T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A DEEP LEARNING BASED TRANSLATION SYSTEM FROM OTTOMAN TURKISH TO MODERN TURKISH

ABDULLAH BAKIRCI

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF COMPUTER ENGINEERING

GEBZE

2019

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A DEEP LEARNING BASED TRANSLATION SYSTEM FROM OTTOMAN TURKISH TO MODERN TURKISH

ABDULLAH BAKIRCI

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR

PROF. DR. YUSUF SİNAN AKGÜL

GEBZE

2019

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OSMANLI TÜRKÇESİNDEN MODERN TÜRKÇEYE DERİN ÖĞRENME TABANLI ÇEVİRİ SİSTEMİ

ABDULLAH BAKIRCI
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
PROF. DR. YUSUF SİNAN AKGÜL

GEBZE
2019

GEBZE TEKNİK ÜNİVERSİTESİ	YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 10/07/2019 tarih ve 2019/31 sayılı kararıyla oluşturulan jüri tarafından 23/07/2019 tarihinde tez savunma sınavı yapılan Abdullah Bakırcı'nın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI) : Prof. Dr. Yusuf Sinan AKGÜL

ÜYE : Prof. Dr. Banu DİRİ

ÜYE : Dr. Öğr. Üyesi Burcu YILMAZ

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve ...../..... sayılı kararı.

SUMMARY

Languages develop and change over time. Consequently, old texts may become unintelligible for today's generation. Ottoman Turkish writings are an example. In modern Turkish, the alphabet has changed, some grammatical structures are no longer in use, and some foreign words and phrases are not actively employed anymore. Moreover, rewriting old documents in today's language requires qualified writers, who are responsible for the translation. Unfortunately, that is costly in both time and resources. Despite the importance of this problem, few researchers have worked on it. Their solutions, mainly, focus on building rule-based-systems to transliterate old texts from the Arabic alphabet to the Latin alphabet. While that is essential, further work needs to be done.

This thesis approaches the problem of translating and transliterating Ottoman Turkish to modern Turkish using Neural Machine Translation techniques. Our contributions are on three folds. First, we develop three parallel corpora; the first one from Ottoman Turkish to old Turkish written with Latin characters, the second one from old Turkish written with Latin characters to modern Turkish and the third one from Ottoman Turkish to modern Turkish. Second, for the development of Ottoman Turkish to old Turkish parallel corpus, we re-implement a morphology-based transliteration tool in the other way around; from old Turkish to Ottoman Turkish. Finally, as our main contribution, we develop three neural-based end-to-end translation systems using the three parallel corpora mentioned.

Key Words: Natural Language Processing (NLP), Neural Machine Translation (NMT), Ottoman Turkish, Turkish Language.

ÖZET

Diller zamanla değişim ve gelişime uğrar. Sonuç olarak eski metinler günümüz nesil için anlaşılmaz hale gelebilir. Osmanlı Türkçe'sinin yazıları bir örnektir. Modern Türkçe'de alfabenin değişmesiyle birlikte, bazı gramer yapıları, eski kelimeler ve ifadeler artık aktif olarak kullanılmamaktadır. Ayrıca, eski belgelerin bugünkü dilde yeniden yazılması, eski ve yeni dile olabildiğince hakim yazarlar gerektirir. Ancak böyle kişilerin bulunması zaman ve kaynak açısında maliyetlidir. Bu problemin önemine rağmen az sayıda araştırmacı bu konuda çalışmıştır. Çalışmaları, çoğunlukla eski metinleri Arap alfabesinden Latin alfabesine çevirmek için kural tabanlı sistemler oluşturmaya odaklanır. Bu gerekli olsa da, daha fazla çalışma yapılması gerekmektedir.

Bu tez, Sinirsel Makine Çevirisi tekniklerini kullanarak Osmanlı Türkçesini modern Türkçeye çevirme ve tercüme etme problemini ele alıyor. Katkılarımız üç bölümden oluşmaktadır. İlk olarak, Osmanlı Türkçe'sinden Latin karakterleri ile yazılmış eski Türkçe'ye, Latin karakterleri ile yazılmış eski Türkçe'den modern Türkçe'ye ve Osmanlı Türkçe'sinden modern Türkçe'ye üç paralel korpusu geliştiriyoruz. İkincisi, Osmanlı Türkçe'sinden eski Türkçe'ye paralel korpusu geliştirmek için aksi yönde yani eski Türkçe'den Osmanlı Türkçe'sine bir biçimbilim tabanlı çeviri aracı yeniden gerçekleştiriyoruz. Son olarak, ana katkımız olarak bahsi geçen üç paralel korpusu kullanarak sinirsel uçtan uca çeviri sistemi geliştiriyoruz.

Anahtar Kelimeler: Doğal Dil İşleme, Sinirsel Makine Çevirisi, Osmanlı Türkçesi, Türk Dili.

ACKNOWLEDGEMENTS

I want to express my deep and sincere gratitude to my supervisor, Prof. Dr. Yusuf Sinan AKGÜL, who not only shared his profound scientific knowledge with me but also taught me great lessons of life. His support, suggestions, and encouragement gave me the drive and will to complete this work.

I am grateful to my parents, brothers, and sisters, for their love and support, especially to my elder brother Beshr AL NAHAS.

Finally, I want to thank Muhammed Enes ALMAHDI, Ahmet SOYYİĞİT, Murat Salih TUNALI, Ahmed ALHASHIMI, Hadi ALİZADEH, Ahmed ALBAHNASAWI, and Mahmud ALDABABSA for their enduring friendship and support.

TABLE of CONTENTS

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		**Explanations**
NLP	:	Natural Language Processing
NMT	:	Neural Machine Translation
RNN	:	Recurrent Neural Networks
LSTM	:	Long Term Short Term Memory
ReLU	:	Rectified Linear Unit
CNN	:	Convolutional Neural Networks

LIST of FIGURES

LIST of TABLES

1. INTRODUCTION

Political events, accepting a new religion, technological advancement are some factors that affect a language [1]. The Turkish language is an example of a dramatically evolving language [2]. For example, during the Ottoman Empire days, the Turkish language (known as the Ottoman Turkish [2]) used to use a modified version of the Arabic alphabet (see Table 1.1 [4]), imported Arabic and Persian words and phrases (see Table 1.2), plus some grammatical structures from Arabic and Persian [3]. On the other hand, today's Turkish uses a modified version of the Latin alphabet, less imported Arabic and Persian words and has got rid of most exotic grammatical structures [2]. These changes in the Turkish language, among others, made the Ottoman Turkish obscure for the period's generation [2]. Furthermore, what makes it challenging to learn the Ottoman script is the lack of a one-to-one correspondence between the Ottoman alphabet and the modern Turkish alphabet [4].

1.1. Some Complications in The Ottoman Script

In Ottoman Turkish, consonants are always written, but vowels are not. For example, the vowel 'he' (see Table 1.1) (called okutucu in Ottoman Turkish) which is represented with 'e' in the modern alphabet is written in open syllables but omitted in both colsed syllables and open syllables if they appear the first in words. That can be demonstrated with the word 'kelebek' (means butterfly). Its first syllable is open 'ke' but 'he' is omitted because it is the first syllable in the word, the second syllable 'le' is open so 'he' is written and the last syllable 'bek' is closed so 'he' is omitted.

Moreover, a problem that makes some loan Arabic words difficult to read is the omission of vowel points (hareke). For example, the two Arabic words hulm (meaning dream) and hilm (meaning mildness) are both written the same way in the Ottoman script if the vowel points are omitted.

In addition, some letters like 'kaf' can mean four different letters; kâf (kef or k in modern Turkish), kâf-ı Fârisî (gef, gâf or g in modern Turkish), kâf-ı nûnî (nef, nâf or n in modern Turkish), kâf-ı yâyî (kâf-ı vâvî or ğ in modern Turkish). Some other

different letters have all the same pronounciation in Turkish, while differ in their original scripts like ‘zel’, ‘zı’, ‘ze’ and ‘dad’ are all pronounced as ‘z’ in Turkish.

The modern Turkish letters ‘o, ö, u, ü’ are all represented by ‘vav’ in Ottoman Turkish. While, most of the times, one can differentiate between back vowels ‘o, u’ and front vowels ‘ö, ü’ according to the consonants, there is no rule to differentiate between ‘o’ and ‘u’ or between ‘ö’ and ‘ü’.

Loan Arabic and Persian words are written not as they are pronounced in Turkish, but as they are written in their original languages.

1.2. Suffixes in The Ottoman Turkish

Some Ottoman Turkish suffixes do not change between back vowels (kalın ünlü) and front vowels (ince ünlü), others change. However, it is not hard to retrieve their modern Turkish counterparts using Turkish vowel harmony rules (ünlü uyumu). In table 1.3 we list the correspondance between modern Turkish and Ottoman Turkish suffixes.

1.3. Correspondence Between The Ottoman Script and the Modern Turkish Script

As we describe in sections 1.2 and 1.3, translating between Ottoman and modern Turkish is not a straight forward process and is infeasible on the morphological level.

Building a finite-state transducer that act on the sentence level [5] would be an intersting research direction but requires extensive expertise in both the Ottoman Turkish and the modern Turkish linguistics. However, that is unworkable for us.

- Translation or Transliteration:

To make the Ottoman legacy accessible, qualified writers have made significant efforts to translate and transliterate some of the Ottoman Turkish documents into modern Turkish. But what is the difference between translation and transliteration? Tranliteration is transferring a word from one alphabet to another in the same or a

different language. “Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text” [6].

In early years of the new Turkish alphabet, some writers made eforts to transliterate Ottoman texts to the new Turkish alphabet, without changing words or sentence structures. Later on, when old language is not used any more, writers started translating old texts into the modern languge.

Although that is helpful, manual translation/transliteration is not always an affordable solution. Hence, the need for a handy computerized tool.

Table 1.1: The Ottoman Turkish alphabet and each letter's corresponding Latin letter. The Ottoman Turkish is written right to left.

Isolated	Final	Medial	Initial	Name	Modern Turkish
ا	ـا	—		elif	a, e
ء	—			hemze	—, ‘
ب	ـب	ـبـ	بـ	be	b (p)
پ	ـپ	ـپـ	پـ	pe	p
ت	ـت	ـتـ	تـ	te	t
ث	ـث	ـثـ	ثـ	se	s
ج	ـج	ـجـ	جـ	cim	c
چ	ـچ	ـچـ	چـ	çim	ç
ح	ـح	ـحـ	حـ	ha	h
خ	ـخ	ـخـ	خـ	hı	h
د	ـد	—		dal	d
ذ	ـذ	—		zel	z
ر	ـر	—		re	r
ز	ـز	—		ze	z
ژ	ـژ	—		je	j
س	ـس	ـسـ	سـ	sin	s
ش	ـش	ـشـ	شـ	şın	ş
ص	ـص	ـصـ	صـ	sad	s
ض	ـض	ـضـ	ضـ	dad	d, z
ط	ـط	ـطـ	طـ	tı	t
ظ	ـظ	ـظـ	ظـ	zı	z
ع	ـع	ـعـ	عـ	ayn	, —
غ	ـغ	ـغـ	غـ	gayn	g, ğ, v
ف	ـف	ـفـ	فـ	fe	f
ق	ـق	ـقـ	قـ	kaf	k
ك	ـك	ـكـ	كـ	kef	k (g if loanword from Uzbek)
گ	ـگ	ـگـ	گـ	gef (1), kāf-ı fārsī	g, ğ, (v if loanword from Azeri or Turkmen)
ڭ	ـڭ	ـڭـ	ڭـ	nef, ñef, sağır kef (1), kāf-ı nūnī	n
ل	ـل	ـلـ	لـ	lam	l
م	ـم	ـمـ	مـ	mim	m
ن	ـن	ـنـ	نـ	nun	n
و	ـو	—		vav	v, o, ö, u, ü
ه	ـه	ـهـ	هـ	he (3)	h, e, a
ى	ـى	ـيـ	يـ	ye	y, ı, i

Table 1.2: Examples of Ottoman Turkish words that were imported from Arabic and Persian.

Ottoman Word/Phrase	Latin Transliteration	Turkish Meaning	Enghlish Meaning
حسن خط	Hüsn-i hat	Güzel yazı	Calligraphy
قیل و قال	Kîl-u kâl	Dedikodu	Gossip
لسان حال	Lisan-ı hal	Hal dili	State speaking for itself
سعادت ابدیه	Saadet-i ebediye	Ebedi saadet	Eternal happiness
دار الفنون	Daru'l-fünun	Üniversite	University
شهور ثلاثه	Şuhur-u selase	Üç aylar	The three month
منفی	Menfi	Olumsuz	Unfavorable
موت	Mevt	Ölüm	Death
بیهوده	Beyhude	Boşu boşuna	Vain
آواره	Avare	Boş gezen	Wanderer
آهسته	Aheste	Sakin sakin	Slowly

To address the need for a fast and inexpensive automated tool, this thesis proposes three machine transliteration/translation models.

- A neural machine transliteration model to transliterate between Ottoman Turkish and old Turkish written with the Latin characters,
- A neural machine translation model to translate between old Turkish written with the Latin characters and modern Turkish, and
- A neural machine translation model to translate between the Ottoman Turkish and the modern Turkish in and end-to-end manner.

To build these machine translation systems, we collect three parallel corpora; from Ottoman Turkish to old Turkish, from old Turkish to modern Turkish and from Ottoman Turkish to modern Turkish. To collect the data, we scan and OCR an old version and a modern version of a book (Nutuk), which provides us with old to modern Turkish corpus. To obtain the Ottoman version of it, we reimplement a morphology-based tool that transliterates the Ottoman script into the old Turkish script.

After this introduction, we review the literature in chapter 2. We mention previous work about translating between Ottoman Turkish, old Turkish, and modern Turkish. Then we go over sequence-to-sequence models of deep neural networks.

Next, in chapter 3, we describe our dataset. After that, in chapter 4, we talk in detail about neural machine translation.

Subsequently, in chapter 5, we describe our method. Finally, in chapter 6, we conclude the thesis with some possible directions for future work that builds upon and extends ours.

Table 1.3: Correspondence between the modern Turkish and Ottoman Turkish suffixes.

Suffiex name	Modern Turkish	Ottoman
Verbal (isim fiil eki)	-mak, mek	مق، مك
Verbal (isim fiil eki)	-ma, me	مـه
Verbal (isim fiil eki)	-ış, -iş, uş, üş	يش، وش
Verbal adjective (sıfat fiil eki)	-an, -en	ان، ن
Verbal adjective (sıfat fiil eki)	-ası, -esi	اسى، مسى
Verbal adjective (sıfat fiil eki)	-maz, -mez	مز، ماز
Verbal adjective (sıfat fiil eki)	-r, ar, er, ır, ir, ur, ür	ر، ار، ير، ور
Verbal adjective (sıfat fiil eki)	-dık, -dik, -duk, -dük	دق، دك
Verbal adjective (sıfat fiil eki)	-acak, -ecek	اجق، هجك
Verbal adjective (sıfat fiil eki)	-mış, -miş, muş, müş	مش
Verbal adverb (zarf fiil)	-arak, -erek	ارق، هرك
Verbal adverb (zarf fiil)	-ıp, -ip, -up, -üp	وب
Verbal adverb (zarf fiil)	-dıkça, -dikçe, -dukça, -dükçce	دقجه، دكجه

Table 1.3: Continuation.

Verbal adverb (zarf fiil)	-alı, -eli	الی، ەلی
Verbal adverb (zarf fiil)	-ınca, -ince, unca, ünce	نجه
Verbal adverb (zarf fiil)	-maktansa, -mektense	مقدنسه، مکدنسه
Verbal adverb (zarf fiil)	-ken	کن
Verbal adverb (zarf fiil)	-madan, -meden	مدن
Verbal adverb (zarf fiil)	-maksızın, -meksizin	مقسزین، مکسزین
Verbal adverb (zarf fiil)	-dığında, -diğinde, -duğunda, -düğünde, -tığında, -tiğinde, -tuğunda, -tüğünde	دیغنده، دیگنده
Denominal noun (isimden isim yapım eki)	-lık, -luk, -lik, -lük	لق، لك
Denominal noun (isimden isim yapım eki)	-cı, -ci, -cu, -cü, -çı, -çi, -çu, -çü,	جی
Denominal noun (isimden isim yapım eki)	-lı, li, -lu, -lü	لی
Denominal noun (isimden isim yapım eki)	-sız, -siz, -suz, -süz	سز
Denominal noun (isimden isim yapım eki)	-cık, -cik, -cuk, -cük	جق، جك
Denominal noun (isimden isim yapım eki)	-daş, -deş, -taş, -teş	داش
Denominal verb (isimden fiil yapım eki)	-la, -le	لا، له
Denominal verb (isimden fiil yapım eki)	-lan, -len	لان، لن
Denominal verb (isimden fiil yapım eki)	-laş, -leş	لاش، لش
Verb-verb derivation (fiilden fiil yapım eki)	-ır, -ir, -ur, -ür, -dır, -dir, -dur, -dür, -tır, -tir, -tur, -tür	یر، دیر
Verb-verb derivation (fiilden fiil yapım eki)	-t	ت

Table 1.3: Continuation.

Verb-verb derivation (fiilden fiil yapım eki)	-l, -n	ل، ن
Verb-verb derivation (fiilden fiil yapım eki)	-ş	ش
Verb conjugation (fiil çekim eki)	-yor	يور
Verb conjugation (fiil çekim eki)	-makta, -mekte	مقده، مكده
Verb conjugation (fiil çekim eki)	-dı, -di, -du, -dü, -tı, -ti, -tu, -tü	دي
Verb conjugation (fiil çekim eki)	-acak, -ecek	اجق، ەجك
Imperative mood (emir kip eki)	-sın, -sin, -sun, -sün	سينـ، سون
Imperative mood (emir kip eki)	-ın, -in, -un, -ün, -yın, -yin, -yun, -yün, -ınız, -iniz, -unuz, -ünüz, -yınız, -yiniz, -yunuz, -yünüz	ڭ، يڭ، يڭز
Imperative mood (emir kip eki)	-sınlar, -sinler, -sunlar, -sünler	سينلر، سونلر
Conditional (şart eki)	-sa, -se	سـه
Optative mood (istek eki)	-a, -e	ا، ه
Necessitive mood (gerklilik kip eki)	-malı, -meli	ملى
Adequacy mood (yeterlik kip eki)	-abil, -ebil	ابيل، ەبيل
Genitive case (Tamlayan eki)	-ın, -in, -un, -ün, -nın, -nin, -nun, -nün	نڭ، ڭ
Accusative case (Belirtme hal eki)	-ı, -i, -u, -ü	ى

Table 1.3: Continuation.

Dative mood (Yönelme hal eki)	-a, -e	ه
Locative (Bulunma hal eki)	-da, -de, -ta, -te	ده
Ablative (Ayrılma hal eki)	-dan, -den, -tan, -ten	دن
Instrumental case (Vasıta hal eki)	-la, -le, -yla, -yle	لـه، يله
Possessive (iyelik eki)	-m, -ım, -im, -um, -üm	م
Possessive (iyelik eki)	-n, -ın, -in, -un, -ün	ڭ
Possessive (iyelik eki)	-ı, -i, -u, -ü	ى
Possessive (iyelik eki)	-mız, -miz, -muz, -müz	مز
Possessive (iyelik eki)	-nız, -niz, -nuz, -nüz	ڭز
Possessive (iyelik eki)	-ları, -leri	لرى
Possessive (iyelik eki)	-ki	كي
Conjunction	-ki	كـه
Plural (çoğul eki)	-lar, -ler	لر
Interrogative (soru eki)	-mı, -mi, -mu, -mü	مى

2. LITERATURE REVIEW

In this chapter, we mention related work that is grouped into three categories; translating and transliterating between Ottoman Turkish and modern Turkish, transliterating between two different scripts of the same language and neural machine translation. We want to note that works on transliterating from the Latin script of Turkish to the Ottoman script are also related, but we could only find this [7] which does not mention any detail.

2.1. Translating/Transliterating from Ottoman Turkish to Modern Turkish

Scientific research in this area is somewhat limited [8] [9] [10] [11], and mainly depends on morphological analysis of the input Ottoman text.

In [8] [9] researchers describe a six-step Ottoman Turkish to modern Turkish transliteration system. Their proposed framework starts by morphological analysis of each word. The first step results in producing possible stem/suffix pairs. Second, their framework looks up each stem/suffix from a dictionary. Third, it synthesizes each word in with the Latin script. Fourth, it proceeds with word disambiguation, where it chooses the most probable words using an n-gram language model. Fifth, it deals with errors caused by previous steps. Those errors could be typographical, wrong segmentation of Ottoman words and unknown words. Finally, it detects noun adjuncts.

In [10], researchers build a word-based analysis tool of Ottoman Turkish that works as follows. First, it takes an Ottoman word as input and checks if it exists in an exception list. If yes, it immediately outputs the Latin mapping of it. If not, it morphologically analyses the word and generates possible stem/suffix pairs. Finally, it maps each Ottoman stem/suffix pair to its Latin counterpart and outputs the most common one of them.

In [11], researchers do very similar work to [8] [9] but without ambiguity handling. When there are multiple outputs of a word, they output them all and tell the user to choose between them.

2.2. Transliterating between two different scripts of the same language

Some languages, other than Turkish, use more than one writing system. For example, some Arabic speakers romanize their written communications in social media and messaging applications and mix them with some English words [12] [13]. The romanized version of the Arabic they use is nonstandard and is called Arabizi. The motivation to transliterate from Arabizi to Arabic is to be able to use the already existing Arabic NLP tools. In [12], researchers build the Arabizi-Arabic transliteration system as following. First, they use Moses [14] to build a statistical phrase-based machine translation model. They treated words as sentences and letters as words. After generating candidate translations of a word, they check if it exists in a large monolingual corpus. Having multiple possible word-to-word translations, they pick the most suitable translation using a trigram language model.

In [13], researchers first train a character-level finite state transducer that generates all possible transliterations of an input Arabizi word. Then, they filter the generated list using a morphological analyzer. Finally, they choose the most appropriate words using a language model.

In [15], researchers design a neural machine transliteration system between the English alphabet and the Vietnamese alphabet. Their purpose is to transliterate English named entities into the Vietnamese alphabet. Their work has two stages. First, they prepare a dataset by segmenting words, then looking them up from a pronunciation dictionary and aligning them on character level, so they have parallel data. Then, they train an RNN-based machine transliteration model using the collected data.

2.3. Neural Machine Translation

Neural machine translation (NMT) is an end-to-end transduction of a source sequence into a target sequence, using neural networks [16] [17]. NMT is used in the state of the art machine translation systems [18] [19].

2.3.1. The Rise of NMT

NMT using the encoder-decoder framework in an end-to-end manner without using any other helper system started from [16], in which the authors proposed to use an RNN encoder to get a sentence representation of the input source sentence, then intialize another RNN, a decoder, with the source sentence vector to start an auto-regressive decoding of the target sentence. They acheived a BLEU score of 34.54 on WMT'14 French to English task. Then in [17] as a development of the work proposed in [16], the autors replaced the RNNs with deep LSTMs and the term NMT was coined. They acheived a BLEU score of 34.81 on the WMT'14 task. The next big advancment in NMT was the proposal of attention mechanisim [20], in which the authors propse to get a different vector representation of the input sentence according to thus far decoded sentence. Their method got 36.15 BLEU score points on WMT'14 task.

2.3.2. Morphology-Aware NMT

To address the out of vocabulary (OOV) problem, a group of scientists suggested building a word-unit-aware NMT [21]. They extract sub-word units from the training corpus using using byte pair encoding (BPE) algorithm. Then, during the test time, they chunck the input words into sub-words using the codes of the BPE model. Their work improve the perfomance of English → German and Enghlish → Russian translation tasks by 1.1 and 1.3 BLEU, respectively. Subsequently, another group of researchers [22] take the work in [21] to the next level, where they do not segment the input sentence into words nor sub-words, instead they feed it as characters to the encoder. Also, the decoder acts on the character level too. There method acheives new state of the art, at the time, on German → English and Czech → Enghlish translation tasks.

2.3.3. NMT for Low-Resource Settings

Another active area of research in NMT is developing NMT models for low-resource language pairs. In [23], researchers fuse a pre-trained recurrent neural network language model (RNNLM) trained on a large monolingual corpus of the target languge into their pre-trained NMT model. They obtain an improvment of 1.96 BLEU points over a base-line in Turkish-English language pair. In [24], authors propose what they call 'back translation' to create synthetic parallel corpora. They bootstrap the process by training an NMT model using the existing parallel corpus. Then starting from monolingual data in language l_1, they translate it to language l_2 using the NMT model they have. After that, they use the model output, which is in l_2, as the source and the original monolingual data, which is in l_1, as the target. Finally, they mix the synthetic data with the human generated data and train a new model. They get BLEU gains in low-resource settings as well as in high-resource settings.

2.3.4. Zero-shot NMT

Zero-shot translation is translating between two languagse during test time, that were not seen together in training time [25]. It first appeared in [26], where the researchers trained an NMT model on multiple language pairs using one shared encoder and one shared decoder for all language pairs. They did not introduce any modifications to the standard NMT architecture but instead introduced an artificial start of sentence token for each language, so the decoder knows the target language. In [25], authors trait the problem of zero-shot translation as a domain adaptation problem [27] [28], where they consider the English language as the source domain and every other language as the target domain. Their approach considerably enhances zero-shot translation quality without declining supervised directions.

2.3.5. Unsupervised NMT

Inspired by the advancements in low-resource and zero-shot NMT, researchers [29] [30] developped NMT models without using any parallel copora. They basically employ three techniques to acheive that; back translation [24], denoising auto-encoding [31] and adverserial domain adaptation [27].

2.3.6. NMT Architectures

Another active area of research in NMT is developing new encoder/decoder architectures while respecting the sequence-to-sequence framework. It started by usign improved RNNs over the vanila RNN. For example, in [17] the researchers propose to use deep LSTMs as the encoder and the decoder. In [32], the authors propose the GRU recurrent network. use residual LSTMs.

An important architecture advancement of NMT is the attention mechanism [20]. Instead of having a fixed length vector representation of the source sentence, the attention mechanism creates a context-based vector representation of the source sentence.

Beyond using any type of RNNs, in [33] the authors propose to use a convolutional neural networks-based sequence-to-sequence model. Further more, in [34], authors invent the Transformer network, which is a non-recurrent neural netowrk that relies on the attention mechanism and feed forward neural networks.

3. DATASET

In this chapter, we describe the preparation of the datasets we use to train our different translation and transliteration models.

3.1. Overview

We need to collect parallel corpora for the following translation directions:

- Ottoman Turkish → Old Turkish,
- Old Turkish → Modern Turkish, and
- Ottoman Turkish → Modern Turkish.

For that purpose, we use a book titled Nutuk, for which we find an Ottoman Turkish, old Turkish and modern Turkish versions. It was originally written with Ottoman Turkish. Then, after the adoption of the Latin-based Turkish alphabet, it was re-written with it. Moreover, because the change of the Turkish language over the years, old Turkish became cryptic. Consequently, qualified writers who are good at both old and modern Turkish re-wrote Nutuk again in nowadays Turkish.

For (Old Turkish → Modern Turkish) dataset, we simply scan and OCR the old and the modern versions of Nutuk to get a parallel corpus. Unfortunately, however, we could not find an accurate OCR tool for Ottoman Turkish documents. Therefore, another work around is needed.

To overcome this, we re-implement an old Turkish-Ottoman Turkish morphology-based transliteration tool. Then, we use the mentioned transliteration tool to transliterate the old Turkish version of Nutuk to Ottoman Turkish. As a result, we get Ottoman Turkish → Old Turkish and Ottoman Turkish → Modern Turkish parallel corpora. See the details in the following subsections.

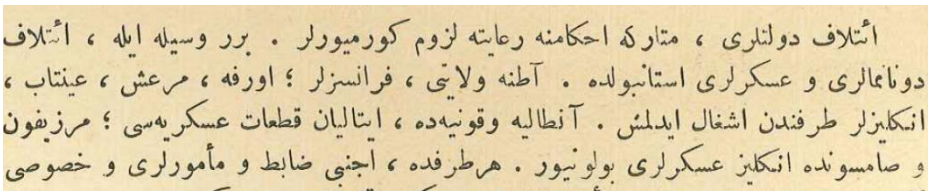
3.2. Nutuk

Nutuk, meaning the speech, is a book compilation of the speeches delivered by the founder of The Republic of Turkey, Mustafa Kemal Atatürk, between the 15th and 20th of October 1927. It talked about the political and the marital situation of the last days of the Ottoman Empire, the Turkish War of Independence and the foundation of the Republic of Turkey. It was delivered at the second congress of Cumhuriyet Halk Partisi; the political party founded by Mustafa Kemal Atatürk.

Nutuk was originally written with Ottoman Turkish. Then, after the adoption of the new Latin-based Turkish alphabet, qualified writers re-wrote it using the new Turkish alphabet. Since the foundation of the Republic of Turkey, the Turkish language has changed dramatically. Consequently, old texts, even those written with Latin alphabet, became cryptic to the period's generation. To make Nutuk more comprehensible, writers who are good at both the old and the modern Turkish re-wrote it again with today's language. Accordingly, there are three versions of Nutuk (see Table 3.1):

- The original, written with Ottoman Turkish,
- A Latin transliteration of the original one, and
- A translation of the original one, written with today's Language.

Table 3.1: Screenshots from Ottoman Nututk, old Nutuk and modern Nutuk.


İtilâf Devletleri, mütareke ahkâmına riayete lüzum görmüyorlar. Birer vesile ile, İtilâf donanmaları ve askerleri İstanbulda. Adana vilâyeti, Fransızlar; Urfa, Maraş, Ayıntap, İngilizler tarafından işgal edilmiş. Antalya ve Konyada, İtalyan kıtaatı askeriyesi; Merzifon ve Samsunda İngiliz askerleri bulunuyor. Her
İtilâf Devletleri, Ateşkes Anlaşmasının hükümlerine uymayı gerekli bulmuyorlar. Birer bahane ile İtilâf donanmaları ve askerleri İstanbul'da, Adana ili Fransızlar; Urfa, Maraş, Gaziantep İngilizler tarafından işgal edilmiş, Antalya ve Konya'da İtalyan askerî birlikleri, Merzifon ve Samsun'da İngiliz askerleri bulunuyor. Her tarafta yabancı subay ve

In Table 3.1, we put screenshots from the different versions of Nutuk. The first row shows a sentence from the original Nutuk in Ottoman Turkish, the second row shows the transliteration of the first row and the third row is the first row's translation into today's Turkish.

3.3. An Old Turkish to Modern Turkish Corpus

For our old Turkish to modern Turkish translation system, we scan and OCR the old and the modern versions of Nutuk. Therefore, we obtain a parallel corpus between the old and modern Turkish.

Table 3.2: Some statistics on our old Turkish – modern Turkish parallel corpus.

	Split	No. Tokens	No. Unique Tokens	No. Sentences	Avg. Sentence Len.
Old	Train	364,493	29,868	29,670	12.28
New	Train	353,984	31,481	29,670	11.93
Old	Validation	3,767	1,991	305	12.35
New	Validation	3,674	1,947	305	12.04
Old	Test	3,753	1,991	312	12.02
New	Test	3,638	1,888	312	11.66

3.4. Monolingual Corpus to Pre-train Word Embeddings

We use a monolingual corpus that is a mix between old Turkish written with Latin characters and modern Turkish. We use this monolingual corpus for pre-training word embedding models that we utilize to initialize the lookup tables of our NMT translation models (refer to chapter 4 for details).

It consists of texts from old and modern Turkish, in addition to texts from the period in between. More specifically we use the Proceedings of the Turkish Parliament between 1923 and 2000.

The overall monolingual corpus consists of 74,973,993 tokens, 1,339,187 of which are unique, 7,175,443 sentences, with an average sentence length of 10.44 tokens.

3.5. Ottoman Turkish to Modern Turkish Dataset

To collect such a dataset, we evaluate three options.

i) Getting a qualified Ottoman reader-writer to transcribe the original Ottoman Nutuk,
ii) To use an optical character recognition (OCR) tool to convert a scanned version of the Ottoman Nutuk to written text document,
iii) To use a tool to transliterate Turkish written with the modern script to the Ottoman script.

The first option is expensive in resources, so we decide to ignore it. For the second one, we try some Arabic and Persian OCR tools; however, they produced very noisy output. That is because such tools' implementations depend on a language model of the target language. Therefore, we decide to go with the third option and re-implement an accurate modern Turkish-Ottoman transliteration tool [35].

We employ our implementation of this tool to transliterate both the old version of Nutuk and the monolingual corpus. As a result, we end up having three versions of Nutuk; Ottoman Turkish Nutuk, old Turkish Nutuk written with Latin characters, and modern Turkish Nutuk. In addition to that, we have an Ottomanized version of the monolingual corpus that we describe in 3.4.

3.5.1. A Morphology-based Old/Modern Turkish to Ottoman Turkish Transliterating Tool

The transliteration from the Ottoman Turkish alphabet to the modern Turkish alphabet is complex. Especially if we try to do it using only morphological analysis. However, the other way around (i.e., from the modern Turkish alphabet to Ottoman alphabet) is more straightforward. Consequently, we re-implement a modern Turkish-Ottoman Turkish transliteration tool [35]. Then we use it to build a parallel corpus to train an NMT model from Ottoman to modern Turkish. We show the old/modern Turkish – Ottoman Turkish transliteration algorithm in Figure 3.1.

```
1: procedure OTTOMANIZE(inputSentence, wordDictionary, suffixDictionary)
2:     sentenceWords ← splitIntoWords(inputSentence)
3:     for word in sentenceWords do
4:         wordIsFound ←FALSE
5:         suffix ←""
6:         while NOT wordIsFound or word ≠ "" do
7:             suffix ← suffix + word.lastChar()
8:             word ← word.removeLastChar()
9:             if word IN dictionary then
10:                wordIsFound ← TRUE
11:                currentWordInOttoman ← wordDictionary.lookup(word)+
   suffixDictionary.lookup(suffix)
12:            end if
13:            if wordIsFound then
14:                ottomanSentence ← ottomanSentence +
   currentWordInOttoman
15:            else
16:                return ""                ▷ Cannot transliterate this sentence.
17:            end if
18:        end while
19:    end for
20:    return ottomanSentence
21: end procedure
```

Figure 3.1: A Morphology-based Old/Modern Turkish to Ottoman Turkish Transliteration Algorithm.

3.5.1.1. A Test Corpus

According to the author of old/modern Turkish - Ottoman Turkish transliterating tool, its accuracy is between 90 - 95%, however, we test it again using a corpus we manually prepare. We manually transcribe 70 sentences from the original scanned Ottoman Nutuk. We get an accuracy of **85%** calculated as the number of correctly transliterated words over the size of the corpus.

In Table 3.3 we show some outputs of the transliteration tool, their original Latin text and their human transcription.

Table 3.3: Latin-written Turkish to Ottoman Turkish transliteration tool sample outputs.

Original Text	1335 senesi Mayısının 19 uncu günü Samsuna çıktım .
Human Transliteration	۱۳۳۵ سنەسی مایسنڭ ۱۹ اونجی گونی صامسونه چیقدم.
Tool Output	۱۳۳۵ سنەسی مایسنڭ ۱۹ اونجی گونی صامسونه چیقدم.
Human Rating	5/5
Original Text	Vaziyet ve manzarai umumiye: Osmanlı Devletinin dahil bulunduğu grup , Harbi Umumîde mağlûp olmuş , Osmanlı ordusu her tarafta zedelenmiş , şeraiti ağır , bir mütarekename imzalanmış .
Human Transliteration	وضعیت و منظرۂ عمومیه : عثمانلی دولتنڭ داخل بولوندیغی غروپ ، حرب عمومیده مغلوب اولمش ، عثمانلی اردوسی هر طرفده زدەلنمش ، شرائطی آغیر ، بر متارکەنامه امضالانمش.
Tool Output	وضعیت و منظرۂ عمومیه : عثمانلی دولتنڭ داخل بولوندیغی غروپ ، حربی عمومیده مغلوب اولمش ، عثمانلی اردوسی هر طرفده زدەلنمش ، شرائطی آغیر ، بر متارکەنامه امضالانمش.
Human Rating	4.7/5
Original Text	Büyük Harbin uzun seneleri zarfında , millet yorgun ve fakir bir halde.
Human Transliteration	بیوك حربڭ اوزون سنەلری ظرفنده ، ملت یورغون و فقیر بر حالده .
Tool Output	بیوك حربیڭ اوزون سنەلری ظرفنده ، ملت یورغون و فقیر بر حالده .
Human Rating	4.8/5
Original Text	Millet ve memleketi Harbi Umumîye sevkedenler , kendi hayatları endişesine düşerek , memleketten firar etmişler .
Human Transliteration	ملت و مملکتی حرب عمومییه سوق ایدنلر ، کندی حیاتلری اندیشەسنه دوشەرك ، مملکتدن فرار ایتمشلر .
Tool Output	ملت و مملکتی حربی عمومیه سوق ایدنلر ، کندی حیاتلری اندیشەسنه دوشرك ، مملکتدن فرار ایتمشلر .
Human Rating	4.5/5
Original Text	İtilâf Devletleri , mütareke ahkâmına riayete lüzum görmüyorlar.
Human Transliteration	ائتلاف دولتلری ، متارکه احکامنه رعایته لزوم گورمیورلر .

Table 3.3: Continuation.

Tool Output	ائتلاف دولتلری ، متارکه احکامینه رعایته لزوم گورمییورلر .
Human Rating	4.8/5
Original Text	Nihayet , mebdei kelâm kabul ettiğimiz tarihten dört gün evvel , 15 Mayıs 1335 de İtilâf Devletlerinin muvafakatile Yunan ordusu İzmire ihraç ediliyor.
Human Transliteration	نهایت ، مبدأ کلام قبول ایتدیگمز تاریخدن درت گون اول ، ۱۵ مایس ۱۳۳۵ ده ائتلاف دولتلرینڭ موافقتیله یونان اردوسی ازمیره اخراج ایدییورلر .
Tool Output	نهایت ، مبدأی کلام قبول ایتدیگمز تاریخدن درت گون اول ، ۱۵ مایس ۱۳۳۵ ده ائتلاف دولتلرینڭ موافقتیله یونان اردوسی ازمیره اخراج ایدیلیور .
Human Rating	4/5
Original Text	Bundan başka , memleketin her tarafında , anasırı hıristiyaniye hafî , celî , hususî emel ve maksatlarının temini istihsaline , devletin bir an evvel , çökmesine sarfı mesai ediyorlar.
Human Transliteration	بوندن بشقه ، مملکتڭ هر طرفنده ، عناصر خرستیانیه خفی ، جلی ، خصوصی أمل و مقصدلرینڭ تأمین استحصالینه ، دولتڭ بر آن اول ، چوکمسنه صرف مساعی ایدییورلر .
Tool Output	بوندن بشقه ، مملکتڭ هر طرفنده ، عناصری خرستیانیه خفی ، جلی ، خصوصی أمل و مقصدلرینڭ تامینی استحصالینه ، دولتڭ بر آن اول ، چوکمسنه صرفی مساعی ایدییورلر .
Human Rating	4/5
Original Text	Bilâhare elde edilen mevsuk malûmat ve vesaik ile teeyyüt etti ki , İstanbul Rum Patrikhanesinde teşekkül eden Mavri Mira heyeti (Vesika:1) , vilâyetler dahilinde çeteler teşkil ve idare etmek , mitingler ve propagandalar yaptırmakla meşgul.
Human Transliteration	بالآخره الده ایدیلن موثوق معلومات و وثایق ایله تأید ایتدی که ، استانبول روم پطریقخانه‌سنده تشکل ایدن ماوری میرا هیئتی (وثیقه : ۱) ، ولایتلر داخلنده چته‌لر تشکیل و اداره ایتمك ، متینغلر و پروپاغندالر یاپدیرمقله مشغول .

Table 3.3: Continuation.

Tool Output	بالآخره ألده ايديلن موثوق معلومات و وثايق ايله تأيد ايتدى كه ، استانبول روم پطريقخانەسنده تشكل ايدن ماورى ميرا هيئتى (وثيقه : ١) ، ولايتلر داخلنده چتەلر تشكيل و اداره ايتمك ، متينغلر و پروپاغندالر ياپديرمقله مشغول .
Human Rating	5/5

4. NEURAL MACHINE TRANSLATION

In this chapter, we describe the neural machine translation framework. Then, we talk about BLEU score; the standard evaluation metric for machine translation systems.

4.1. Sequence-to-Sequence models

A sequence-to-sequence model is a neural network that transduces a variable length input sequence into a variable length output sequence [17]. They are trained to maximize the probability of the output sequence given the input sequence [17]. Their first implementation [16] depends on the recurrent neural network (RNN) [36]. Can be summarized as follows.

- An RNN encoder takes the input sequence and produces a context vector,
- An RNN decoder gets initialized with the context vector from the encoder, then produces the output sequence.

After that, a group of scientists [17] proposed to use a long-short-term memory (LSTM) neural network instead of the RNN. In addition, stacking multiple LSTMs in both the encoder and the decoder.

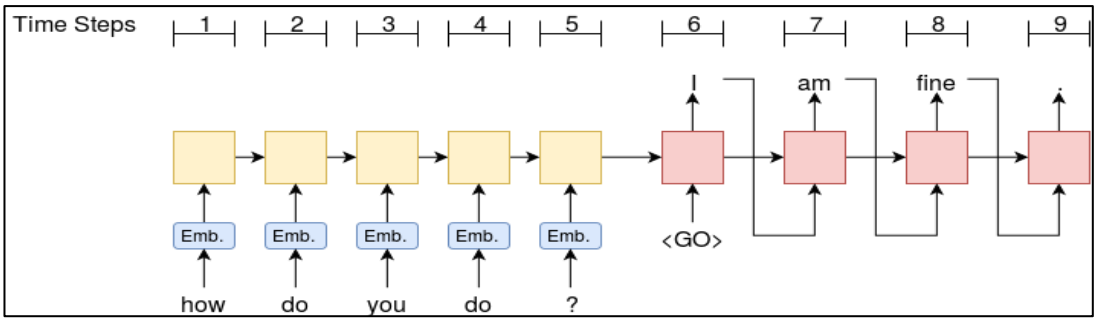


Figure 4.1: An illustration of sequence-to-sequence models.

4.1.1. Overview

Given an input sequence $\boldsymbol{x} = (x_1, x_2, .., x_n)$, each input x_i is mapped to a vector v_{xi} using a word vector embedding lookup table, then the RNN encoder maps $\boldsymbol{v_x} = (v_{x1}, .., v_{xn})$ to another sequence $\boldsymbol{x'}$, and because x'_n is a function of $x'_1, .., x'_{n-1}, x_n$, it could be thought of as a summary of the whole input sequence $\boldsymbol{x}$. Let us call x'_n the context vector. After that, the RNN decoder gets initialized with the context vector and is given a *go token* to trigger it to start generating the output sequence, one at each time step. See Figure 4.1.

4.1.2. Word Vectors

Word vectors are dense, fixed-length and distributed representations of words [37]. In NMT, they are usually called the embedding layer or the lookup table. In scenarios where not enough training data is available, word vectors are pre-trained with monolingual corpus, then used as initial values for the embedding layer [38]. In other scenarios, however, they are frozen during the NMT training iterations [39].

Word vectors produce meaningful results when added or subtracted from each other. For example, vec("Berlin") - vec("Germany") + vec("France") would equal a vector that is most near to vec("Paris").

Some algorithms to train word vectors using monolingual corpora include; skip-gram [37], continuous bag of words [37], fasttext [40], and GloVe [41].

4.1.3. Recurrent Neural Networks

Unlike feed forward neural network, recurrent neural network does not only produce output according to its input at time step t, but it also takes into account the output from a previous step $t-1$. That means can map an input $(x_1, x_2, .., x_T)$ sequence into another sequence $(y_1, y_2, .., y_T)$ by using the following equations iteratively.

$$h_t = sigm(W^{hx} x_t + W^{hh} h_{t-1}) \tag{4.1}$$

$$y_t = W^{yh} h_t \tag{4.2}$$

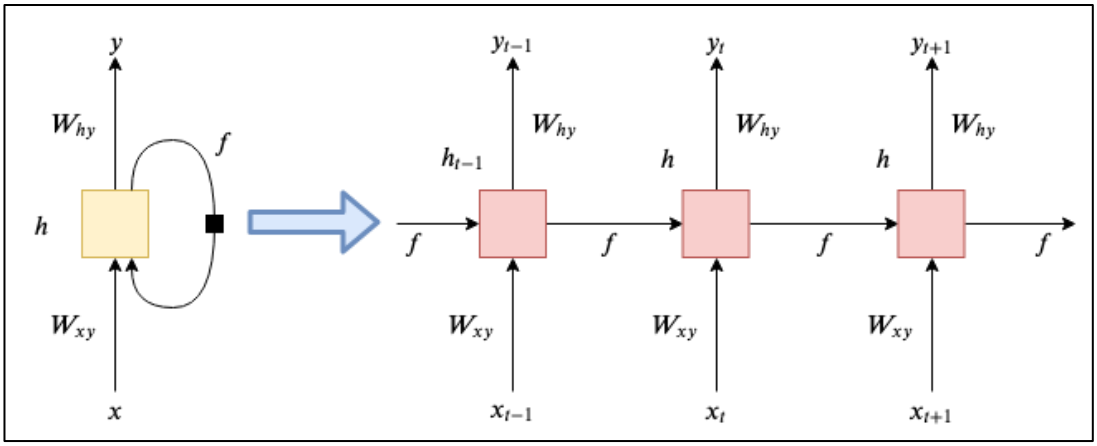


Figure 4.2: An ilustration of the recurrent neural netwok.

4.1.4. Long-Short Term Memory Neural Networks

Although RNNs have the benefit of processing sequences, they are tricky to train on long sequences, because of the vanishing gradient or exploding gradient problems [42]. To overcome these problems, researchers [43] have developed better RNNs. The most famous of them is the long-short-term memory recurrent neural network (LSTM) [44]. The history or memory cell in LSTM is a function of the input, output and forget gates, see Figure 4.3. The input gate decides how much new input need to be included in the memory cell. The forget gate decides what previous memory need to be erased. The output cell decides what information should be output at the current time step. LSTM work by iteratively running the following equations.

$$f_t = \sigma\left(W_{fx} x_t + W_{fh} h_{t-1} + w_{fc} * c_{t-1} + b_f\right) \tag{4.3}$$

$$i_t = \sigma(W_{ix} x_t + W_{ih} h_{t-1} + w_{ic} * c_{t-1} + b_i) \tag{4.4}$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{cx} x_t + W_{ch} h_{t-1} + b_c) \tag{4.5}$$

$$o_t = \sigma(W_{ox} x_t + W_{oh} h_{t-1} + w_{oc} * c_t + b_o) \tag{4.6}$$

$$h_t = o_t * \tanh(c_t) \tag{4.7}$$

where $tanh$ is the hyperbolic tangent function, σ is the logistic function, and $*$ is element-wise multiplication. o, f and i are the output, forget and input gates, c indicates the cell vector, and h is the history state or hidden state vector. The cell vector and all gate vectors must have the same dimensionality as the hidden state vector. Upper case letters represent matrices, lowercase letters denote vectors, and subscripts show the connection (e.g., W_{ox}: input to output gate weight matrix).

4.1.5. Bidirectional RNNs

The unidirectional RNN's history cell at time step t has information only about the past, which makes its processing power limited. To overcome that, [45] propose a bidirectional version of RNN, which is basically two RNN units; one processes the input from right to left, and the other from left to right. The history vector at time step t is the concatenation of the left to right history vector $h_t^{\rightarrow}$and the right to left history vector $h_t^{\leftarrow}$. See Figure 4.3.

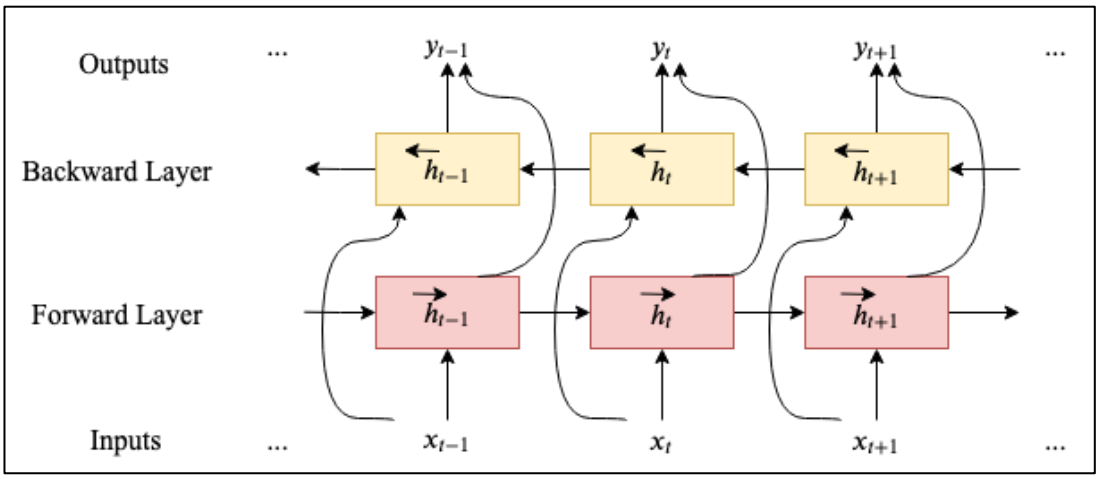


Figure 4.3: An illustruation of the bidirectional RNNs.

4.1.6. Stacked RNNs

In [46], researchers suggest stacking RNNs on top of each other. That is deep RNNs, to increase the representation capability of the RNN, which have proved effective in deep feedforward neural networks. See Figure 4.4.

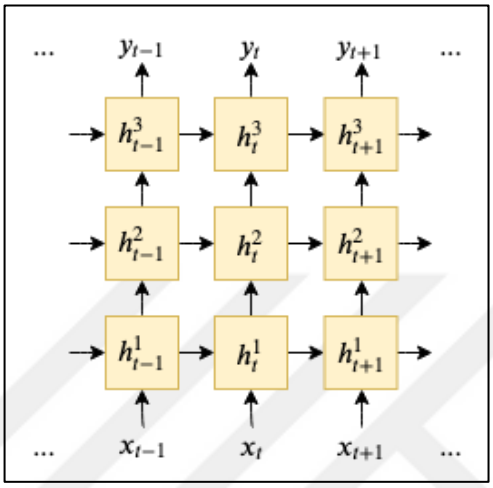


Figure 4.4: An illustration of deep RNN.

4.1.7. Sharing the Decoder's Input and Output Embeddings

In [47], the authors investigate the topmost layer of a neural language model, and find out that it is a valid word embedding layer; let's call it the output embedding layer. By tying the weights of the input and output embedding layers, they observe reduction in the perplexity. We investigate and apply this weight tying for our NMT models. See Figure 4.5.

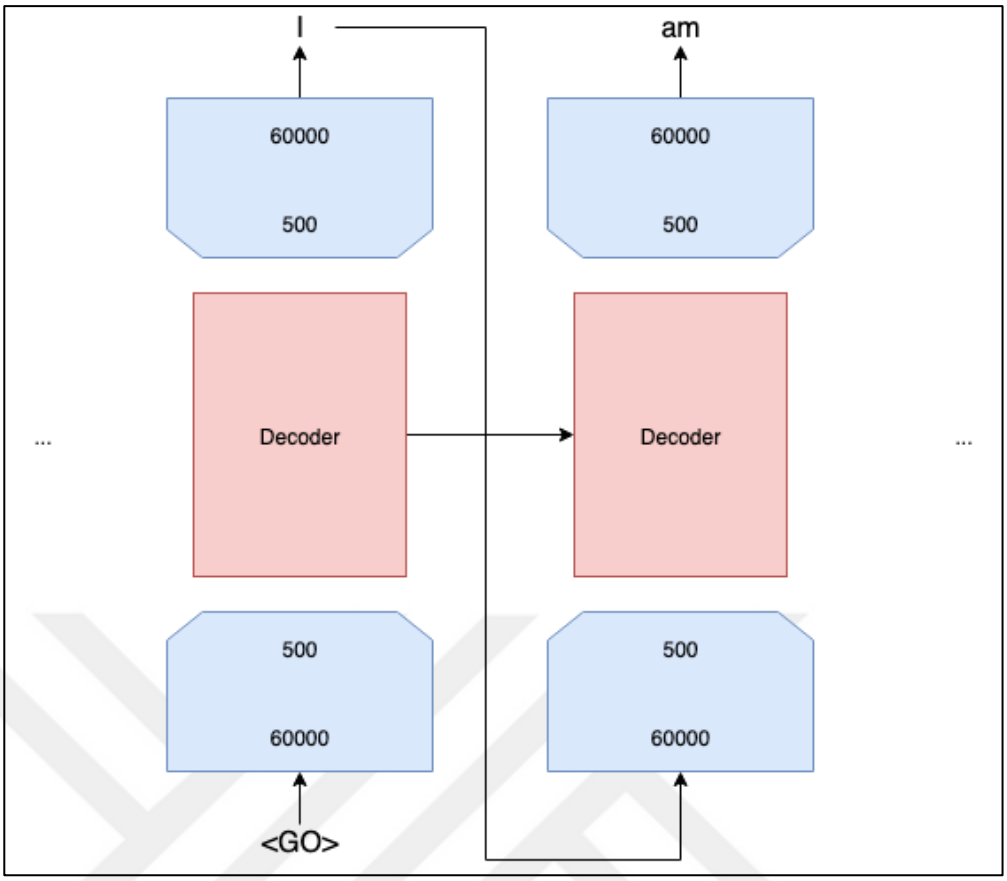


Figure 4.5: An illustration of the decoder's input and output embeddings weight tying.

4.1.8. Attention Mechanism

As we mention in 4.1.1, the decoder is initialized with a context vector. This context vector could be the last hidden state of the encoder [16], the elementwise max of the encoder's hidden states across all time step [48], or instead of using a single sentence representation for the entire input sequence, we could calculate a new context vector every decoding step. This context vector should consider information from the source that is relevant to the current decoding step [49]. That is called the 'attention mechanism'. In the following subsections, we will talk about one conventional model of the attention mechanism. However, usually, attention models have the same outline we mention in the following subsection, and demonstrate in Figure 4.6.

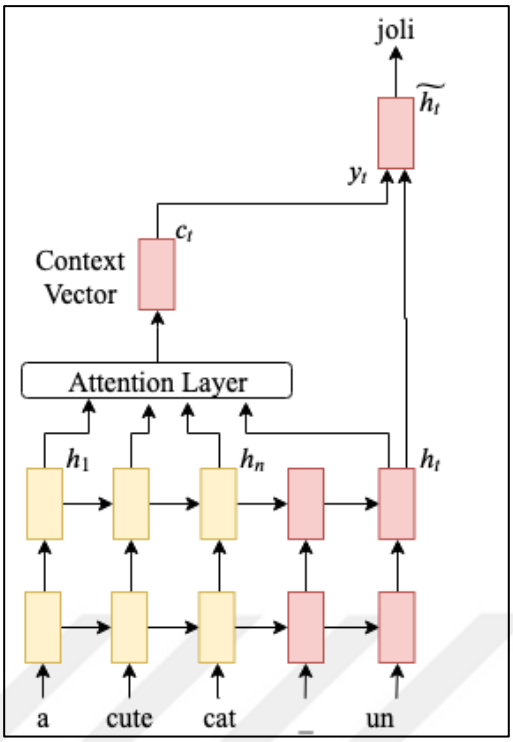


Figure 4.6: Attention mechanism.

4.1.8.1. An Outline

Let us say we use some attention model to get the context vector, denoted by c_t. After that, c_t is concatenated with the target-side hidden vector h_t. Then together, they are passed to a non-linear layer to produce the attentional target hidden state h'_t.

$$h'_t = \tanh(W_c[c_t; h_t]) \tag{4.8}$$

Finally, the attentional hidden state h'_t is passed to a softmax layer to produce the next predictive distribution over the output vocabulary.

$$p(y_t|y_{<t}, x) = softmax(W_s h'_t) \tag{4.9}$$

4.1.8.2. Global Attention

One popular attention model is the global attention model [49]. This model takes into account all of the encoder's hidden states when obtaining the context vector. In other contexts, it is called soft attention. It calculates an attention score (see equation 4.10) for each encoder's hidden state using a function of the current decoder's hidden state and the considered encoder's hidden state. Then, the attention scores are normalized using the softmax function.

$$score(h_t, h_s)\begin{cases} h_t^T h_s & dot \\ h_t^T W_a h_s & general \\ v_a^T \tanh(W_a[h_t; h_s]) & concat \end{cases} \tag{4.10}$$

Finally, having the attention scores as weights, the context vector is calculated as the weighted sum of the encoder's hidden states.

4.2. BLEU As an Evaluation Metric for NMT

BLEU score has been the standard evaluation metric for NMT model. But what is BLEU score? BLEU stands for 'Bilingual Evaluation Understudy' [50]. It is a score for comparing a candidate translation of text to one or more reference translations. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

The score was developed for evaluating the predictions made by automatic machine translation systems. It is not perfect, but does offer 5 compelling benefits:

- It is quick and inexpensive to calculate.
- It is easy to understand.
- It is language independent.
- It correlates highly with human evaluation.
- It has been widely adopted.

The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order. "The primary programming task for a BLEU implementer is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is" [50].

The counting of matching n-grams is modified to ensure that it takes the occurrence of the words in the reference text into account, not rewarding a candidate translation that generates an abundance of reasonable words. This is referred to in the paper as modified n-gram precision.

The score is for comparing sentences, but a modified version that normalizes n-grams by their occurrence is also proposed for better scoring blocks of multiple sentences. A perfect score is not possible in practice as a translation would have to match the reference exactly. This is not even possible by human translators. The number and quality of the references used to calculate the BLEU score means that comparing scores across datasets can be troublesome.

We show the formula for calculating BLEU score in equation 4.11.

$$BLEU = \min\left(1, \frac{output\ length}{reference\ length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}} \qquad (4.11)$$

where $precision_i$ is the precision for n-gram i and the first term is a brevity penalty for short translations.

5. METHOD

We aim to accomplish two goals. The first is to transliterate a text from the Ottoman alphabet to the modern Turkish alphabet. The second is to translate from old Turkish to modern Turkish. We reach both goals by training NMT models.

In the following subsections, we describe our translation and transliteration models. Then report our experiments and results.

5.1. Ottoman Turkish to Old Turkish Transliteration

To achieve this, we first do it the other way around. That is, we transliterate from Turkish written with the Latin alphabet to the Ottoman Turkish using a rule-based tool, see section 3.5.1. By doing that we get a parallel corpus. Then, we train an NMT model.

5.2. Old Turkish to Modern Turkish Translation

We mean by ‘old Turkish’ the Turkish that was used during the Ottoman Empire days and the early days of the Republic of Turkish but written with the *Latin alphabet*. Old Turkish texts are, in most cases, difficult to grasp by today’s Turkish generation. Some reasons are the extensive use of foreign vocabulary and grammar that are not used anymore these days.

5.3. An End-to-End Ottoman Turkish to Modern Turkish Translation Model

We train end-to-end translation models, then compare them with two step models. That is, compare end-to-end models with transliterating from Ottoman Turkish to old Turkish, then translating from old Turkish to modern Turkish.

The old version of Nutuk has the original wording but transliterated from Ottoman alphabet to the Latin alphabet by a qualified writer. Therefore, we transliterate it back to the Ottoman alphabet to get an Ottoman – modern Turkish

parallel corpus. Our monolingual corpus is, again, the Ottoman transliteration of the same monolingual corpus we mention in section 3.2.

5.4. Training NMT Models

We develop NMT models to

- Transliterate from Ottoman Turkish to old Turkish,
- Translate from old Turkish to modern Turkish, and
- Translate from Ottoman Turkish to modern Turkish in an end-to-end manner.

To realize that, we use NMT techniques (see chapter 4). Mainly, we use sequence-to-sequence models with LSTMs as the encoder and decoder with global attention.

We pre-train word embeddings and use them to initialize the embedding tables of the encoder and the decoder.

5.4.1. Pre-training Word Vectors

Due to the scarcity of our parallel training data, we used pre-trained word embeddings to initialize the encoder's and the decoder's embedding tables. As discussed in [38], pre-trained word embeddings are useful in low-resource settings.

5.4.2. Aligning Pre-trained Word Vectors of the Source and Target Languages

For old Turkish to modern Turkish word vectors, we apply a data-driven method for aligning source and target word embeddings. We compile an unsupervised corpus that consists of texts written between 1928 and 2015, see chapter 3 for details. This corpus allows aligning the old with the modern Turkish words automatically without using any special aligning algorithm. We simply use an off-the-shelf word embeddings algorithm on the mixed corpus.

5.5. Experiments

The baseline we use for our experiments has two LSTM layers for both the encoder and the decoder with size of 250 and a word embeddings size of 500. We train the baseline for 40 epochs and get validation set BLEU score of 28 points.

As an NMT implementation, we use Open-NMT-py [51]. We initialize our weights using a uniform random distribution in the range (-0.1, 0.1). As optimization algorithm, we use Adam [52], an initial learning rate of 0.001, and Noam learning rate schedule [53]. To prevent overfitting, we use dropout [54] in the LSTM stacks with a probability of 0.65. Our batch size is 32.

We do parameter search to determine good configurations for

- The number of training epochs,
- Word embedding-vector sizes,
- LSTM cell sizes,
- The global attention type,
- Unidirectional vs. bidirectional LSTM encoder,
- Sharing vs not sharing decoder's input and output embeddings, and
- Using pre-trained word embedding-vectors (don't use, use and freeze, use and modify while training the NMT).

We use Ottoman Turkish → modern Turkish parallel corpus while doing parameter search. Then, we apply the best configurations we find remaining corpora.

5.5.1. Parameter Search for Number of Training Epochs

To select the right number of epochs for our dataset, we train our baseline model for 170 epochs, while evaluating it on the validation set and getting the corresponding BLEU score. We draw Figure 5.1 as number of epochs vs BLEU score of the model on the validation set. From Figure 5.1, we realize that after epoch 40, the model almost saturates, where it gets a BLEU score 28.28. It reaches a BLEU score of 29.07 at the 48^{th} epoch. While the max BLEU score of 29.80 is at the 160^{th} epoch, we will choose to train our models for 48 epochs from now on.

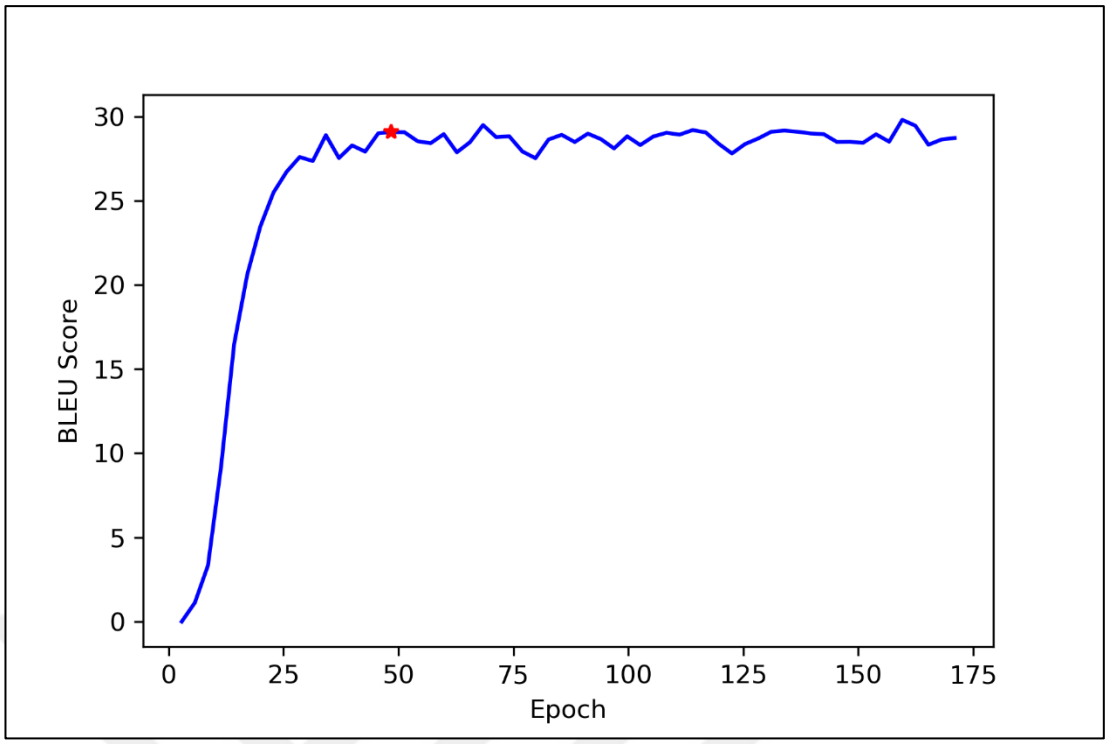


Figure 5.1: Parameter search for the number of epochs.

5.5.2. Parameter Search for Word Embeddings Size

To find most suitable word embeddings size, we train several models with the baseline settings while differing word embeddings sizes, and evaluate them on the validation set. We start by training a model with 100-dimension-word-embeddings, increase it 100 dimensions at a time, till we reach 2000-dimension-word-embeddings. In Figure 5.2, we draw the relation between word-embeddings size vs validation set BLEU score. From Figure 5.2, we realize that between 100-dimension-word-embeddings and 500- dimension-word-embeddings the validation BLEU score increases with the increment of the word-embeddings dimensions. After that, it becomes instable. Therefore, we decide to have 500-word-embeddings dimensions.

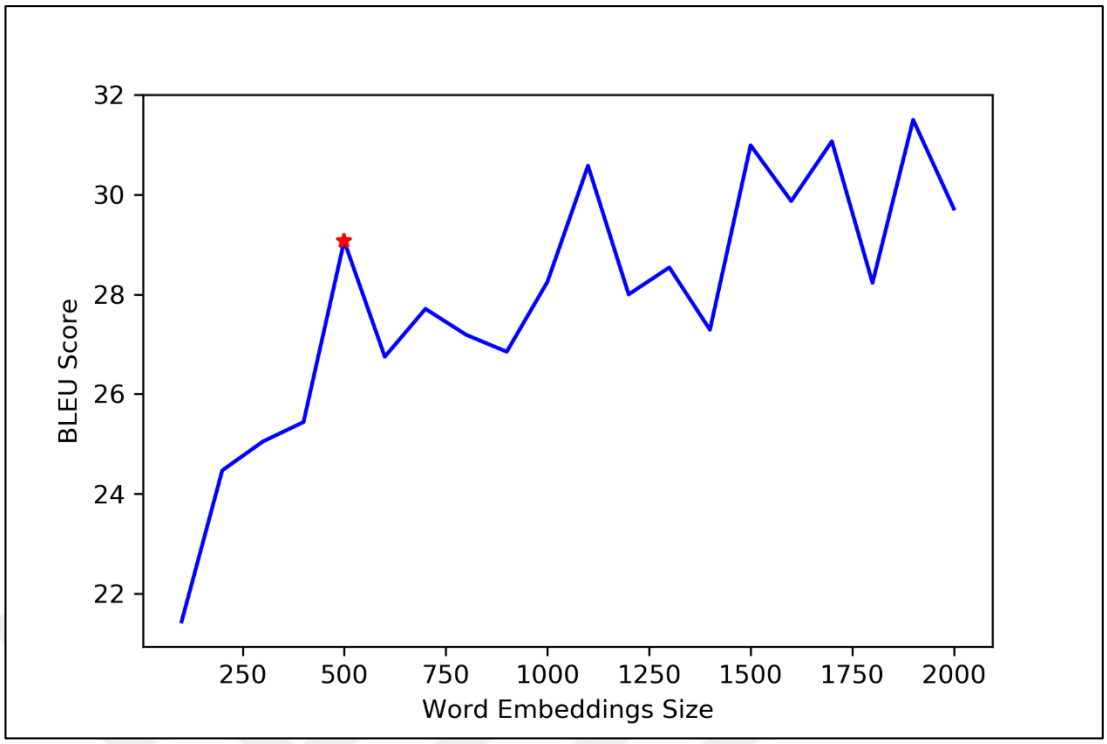


Figure 5.2: Parameter search for word embeddings size.

5.5.3. Parameter Search for LSTM Cell Size

To find most suitable word LSTM cell size, we train several models with the baseline settings while differing the LSTM cell sizes, and evaluate them on the validation set. We start by training a model with 100-size-LSTM cell, increase it 100 dimensions at a time, till we reach 1000-size-LSTM cell. In Figure 5.3, we draw the relation between LSTM cell size vs validation-set-BLEU score. From Figure 5.3, we realize that between 100-dimension-LSTM-cell and 500-LSTM-cell the validation BLEU score increases with the increment of the LSTM-cell dimensions. After that, it becomes instable. Therefore, we decide to have 500-size-LSTM cell.

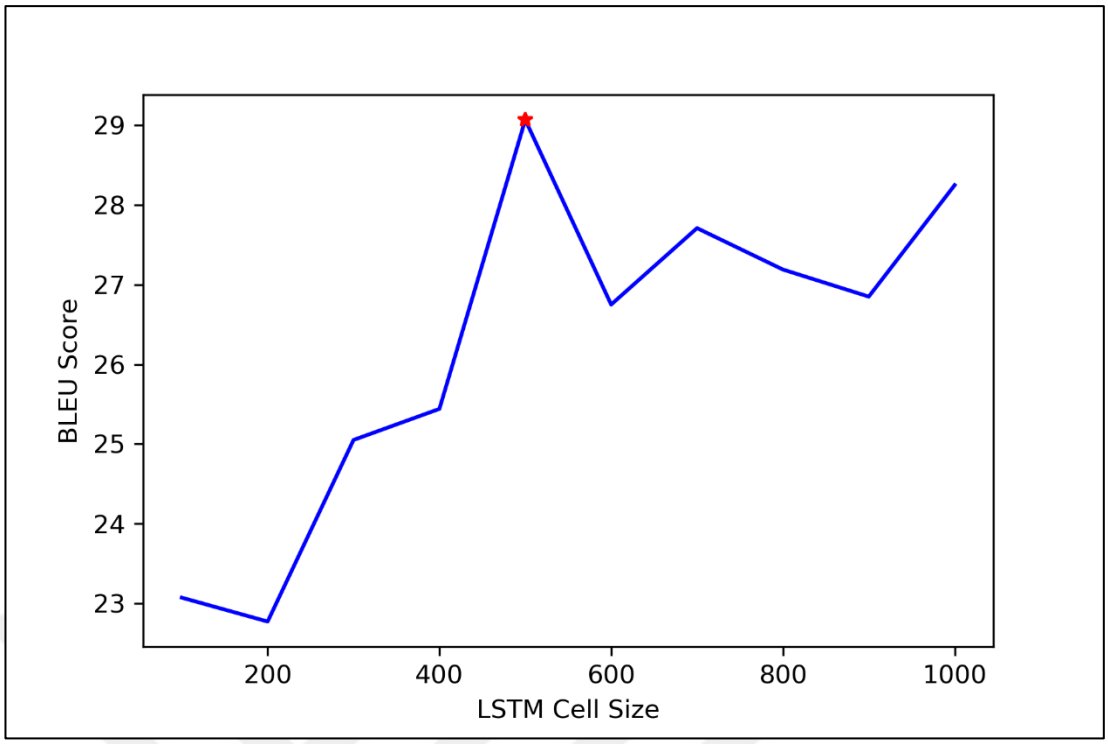


Figure 5.3: Parameter search for LSTM cell size.

5.5.4. Parameter Search for the Global Attention Type

We train several NMT models with the baseline settings while differing the global attention type (dot product, general, MLP), and evaluate them on the validation set. We list the validation set BLEU scores in Table 5.1. We see no significant difference between the different types. Consequently, we opt for ‘dot product’ global attention, because it does not use any extra parameters.

Table 5.1: General attention type parameter search.

Global Attention Type	Validation set BLEU score
Dot product	31.07
General	31.13
MLP	31.65

5.5.5. Unidirectional vs. Bidirectional Encoder

For this parameter, we train an NMT model with the baseline settings except for the LSTM encoder, which we set as bidirectional LSTM. We get 4+ BLEU score points gain over the baseline.

5.5.6. Sharing the Decoder's Input and Output Embeddings

For this parameter, we train an NMT model with the baseline settings except for sharing the decoder's input and output embeddings. We get 3+ BLEU score points gain over the baseline.

5.5.7. Pre-trained Word Embedding-Vectors

To determine our word embeddings strategy, we train three NMT model with the baseline settings except for the encoder's and the decoder's word embeddings look-up tables. In Table 5.2 we list the results. When we use pre-trained word vectors and freeze them during the training of the NMT model, we get 2 BLEU score points loss on the validation set. When we use pre-trained word-vectors and fine-tune them while training the NMT model, we get 3+ BLEU score points gain over the baseline.

Table 5.2: Using pre-trained word vectors.

Pre-trained word vectors	Fine-tuning	BLEU score
No	_	28
Yes	No	26.03
Yes	Yes	31.02

5.5.8. NMT Model's Final Configurations

According to our experiments in the previous subsections, we decide to use the following configurations for all our NMT models.

- We train our models for 48 training epochs,

- 500-dimensional word vectors,
- 500-dimensional LSTM cells,
- Dot product global attention,
- A bidirectional LSTM encoder,
- We tie (share) the decoder's input and output embeddings, and
- We use pre-trained word vectors for both the encoder's and the decoder's look-up tables, and fine-tune them while training the NMT model.

5.5.9. Best Configurations BLEU Scores

We apply the configurations mentioned in 5.5.8 to all transliteration and translation tasks. Our BLEU score results on test sets are in Table 5.3.

Table 5.3: Best configurations results on all tasks.

Task	Test set BLEU score
Ottoman → Old	73.53
Old → Modern	42.34
Ottoman → Old → Modern	42.10
Ottoman → Modern	41.84

Note that the two-step model for Ottoman Turkish to modern Turkish task outperforms the end-to-end model. That is because the Ottoman Turkish to old Turkish transliteration model generates near perfect transliterations, and the old Turkish to modern Turkish model is 'better' than the end-to-end Ottoman Turkish to modern Turkish model.

5.5.10. Sample Outputs from the Best Models

We show sample outputs from our best models in the following tables.

Table 5.4: Ottoman Turkish to old Turkish model output samples.

Original	عثمانلی حکومتینه ، عثمانلی پادشاهینه و مسلمینڭ خلیفه‌سنه عصیان ایتمك و بتون ملتی و اوردویی عصیان ایتدیرمك لازم گلییوردی .
Human	osmanlı hükumetine , osmanlı padişahına ve müsliminin halifesine isyan etmek ve bütün 42illet ve orduyu isyan ettirmek lazımgeliyordu .
Output	osmanlı hükumetine , osmanlı padişahına ve <s> vapesinini isyan etmek ve bütün 42illet ve orduyu isyan ettirmek lazımgeliyordu .
Rating	4.5/5
Original	هرکس ساکنانه عیش و گوچلریله مشغولدر .
Human	herkes sakinane iş ve güçlerile meşguldur .
Output	herkes sakitane iş ve güçlerile aleyhtarımızdır .
Rating	3.5/5
Original	- بونڭ ایچون تکمیل ولایتلرڭلر هر لواسندن ملتڭ اعتمادینه مظهر اوچ مرخصین سرعتی ممکنه ایله یتیشمك اوزره همان یوله چیقاریلمسی ایجابات ایتمکده‌در .
Human	- bunun için tekmil vilayetlerin her livasından milletin itimadına mazhar üç murahhasın sürati mümkine ile yetişmek üzere hemen yola çıkarılması icap etmektedir .
Output	- bunun için tekmil vilayetlerin her livasından milletin itimadına mazhar üç murahhasın sürati mümkine ile yetişmek üzere hemen yola çıkarılması icap etmektedir .
Rating	5/5
Original	یالڭز ، ملتڭ اعتمادنی حایز بر قابینه تشکیل ایتمك ایچون اوّلا او قابینه‌نڭ استناد ایده‌بیله‌جگی بر قویتی وجوده گتیرمك لازمدر .
Human	yalnız , milletin itimadını haiz bir kabine teşkil etmek için evvela o kabinenin istinat edebileceği bir kuvveti vücuda getirmek lazımdır .
Output	yalnız , milletin itimadını haiz bir kabine teşkil etmek için evvela o kabinenin istinat edebileceği bir kuvveti vücuda getirmek lazımdır .
Rating	5/5
Original	اشخاصی مرقومه ۱۰ ایلول گونی علی الصباح قاچدقلری حالده جمال بگ ، بو معلوماتی ، انجق ، الیاس بگ مفرزه‌سنڭ مواصلتندڭ و الیاس بین راپورندڭ صوڭره بیلدیرییور .

Table 5.4: Continuation.

Human	eşhası merkume 10 eylul günü alessabah kaçtıkları halde cemal bey , bu malumatı , ancak , ilyas bey müfrezesinin muvasalatından ve ilyas beyin raporundan sonra bildiriyor .
Output	eşhası islaliyenin 10 eylul günü alessabah selahiyettar halde cemal bey , bu malumatı , ancak , ilyas bey müfrezesinin müşar ve ilyas beyin telgrafından sonra bildiriyor .
Rating	4/5
Original	پادشاه ، وضعیتدن اگاه اولدیغی تقدیرده درحال کندیسنی اغفال ایدنلره لایق اولدقلری معاملهیی تطبیق ایدهجگینه امنیتمز اولدیغنی ایلری سوردك و حکومتك ثابت اولان جنایتی اوزرینه کندیسنه اعتمادك انسلابی طبیعی اولدیغندك حقیقتی حالی یالکز و انجق طوغریدن طوغری یه پادشاهه عرض ایتمکله وضعیتك اصلاحی ممکن اولاجغنی تشبثاتمز ایچون نقطهی عزیمت عد ایتدك .
Human	padişah , vaziyetten agah olduğu takdirde derhal kendisini iğfal edenlere layık oldukları muameleyi tatbik edeceğine emniyetimiz olduğunu ileri sürdük ve hükumetin sabit olan cinayeti üzerine kendisine itimadın insilabı tabii olduğundan hakikati hali yalnız ve ancak doğrudan doğruya padişaha arzetmekle vaziyetin ıslahı mümkün olacağını teşebbüsatımız için noktai azimet addettik .
Output	padişah , vaziyetten agah olduğu takdirde derhal kendisini iğfal edenlere layık oldukları muameleyi tatbik edeceğine emniyetimiz olduğunu ileri akıl ve hükumetin sabit olan imtizaçları üzerine kendisine itimadın alışkınsınız tabii olduğundan hakikati hali yalnız ve ancak doğrudan doğruya padişaha arzetmekle vaziyetin ıslahı mümkün olacağını arzedilmesini için noktai azimet addettik .
Rating	4/5
Original	تلغرافجی — یونس نادی بگ و یاننده ناظر پاشانك یاوری جواد رفعت بگ واردر افندیم .
Human	telgrafçı — yunus nadi bey ve yanında nazır paşanın yaveri cevat rifat bey vardır efendim .
Output	telgrafçı — yunus nadi bey ve yanında nazır paşanın yaveri cevat rifat bey vardır efendim .
Rating	5/5

Table 5.5: Old Turkish to Modern Turkish model output samples.

Original	meclisi meb'usan içtima ile bilfiil murakabeye başlayacağı güne kadar , heyeti temsiliyenin mukadderatı millet ve memleketle alakadar bulunması zaruridir .
Human	meb'uslar meclisi toplanıp etkin olarak denetime başlayacağı güne kadar , temsilci kurulun ulus ve ülkenin kaderi ile ilgilenmesi zorunludur .
Output	meb'uslar meclisi toplanıp edimli olarak edimli olarak güne kadar , temsilci kurulun ulusun kaderini ve çok ilgilenmesi zorunludur .
Rating	3/5
Original	hükumeti , ya fethi bey veyahut kara bekir paşa teşkil edecektir .
Human	hükumeti , ya fethi bey ya da kara bekir paşa kuracaktır .
Output	hükumeti , ya fethi bey veya kara bekir paşa olacaktır .
Rating	4/5
Original	bunu ehemmiyetle arzediyorum .
Human	bunu önemle bilginize sunuyorum .
Output	bunu önemle sunuyorum .
Rating	4.5/5
Original	yenihan isyanı , orta anadolunun diğer yerlerindeki erbabı fesadı da harekete getirdi .
Human	yenihan ayaklanması , orta anadolunun öbür yerlerindeki kışkırtıcıların da başkaldırmalarına yol açtı .
Output	yerel ayaklanması , orta anadolunun öbür ortalığı karıştıranları getirdi .
Rating	3/5
Original	bursaya azimet ve orada ikametiniz meselesine gelince ; müntesip bulunduğunuz mesleki askeri dolayısile bu bapta müdafaai milliye vekaleti celilesine alelusul müracaat buyurmanız lüzumu tebliğ olunur efendim .
Human	Bursaya gidip orada oturmak konusuna gelince ; içinde bulunduğunuz askerlik mesleği dolayısile bu konuda yüksek milli savunma bakanlığına usulünce başvurmanız zorunluğu bildirilir efendim .

Table 5.5: Continuation.

Output	bursaya yola çıkmak ve orada savaşın konusuna gelince ; sizin bulunduğunuz askerlik askeri dolayısile bu konuda milli savunma bakanlığı başkanlığına bilgi vermeniz gereği bildirilir efendim .
Rating	3.5/5
Original	alay 3 ten tasavvur olunan kuvvetin izamı yapılacaktır .
Human	alay 3 ten düşünülen kuvvet gönderilecektir .
Output	alay 3 den sayılan kuvvetin gönderilmesi yapılacaktır .
Rating	4.5/5
Original	cünkü refet paşa , muhtelif zamanlarda muhtelif sebeplerle konyaya , denizliye gitmiş , garp cephesinin cenup kısmile alakadar olmuş ve o kısımla münasebettar mıntıkaları tanımış bulunuyordu .
Human	cünkü refet paşa , değişik zamanlarda değişik nedenlerle konyaya , denizliye gitmiş , batı cephesinin güney bölümüyle ilgilenmiş ve o bölümle ilişkili bölgeleri tanımış bulunuyordu .
Output	cünkü refet paşa , değişik zamanlarda değişik nedenlerle konyaya , alayunda gitmiş , batı cephesinin güney güvenliği ilgili olmuş ve o yöredeki ilişkili bölgeleri tanimiş bulunuyordu .
Rating	4.5/5

Table 5.6: End-to-end Ottoman Turkish to Modern Turkish model output samples.

Original	بز ، بو پرنسيبڭ ، اصله مروجی اولامازدق .
Human	biz , bu prensibi , hiçbir zaman uygun bulamazdık .
Output	biz , bu cloruldar , hiç yetencği biriyim .
Rating	2/5
Original	بو تلغرافڭ بر تورلی معناسینی اڭلامییوردم .
Human	bu telgrafın anlamını bir türlü kavrıyamıyordum .
Output	bu telgrafın bir türlü anlamını gerektiriyordu .
Rating	3/5
Original	یدنجی مکتوب استادم ؛ علی کمال بگ ، دون او ذاتله گوروشمش .
Human	yedinci mektup ustadım ; ali kemal bey dün o kişiyle görüşmüş .
Output	yedinci mektup ustadım ; ali kemal bey , dün o kişiyle görüşmüş .
Rating	5/5
Original	او موجود اولونجه بوگونکی وضعیت ، شکل ، صلاحیت موقتدر ، مقامی خلافت و سلطنت ، اجرایی فعالیته فرصت بولونجه ، تشکیلاتی سیاسی یه و اساسی یەنڭ نه اولدیغی معیندر ، معلومدر .
Human	o var olunca bugünkü durum , biçim , yetki geçicidir , halife ve padişah çalışma fırsatı bulunca , yasal ve anayasal kuruluşun ne olacağı bellidir , bilinmektedir .
Output	o öğrenilince bugünkü durum , biçim , yetki geçicidir , halifelik ve padişahlık , iş çalışmaya fırsat bulunca , siyasi ve anayasanın ne olduğu belirlidir , bilinir .
Rating	4/5
Original	تبلیغ اولونان نوطه و بیاناتڭز واضحا گوسترمکدەدر که ، دولی ایتلافیه لوندرا قونفەرانسینه یالڭز اولارق اناطولی مرخصلرینی قبول ایتمەمکدەدر .
Human	verilen nota ve sizin sözleriniz açıkça göstermektedir ki , itilaf devletleri londra konferansına anadolu delegelerini tek başlarına kabul etmemektedir .
Output	bildirilen nota ve açıkça göstermektedir ki , itilaf devletleri londra konferansına yalnız olarak anadolu adamlari kabul edilebilir .
Rating	3/5

Table 5.6: Continuation.

Original	بز ، مملكتڭ ، اشر ايديلمش ، اختيارنى غايب ايتمش پارچەسنى حر و مستقل قسمه الحاق ايتمك ايستييورز .
Human	biz memleketin tutsak edilmiş , özerkliğini yitirmiş parçasını özerk ve bağımsız bölüme bağlamak istiyoruz .
Output	biz , memleketin , tutsak edilmiş , gerisindeki yitirmiş parçasını hür ve bağımsız bölüme bağlamak istiyoruz .
Rating	4.5/5
Original	رافت پاشانڭ تختى قوماندەسنه ويريلن قويتلر ، تعرضلركده موفق اولامديلر ، بالعكس فضله ضايعات ويريلدى .
Human	refet paşanın komutası altına verilen kuvvetler , saldırılarında başarılı olamadılar , tersine kayıp verildi .
Output	refet paşanın komutasına verilen kuvvetler , durmadan başarili olamadilar , tersine fazla yitik verildi .
Rating	3.5/5

Table 5.7: Two-step Ottoman Turkish to Modern Turkish model output samples.

Original	بکر سامی بین ییدکده بولونان نسخەنڭ محتویاتینه مطلع ایدیلدیگمی درخاطر ایتمییورم .
Human	bekir sami beyin elinde bulunan kopyada yazılanlardan bana bilgi verildiğini anımsamıyorum .
Output	bekir sami beyin elinde bulunan bulunanların içeriğini öğrendim .
Rating	3/5
Original	جریانی ملی علیهنده ، حرکاتی خاینانەده بولوندیغی تحقق ایدن انقره والیسی محی الدین پاشا ، مقصدی مخصوصله دوره چیقمش ایدی .
Human	ulusal akıma karşı , haince davranışlarda bulunduğu kesinlikle anlaşılan ankara valisi muhittin paşa , özel amaçla görev gezisine çıkmış idi .
Output	akıma karşı akıma karşı , savaşmaya verildiğini kanıtlanan ankara valisi muhittin paşa , özel amaçla devre çıkmış idi .
Rating	3/5
Original	بو تلغرافه تجارت و زراعت نازڭ هادی پاشا وساطتیله و عین شفره ایله جواب انتظارندەدر .
Human	bu telgrafa ticaret ve tarım bakanı hadi paşa aracılığıyla ve aynı şifre ile yanıt beklemektedir .
Output	bu telgrafa kurula ve tarım ahmet hadi paşa elile ve aynı şifre ile yanıt geldikte .
Rating	3/5
Original	نظری دقته المەلیدر که ، بو یازیلری بن یازمش و صدری سابق ایله پادشاهمز افندیمز حضرتلری ، بونڭ جریانی کاملندڭ صوكره ، نتایجینه اطلاع ایله درجاتی محکمەسی قارشیسنده اتحادی قرار قیلمشلردر . . .
Human	dikkat edilmelidir ki , bu yazıları ben yazmış ve eski başbakan ile padişahımız efendimiz hazretleri , bunun sonucu alındıktan sonra , öğrenmek ve değerlendirerek karar almışlardır . . . .
Output	surası göz önünde bulundurulmalıdır ki , bu yazıları ben yazmamdan sonra ve eski başbakanla padişahımız efendimiz hazretleri , bunun sonucunu öğrenip değerlendirmeleri üzerine yazılanların sağlam dayanaklarını görünce kararlarını vermişlerdir . . .
Rating	3/5
Original	ارتیق ، انجق عوامڭ قیل و قالندڭ باشقه بر ماهیتی اولامایان دیدی قودیلرڭ ، ایتایی قرار خصوصنده ، موثر اولابیلەجگنه امکان تصور ایتمییورز .

Table 5.7: Continuation.

Human	artık , ancak ayaktakımının söylentilerinden başka bir niteliği olamayan dedikoduların , karar vermekte , etken olabileceğine olanak düşünemiyoruz .
Output	artık , ancak basit asilerin kıl ve yapmaları başka bir anlamı isteklerin dedikoduların , bilgi konusunda , etkili olabileceğine olanak dışıdır .
Rating	4/5
Original	اتم بين ارقداشى طرفندن يازيلان بو تلغراف نامه مفادينڭ ، يالڭز ارقداشينڭ مطالعەسى و حقيقته مطابق اولديغى البته قبول ايديلەمزدى .
Human	etem beyin arkadaşı tarafından yazılan bu telgrafda bildirilenlerin , sadece arkadaşının düşüncesi olduğu ve gerçeğe uygun bulunduğu elbette kabul edilemezdi .
Output	etem beyin arkadaşı tarafından yazılan bu telgraf verilmesi , yalnız arkadaşının düşüncesi ve gerçeğe uygun olduğu elbette kabul edilemezdi .
Rating	4.5/5
Original	مملكتمزده بولونان دوشمانلرى سلاح قويتيله چيقارمادقجه ، چيقارابيلەجك موجوديت و قدرتى مليەمزى فعلا اثبات ايتمەدكجه ، ديپلوماسى صخاسنده اميده قاپيلمەنڭ جايز اولمديغى حقندەكى قناعتمز قاتى و دايمى ايدى .
Human	memleketimizde bulunan düşmanları silah kuvveti ile çıkarmadıkça , çıkarabilecek ulusal varlık ve gücümüzü edimli olarak kanıtlamadıkça , diplomasi alanında umuda kapılmanın yeri olmadığı yolundaki inancımız kesin ve sürekli idi .
Output	memleketimizde bulunan düşmanları silah kuvvetile gücün , çıkarabilecek varlık ve ulusal egemenliğimize edimli olarak kanıtlamak , siyaset güvensizliğin ümide çok uygun olmadığı hakkındaki geldiğini kesin ve sürekli idi .
Rating	3.5/5

5.5.11. Our Training Environment

We use one of GTU-NLP-Lab's deep learning servers to run our experiments. The model of the GPU we use is Nvidia GEFORCE GTX 1080 Ti. The operating system it runs on is Ubuntu Server 18.04.3.

6. CONCLUSION

As the Turkish language has changed to a great extent since a hundred years, both the old Turkish written with Latin characters and the Ottoman Turkish are cryptic to nowadays' generation. In order to make the Ottoman written heritage relatable, qualified writers who are good at both the modern Turkish and the old Turkish are manually rewriting old documents into modern Turkish. That is a vital but a resource-expensive workaround. To work out this problem, we develop multiple translation models. The first one transliterates from Ottoman Turkish to old Turkish that is written with the Latin alphabet. The second one translates from old Turkish that is written with Latin characters to modern Turkish. The last one translates and transliterates from Ottoman Turkish to modern Turkish in an end-to-end fashion. We collect, clean and use parallel data for these purposes. Our parallel data for these models are parallel versions of Nutuk, which is a book compilation of the speech given by Mustafa Kemal Atatürk at the second congress of Cumhuriyet Halk Partisi.

Additionally, we obtain an Ottoman version of Nutuk by transliterating the old version of it that is with Latin characters to the Ottoman alphabet. For the Latin-Ottoman transliteration, we re-implement an existing tool. Furthermore, we use monolingual corpora to improve our models' performance. More specifically, we pre-train word embeddings models and initialize the NMT models' lookup tables with them. That yields in a considerable performance gain.

We summarieze our contributions as follows.

- We develop three parallel corpora; from Ottoman Turkish to old Turkish written with Latin characters, from old Turkish written with Latin characters to modern Turkish and from Ottoman Turkish to modern Turkish,
- For the development of Ottoman Turkish to old Turkish parallel corpus, we re-implement a morphology-based transliteration tool in the other way around; from old Turkish to Ottoman Turkish, and
- As our main contribution, we develop three neural-based end-to-end translation systems using the three parallel corpora mentioned.

Our future work will consider the following directions:

- Build an OCR model for Ottoman Turkish documents.
- Ensemble NMT models with a morphology-based system for the translation and transliteration of Ottoman Turkish texts.

REFERENCES

[1] Lightfoot D., (2006), "How new languages emerge", 1st Edition, Cambridge University Press.

[2] Lewis G., (1999), "The Turkish Language Reform: A Catastrophic Success", 1st Edition, OUP Oxford.

[3] Ensari M. A., (2015), "Osmanlıca İmla Müfredatı", 1st Edition, Istanbul: Hayrat Vakfı Yayınları.

[4] Hagopian V.H., (1907), "Ottoman-Turkish conversation-grammar: a practical method of learning the Ottoman-Turkish language", 1st Edition, J. Groos.

[5] Bangalore S., Riccardi G., (2001), "A finite-state approach to machine translation", In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, 246-253, Orlando, FL, USA, 18-21 August.

[6] Web 1, (2001), https://en.wikipedia.org/wiki/Translation, (Access Date 16/07/2019).

[7] Dolek I., Kazan S., (2016), "Translating Contemporary Turkish To Ottoman Turkish By Using Artificial Neural Network Based Optical Character Recognition", International Journal of Information Technology & Mechanical Engineering, 12 (2), 5-10.

[8] Kurt A., Bilgin E. F., (2012), "The Outline of an Ottoman-to-Turkish Automatic Machine Transliteration System", First Workshop on Language Resources and Technologies for Turkic Languages, 45-50, Istanbul, Turkey, 10-15 October.

[9] Web 2, (2012), https://en.academicresearch.net/machine-transliteration-of-ottoman-turkish-texts-to-modern-turkish/, (Access Date 16/06/2019).

[10] Ergişi A., Şahin İ. E., (2017), "Dervaze: A Spelling Dictionary for Digital Translation , International Journal of Languages' Education and Teaching, 12 (5), 30-40.

[11] Web 3, (2019), http://www.cs.princeton.edu/~ckorkut/papers/ottoman.pdf, (Access Date 10/06/2019).

[12] Darwish K., (2014), "Arabizi Detection and Conversion to Arabic," Proceedings of the EMNLP Workshop on Arabic Natural Language Processing, Doha, Qatar, 1-5 October, .

[13] Al-Badrashiny M., Eskander R., Habash N., Rambow O., (2014), "Automatic transliteration of romanized dialectal Arabic", Proceedings of the eighteenth conference on computational natural language learning, Baltimore, Maryland, 7-13 June.

[14] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., ... Dyer C., (2007), "Moses: Open source toolkit for statistical machine translation", In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, Prague, Czech Republic, 7-13 June.

[15] Le N. T., Sadat F., Menard L., Dinh, D., (2019), "Low-Resource Machine Transliteration Using Recurrent Neural Networks", ACM Transactions on Asian and Low-Resource Language Information Processing, 18 (2), 13, 380-389.

[16] Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y., (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1723-1734, Doha, Qatar, 1-5 October.

[17] Sutskever I., Vinyals O., Le Q. V., (2014), "Sequence to sequence learning with neural networks", In Advances in neural information processing systems, 100-109, Montréal, Canada, 8-13 September.

[18] Edunov S., Ott M., Auli M., Grangier D., (2018), "Understanding back-translation at scale", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 489–500, Brussels, Belgium, 31 October - 4 November.

[19] Ding L., Tao D., (2019), "The University of Sydney's Machine Translation System for WMT19", Proceedings of the Fourth Conference on Machine Translation (WMT), 175–182, Florence, Italy, 1-2 August, 2019.

[20] Bahdanau D., Cho K., Bengio Y., (2014), "Neural machine translation by jointly learning to align and translate", International Conference on Learning Representations, 80-90, San Diego, CA, USA, 7-9 May.

[21] Sennrich R., Haddow B., Birch A., (2015), "Neural machine translation of rare words with subword units", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725, Berlin, Germany, 7-12 August.

[22] Lee J., Cho K., Hofmann T., (2017), "Fully character-level neural machine translation without explicit segmentation", Transactions of the Association for Computational Linguistics, 100 (5), 365–378.

[23] Web 4, (2015), https://arxiv.org/pdf/1503.03535, (Access Date 10/06/2019).

[24] Sennrich R., Haddow B., Birch A., (2015), "Improving neural machine translation models with monolingual data", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 86–96, Berlin, Germany, 7-12.

[25] Web 5, (2019), https://arxiv.org/pdf/1903.07091, (Access Date 10/06/2019).

[26] Johnson M., Schuster M., Le Q. V., Krikun M., Wu Y., Chen Z., ... Hughes M., (2017), "Google's multilingual neural machine translation system: Enabling zero-shot translation", Transactions of the Association for Computational Linguistics, 100 (5), 339–351.

[27] Ganin Y., Ustinova E., Ajakan H., Germain P., Larochelle H., Laviolette F., ..., Lempitsky V., (2016), "Domain-adversarial training of neural networks", The Journal of Machine Learning Research, 17 (1), 2096-2030.

[28] Web 6, (2016), https://arxiv.org/pdf/1604.00117, (Access Date 06/10/2019).

[29] Web 7, (2017), https://arxiv.org/pdf/1711.00043, (Access Date 06/10/2019).

[30] Web 8, (2017), https://arxiv.org/pdf/1710.11041, (Access Date 06/10/2019).

[31] Hill F., Cho K., Korhonen A., (2016), "Learning distributed representations of sentences from unlabelled data", Proceedings of NAACL-HLT 2016, 1367–1377, San Diego, California, 12-17 June.

[32] Web 9, (2014), https://arxiv.org/pdf/1412.3555, (Access Date 06/10/209).

[33] Gehring J., Auli M., Grangier D., Yarats D., Dauphin Y. N., (2017), "Convolutional sequence to sequence learning", In Proceedings of the 34th International Conference on Machine Learning, 70 (4) , 99-105.

[34] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., ..., Polosukhin I., (2017), "Attention is all you need", In Advances in neural information processing systems, 220-229, LA, USA, 4-9 December.

[35] Web 10, (2015), https://play.google.com/store/apps/details?id=com.tahirhoca.osmanlicaceviri&hl=tr, (Access Date 16/06/2019).

[36] Jordan M., (1986), "Attractor dynamics and parallelism in a connectionist sequential machine", In Proc. of the Eighth Annual Conference of the Cognitive Science Society, 100-112, Orlando, FL, USA, 18-21 August.

[37] Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., (2013), "Distributed representations of words and phrases and their compositionality", In Advances in neural information processing systems, 50-64, Orlando, FL, USA, 18-21 August.

[38] Web 11, (2018), https://arxiv.org/pdf/1804.06323, (Access Date 16/06/2019).

[39] Lample G., Ott M., Conneau A., Denoyer L., Ranzato M. A, (2018), "Phrase-based & neural unsupervised machine translation", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 5039–5049, Brussels, Belgium, 31 October - November 4.

[40] Bojanowski P., Grave E., Joulin A., Mikolov T., (2017), "Enriching word vectors with subword information", Transactions of the Association for Computational Linguistics 5., 138 (5), 135-146.

[41] Pennington J., Socher R., Manning C., (2014), Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing, 20-33, Doha, Qatar, 1-5 October.

[42] Bengio Y., Simard P., Frasconi P., (1994), "Learning long-term dependencies with gradient descent is difficult", IEEE transactions on neural networks, 5 (2), 157-166.

[43] Greff K., Srivastava R. K., Koutník J., Steunebrink B. R., Schmidhuber J., (2016), "LSTM: A search space odyssey", IEEE transactions on neural networks and learning systems, 28 (10), 2222-2232.

[44] Hochreiter S., Schmidhuber J., (1997), "Long short-term memory", Neural computation, 9 (8), 1735-1780.

[45] Graves A., Schmidhuber J., (2005), "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", Neural networks, 18 (5-6), 602-610.

[46] Graves A., Mohamed A. R., Hinton G., (2013), "Speech recognition with deep recurrent neural networks", In 2013 IEEE international conference on acoustics, speech and signal processing,, 200-208, Vancouver, Canada, 26-31 May.

[47] Press O., Wolf L., (2016), "Using the output embedding to improve language models", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 157–163, Valencia, Spain, 3-7 April.

[48] Conneau A., Kruszewski G., Lample G., Barrault L., Baroni M., (2018), "What you can cram into a single vector: Probing sentence embeddings for linguistic properties", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2126–2136, Melbourne, Australia, 15-20 July.

[49] Luong M. T., Pham H., & Manning C. D, (2015), "Effective approaches to attention-based neural machine translation", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1412–1421, Lisbon, Portugal, 17-21 September.

[50] Papineni K., Roukos S., Ward T., & Zhu W. J, (2002), "BLEU: a method for automatic evaluation of machine translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318, Philadelphia, Pennsylvania, USA, 10-14 July.

[51] Klein G., Kim Y., Deng Y., Senellart J., Rush A. M., (2017), "Opennmt: Open-source toolkit for neural machine translation", Proceedings of AMTA, 103 (1), 177-184.

[52] Web 12, (2014), https://arxiv.org/pdf/1412.6980, (Access Date 10/06/2019).

[53] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., ... Polosukhin I., (2017), "Attention is all you need", In Advances in neural information processing systems, 220-229, LA, USA, 4-9 December.

[54] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., (2014), "Dropout: a simple way to prevent neural networks from overfitting", The journal of machine learning research, 15 (1), 1929-1958.

[55] Web 13, (2006), https://en.wikipedia.org/wiki/Ottoman_Turkish_alphabet, (Access Date 06/07/2019).

[56] Shu R., Miura A., (2016), "Residual stacking of rnns for neural machine translation", In Proceedings of the 3rd Workshop on Asian Translation, 200 (4), 223-229.

BIOGRAPHY

Abdullah BAKIRCI (AL NAHAS) was born in Damascus, Syria in 1992. He graduated from the Department of Artificial Intelligence and Natural Language Processing, Faculty of Information Technology Engineering, Damascus University in 2015. He started his Master's at the Department of Computer Engineering, Engineering Faculty, Gebze Technical University at 2016. During his Master's, he worked on a project about social media analysis and anomaly detection. His research interests include natural language processing, machine learning, and deep learning.

APPENDICES

Appendix A: Publications on the thesis

AL NAHAS A., TUNALI M. S., AKGÜL Y. S., (2019), "Supervised Text Style Transfer Using Neural Machine Translation: Converting between Old and Modern Turkish as an Example", IEEE Signal Processing and Communications Applications Conference (SIU), Sivas, 24-26 April.