence from *Gaj* with the CBNMT approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original: mislili so povsod, de nihče iz zlate vasí beračevati ne more.
Normalized: mislili so povsod, da nihče iz zlate vasi berači␣␣ti ne more.

En. CBNMT$_{\text{LSTM}}$: mislili so povsod, da nihče iz zlate vasi beračevati ne more.
En. CBNMT$_{\text{Trans.}}$: mislili so povsod, da nihče iz zlate vasi beračevati ne more.

En. SubChar$_{\text{LSTM}}$: mislili so povsod, da nihče iz zlate vasi beračevati ne more.
En. SubChar$_{\text{Trans.}}$: zlate mislili so ␣␣␣␣␣␣, da nihče iz zlate vasi berača␣␣ti ne more.

En. CharSub$_{\text{LSTM}}$: mislili so povsod, da nihče iz zlate vasi beračevati ne more.
En. CharSub$_{\text{Trans.}}$: mislili so povsod, da nihče iz zlate vasi beračevati ne more.

Finally, in the example from *Gaj* (see Example 5.8), with one exemption, all approaches behave equally: they successfully normalized two characters and left the other three untouched. The exemption was the SubChar approach with a Transformer architecture, which succeeded in normalizing four of the five characters but introduced several errors (including a missing word).

5.6 Conclusions

In this chapter, we have presented our contributions to the spelling normalization task. In order to account for the lack of spelling conventions by updating a document's orthography according to modern standards, we proposed several normalization approaches based on MT and CBMT, which included the use of modern documents to enrich the neural models.

We tested our approaches on different datasets from different time periods and languages, reaching the conclusion that the CBSMT approach is the most suitable for this task. We believe that this is mostly due to the scarce availability of parallel training data when working with historical documents (Bollmann and Søgaard, 2016).

Finally, we evaluated our best approaches in a similar but more challenging task from the field of HTR, achieving very encouraging results. While HTR has the advantage of generating modern transcripts from a manuscript's image, our approach is more suitable for documents that have already been transcribed and only need to be normalized.

As a future work, we would like to further research the use of modern documents to enrich the neural systems and to conduct a human evaluation to verify our automatic results. Additionally, we would like to try another neural architectures, such as convolutional neural networks. Moreover, we want to address these particularities of the HTR task to try to improve our system's performance.

5.7 Publications

Some of our contributions to the spelling normalization task were accepted for publication at international conferences and journals:

- Miguel Domingo and Francisco Casacuberta. "A comparison of character-based neural machine translations techniques applied to spelling normalization". In *Proceedings of the International Conference on Pattern Recognition. International Workshop on Pattern Recognition for Cultural Heritage*, pages 326–338, 2021.

  Contributions: *Additional CBNMT and enriched CBNMT approaches.*

- Miguel Domingo and Francisco Casacuberta. "Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents". In *Proceedings of the International Conference on Image Analysis and Processing. International Workshop on Pattern Recognition for Cultural Heritage*, pages 59–69, 2019.

  Contributions: *Main enriched CBNMT approach.*

- Miguel Domingo and Francisco Casacuberta. "Spelling normalization of historical documents by using a machine translation approach". In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 129–137, 2018. CORE B.

  Contributions: *NMT and main CBMT approaches.*

- Erik Tjong Kim Sang, Marcel Bollmann, Remko Boschker, Francisco Casacuberta, Feike Dietz, Stefanie Dipper, Miguel Domingo, Rob van der Goot, van Marjo Koppen, Nikola Ljubešić, Robert Östling, Florian Petran, Eva Pettersson, Yves Scherrer, Marijn Schraagen, Leen Sevens, Jörg Tiedemann, Tom Vanallemeersch and Kalliopi Zervanou. "The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation". *Computational Linguistics in the Netherlands Journal*, 7:53–64, 2017.

  Contributions: *SMT approach.*

Chapter 6
Interactive Machine Translation for the Processing of Historical Documents

Paladins de l'Art t'ofrenen
ses victorias gegantines;
i als teus peus, Sultana, tots jardins extenen
un tapís de murta i de roses fines.

(***Himne de l'Exposició***. Maximilià Thous i Orts.)

Paladins of Art offer you
its gigantic victories;
and at your feet, Sultana, all gardens extend
a tapestry of myrtle and fine roses.

(***Hymn of the Exhibition***. Google Translate.)

Contents

With the aim of helping scholars in their work, in this chapter we applied the interactive machine translation (IMT) framework (see Chapter 3) to the tasks related to the processing of historical documents studied in this thesis: language modernization (see Chapter 4) and spelling normalization (see Chapter 5). Additionally, we developed an online demonstrator that showcases this workflow.

6.1 Language modernization

While the goal of language modernization is limited to helping non-experts to understand historical documents, scholars often need to generate error-free language modernization (e.g., for creating modern versions of classic literature). Therefore, by applying IMT to language modernization we can help scholars leverage this process. By using this framework, instead of modernizing from scratch, they will work with the system to generate the final modernizations with less effort.

With this aim, we propose to apply the classical prefix-based and the segment-based approach developed in Chapter 3 to our best language modernization proposals (see Chapter 4): statistical machine translation (SMT), enriched neural machine translation (NMT) with an long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) architecture and enriched NMT with a Transformer architecture.

6.1.1 Experimental framework

In this section, we describe the corpora, systems and metrics used in our experimental session.

Corpora

The corpora selected for assessing our proposals is the same used in Chapter 4:

- Dutch Bible: a collection of different versions of the Dutch Bible.
- El Quijote: the well-known 17th century Spanish novel by Miguel de Cervantes and a recent modern translation.
- OE-ME: the original 11th century English text *The Homilies of the Anglo-Saxon Church* and a 19th century version.

Systems

The SMT systems are the same ones used in Chapter 4. Then, we used the prefix-based and segment-based implementations from Chapter 3 to run the IMT sessions (refer to the aforementioned chapter for more details about the implementations).

The NMT systems were trained with *NMT-Keras* (Peris and Casacuberta, 2018) under the same conditions as in the systems used in Chapter 4 (refer to this chapter for more details). The reason for retraining these systems is that we are using Peris et al. (2017)'s prefix-based and segment-based protocols, which are implemented in *NMT-Keras*—including the user simulations.

Metrics

In order to evaluate our proposals, we made use of the same metrics used in Chapter 3:

- word stroke rate (WSR): to quantify the typing effort.
- mouse action rate (MAR): to quantify the effort of using the mouse.

Additionally, to assess the initial quality of the modernization systems, we made use of the metrics from Chapter 4:

- translation error rate (TER) (Snover et al., 2006).
- bilingual evaluation understudy (BLEU) (Papineni et al., 2002).

6.1.2 Evaluation

Table 6.1 presents the results of our experimental sessions. It presents the initial quality of each modernization system, and compares our best modernization approaches in a prefix-based or a segment-based framework.

In all cases, the SMT approach yielded the best results with a great difference. The prefix-based protocol successfully reduces the human effort of creating error-free modernizations. More over, the segment-based protocol proposed in this thesis reduced the typing effort even more, at the expenses of a small increase in the use of the mouse, which we are assuming that has a smaller impact in the human effort—we shall corroborate this in a future work.

Table 6.1: Experimental results of our language modernization IMT approaches. The initial modernization quality is meant to be a starting point comparison of each system. *En.* stands for enriched. All results are significantly different between all approaches except those denoted with †. Given the same approach, all results are significantly different between the different IMT protocols except those denoted with ‡. [↓] indicates that the lowest the value the highest the quality. [↑] indicates that the highest the value the highest the quality. Best results are denoted in **bold**.

Corpus	Approach	Modernization quality		Prefix-based		Segment-based	
		TER [↓]	**BLEU** [↑]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
Dutch Bible	SMT	11.5	77.5	14.3	4.4	**9.0**	**10.8**
	En. NMT_{LSTM}	$50.7^{\dagger}$	43.4	$42.6^{\ddagger}$	9.2	$42.6^{\ddagger}$	50.9
	En. $\text{NMT}_{\text{Transformer}}$	$50.3^{\dagger}$	35.8	$49.2^{\ddagger}$	10.4	$49.2^{\ddagger}$	48.3
El Quijote	SMT	30.7	58.3	38.8	10.9	**22.0**	**19.7**
	En. NMT_{LSTM}	42.9	50.4	$68.9^{\ddagger}$	11.8	$68.9^{\ddagger}$	47.8
	En. $\text{NMT}_{\text{Transformer}}$	47.3	46.1	$73.2^{\ddagger}$	13.4	$73.2^{\ddagger}$	50.5
OE-ME	SMT	39.6	39.6	58.2	15.5	**28.2**	**26.1**
	En. NMT_{LSTM}	56.4	30.3	$72.1^{\ddagger}$	$12.8^{\dagger}$	$72.1^{\ddagger}$	59.5
	En. $\text{NMT}_{\text{Transformer}}$	58.9	28.2	$73.5^{\ddagger}$	$13.3^{\dagger}$	$73.5^{\ddagger}$	49.5

Regarding the NMT approaches, while all of them also successfully decreased the human effort, this diminishment is significantly smaller than with the SMT approach. Moreover, unlike in interactive neural machine translation (INMT), the segment-based protocol does not offer any benefit with respect to the prefix-based: both protocols have the same typing effort and the segment-based protocol has a significant increase in the mouse usage. Finally, is it worth noting how the initial quality of the systems is considerably lower than the qualities reported at Table 4.2—specially in the case of *Ductch Bible*. Most likely, this is due to *NMT-Keras* not having implemented a strategy for dealing with unknown words.

6.1.3 Qualitative analysis

Example 6.1 showcase some strengths and weakness of our proposed segment-based protocol applied to language modernization. On the one hand, the user only needs to type six words to generate the modernization, while generating it from scratch would have taken them to type twenty-four words and post-editing the system's initial hypothesis would have taken them twenty-one corrections. On the other hand, the spurious words' problem (see Section 3.3.1) is present, resulting in the same words appearing over and over, and an increase of the number of mouse actions.

Example 6.1: Example of modernizing a sentence using the segment-based IMT protocol. Words in *italic* represent undesired translations produced by the system.

source ($\mathbf{x}$): He forfleah þone woruldlican wurðmynt, þaþa he wæs to cyninge gecoren; ac he ne forfleah na þæt edwit and ðone hosp, þaþa ða Iudeiscan hine woldon on rode ahón.

target translation ($\tilde{\mathbf{y}}$): He fled from worldly honour, when he was chosen king; but he fled not from reproach and scorn, when the Jews would hang him on a cross.

IT-0	**MT**	He fled from the woruldlican honour to the king, when he was not chosen; but he fled from not that edwit and the mockery, when the Jews him would hanged on a cross.
IT-1	**User**	$\boxed{\text{He fled from}}$ the **worldly** $\boxed{\text{honour}}$ to the king, $\boxed{\text{when he was}}$ not $\boxed{\text{chosen}}$; $\boxed{\text{but he fled}}$ $\boxed{\text{from}}$ not that edwit and the mockery, $\boxed{\text{when the Jews}}$ him $\boxed{\text{would}}$ hanged $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from}}$ *the woruldlican* $\boxed{\text{worldly}}$ $\boxed{\text{honour}}$ to the king, $\boxed{\text{when he was}}$ $\boxed{\text{chosen}}$ nor not that edwit; $\boxed{\text{but he fled}}$ $\boxed{\text{from}}$ the mockery and him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
IT-2	**User**	$\boxed{\text{He fled from worldly honour}}$ to the king, $\boxed{\text{when he was chosen}}$ nor not that **king**; $\boxed{\text{but he fled}}$ $\boxed{\text{from}}$ the mockery $\boxed{\text{and}}$ him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from worldly honour}}$ $\boxed{\text{when he was chosen}}$ *the woruldlican* to the $\boxed{\text{king}}$; $\boxed{\text{but he fled}}$ nor not that edwit $\boxed{\text{from}}$ the mockery $\boxed{\text{and}}$ him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
IT-3	**User**	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ **reproach** mockery $\boxed{\text{and}}$ him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ *woruldlican to not that edwit* $\boxed{\text{reproach}}$ the mockery $\boxed{\text{and}}$ him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
IT-4	**User**	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and}}$ **scorn**, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and}}$ *the woruldlican to not that edwit the mockery* $\boxed{\text{scorn}}$ him, $\boxed{\text{when the Jews}}$ $\boxed{\text{would}}$ hang $\boxed{\text{on a cross.}}$
IT-5	**User**	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and scorn when the Jews would}}$ **hang** $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and scorn when the Jews would}}$ *slay the woruldlican to not that edwit the mockery* him to $\boxed{\text{hang}}$ $\boxed{\text{on a cross.}}$
IT-6	**User**	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and scorn when the Jews would hang}}$ **him** $\boxed{\text{on a cross.}}$
	MT	$\boxed{\text{He fled from worldly honour when he was chosen king but he fled not from}}$ $\boxed{\text{reproach and scorn when the Jews would hang}}$ $\boxed{\text{him}}$ *woruldlican edwit mockery* $\boxed{\text{on a cross.}}$
END	**User**	He fled from worldly honour when he was chosen king but he fled not from reproach and scorn when the Jews would hang him on a cross.

6.2 Spelling normalization

Similarly to machine translation (MT), spelling normalization is still not able to generate error-free normalizations. Therefore, a scholar needs to either correct the system's mistakes or to generate normalizations from scratch. To reduce their effort and increase their productivity, we propose to apply the IMT field to spelling normalization so that scholars work together with the system to generate error-free normalizations.

With this aim, we propose to apply the classical prefix-based and the segment-based approach developed in Chapter 3 to our best normalization systems (see Chapter 5): character-based statistical machine translation (CBSMT), enriched character-based neural machine translation (CBNMT) with an LSTM architecture and enriched CBNMT with a Transformer architecture.

Corpora

The corpora selected for assessing our proposals is the same used in Chapter 5:

- Entremeses y Comedias: a 17th century Spanish collection of comedies by Miguel de Cervantes.
- Quijote: the 17th century Spanish two-volumes novel by Miguel de Cervantes.
- Bohorič: a collection of 18th century Slovene texts written in the old Bohorič alphabet.
- Gaj: a collection of 19th century Slovene texts written in the Gaj alphabet.

Systems

The SMT systems are the same ones used in Chapter 5. Then, we used the prefix-based and segment-based implementations from Chapter 3 to run the IMT sessions (refer to the aforementioned chapter for more details about the implementations).

The NMT systems were trained with *NMT-Keras* (Peris and Casacuberta, 2018) under the same conditions as in the systems used in Chapter 5 (refer to this chapter for more details). The reason for retraining these systems is that we are using Peris et al. (2017)'s prefix-based and segment-based protocols, which are implemented in *NMT-Keras*—including the user simulations.

Metrics

In order to evaluate our proposals, we made use of the same metrics used in Chapter 3. However, since our spelling normalization systems work at a character level, we replaced WSR with key stroke rate (KSR) (Tomás and Casacuberta, 2006):

- KSR: To quantify the typing effort. Measures the number of characters edited by the user, normalized by the number of characters in the final translation.
- MAR: to quantify the effort of using the mouse.

Additionally, to assess the initial quality of the modernization systems, we made use of the metrics from Chapter 5: character error rate (CER), TER and BLEU.

6.2.1 Evaluation

Table 6.2 presents the results of applying IMT to spelling normalization. It presents the initial quality of each normalization system and compares our best normalization approaches in a prefix-based or a segment-based framework.

Except for *Quijote*, in which there were no significant differences in the typing effort but an increase in the mouse effort, the CBSMT systems yielded the best results. We should note that, due to the great differences in normalization quality of the CBNMT approaches for *Bohorič* and *Gaj* (these systems were not able to improve the baseline; refer to Chapter 5 for more details), we decided not to apply the INMT framework to these approaches for those datasets.

In the case of *Entremeses y Comedias* and *Quijote*, there were no significant differences in the typing effort between the classic prefix-based protocol and our proposed segment-based one. Most likely, this is due to the highest initial quality of the systems: since there are fewer errors to correct, using one methodology over the other one is not so relevant as when there are more errors. However, our segment based protocol comes with a small increase in the mouse effort, since more user actions are defined for this protocol. In the case of *Bohorič* and *Gaj*, the segment-based protocol achieves a small reduction of the typing effort, at the expenses of a small increase of the mouse effort. We consider the typing reduction

Table 6.2: Experimental results of our spelling normalization IMT approaches. The initial modernization quality is meant to be a starting point comparison of each system. *En.* stands for enriched. All results are significantly different between all approaches except those denoted with †. Given the same approach, all results are significantly different between the different IMT protocols except those denoted with ‡. [↓] indicates that the lowest the value the highest the quality. [↑] indicates that the highest the value the highest the quality. Best results are denoted in **bold**.

		Normalization quality			**Prefix-based**		**Segment-based**	
Corpus	**Approach**	**CER** [↓]	**TER** [↓]	**BLEU** [↑]	**KSR** [↓]	**MAR** [↓]	**KSR** [↓]	**MAR** [↓]
Entremeses y Comedias	CBSMT	$1.3^{\dagger}$	4.4	91.7	**0.9**‡	**4.1**	$0.7^{\ddagger}$	6.7
	En. CBNMT$_{\text{LSTM}}$	3.5	9.4	84.9	$1.9^{\ddagger}$	$2.1^{\dagger}$	$1.9^{\ddagger}$	3.3
	En. CBNMT$_{\text{Transformer}}$	$1.5^{\dagger}$	6.5	87.2	$1.4^{\ddagger}$	$2.1^{\dagger}$	$1.4^{\ddagger}$	3.4
Quijote	CBSMT	$2.5^{\dagger}$	$3.0^{\dagger}$	$94.4^{\dagger}$	$1.4^{\dagger\ddagger}$	3.7	$1.1^{\dagger\ddagger}$	5.3
	En. CBNMT$_{\text{LSTM}}$	$2.6^{\dagger}$	4.3	$93.9^{\dagger}$	**1.4**†	**1.4**†‡	$1.4^{\dagger\ddagger}$	2.1
	En. CBNMT$_{\text{Transformer}}$	$2.2^{\dagger}$	$3.7^{\dagger}$	$94.4^{\dagger}$	**1.5**†‡	**1.4**†	$1.5^{\dagger\ddagger}$	2.1
Bohorič	CBSMT	2.4	8.7	80.4	2.5	3.7	**2.0**	**8.6**
Gaj	CBSMT	1.4	5.1	88.3	1.9	3.0	**0.9**	**5.2**

beneficial despite the mouse increase. However, we should verify this assumption in a future work with actual humans.

6.2.2 Qualitative analysis

Due to the high quality of the systems, the IMT sessions consists on a few user interactions (exemplified by Example 6.2). It is worth noting, however, that working at a character level has significantly decreased the occurrence of spurious words[1] (see Section 3.3.1). Recall that this problem was mainly caused by the systems failing to deal with the user corrections when they were out-of-vocabulary words. At a character level, this type of corrections happen significantly less frequently. We should further investigate this phenomenon as a solution to this problem.

[1] In this case, spurious characters.

Example 6.2: Example of normalizing a sentence using the segment-based IMT protocol.

source (x): silni stresaj mu je roko umrtvil, da mu ni rabila.
target translation (ŷ): silni stresaj mu je roko omrtvil, da mu ni rabila.

IT-0	**MT**	silni stresaj mu je roko umrtvil, da mu ni rabila.
IT-1	**User**	$\boxed{\text{silni stresaj mu je roko}}$ **u** $\boxed{\text{mrtvil, da mu ni rabila.}}$
	MT	$\boxed{\text{silni stresaj mu je roko}}$ $\boxed{\text{u}}$ $\boxed{\text{mrtvil, da mu ni rabila.}}$
END	**User**	silni stresaj mu je roko omrtvil, da mu ni rabila.

6.3 Online demonstrator

In this section we present the online demonstrator[2] that we developed to showcase the advantages of working on an IMT workflow. In order to simplify the user interface, we decided to only implement the prefix-based protocol. Additionally, while our best results were obtained using our proposed segment-based protocol with SMT systems, we chose to limit our demonstrator to INMT systems for logistic reasons.

Our system is composed of two main elements: the client and the server. The client is an HTML website, which uses javascript to interact with the user and the HTTP protocol with the PHP curl tool to communicate with the server. The server is deployed as a Python HTTP server that handles the client's requests. It is the core element and contains the NMT systems, which were developed with *NMT-Keras* (Peris and Casacuberta, 2018). All code is open-source and publicly available[3].

Initially, old sentences are presented to the user in the client website. When the user requests an automatic modernization/normalization, the client communicates the server via PHP. Then, the server queries the NMT system, which generates an initial hypothesis applying Eq. (2.9). After that, the hypothesis is sent back to the client website.

At this point, the interactive-predictive process starts: The user reviews the hypothesis and introduces a correction with the keyboard (writing one or more characters). Then, the client reacts to this feedback by sending a request to the server, which contains the old sentence and the user feedback (the sequence of

[2] `http://demosmt.prhlt.upv.es/mthd/`.

[3] `https://github.com/midobal/mthd`.

characters that conform the prefix). Then, the NMT system applies Eq. (3.6) to produce an alternative hypothesis coherent with the user's feedback and sends it back to the client website. This process is repeated until the user is satisfied with the system's hypothesis. Fig. 6.1 illustrates one step of this process.

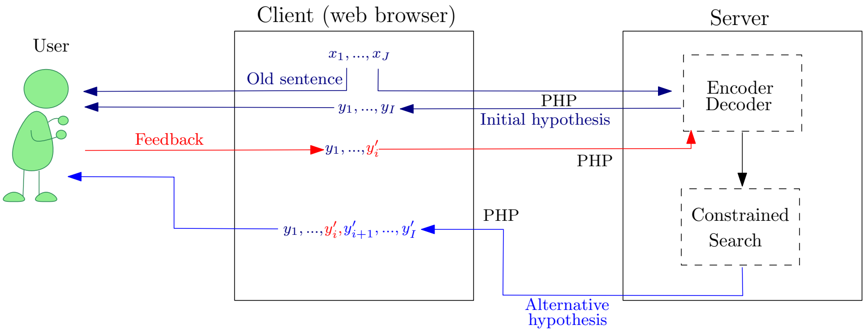
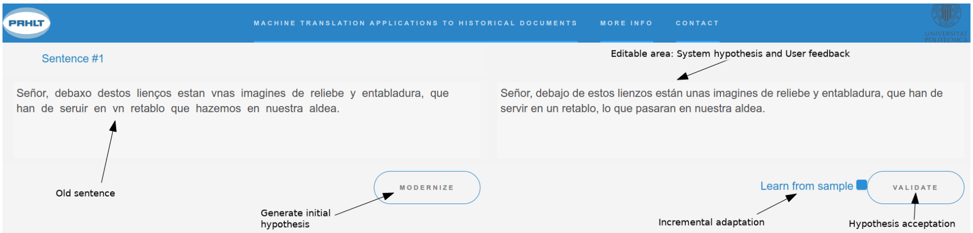


Figure 6.1: System architecture. The client presents the user an old sentence and a prediction. Then, the user introduces a feedback signal for correcting this prediction (in this example, they are validating the prefix $y_1, .., y_{i-1}$ and correcting the word y'_i). After that, the old sentence and the user's feedback is sent to the server, which generates an alternative hypothesis that takes into account the user corrections (in this example, a new suffix $y'_{i+1}, .., y'_I$ that completes the user's feedback).


PRHLT
MACHINE TRANSLATION APPLICATIONS TO HISTORICAL DOCUMENTS
MORE INFO
CONTACT
Sentence #1
Editable area: System hypothesis and User feedback
Old sentence
MODERNIZE
Generate initial hypothesis
Learn from sample
VALIDATE
Incremental adaptation
Hypothesis acceptation

Señor, debaxo destos lienços estan vnas imagines de reliebe y entabladura, que han de seruir en vn retablo que hazemos en nuestra aldea.

Señor, debajo de estos lienzos están unas imagines de reliebe y entabladura, que han de servir en un retablo, lo que pasaran en nuestra aldea.

Figure 6.2: Frontend of the client website. As the button "Modernize" is clicked (or "Normalize", depending on the task you are performing), an initial hypothesis for the old sentence appears in the right area. Then, the user can introduce corrections of this text. The system will react to each correction, producing alternative hypotheses coherent with the user feedback. Once the user is satisfied with the modernization hypothesis, they can click in the "Validate" button to accept the hypothesis.

Once the user is satisfied with the system's hypothesis, they can validate it. When this happens, the system is updated with this new sample following an incremental learning setup (see Section 7.4.2). Thus, in future interactions, the system will be progressively updated and able to generate better hypothesis.

Fig. 6.2 illustrates an example of how to perform a task using the client server. After having selected the task to perform, a list of old sentences will appear. When you click on "Modernize/Normalize", the system will generate an initial hypothesis. If you desire to improve this hypothesis, you can click on the left box and type a correction. The system will, then, generate a new hypothesis to take that correction into account. You can repeat this process for as many corrections as you desire to make. Finally, you can click "Validate" to tell the system that you are happy with the modernization/normalization and, if the *Learn from sample* option is activated (in blue), the system will use the sample to improve its model.

6.4 Conclusions

In this chapter, we have applied IMT to the studied tasks for processing historical documents in order to create a workflow in which system and scholar work together to generate error-free modernizations/normalizations.

We have combined the classic prefix-based and our proposed segment-based protocol with our best proposals for each task, reaching the conclusion that our segment-based protocol significantly reduces the human effort. The only exception were two cases in which, due to the high quality of the systems, both protocols had similar behaviors. Nonetheless, they succeeded in reducing the human effort. Additionally, we created an online demonstrator to showcase the proposed workflow.

We evaluated our proposals under a simulated setting. As a future work, we should conduct a human evaluation with the help of scholars specialized on these tasks. Additionally, we observed that working at a character level mitigated the undesired translations problem present in the segment-based approach. We would like to further research this issue.

6.5 Publications

Some of our contributions to applying an IMT framework to historical documents were accepted for publication at international conferences and journals:

- Miguel Domingo and Francisco Casacuberta. "Two demonstrations of the machine translation applications to historical documents". *arXiv preprint arXiv:2102.01417*, 2021. Presented at the Demos session of ICPR 2020: `https://www.micc.unifi.it/icpr2020/index.php/demos/`. CORE B.

  Contributions: *Online demonstrator.*

Chapter 7

conclusions

¿Qué es la vida? Un frenesí.
¿Qué es la vida? Una ilusión,
una sombra, una ficción,
y el mayor bien es pequeño;
que toda la vida es sueño,
y los sueños, sueños son.

(***La vida es sueño***. Calderón de la Barca.)

What is life? A frenzy.
What is life? An illusion,
a shadow, a fiction,
and the greatest good is small;
that all life is a dream,
and dreams are dreams.

(***The life is dream***. Google Translate.)

Contents

We conclude this dissertation by summarizing the goals achieved. Then, we list all the publications that derived from this thesis and the software contributions that were made. Finally, we offer some lines of future work which we consider to be interest steps to take.

7.1 Scientific contributions

In this thesis, we worked on improving the interactive machine translation (IMT) framework to reduce the human effort for generating high-quality translations. Then, we applied the machine translation (MT) field to language modernization and spelling normalization, which are two tasks related with the language properties of historical documents. Finally, we brought the IMT framework into these two tasks in order to help scholar to generate error-free modernizations/normalizations.

7.1.1 Interactive machine translation

We developed a new IMT protocol that allows the user to validate the correct parts of a translation hypothesis, breaking the left-to-right constrains present in most protocols. We conducted a wide experimentation, which showed that the segment-based methodology successfully takes advantage of the correct parts of a translation hypothesis, achieving a substantial decrease of the typing effort, at the expenses of an increase in the number of mouse actions. This increase is mostly due to the system's main weaknesses, which are related with the user corrections being out-of-vocabulary words.

Additionally, we tested an active prediction protocol to assist the user in the correction step of the process. This protocol was based on the use of different confidence measures techniques and consisted in the system guiding the user to make first the corrections which would benefit the system most. However, results did not present statistical differences between each approach. Therefore, we concluded that changing the order in which words are corrected had no effect in the overall user effort.

Finally, we applied the IMT field to two tasks related with the processing of historical documents, creating a workflow in which a scholar collaborates with a modernization/normalization system to generate error-free modernizations/normalizations. We combined the classic prefix-based and our proposed segment-based protocol with our best proposals for each task, reaching the conclusion that our segment-based protocol significantly reduces the human effort. Additionally, we created an online demonstrator to showcase this workflow.

7.1.2 Machine translation applications to historical documents

We developed several machine translation applications for two tasks related with the language properties of historical documents: language modernization and spelling normalization.

Language modernization

With the aim of making historical documents easier to understand and more accessible to a broader audience, we proposed several modernization approaches based on MT.

We conduced a wide experimentation, which counted with the help of 4 scholars and 42 volunteers. While being far from perfect, results showed that our approaches succeeded in making historical documents easier to comprehend by a more general audience.

Spelling normalization

In order to account for the lack of spelling conventions by updating a document's orthography according to modern standards, we proposed several normalization approaches based on MT and character-based machine translation (CBMT).

We evaluated our approaches using different datasets from different time periods and languages, obtaining satisfactory results and observing how the CBMT approaches were more suitable for this task.

Finally, we tested our best approaches in a similar—although more challenging—task from the field of handwritten text recognition (HTR), achieving very encouraging results. While HTR has the advantage of generating modern transcripts from a manuscript's image, our approach is more suitable for documents that have already been transcribed and only need to be normalized.

7.2 Publications derived from the thesis

Throughout the development of the thesis, several works were accepted for publication at international conferences and journals. At each chapter, we have presented the related publications and their contributions. Here we provide a general overview of all publications together with their related chapter.

- JCR-ranked journals:
  - Miguel Domingo and Francisco Casacuberta. “Modernizing historical documents: A user study”. *Pattern Recognition Letters*, 133:151–157, 2020. JCR Q2. ***Chapter 4 - Language Modernization***.

    Contributions: *human evaluation and user study.*
  - Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. “Interactive neural machine translation”. *Computer Speech & Language*, 45:201–220, 2017. JCR Q2. ***Chapter 3 - Interactive Machine Translation***.

    Contributions: *I helped in the design of the segment-based interactive neural machine translation (INMT) protocol.*
- Other journals (peer-reviewed):
  - Miguel Domingo and Álvaro Peris and Francisco Casacuberta. “Segment-based interactive-predictive machine translation”. *Machine Translation Journal*, 31:163–185, 2017. ***Chapter 3 - Interactive Machine Translation***.

    Contributions: *extension and in-depth study of the segment-based protocol and use of confidence measures (CM).*
  - Erik Tjong Kim Sang, Marcel Bollmann, Remko Boschker, Francisco Casacuberta, Feike Dietz, Stefanie Dipper, Miguel Domingo, Rob van der Goot, van Marjo Koppen, Nikola Ljubešić, Robert Östling, Florian Petran, Eva Pettersson, Yves Scherrer, Marijn Schraagen, Leen Sevens, Jörg Tiedemann, Tom Vanallemeersch and Kalliopi Zervanou. “The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation”. *Computational Linguistics in the Netherlands Journal*, 7:53–64, 2017. ***Chapter 5 - Spelling Normalization***.

    Contributions: *statistical machine translation (SMT) approach.*

- CORE-ranked conferences:

  - Miguel Domingo and Francisco Casacuberta. “Two demonstrations of the machine translation applications to historical documents”. *arXiv preprint arXiv:2102.01417*, 2021. Presented at the Demos session of ICPR 2020: `https://www.micc.unifi.it/icpr2020/index.php/demos/`. CORE B. ***Chapter 6 - Interactive Machine Translation for the Processing of Historical Documents***.

    Contributions: *Online demonstrator.*

  - Miguel Domingo and Francisco Casacuberta. “Spelling normalization of historical documents by using a machine translation approach”. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 129–137, 2018. CORE B. ***Chapter 5 - Spelling Normalization***.

    Contributions: *neural machine translation (NMT) and main CBMT approaches.*

  - Miguel Domingo, Mara Chinea-Rios, and Francisco Casacuberta. Historical documents modernization. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 295–306, 2017. CORE B. ***Chapter 4 - Language Modernization***.

    Contributions: *SMT approach.*

  - Miguel Domingo, Álvaro Peris, and Francisco Casacuberta. “Interactive-predictive translation based on multiple word-segments”. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 282–291, 2016. CORE B. **Best paper award**. ***Chapter 3 - Interactive Machine Translation***.

    Contributions: *segment-based protocol.*

- Workshops (peer-reviewed):

  - Miguel Domingo and Francisco Casacuberta. “A machine translation approach for modernizing historical documents using back translation”. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 39–47, 2018. ***Chapter 4 - Language Modernization***.

Contributions: *NMT and enriched NMT approaches.*

- Miguel Domingo and Francisco Casacuberta. “A comparison of character-based neural machine translations techniques applied to spelling normalization”. In *Proceedings of the International Conference on Pattern Recognition. International Workshop on Pattern Recognition for Cultural Heritage*, pages 326–338, 2021. ***Chapter 5 - Spelling Normalization***.

  Contributions: *Additional character-based neural machine translation (CBNMT) and enriched CBNMT approaches.*

- Miguel Domingo and Francisco Casacuberta. “Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents”. In *Proceedings of the International Conference on Image Analysis and Processing. International Workshop on Pattern Recognition for Cultural Heritage*, pages 59–69, 2019. ***Chapter 5 - Spelling Normalization***.

  Contributions: *Main enriched CBNMT approach.*

7.3 Software contributions

All the software developed in this thesis has been open-sourced and it is publicly available:

- Segment-based IMT (`https://github.com/midobal/sb-imt`): Implementation of the segment-based IMT protocol. ***Chapter 3 - Interactive Machine Translation***.

- Prefix-based IMT (`https://github.com/midobal/pb-imt`): Extension of the implementation of the prefix-based IMT protocol, to give compatibility with *Moses* version 3. ***Chapter 3 - Interactive Machine Translation***.

- Statistical dictionary (`https://github.com/midobal/sd`): Implementation of the statistical dictionary used in this thesis. ***Chapter 5 - Spelling Normalization***.

- Online demonstrator (`https://github.com/midobal/mthd`): Implementation of the online demonstrator. ***Chapter 6 - Interactive Machine Translation for the Processing of Historical Documents***.

- Active learning (`https://github.com/midobal/OpenNMT-py/tree/OnlineLearning`): Implementation of the active learning IMT protocol. ***Chapter 3 - Interactive Machine Translation***.
- INMT (`https://github.com/PRHLT/nmt-keras/tree/interactive_NMT`): Toolkit used for the INMT experiments. This toolkit was developed by Álvaro Peris as part of their thesis (Peris, 2019). I merely made some contributions to it such as migrating the source code from python 2 to python 3. ***Chapter 3 - Interactive Machine Translation***.

7.4 Other contributions

During the development of the thesis I worked on other MT-related tasks which were left out from the thesis for not being directly related with it. Nonetheless, this section includes a brief description of those contributions. Most of them were conducted in collaboration with Pangeanic[1], a professional translation company. Their main aim was to improve their production system by offering them tailored solutions to their industrial needs.

7.4.1 Tokenization

In this work, we conducted a study comparing several tokenization methods and their impact on NMT for the language pairs most frequently used by Pangeanic. While not directly related to the thesis, it had an influence in the NMT systems built as part of this thesis. The results of this study derived in the following publication:

- Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. “How much does tokenization affect neural machine translation?”. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, 2019. In press. CORE B.

[1] `https://pangeanic.com/`.

7.4.2 Incremental learning

In this work, we studied the use of incremental learning techniques in order to train adaptive NMT systems. Our goal was to take profit from the post-editing process from Pangeanic's workflow to improve the NMT systems, as well as to tailor those systems to translator's personal preferences. We conducted a user study with the help of several professional translators working for Pangeanic, observing a significant increase of the translator's productivity. Thus, we developed and integrated a first prototype into the company's production system. While not having a direct relation with the protocols studied at Chapter 3, the incremental learning workflow is straightly related with IMT. Furthermore, we incorporated this workflow into the demonstrator developed at Section 6.3. Several publications were derived from this work:

- Miguel Domingo, Mercedes García-Martínez, Álvaro Peris, Alexandre Helle, Amando Estela, Laurent Bié, Francisco Casacuberta, and Manuel Herranz. "A user study of the incremental learning in NMT". In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 319–328, 2020. CORE B.

  Contributions: *extended study of incremental learning on an industrial setting.*

- Miguel Domingo, Mercedes García-Martínez, Amando Estela, Laurent Bié, Alexandre Helle, Álvaro Peris, Francisco Casacuberta, and Manuel Herranz. "Demonstration of a neural machine translation system with online learning for translators". In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 70–74, 2019. CORE A+.

  Contributions: *demonstrator of an industrial production system with incremental learning.*

- Miguel Domingo, Mercedes García-Martínez, Álvaro Peris, Alexandre Helle, Amando Estela, Laurent Bié, Francisco Casacuberta, and Manuel Herranz. "Incremental adaptation of NMT for professional post-editors: A user study". In *Proceedings of the Machine Translation Summit*, pages 219–227, 2019. CORE B.

  Contributions: *user study of incremental learning on an industrial setting.*

7.5 Future works

Finally, we present some lines of work which we would like to address in a near future.

7.5.1 Interactive machine translation

We need to improve the way in which the system deals with the user corrections. An interesting point that arose when integrating our protocol for the spelling normalization task is that, since working at a character level reduces the vocabulary problems, our system is able to find the corresponding source words of a user correction more easily. Thus, we would like to research this line of work.

Additionally, we want to develop new protocols to assist the user in the segment validation step of the process and add newer features such as reordering segments.

Finally, we assumed that making a mouse action has a smaller effort than typing a word and, therefore, that the increase in the mouse effort pays off with respect to the great reduction of the typing effort. However, we should test our proposal with real users to obtain actual measures of the effort reduction. Additionally, we should conduct a human evaluation with the help of scholars specialized on the tasks related with historical documents.

7.5.2 Language modernization

We would like to benefit from the feedback received during the scholar evaluation and the user study and tackle the main problems that were pointed out. Mainly, punctuation, diacritical marks, the introduction of non-existent words and loosing parts of the given sentence. We would also like to conduct a new evaluation involving more scholars and more languages and datasets, and a new user study for different languages and datasets.

7.5.3 Spelling normalization

We would like to further research the use of modern documents to enrich the neural systems and to conduct a human evaluation to verify our automatic results. Additionally, we would like to try another neural architectures, such as convolutional neural networks. Moreover, we want to address these particularities of the HTR task to try to improve the performance of our techniques.

Appendix A

Tuning the Number of Merge Operations for the BPE Algorithm

During the evaluation process, one of the reviewers pointed out that the number of merge operations can be critical for low-resource languages, especially when using the Transformer model (e.g., Ding et al., 2019). Thus, the comparison of our neural machine translation and statistical machine translation proposals could not be fair enough. To better study this, we decided to run new experiments exploring the tuning of the merge operations for our main Transformer experiments. Fig. A.1 illustrates the results.

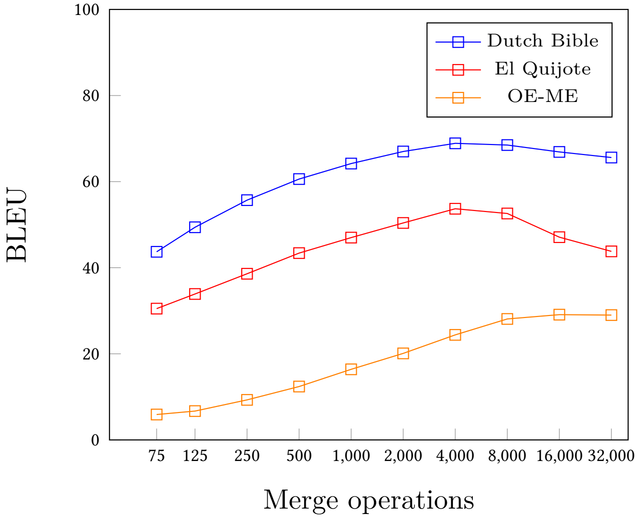


Figure A.1: Results of the BPE experiment for studying the relevance of tuning the number of merge operations.

From Fig. A.1, we selected the best configuration for each experiment. Then, we compared them against our current models (using the default number of merge operations). We studied the statistical differences between each system using approximate randomization testing (ART) (Riezler and Maxwell, 2005), observing that results were not significantly different among them according to any metric. Thus, we concluded than tuning this hyperparameter is not as critical for language modernization as it can be for other low-resource tasks. Table A.1 presents these results.

Table A.1: Comparison of tuning the number of merge operations against using the default value. Results are not significantly different between them. [↓] indicates that the lowest the value the highest the quality. [↑] indicates that the highest the value the highest the quality.

Approach	**Dutch Bible**		**El Quijote**		**OE-ME**	
	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]
Hyperparameter tuned	18.4	68.9	38.7	53.7	54.0	29.1
Default value	18.8	67.9	40.4	51.5	53.6	28.7

Appendix B
Language Modernization: Questionnaire from the User Study

Here we present a simplified copy of the questionnaire to the participants of the user study.

Instructions

The goal of this questionnaire is to evaluate if language modernization helps in the comprehension of old documents. To this effect, several questions in which you shall select the sentence which is easier for you to read and comprehend are asked. If you consider both sentences to be equally complex, you should select the option *Indifferent*. Finally, if you consider that the content of the sentences is different, then you should select the option *Both sentences do not have the same meaning*.

Personal data

Name: ....

Age:

○ Less than 20 years old.

○ 21 to 30 years old.

○ 31 to 40 years old.

○ 41 to 50 years old.

○ 51 to 60 years old.

○ 61 to 70 years old.

○ More than 70 years old.

Which is your degree of familiarity with *El Quijote*?

○ I am not familiarized at all.

○ I know what it is about, but I have never read it.

○ I have read fragments of an adapted version.

○ I have read an adapted version.

○ I have read fragments of the original novel.

○ I have read the original novel.

○ I have read the novel modernized by Andrés Trapiello.

○ I have read the original novel and the novel modernized by Andrés Trapiello.

Questions

This section was composed of 100 questions—50 in which modernization had been generated with the statistical machine translation approach and 50 generated with the neural machine translation approach—following this scheme:

Select the sentence which is easier for you to read and comprehend:

○ *Sentence 1.*

○ *Sentence 2.*

○ Indifferent.

○ Both sentences do not have the same meaning.

where *Sentence 1* and *Sentence 2* were either the original sentence or its modernized version (the order of appearance was randomized to avoid any bias). For the same reason, no information about the modernization systems was given to the users.

Appendix C
Post-editing the abstracts

In this appendix, we show the original translations and the modifications done during the post-editing process of each abstract. Post-editions are denoted in red. ~~Stroked~~ words denote words that have been deleted (e.g., *~~patrimonio~~ herencia* denotes that the word *patrimonio* has been replaced by the word *herencia*). ~~Stroked~~ characters denotes partial changes in a word (e.g., *nuestr~~o~~a* denotes that the word *nuestro* has been changed into *nuestra*). Finally, red characters denote additions to a word (e.g., *tareas* denotes that the word *tarea* has been changed into the word *tareas*).

Resumen

Los documentos históricos son una parte importante de nuestr~~o~~a ~~patrimonio~~ herencia cultural. Sin embargo, debido a la barrera ~~del~~ idiomática inherente ~~al~~ en el lenguaje humano y a las propiedades lingüísticas de estos documentos, su accesibilidad ~~se limita~~ está principalmente restringida a los académicos. Por un lado, el lenguaje humano evoluciona con el paso del tiempo. Por otro lado, las convenciones ortográficas no se crearon hasta hace poco y, por ~~lo~~ tanto, la ortografía cambia según el período ~~de tiempo~~ temporal y el autor. Por estas razones, el trabajo de los académicos es necesario para que los no expertos ~~obtengan~~ puedan obtener una comprensión básica de un documento ~~dado~~ determinado.

En esta tesis abordamos dos tareas relacionadas con el procesamiento de documentos históricos. La primera tarea es la *modernización del lenguaje* que, ~~con el~~ a fin de hacer que los documentos históricos ~~sean~~ estén más accesibles para los no expertos, tiene como objetivo reescribir un documento utilizando la versión moderna del idioma original del documento. La segunda tarea es la *normalización ortográfica*. Las propiedades lingüísticas ~~antes mencionadas~~ de los documentos históricos mencionadas con anterioridad suponen un desafío adicional para ~~el procesamiento~~ la aplicación efectiv~~o~~a del procesado del lenguaje natural ~~de~~ en estos documentos. Por lo tanto, esta tarea tiene como objetivo adaptar la ortografía de un documento a los estándares modernos ~~para~~ a fin de lograr una ~~coherencia~~ consistencia ortográfica.

~~Afrontamos tanto la~~ Ambas tareas las afrontamos desde una perspectiva de traducción automática, considerando el idioma original de un documento como el idioma ~~de origen~~ fuente, ~~como~~ y su ~~contraparte~~ homólogo modern~~a~~o/normalizad~~a~~o como el idioma ~~de destino~~ objetivo. Proponemos varios enfoques basados en la traducción automática estadística y neuronal, y llevamos a cabo una amplia experimentación que ~~muestra~~ ratifica el potencial de nuestras contribuciones –~~con~~ en donde los enfoques estadísticos ~~que~~ arrojan resultados iguales o mejores que los enfoques neuronales ~~en~~ para la mayoría de los casos–. ~~Para~~ En el caso de la tarea de modernización del lenguaje, esta experimentación incluye una evaluación humana realizada con la ayuda de académicos y un estudio ~~de~~ con usuarios que verifica que nuestras propuestas pueden ayudar a los no expertos a obtener una comprensión básica de un documento histórico sin la intervención de un académico.

Como ocurre con cualquier problema de traducción automática, nuestras aplicaciones no están libres de errores. Por lo tanto, para obtener modernizaciones/normalizaciones perfectas, un académico debe supervisar y corregir los errores. Este es un procedimiento común en la industria de la traducción. ~~El marco interactivo~~ La metodología de traducción automática interactiva tiene como objetivo reducir el esfuerzo necesario para obtener traducciones de alta calidad ~~integrando~~ uniendo ~~e~~al agente humano y ~~e~~al sistema de traducción en un proceso de corrección cooperativo. Sin embargo, la mayoría de los protocolos interactivos siguen una estrategia de izquierda a derecha. En esta tesis desarrollamos un nuevo protocolo interactivo que rompe con esta barrera de izquierda a derecha. ~~Evaluamos~~ Hemos evaluado este nuevo protocolo en un entorno de traducción automática, obteniendo grandes reducciones del esfuerzo humano. Finalmente, dado que este marco interactivo es de aplicación general ~~para~~ a cualquier problema de traducción, lo hemos aplica~~mos~~do –nuestro nuevo protocolo junto con uno de los protocolos clásicos de izquierda a derecha– ~~para~~ a la modernización del lenguaje y a la normalización ortográfica. Al igual que ~~con la~~ en traducción automática, el marco interactivo ~~disminuyó~~ logra disminuir el esfuerzo requerido para corregir los resultados de un sistema automático.

Resum

Els documents històrics són una ~~comunicat~~ part important de la nostra herència cultural. ~~Tanmateix~~ No obstant això, degut a la barrera idiomàtica inherent en el llenguatge humà i a les propietats lingüístiques d'aquests documents, ~~e~~la seua accessibili~~d~~ta~~d~~t està principalment restringida als acadèmics. D'una banda, el llenguatge humà evoluciona amb el pas del temps. D'altra banda, les convencions ortogràfiques no es van crear fins fa poc i, per tant, l'ortografia canvia segons el període temporal i l'autor. Per aquestes raons, el treball dels acadèmics és necessari perquè els no experts puguen obt~~enir~~indre una comprensió bàsica d'un document determinat.

En aquesta tesi ~~vam~~ abord~~ar~~em dues tasques relacionades amb el processament de documents històrics. La primera tasca és la *modernització del llenguatge* que, a fi de fer que els documents històrics estiguen més accessibles per als no experts, té per objectiu reescriure un document utilitzant la versió moderna de l'idioma original del document. La segona tasca és la *normalització ortogràfica.* Les propietats lingüístiques dels documents històrics mencionades amb anterioritat suposen un desafiament addicional per a l'aplicació efectiva del processat del llenguatge natural en aquests documents. Per tant, aquesta tasca té per objectiu adaptar l'ortografia d'un document als estàndards moderns a fi d'aconseguir una consistència ortogràfica.

~~Ambd~~Dues tasques les ~~vam~~ afront~~ar~~em des d'una perspectiva de traducció automàtica, considerant l'idioma original d'un document com a l'idioma font, i el seu homòleg modern/normalitzat com a l'idioma objectiu. Proposem diversos enfocaments basats en la traducció automàtica estadística i neuronal, i ~~vam~~ port~~ar~~em a terme una àmplia experimentació que ratifica el potencial de les nostres contribucions –on els enfocaments estadístics ~~llancen~~ obtenen resultats iguals o millors que els enfocaments neuronals per a la majoria dels casos–. En el cas de la tasca de modernització del llenguatge, aquesta experimentació inclou una avaluació humana realitzada amb l'ajuda d'acadèmics i un estudi amb usuaris que verifica que les nostres propostes poden ajudar ~~e~~als no experts a obt~~enir~~indre una comprensió bàsica d'un document històric sense la intervenció d'un acadèmic.

Com ~~acudeix~~ ocurreix amb qualsevol problema de traducció automàtica, les nostres aplicacions no estan lliures d'errades. Per tant, per obt~~enir~~indre modernitzacions/normalitzacions perfectes, un acadèmic ha de supervisar i corregir les errades. Aquest és un procediment comú en la indústria de la traducció. La metodologia de traducció automàtica interactiva té per objectiu reduir l'esforç necessari per obt~~enir~~indre traduccions d'alta qualitat unint a l'agent humà i al sis-

tema de traducció en un procés de correcció cooperatiu. ~~Tanmateix~~ Tot i això, la majoria dels protocols interactius ~~continuen~~ segueixen una estratègia d'esquerra a dreta. En aquesta tesi ~~vam~~ desenvolup~~ar~~em un nou protocol interactiu que trenca amb aquesta barrera d'esquerra a dreta. Hem avaluat aquest nou protocol en un entorn de traducció automàtica, obtenint grans reduccions de l'esforç humà. Finalment, atès que aquest marc interactiu és ~~aplicable~~ d'aplicació general a qualsevol problema de traducció, l'hem aplicat –el nostre nou protocol junt amb un dels protocols clàssics d'esquerra a dreta– a la modernització del llenguatge i a la normalitzaciò ortogràfica. De la mateixa manera que en traducció automàtica, el marc interactiu aconsegueix disminuir l'esforç requerit per corregir els resultats d'un sistema automàtic.

List of Figures

List of Tables

List of Equations

List of Examples

Bibliography

Agrawal, R. R., Turchi, M., and Negri, M. (2018). Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 11–20.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González-Rubio, J., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Sanchis-Trilles, G., and Tsoukala, C. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Alabau, V., Rodríguez-Ruiz, L., Sanchis, A., Martínez-Gómez, P., and Casacuberta, F. (2011). On multimodal interactive machine translation using speech recognition. In *Proceedings of the International Conference on Multimodal Interaction*, pages 129–136.

Alabau, V., Sanchis, A., and Casacuberta, F. (2014). Improving on-line handwritten recognition in interactive machine translation. *Pattern Recognition*, 47(3):1217–1228.

Apostolico, A. and Guerra, C. (1987). The longest common subsequence problem revisited. *Algorithmica*, 2:315–336.

Azadi, F. and Khadivi, S. (2015). Improved search strategy for interactive machine translation in computer-asisted translation. In *Proceedings of Machine Translation Summit XV*, pages 319–332.

Bahar, P., Brix, C., and Ney, H. (2018). Towards two-dimensional sequence to sequence model in neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. *Postgraduate conference in corpus linguistics.*

Baron, A., Rayson, P., and Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35:3–28.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors (2020). *Proceedings of the Fifth Conference on Machine Translation.*

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transaction of the Royal Society of London*, pages 370–418.

Biçici, E. and Yuret, D. (2015). Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(2):339–350.

Bollmann, M. (2018). *Normalization of Historical Texts with Neural Network Models.* PhD thesis, Sprachwissenschaftliches Institut, Ruhr-Universität.

Bollmann, M. and Søgaard, A. (2016). Improving historical spelling normalization with bi-directional lstms and multi-task learning. In *Proceedings of the International Conference on the Computational Linguistics*, pages 131–139.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Casacuberta, P. and Vidal, E. (2007). Learning finite-state models for machine translation. *Machine Learning*, 66:69–91.

Chatterjee, R., Farajian, M. A., Negri, M., Turchi, M., Srivastava, A., and Pal, S. (2017). Multi-source neural automatic post-editing: Fbk's participation in the wmt 2017 ape shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638.

Cheng, S., Huang, S., Chen, H., Dai, X., and Chen, J. (2016). Primt: A pick-revise framework for interactive machine translation. In *Proceedings of the*

North American Chapter of the Association for Computational Linguistics, pages 1240–1249.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Chinea-Rios, M. (2019). *Advanced techniques for domain adaptation in Statistical Machine Translation*. PhD thesis, Universitat Politècnica de València.

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1693–1703.

Costa-Jussà, M. R., Aldón, D., and Fonollosa, J. A. (2017). Chinese–spanish neural machine translation enhanced with character and word bitmap fonts. *Machine Translation*, 31:35–47.

Costa-Jussà, M. R. and Fonollosa, J. A. (2016). Character-based neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 357–361.

Crowther, J. (2003). *No Fear Shakespeare: Hamlet*. SparkNotes.

Daems, J. and Macken, L. (2019). Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33:117–134.

Descartes, R. (1970). *Descartes to Mersenne, 20 November 1629*. Descartes: Philosophical Letters. Oxford Clarendon Press.

Ding, S., Renduchintala, A., and Duh, K. (2019). A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of the Machine Translation Summit*, pages 204–213.

Domingo, M. and Casacuberta, F. (2018). Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 129–137.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Ernst-Gerlach, A. and Fuhr, N. (2006). Generating search term variants for text collections with historic spellings. In *European Conference on Information Retrieval*, pages 49–60.

F. Jehle, F. (2001). *Works of Miguel de Cervantes in Old- and Modern-spelling.* Indiana University Purdue University Fort Wayne.

Federico, M., Bentivogli, L., Paul, M., and Stüker, S. (2011). Overview of the IWSLT 2011 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 11–27.

Fiebranz, R., Lindberg, E., Lindström, J., and Ågren, M. (2011). Making verbs count: the research project 'gender and work'and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Flood, A. (2015). Modern version of Don Quixote declared 'crime against literature'. https://www.theguardian.com/books/2015/aug/19/modern-version-of-don-quixote-declared-against-literature.

Foster, G., Isabelle, P., and Plamondon, P. (1997). Target-text mediated interactive machine translation. *Machine Translation*, 12:175–194.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proceedings of the Association for Computational Linguistics Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the International Conference on Machine Learning*, page 1243–1252.

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 228–235.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471.

Given, M. D. (2015). A discussion of bible translations and biblical scholarship. `http://courses.missouristate.edu/markgiven/rel102/bt.htm`.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2010). On the use of confidence measures within an interactive-predictive machine translation system. In *Proceedings of the Annual Conference of the European Association for Machine Translation.*

Gupta, K. K., Haque, R., Ekbal, A., Bhattacharyya, P., and Way, A. (2020). Syntax-informed interactive neural machine translation. In *International Joint Conference on Neural Networks*, pages 1–8.

Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., and Mäkelä, E. (2018). Normalizing early english letters to present-day english spelling. In *Proceedings of the Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 87–96.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hundt, M., Denison, D., and Schneider, G. (2011). Retrieving relatives from historical data. *Literary and Linguistic Computing*, 27(1):3–16.

Hutchings, J. (2004). Two precursors of machine translation: Artsrouni and trojanskij. *International Journal of Translation*, 16(1):11–31.

Ilonka, K. (2018). Is it possible for machines to translate poetry, when humans can barely do it? https://electricliterature.com/is-it-possible-for-machines-to-translate-poetry-when-humans-can-barely-do-it/.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition.* MIT Press.

Jordan, M. I. (1990). *Attractor Dynamics and Parallelism in a Connectionist Sequential Machine*, page 112–127. IEEE Press.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Kay, M. (1973). Automatic translation of natural languages. *Daedalus*, 1602(3):217–230.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the Association for Computational Linguistics: System Demonstration*, pages 67–72.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational linguistics*, 25(4):607–615.

Knowles, R. and Koehn, P. (2016). Neural interactive translation prediction. In *Proceedings of the Association for Machine Translation in the Americas*, pages 107–120.

Knowles, R. and Koehn, P. (2018). Lightweight word-level confidence estimation for neural interactive translation prediction. In *Proceedings of the AMTA Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 35–40.

Knowles, R., Sanchez-Torron, M., and Koehn, P. (2019). A user study of neural interactive translation prediction. *Machine Translation*, 33:135–154.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit*, pages 79–86.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P., Tsoukala, C., and Saint-Amand, H. (2014). Refinements to interactive translation prediction based on search graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 574–578.

Korchagina, N. (2017). Normalizing medieval german texts: from rules to deep learning. In *Proceedings of the Nordic Conference on Computational Linguistics Workshop on Processing Historical Language*, pages 12–17.

Kreutzer, J. and Riezler, S. (2019). Self-regulated interactive sequence-to-sequence learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 303–315.

Laing, M. (1993). The linguistic analysis of medieval vernacular texts: Two projects at edinburgh'. In *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, edited by M. Rissanen, M. Kytd, and S. Wright. St Catharine's College Cambridge*, volume 25427, pages 121–141.

Lam, T. K., Schamoni, S., and Riezler, S. (2019). Interactive-predictive neural machine translation through reinforcement and imitation. In *Proceedings of Machine Translation Summit*, pages 96–106.

Lin, H., Liu, L., Huang, G., and Shi, S. (2021). GWLAN: General word-level autocompletion for computer-aided translation. In *Proceedings of theJoint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing.* In Press.

Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015). Character-based neural machine translation. *arXiv preprint arXiv:1511.04586.*

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the International Conference on Language Resources Association*, pages 923–929.

Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Dataset of normalised slovene text KonvNormSl 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1068.

Ljubešic, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Normalising slovene data: historical texts vs. user-generated content. In *Proceedings of the Conference on Natural Language Processing*, pages 146–155.

Lowerre, B. T. (1976). *The Harpy Speech Recognition System.* PhD thesis, Carnegie Mellon University.

Marie, B. and Max, A. (2015). Touch-based pre-post-editing of machine translation output. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1040–1045.

Monk, C. (2018). Custumale Roffense: An overview of the thirteenth-century custumal of Rochester Cathedral priory. https://www.themedievalmonk.com/the-rochester-customs-book.html.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 301–305.

Navarro, A. and Casacuberta, F. (2021a). Confidence measures for interactive neural machine translation. In *Proceedings of the IberSPEECH conference*, pages 195–199.

Navarro, A. and Casacuberta, F. (2021b). Introducing mouse actions into interactive-predictive neural machine translation. In *Proceedings of the Machine Translation Summit*. In press.

Nepveu, L., Lapalme, G., Langlais, P., and Foster, G. (2004). Adaptive language and translation models for interactive machine translation. In *Proceedings of the Conference on Empirical Method in Natural Language Processing*, pages 190–197.

Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc.

Och, F. J. (2002). *Statistical machine translation: from single-word models to alignment templates*. PhD thesis, RWTH Aachen.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Ortiz-Martínez, D. (2011). *Advances in fully-automatic and interactive phrase-based statistical machine translation*. PhD thesis, Universitat Politècnica de València.

Ortiz-Martínez, D. (2016). Online learning for statistical machine translation. *Computational Linguistics*, 42(1):121–161.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2013). How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.

Peng, X., Zheng, Y., Lin, C., and Siddharthan, A. (2021). Summarising historical text in modern languages. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 3123–3142.

Peris, Á. (2019). *Interactivity, Adaptation and Multimodality in Neural Sequence-to-sequence Learning*. PhD thesis, Universitat Politècnica de València.

Peris, A. and Casacuberta, F. (2018). NMT-Keras: a Very Flexible Toolkit with a Focus on Interactive NMT and Online Learning. *The Prague Bulletin of Mathematical Linguistics*, 111:113–124.

Peris, Á. and Casacuberta, F. (2019). Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language*, 58:98–126.

Peris, Á., Domingo, M., and Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Pettersson, E., Megyesi, B., and Tiedemann, J. (2013). An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics*, pages 54–69.

Poncelas, A., Shterionov, D., Way, A., Maillette de Buy Wenniger, G., and Passban, P. (2018). Investigation backtranslation in neural machine translation. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 249–258.

Porta, J., Sancho, J.-L., and Gómez, J. (2013). Edit transducers for spelling variation in old spanish. In *Proceedings of the workshop on computational historical linguistics*, pages 70–79.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191.

Rajvanshi, T. (2015). The poetics of poetry translation. `https://unravellingmag.com/articles/poetics-poetry-translation/`.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Rodríguez Marcos, J. (2015). Un ‘Quijote’ moderno. `https://elpais.com/cultura/2015/05/27/babelia/1432726379_211033.html`.

Rogers, H. J. and Willett, P. (1991). Searching for historical word forms in text databases using spelling-correction methods: Reverse error and phonetic coding methods. *Journal of Documentation*, 47(4):333–353.

Romero, V., Toselli, A. H., Vidal, E., Sánchez, J. A., Alonso, C., and Marqués, L. (2019). Modern vs diplomatic transcripts for historical handwritten text recognition. In *Proceedings of the International Workshop on Pattern Recognition for Cultural Heritage*, pages 103–114.

Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., and Hoang, H. (2008). Improving interactive machine translation via mouse actions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 485–494.

Santy, S., Dandapat, S., Choudhury, M., and Bali, K. (2019). INMT: Interactive neural machine translation prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 103–108.

Scherrer, Y. and Erjavec, T. (2013). Modernizing historical slovene words with character-based smt. In *Proceedings of the Workshop on Balto-Slavic Natural Language Processing*, pages 58–62.

Sen, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P., and Way, A. (2019). Take help from elder brother: Old to modern english nmt with phrase pair feedback. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*. In press.

Sennrich, R. (2016). Neural machine translation: Breaking the performance plateau. Keynote presentation at the Multilingual Europe Technology Alliance Forum.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nădejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 257–286.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Tang, G., Cap, F., Pettersson, E., and Nivre, J. (2018). An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the International Conference on Computational Linguistics*, pages 1320–1331.

Tiedemann, J. (2009a). Character-based PSMT for closely related languages. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 12–19.

Tiedemann, J. (2009b). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248.

Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational linguistics*, 29(4):97–133.

Tjong Kim Sang, E., Bollmann, M., Boschker, R., Casacuberta, F., Dietz, F., Dipper, S., Domingo, M., van der Goot, R., van Koppen, M., Ljubešić, N., Östling, R., Petran, F., Pettersson, E., Scherrer, Y., Schraagen, M., Sevens, L., Tiedemann, J., Vanallemeersch, T., and Zervanou, K. (2017). The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, 7:53–64.

Tomás, J. and Casacuberta, F. (2006). Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the International Conference on Computational Linguistics/Association for Computational Linguistics*, pages 835–841.

Torregrosa, D., Forcada, M. L., and Pérez-Ortiz, J. A. (2014). An open-source web-based tool for resource-agnostic interactive translation prediction. *Prague Bulletin of Mathematical Linguistics*, 102:69–80.

Toselli, A. H., Leiva, L. A., Bordes-Cabrera, I., Hernández-Tornero, C., Bosch, V., and Vidal, E. (2017). Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription. *Digital Scholarship in the Humanities*, 33(1):173–202.

Toselli, A. H., Romero, V., Pastor, M., and Vidal, E. (2010). Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825.

Trapiello, A. (2015). *Don Quijote de la Mancha Puesto en castellano actual íntegra y fielmente por Andrés Trapiello*. Ediciones Destino.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of the Special Interest Group of the Association for Computational Linguistics Workshop on Chinese Language Processing*, pages 168–171.

Ueffing, N. and Ney, H. (2005). Application of word-level confidence measures in interactive statistical machine translation. In *Proceedings of the European Association for Machine Translation*, pages 262–270.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 590–596.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vauquois, B. (1971). Modèles pour la traduction automatique. *Mathématiques et Sciences humaines*, 34:61–70.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the Conference on Computational Linguistics*, volume 2, pages 836–841.

Wang, Y.-Y. (1998). *Grammar inference and statistical machine translation*. PhD thesis, Carnegie Mellon University.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Wuebker, J., Green, S., DeNero, J., Hasan, S., and Luong, M.-T. (2016). Models and inference for prefix-constrained machine translation. In *Proceedings of the Annual Meeting of the Association for the Computational Linguistics*, pages 66–75.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Proceedings of the Annual German Conference on Advances in Artificial Intelligence*, volume 2479, pages 18–32.

Zhao, T., Liu, L., Huang, G., Li, H., Liu, Y., GuiQuan, L., and Shi, S. (2020). Balancing quality and human involvement: An effective approach to interactive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9660–9667.