T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE-İNGİLİZCE SİNİRSEL MAKİNE ÇEVİRİ SİSTEMİ

DOKTORA TEZİ

İlhami SEL

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Davut HANBAY

Temmuz 2024

T.C
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE-İNGİLİZCE SİNİRSEL MAKİNE ÇEVİRİ SİSTEMİ

DOKTORA TEZİ

İlhami SEL
(36183619044)

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Davut HANBAY

Temmuz 2024

İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ

Türkçe-İngilizce Sinirsel Makine Çeviri Sistemi

Tezi Hazırlayan: İlhami SEL
Doktora Tezi

TEŞEKKÜR VE ÖNSÖZ

Öncelikle, akademik hayatım boyunca bana yol gösteren, bilgi ve tecrübeleriyle her adımda beni destekleyen danışman hocam Sayın Prof. Dr. Davut Hanbay'a sonsuz teşekkürlerimi sunarım. Kendisi, vizyonu ve rehberliği ile bu çalışmanın şekillenmesinde büyük rol oynamış ve akademik gelişimimde bana ilham kaynağı olmuştur.

Bu zorlu ve uzun süreçte yanımda olan, sabır ve anlayışıyla beni her daim destekleyen sevgili eşim Feride SEL'e en derin teşekkürlerimi iletmek istiyorum. Onun sevgi dolu desteği ve anlayışı, bu yolculuğu benim için çok daha katlanılabilir ve anlamlı kılmıştır.

Ayrıca, hayatımın neşe kaynağı ve en büyük motivasyon kaynağı olan sevgili oğullarım Alperen Ege ve Ediz Timur'a da minnettarlığımı ifade etmek istiyorum. Onların gülüşleri ve varlıkları, bana her daim güç ve enerji vermiştir.

Ve elbette, bu süreç boyunca desteklerini esirgemeyen, varlıklarıyla her zaman yanımda olan sevgili aileme de en içten teşekkürlerimi sunarım. Onların koşulsuz destekleri ve inançları, bu başarının ardındaki en büyük güç olmuştur.

ONUR SÖZÜ

Doktora tezi olarak sunduğum “Türkçe İngilizce Sinirsel Makine Çeviri Sistemi” başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığına ve yararlandığım bütün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

İlhami SEL

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SEMBOLLER VE KISALTMALAR

BLEU : Bilingual Evaluation Understudy (İkili Dil Değerlendirme)

BPE : Bayt Pair Encoding (Bayt çifti kodlaması)

CNN : Convolutional Neural Network (Evrişimsel Sinir Ağı)

DDİ : Doğal Dil İşleme

ESA : Evrişimsel Sinir Ağı

GRU : Gated Recurrent Unit

IWSLT : The International Workshop on Spoken Language Translation (Uluslararası Konuşma Dili Çevirisi Çalıştayı)

İÇS : İstatistiksel Çeviri Sistemi

K : Bin (x1000)

LRL : Low Resources Language (Düşük kaynaklı diller)

LSTM : Long Short Term Memory (Uzun-Kısa Süreli Bellek)

M : Milyon (x1.000.000)

MÇ : Makine Çevirisi

MHA : Multi-Head Attention (Çok Başlı Dikkat)

MLP : Multi Layer Perceptron (Çok Katmanlı Algılayıcı)

MT : Machine Translation (Makine Çevirisi)

NLP : Natural Language Processing (Doğal Dil İşleme)

NMT : Neural Machine Translation (Sinirsel Makine Çevirisi)

PBSMT : Phrase-Based Statistical Machine Translation (İfade Tabanlı Makine Çevirisi)

RNN : Recurrent Neural Network (Yinelenen sinir ağları)

SMÇ : Sinirsel Makine Çevirisi

TB : Tree Bank (Kelime Ağaç Bankası)

TÖ : Transfer Öğrenme

WMT : World Machine Translation Conference (Dünya Makine Çeviri Konferansı)

YSA : Yapay Sinir Ağı

ÖZET

Doktora Tezi

TÜRKÇE-İNGİLİZCE SİNİRSEL MAKİNE ÇEVİRİ SİSTEMİ

İlhami SEL

İnönü Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

110+IX sayfa

2024

Danışman: Prof. Dr. Davut HANBAY

Düşük kaynaklı dil çiftlerinde çeviri sistemlerinin geliştirilmesi, dil veri setlerinin yetersizliği nedeniyle önemli zorluklar içerir. Bu tür dillerde yeterli miktarda ve çeşitlilikte paralel külliyatın bulunmaması, modellerin eğitimi ve doğruluğu üzerinde olumsuz etkiler yaratır. Özellikle, Türkçe-İngilizce gibi dil çiftlerinde, morfolojik zenginlik ve dil bilgisel yapı farklılıkları, çeviri sistemlerinin hassasiyetini artıran faktörlerdir. Çeviri sistemlerinin bu zorlukların üstesinden gelebilmesi için ileri seviye makine öğrenimi tekniklerine ve veri artırma yöntemlerine ihtiyaç vardır. Bununla birlikte, düşük kaynaklı diller için transfer öğrenimi ve sıfırdan öğrenme gibi yöntemler de giderek önem kazanmaktadır. Bu alandaki ilerlemeler, çeviri kalitesini artırarak dil bariyerlerini aşmayı ve bilgiye erişimi küresel ölçekte yaygınlaştırmayı hedeflemektedir.

Bu tez çalışmasında, Türkçe-İngilizce dil çiftinde düşük kaynak sorununu ortadan kaldırabilmek için ilk olarak web kazıma yöntemleri ve cümle hizalama algoritmalarıyla 1.2 milyon paralel cümleye sahip büyük bir paralel külliyat oluşturulmuştur. Bu külliyat, Türkçe ve İngilizce arasındaki çeviri sistemlerinin eğitimi ve doğruluğu açısından önemli bir temel sağlamaktadır. Ayrıca külliyatın oluşturulması diğer düşük kaynaklı diller için de önemli bir yol önermektedir.

Ön eğitimli dil modelleri güncel doğal dil işleme uygulamalarında aktif olarak kullanılmaktadır. Bu sebeple sinirsel makine çeviri görevlerine dahil edilmesi için ek çalışmalar yapılmıştır. Bu çalışmalardan ilki Türkçe dil anlama görevlerinde farklı stratejilerle oluşturulmuş ön eğitimli dil modellerinin başarısının test edilmesidir. Oluşturulan farklı mimarilerle yapılan karşılaştırmalar sonucunda Bert dil modelinin Türkçe için başarılı sonuçlar elde ettiği görülmüştür. Dil modelleri üzerine yapılan ikinci çalışma da ise çok dilli ön eğitimden geçirilmiş modellerin çeviri sistemlerine uyarlanması üzerine deneyler yapılmıştır. Transfer öğrenme için parametre verimli olarak oluşturulan çeviri sistemi hesaplama maliyeti ve çeviri kalitesi açısından başarılı sonuçlar elde etmiştir.

Son olarak çeviri sisteminin daha doğal, akıcı ve dil bilgisel doğruluğu artırabilmek için çeviri sistemi oluşturulmuştur. Bu sistem de öz dikkat mimarisine sahip kodlayıcı kod çözücü mimarisi Türkçe-İngilizce çeviriler yapmak için kullanılmıştır. Ön eğitimli dil modeli ise çevirilerde akıcılığı artırmak için kullanılmıştır. Bu çeviri sistemi için yeni bir sığ füzyon yöntemi önerilmiştir. Önerilen yöntemde ilk çalışma da oluşturulan paralel külliyat ve sonraki çalışmalarda kullanılan dil modelleri ile geliştirilmiştir.

Anahtar Kelimeler: Düşük Kaynaklı Diller, Türkçe-İngilizce, Makine Çevirisi, Paralel Külliyat, Dikkat Mekanizması, Ön Eğitimli Dil Modeli

ABSTRACT

Phd. Thesis

TURKISH-ENGLISH NEURAL MACHINE TRANSLATION SYSTEM

İlhami SEL

Inonu University

Graduate School of Nature and Applied Sciences
Department of Computer Engineering

110+IX page

2024

Supervisor: Prof. Dr. Davut HANBAY

Developing translation systems for low-resource language pairs presents significant challenges due to the lack of adequate language datasets. The absence of a sufficient amount and variety of parallel corpora in such languages negatively impacts the training and accuracy of the models. Especially for language pairs like Turkish-English, morphological richness and differences in grammatical structures are factors that increase the sensitivity of translation systems. Advanced machine learning techniques and data augmentation methods are needed for translation systems to overcome these challenges. Moreover, methods such as transfer learning and zero-shot learning for low-resource languages are gaining importance. Advances in this field aim to improve translation quality, overcome language barriers, and disseminate access to information on a global scale.

In this thesis, to address the low-resource issue for the Turkish-English language pair, a large parallel corpus with 1.2 million parallel sentences was created using web scraping methods and sentence alignment algorithms. This corpus provides a significant foundation for the training and accuracy of translation systems between Turkish and English. Additionally, the creation of this corpus proposes an important path for other low-resource languages.

Pre-trained language models are actively used in current natural language processing applications. Therefore, additional studies have been conducted to incorporate them into neural machine translation tasks. The first of these studies is to test the success of pre-trained language models created with different strategies in Turkish language understanding tasks. Comparisons made with different architectures showed that the BERT language model achieved successful results for Turkish. In the second study on language models, experiments were conducted on adapting models pre-trained in multiple languages to translation systems. The translation system created with parameter-efficient transfer learning achieved successful results in terms of computational cost and translation quality.

Finally, to enhance the translation system's naturalness, fluency, and grammatical accuracy, a translation system was developed. This system used an encoder-decoder architecture with a self-attention mechanism for Turkish-English translations. The pre-trained language model was used to increase fluency in translations. A new shallow fusion method was proposed for this translation system. The proposed method was developed with the parallel corpus created in the first study and the language models used in the subsequent studies.

Keywords: Low-Resource Languages, Turkish-English, Machine Translation, Parallel Corpus, Attention Mechanism, Pre-trained Language Model.

1. GİRİŞ

Zamanın başlangıcından beri, dil insanlar arasında etkileşimin en önemli kaynağı olmuştur. Farklı toplumlar ve dolayısıyla farklı diller arasında çeviri yapılması bilimsel ve kültürel ilerlemelerde çok önemli katkılar sağlamıştır. Günümüzde internetin yaygınlaşmasıyla beraber her türden ve her dilden bilgiye ulaşmak oldukça kolaylaşmıştır. Bu durum bir avantaj olarak görülse de makine çevirisine olan ihtiyacı artırmıştır. Makine çevirisi (MÇ), insan müdahalesi olmadan bir konuşmanın veya metnin bir dilden başka bir dile çevrilmesidir[1]. Makine çeviri görevi ise Doğal Dil İşlemenin (DDİ) yoğun olarak çalışılan alt alanlarından birisidir. Makine çevirisi özellikle global internet sağlayıcıları ve sosyal medya şirketlerinin farklı dillerden bireyler arasında etkileşimi artırma ve dolayısıyla kullanıcı sayısı artırma gibi hedefleri doğrultusunda son yıllarda çok büyük ilerlemeler kaydetmiştir. İngilizce-Fransızca, İngilizce-Almanca gibi yüksek kaynaklı dil çifti olarak gösterilen dillerde çeviri kalitesi insan çevirisine yaklaşmış durumdadır. Fakat Düşük kaynaklı dil çiftleri [2–6] arasında çeviri kalitesinin artırılması için akademik çalışmalar yoğun bir şekilde devam etmektedir. Bu tez çalışmasında düşük kaynaklı diller arasında gösterilen Türkçe-İngilizce dil çifti üzerinde deneysel çalışmalar yapılmıştır.

Doğal dil işleme, yapay zekâ ve dilbilim çalışmalarının bir alt alanıdır. Doğal dil işleme de doğal dillerin kural ve yapılarının anlamlandırılması veya tekrardan üretilmesi amaçlanmaktadır. Metin sınıflandırma, duygu analizi, bilgi çıkarma, metin özetleme, makine çevirisi doğal dil işleme görevlerinden bazılarıdır. Özellikle son 10 yılda Evrişimsel Sinir Ağları (ESA), Yinelenen Sinir Ağları ve Öz dikkat mimarisi gibi derin öğrenme algoritmalarının geliştirilmesiyle doğal dil işleme görevlerinde yüksek başarımlar elde edilmiştir. Diziden diziye (Sequence to Sequence) metnin işlenmesi görevi olan makine çevirisi ise doğal dilin tekrar üretilmesi problemlerinden biridir. Bir makine çeviri sistemi oluşturabilmek için birbirinin çevirisi olan farklı dillerde cümlelere yani dizilere ihtiyaç vardır. Bu tür veri setleri Paralel külliyat (Corpus veya Corpora) olarak adlandırılmaktadır. Makine çevirilerinde kaliteyi belirleyen en önemli unsur paralel külliyattır. Yüksek kaynaklı dil çifti olarak adlandırılan dillerde paralel külliyatın boyutu 100 Milyon sayısına ulaşabilmektedir. Fakat dünya üzerinde 7000 üzerinde dil bulunmaktadır [7]. Her dil çifti için bu oranda paralel külliyatın bulunması mümkün değildir. Bu sebeple özellikle düşük kaynaklı diller için makine çeviri kalitesini artırmanın farklı yolları aranmaktadır. Paralel külliyat dışında o dile ait tek dilli metinlerde çevirilerde kullanılmaya çalışılmaktadır. Tek

bir dilde büyük miktarda metin elde edildiğinde dilde kullanılan kelimeler, bu kelimelerin nasıl kullanıldığı, cümlelerin yapısı vb. gibi bilgiler elde edilebilir. Denetimsiz makine çevirisi adı verilen yöntemle paralel veri olmadığında tek dilli veriler kullanılarak çeviri yapılabilmesi de farklı çalışma alanlarından birisidir[8].

Düşük kaynaklı diller dışında yüksek kaynaklı diller içinde makine çevirilerinde bazı temel problemler vardır. Bir dilde belirtilen bir deyim veya atasözü çoğu zaman gerçek anlamının dışında ek anlamlarda barındırır. Alaycı bir alt anlamı olan bir deyimin çevirisi yapılacağı zaman ekstra bilgilere de ihtiyaç duyulmaktadır. Bunun dışında doğal dil işlemedeki en büyük problem belirsizliktir. Sözcük anlamı, morfoloji, sözdizimsel özellikler ve roller, bir metnin farklı bölümleri arasındaki ilişki, bir çevirinin kalitesini belirleme de önemli unsurlardır. İnsanlar daha geniş bağlamı ve arka plan bilgisini göz önünde bulundurarak bu belirsizlikle başa çıkabilmektedir. Çeviri sistemleri için ise bu alanda çalışmalar hala devam etmektedir.

Mevcut makine çevirisi araştırmalarının amacı, mükemmel çeviri elde etmek değil, makine çevirisi sistemlerinin hata oranlarını azaltmaktır. Bir dilden başka bir dile çeviri çoğu zaman yaklaşık bir çeviridir. Çevirmenler seçim yapmak zorundadır. Farklı çevirmenler farklı seçimler yapabilir. Çeviri de ana hedef yeterlilik (Adequacy) ve akıcılıktır (fluency) [9]. Yeterlilik orijinal metnin anlamını korumak, akıcılık ise hedef dil de yazılmış herhangi bir metin gibi okunan bir çıktı metni üretmeyi gerektirir. Şarkı sözleri gibi metinlerde orijinal cümlenin anlamını tamamen korumak çeviriyi kalitesiz hale getirebilir. Yasal bir metin veya kullanım kılavuzu gibi çevirilerde ise akıcılık ile ilgili endişe duymak anlamsız olabilmektedir [9].

Makine çevirilerinde kaliteyi belirleyen en önemli unsur paralel külliyattır. Çevirmen yardımıyla bir dil çiftine ait büyük miktarda paralel külliyat oluşturmak işlem gücü bakımından oldukça zordur. Bunun yerine paralel külliyat oluşturmak için genellikle resmi hükümet yayınları, açık kaynak yazılım belgeleri, dini kitaplar gibi birbirinin çevirisi olan metinler kullanılmaktadır. Cümle hizalama algoritmaları ile bu metinler cümle cümle eşleştirilip paralel külliyat oluşturulmaya çalışılmaktadır. Web üzerinde açık kaynaklı olarak Europarl [10], OPUS [11], ve Paracrawl [12] gibi platformlar bazı dil çiftleri için oluşturduğu veri setlerini kullanıma açmaktadır. Bunun yanında Dünya Makine Çevirisi (World Machine Translation – WMT) [2,3,13] konferansları da her yıl düzenlenerek farklı dil çiftlerinde çeviri kalitesinin ölçülmesi için yarışmalar düzenlemektedirler. 2016, 2017 ve

2018 yılında Türkçe-İngilizce dil çifti de düşük kaynaklı diller arasında gösterilerek yarışmaya dahil edilmiştir.

1.1 Tezin Amacı

Bu tez çalışmasının amacı, düşük kaynaklı olarak kabul edilen Türkçe-İngilizce dil çifti için makine çevirilerinde kaliteyi artırmaktır. Yapılan çalışmalar doğrultusunda öncelikle yeni ve büyük veri seti oluşturarak düşük kaynak sorununa çözüm bulunması hedeflenmektedir. Bu doğrultuda makine çevirisi için mevcut yöntemler detaylı olarak incelenmiştir. Oluşturulacak çeviri sistemi için yinelenen sinir ağları, evrişimsel sinir ağları ve transformer tabanlı sinir ağları kullanılarak mimariler geliştirilmesi, geliştirilen modellerin yanı sıra akıcılığı artırmak için ön eğitimli dil modellerinin çeviri sistemlerine dahil edilmesi amaçlanmıştır.

Ayrıca çok dilli olarak eğitilen modellerin düşük kaynaklı dillerde çeviri sistemi olarak kullanılabilmesi için ince ayar stratejilerinin geliştirilmesi, bu süreçte geliştirilen yöntemlerin, mimarilerin ve veri setlerinin açık kaynak olarak literatüre kazandırılması hedeflenmiştir.

1.2 Tezin Gerekçeleri

Bu tez kapsamında düşük kaynaklı Türkçe-İngilizce dil çifti için çeviri kalitesinin ve akıcılığının artırılması maksadıyla çalışmalar yapılmıştır. Tez çalışmasının gerekçeleri şu şekildedir:

Düşük Kaynaklı Dil Çiftleri ve Veri Eksikliği: Türkçe-İngilizce dil çiftinde yeterli miktarda ve çeşitlilikte paralel külliyatın bulunmaması, çeviri sistemlerinin eğitimi ve doğruluğu üzerinde olumsuz etkiler yaratmaktadır. Düşük kaynaklı dillerde paralel veri eksikliği, makine çevirisi modellerinin dilin karmaşıklıklarını öğrenmesini zorlaştırmakta ve çeviri kalitesini düşürmektedir. Bu eksiklik, modellerin geniş dil bilgisi yelpazesinde genel performansını etkileyen önemli bir engel teşkil etmektedir.

Morfolojik ve Dilbilgisel Zorluklar: Türkçe'nin morfolojik zenginliği ve dilbilgisel yapısının İngilizce ile farklılıkları, çeviri sistemlerinin geliştirilmesinde ek zorluklar oluşturmaktadır. Türkçe, eklemeli bir dil olarak, kelime köklerine çeşitli ekler eklenerek anlamın ve yapının değiştirilmesini sağlar. Bu durum, İngilizce gibi daha az morfolojik karmaşıklığa sahip olan diller ile karşılaştırıldığında çeviri sistemlerinin doğru anlam ve yapı

analizini zorlaştırır. Ayrıca, Türkçe'nin serbest kelime sırası sebebiyle, çevirinin anlamını önemli ölçüde etkilediği ek zorluklar vardır.

İleri Makine Öğrenimi Tekniklerine İhtiyaç: Bu zorlukların üstesinden gelebilmek için ileri seviye makine öğrenimi tekniklerine ve veri artırma yöntemlerine ihtiyaç duyulmaktadır. Makine çevirisi modellerinin, düşük kaynaklı dil çiftlerinde başarılı olabilmesi için transfer öğrenme, veri sentezi ve veri çoğaltma gibi güncel tekniklerin kullanılması gerekmektedir. Bu teknikler, sınırlı veri setlerinden daha fazla bilgi çıkarılmasını sağlayarak modellerin performansını artırır ve daha doğru çeviriler üretmesine olanak tanır.

Transfer Öğrenme ve Sıfırdan Öğrenme Yöntemleri: Düşük kaynaklı dillerde transfer öğrenme ve sıfırdan öğrenme (zero-shot learning) gibi yöntemlerin önemi giderek artmaktadır. Transfer öğrenme, yüksek kaynaklı dillerde eğitilmiş modellerin, düşük kaynaklı diller için uyarlanması sürecidir ve bu sayede verimlilik artırılabilir. Sıfırdan öğrenme yöntemleri ise, çeviri sistemlerinin hiç görülmemiş dil çiftleri için bile başarılı çeviriler yapabilmesini sağlar. Bu yöntemler, düşük kaynaklı dillerdeki veri eksikliğini gidermek ve çeviri kalitesini iyileştirmek için umut verici çözümler sunmaktadır.

1.3 Tezin Çıktıları

Bu tez çalışmasında Türkçe-İngilizce makine çevirisi için paralel külliyat ve derin öğrenme tabanlı sistemler geliştirilmiştir. Oluşturulan veri setleri ve mimariler açık kaynak olarak literatüre ve araştırmacılara sunulmuştur. Hazırlanan bilimsel çalışmalar aşağıda verilmiştir.

- Sel, İ., Üzen, H., & Hanbay, D. (2021). Creating a Parallel Corpora for Turkish-English Academic Translations. Computer Science, (Special), 335-340.
- Sel, İ., & Hanbay, D. (2021). Gender Identification from Turkish Tweets Using Pre-Trained Language Models. Firat University Journal of Engineering Science, 33(2), 675-684. https://doi.org/10.35234/fumbd.929133
- Sel, İ., & Hanbay, D. (2022). Fully attentional network for low-resource academic machine translation and post editing. Applied Sciences, 12(22), 11456.
- Sel, İ., & Hanbay, D. (2024). Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation, (Hakem değerlendirmesinde)

1.4 Tezin Yenilikçi Yönü ve Ar-Ge Niteliği

Bu tez çalışması kapsamında literatürde en yüksek paralel veri içeren veri setlerinden biri oluşturulmuştur. Oluşturulan bu veri seti kullanılarak üç farklı derin öğrenme mimarisi geliştirilmiştir. Geliştirilen bu yöntemlerin sonuçları literatürde yer alan diğer çalışmalar ile karşılaştırılmıştır. Tam dikkat katmanıyla oluşturulmuş sığ füzyon yöntemi çeviri sistemindeki doğruluğu ve akıcılığı artırdığı Bleu puanı ile gösterilmiştir.

1.5 Tezin Organizasyonu

Bu tez çalışması beş bölümden oluşmaktadır. Bölüm 1'de, tez konusu hakkında bilgiler, tezin amacı, gerekçeleri, çıktıları ve literatüre yapılmış olan katkılar ve tezin organizasyonu verilmiştir. Bölüm 2'de Türkçe çeviri sistemleri için literatürde yapılmış çalışmalar detaylıca incelenmiştir. Bölüm 3'te Makine çevirisi ile ilgili güncel yapılan çalışmalar, mimariler ve değerlendirme ölçütleri verilmiştir. Bölüm 4'te doğal dil işleme de son teknoloji olarak gösterilen Transformer mimarisi detaylıca anlatılmıştır. Bölüm 5'te ise tez sürecinde yapılan çalışmalar sunulmuştur. Son olarak 6. Bölümde Sonuçlar ve gelecek çalışmalara yön verecek önemli noktalar belirtilmiştir.

2. İLGİLİ ÇALIŞMALAR

Türkçe morfolojik açıdan oldukça zengin bir dildir. Özellikle sondan eklemeli yapısı sebebiyle kelime sayısı eklerle artırılarak çeviri sistemlerinde zorluklara yol açmaktadır. Literatür incelendiğinde Türkçe dili üzerine yapılan çalışmalar genel olarak iki başlık altında toplanabilir. Bunlar İstatistiksel Çeviri Sistemleri (İÇS) ve Düşük kaynak dil çifti olarak test edilen Sinirsel Makine Çeviri (SMÇ) sistemleridir. İstatistiksel yöntemler makine çevirilerinin ilk zamanlarında oluşturulmaya başlanmıştır ve başarımları güncel sinirsel yöntemlere göre düşüktür. Sinirsel makine çeviri sistemleri kullanılarak oluşturulan Türkçe-İngilizce (Tr-En) çeviri sistemleri genel olarak düşük kaynaklı dillerde çeviri kalitesini artırmaya yöneliktir. IWSLT2013 [14], WMT16-17-18 [2,3,15] konferanslarında Tr-En dilleri arasında çeviri için yarışmalar düzenlenmiştir. Genel olarak çalışmalarda veri seti olarak eğitim için 200K veri çiftine sahip paralel külliyat ve 3K veri çiftine sahip test veri seti kullanılmıştır. Yüksek kaynaklı dil çiftleri İngilizce-Almanca (En-De) veya İngilizce-Fransızca (En-Fr) dilleri için veri seti boyutunun yaklaşık 50M paralel cümleden oluştuğu bilinmektedir. Tr-En dil çifti için yüksek kaynaklı dil çiftlerinin yakaladığı çeviri kalitesine ulaşabilmek için hem veri seti artırma hem de farklı mimari çalışmaları yoğun olarak sürmektedir. Bu tez çalışmasında da Türkçe-İngilizce dil çiftleri düşük kaynaklı dil çifti olarak ele alınmış ve sinirsel makine çeviri yöntemleri geliştirilerek çeviri kalitesi artırılmaya çalışılmıştır.

2.1 TR-EN İstatistiksel Çeviri Sistemleri

Oflazer ve Ark. (2007)[16] ve Bisazza ve Ark. (2009)[17] istatistiksel çeviri uygulamaları ile Türkçe-İngilizce makine çevirisi için ilk büyük çalışmaları gerçekleştirmişlerdir. Bu çalışmalarda morfolojik kelime bölütleme uygulayarak cümle sıralaması ve nadir kelime problemlerini çözmeye çalışmışlardır. Mermer ve Ark. (2010) [18] Moses "Statistical Machine Translation" aracını kullanarak, Uluslararası Konuşma Dili Çevirisi Konferansı (IWSLT2010) değerlendirme yarışmasında hiyerarşik ifade tabanlı modeller sunmuşlardır. Yeniterzi ve Ark. (2010) [19] Türkçe kelimeleri morfemlere ayırmadan kelime formuna eşdeğer bir gösterim kullanarak istatistiksel çeviri yöntemini oluşturmuş ve 23.78 Bleu puanı elde etmiştir. IWSLT 2013 konferansı için hazırlanan

çalışma, morfem düzeyinde ve hiyerarşik deyim tabanlı istatistiksel çeviri sistemi ile 17.14 BLEU puanına ulaşmayı başarmıştır [14]. Bakay ve Ark. 2019 [20] yılında ağaç tabanlı İngilizce-Türkçe istatistiksel çeviri sistemi getirilmiştir ve hazırlanan sistem 21.4 Bleu puanına ulaşmıştır.

2.2 TR-EN Sinirsel Makine Çeviri Sistemleri

Yılmaz ve Ark. (2014) [21] Biçim birimsel (morphem) temsil kullanarak ve hiyerarşik öbek tabanlı (Phrase Based Hierarchical SMT) makine çevirisi ile en iyi n adet çeviri oluşturmuştur. Sonrasında yinelemeli sinir ağları yardımıyla yeniden skorlama ve sistem birleştirme kullanarak çeviri kalitesini artırmaya çalışmıştır. Hibrit olarak oluşturulan bu yöntem yinelemeli ağların Türkçe çeviri sistemlerinde ilk kullanım örneklerinden birisidir.

Gulcehre ve Ark. (2015) [22] tarafından önerilen sinirsel makine çeviri sistemi çeviri kalitesini artırabilmek için ön eğitimli dil modellerini kullanan ilk çalışmadır. Çalışma da RNNLM dil modelleri hedef dil tarafında çeviri sistemine dahil edilmiştir. Shallow fusion ve Deep Fusion adını verdikleri yöntemleri düşük kaynaklı TR-EN, Çekçe-İngilizce (CS-EN) ve yüksek kaynaklı (DE-EN) gibi dil çiftleri üzerinde test etmişlerdir.

Sennrich ve Ark. (2016) [23] tarafından hazırlanan çalışma da hedef tarafında kullanılan dil modeli yerine tek dilli veriler kullanarak çeviri kalitesini artırmaya çalışmıştır. Ayrıca tek dilli eğitim verilerini otomatik geri çeviri yöntemiyle artırarak ek paralel eğitim verisi olarak kullanmışlardır. Testlerde düşük kaynaklı TR-EN ve yüksek kaynaklı En-De dil çiftlerini kullanmışlardır.

Zoph ve Ark. (2016) [24] düşük kaynak dil çiftleri için transfer öğrenme yöntemini kullanarak sinirsel makine çeviri sisteminin kalitesini artırmak istemişlerdir. Önerilen yöntem 2 adımdan oluşur. Önce yüksek kaynaklı bir dil çifti (ana model) eğitilir. Sonrasında düşük kaynaklı dil çiftinin eğitimini başlatmak ve sınırlandırmak için ana modelden öğrenilen parametreler kullanılır. Çalışmalarında ana model olarak FR-EN, İspanyolca-İngilizce (Sp-En) ve De-En dil çiftini kullanmışlardır. TR-EN dil çiftini ise düşük kaynaklı dil çifti olarak kullanmışlardır. Ayrıca SMÇ mimarisinde Bahdanau ve Ark. [25] tarafından önerilen dikkat mekanizmasını da kullanmışlardır.

Türkçe gibi morfolojik olarak zengin diller makine çevirilerinde sözcük hizalama hataları, veri seyrekliği ve birden çok ek olması sebebiyle çeviri kalitesinin geliştirilmesi

gereken diller arasında gösterilmektedir. Sözcük seviyesinde hizalamalar sözcükleri ve morfemleri ayırt edemediğinden düşük kaliteli hizalama sağlamaktadır. Bu sorunun üstesinden gelebilmek için morfem (biçim birim) kullanarak çeviri sistemleri oluşturmak sözcük dağarcığını azaltarak veri seyrekliğinin üstesinden gelir ve daha kaliteli çeviriler üretebilir. Li ve Ark. (2016) [26] tarafından yapılan çalışmada morfolojik açıdan zengin diller incelenerek morfem düzeyinde Özbekçe-İngilizce ve Türkçe-İngilizce dilleri arasında hizalanmış veri setleri oluşturmuşlardır.

Nguyen ve Ark. (2017) [27] düşük kaynaklı dil çiftlerinde çeviri kalitesini artırmak için Zoph ve Ark. [24] tarafından daha önce önerilen yöntemi geliştirmişlerdir. Transfer öğrenme yoluyla yapılan çalışmada, yüksek kaynaklı çeviri sistemi kullanılarak düşük kaynaklı çeviri sistemine bilgi aktarımı sağlanmıştır. Giriş temsili olarak Byte Pair Encoding (BPE) kullanılarak önceki çalışma geliştirilmiştir. Model, dikkat mekanizmasına sahip sinirsel makine çeviri sistemi kullanılarak oluşturulmuştur. Düşük kaynak dil çifti olarak Türkçe-İngilizce ve Özbekçe-İngilizce üzerine deneyler yapılmıştır.

Currey ve Ark. (2017) [28]düşük kaynaklı dil çiftlerinde çeviri sistemine tek dilli verilerin dahil edilmesi üzerine çalışmalar yapmışlardır. Önerdikleri yöntemde hedef dildeki tek dilli metinden bir çift metin oluşturulur, böylece her bir kaynak cümle hedef cümle ile aynı olur. Bu kopyalanan veriler daha sonra paralel külliyat ile karıştırılır ve SMÇ sistemi, iki giriş dilini ayırt edecek herhangi bir meta veri olmaksızın normal şekilde eğitilir. TR-EN ve Romence-İngilizce (Ro-En) dil çiftleri üzerinde deneyler yapılmıştır ve Bahdanau [25] tarafından önerilen dikkat tabanlı sinirsel çeviri modeli ile BPE [29] kullanılmıştır.

Kodlayıcı-kod çözücü tabanlı diziden diziye modellemede, en yaygın uygulama, kodlayıcı ve kod çözücüde bir dizi tekrarlayan, evrişimli veya ileri beslemeli katmanları artırmaktır. Her yeni katmanın eklenmesi, dizi oluşturma kalitesini artırırken, aynı zamanda parametre sayısında da önemli bir artışa yol açmaktadır. Dabre ve Ark. (2019) [30] parametreleri tüm katmanlar arasında paylaşmayı ve böylece tekrarlayan bir diziden diziye modele yol açmayı önermişlerdir. Deneysel olarak, önemli ölçüde daha az parametreye sahip olmasına rağmen, tekrar tekrar tek bir katmanı 6 kez istifleyen bir modelin çeviri kalitesinin, 6 farklı katmanı istifleyen bir modelinkine yaklaştığını göstermişlerdir. Temel öz dikkat mimarisini kullanan bu çalışma, Türkçe-İngilizce dil çifti üzerinde deneyler yapmıştır.

Luo ve ark. (2019) [31] düşük kaynaklı dil çiftleri için sinirsel makine çeviri (SMÇ) sistemlerinde transfer öğrenme yöntemlerini geliştirmeye çalışmışlardır. Önerdikleri

mimaride, SMÇ modeli sırasıyla ilgisiz yüksek kaynak dil çifti, benzer ara dil çifti ve düşük kaynak dil çiftinde eğitilir. Parametreler, katman katman aktarılır ve ince ayar yapılır. Bu hiyerarşik transfer öğrenme mimarisi, yüksek kaynaklı dillerin veri hacmi avantajlarını ve aynı kökenli dillerin sözdizimsel benzerlik avantajlarını birleştirir. Öz dikkat mimarisini kullanarak sırasıyla İngilizce-Çince, Türkçe-İngilizce ve Uygurca-Çince dil çiftlerini kullanmışlardır. Giriş temsil yöntemi olarak BPE kullanmışlardır.

Liu ve ark. (2020)[32] çok dilli gürültü giderme ön eğitiminin çeşitli makine çevirisi görevlerinde önemli performans artışları sağladığını göstermiştir. "mBART" olarak bilinen modelleri, birçok dilde büyük ölçekli tek dilli külliyat üzerinde önceden eğitilmiş diziden diziye gürültü giderici bir otomatik kodlayıcıdır. Bu model, göreve özgü herhangi bir değişiklik yapılmadan, denetimli (hem cümle düzeyinde hem de belge düzeyinde) ve denetimsiz makine çevirisi için doğrudan ince ayar yapılmasına olanak tanır. Düşük kaynaklı makine çevirisi için 12'ye kadar Bleu puanı ve birçok belge düzeyinde ve denetlenmeyen model için 5'in üzerinde Bleu puanı elde etmişlerdir.

Sun ve ark. (2021) [33] mevcut makine çevirisi (SMÇ) yöntemlerinin yüksek kaynaklı dillerde güçlü olduğunu, ancak düşük kaynaklı dillerde ciddi zorluklar yaşadığını belirtmişlerdir. Çalışmalarında, düşük kaynaklı dillerdeki cümle çiftlerinin eksikliğini azaltmak için rakip ve transfer öğrenme tekniklerini kullanmışlardır. Yeni bir düşük kaynaklı, tartışmalı, diller arası model önermişlerdir. Bu model, bir üreteç ve ayırıcıdan oluşur. Üreteç, kaynaktan hedef dillere çeviriler üreten bir diziden diziye modeldir. Ayırıcı ise makine ve insan çevirileri arasındaki farkı ölçer.

Maimati ve ark. (2022) [34] düşük kaynaklı dillerde veri artırma yöntemleri üzerinde çalışmışlardır. Veri artırma çalışmalarında en sık kullanılan yöntemler, metindeki bazı kelimeleri rastgele örnekleyerek, çıkararak veya değiştirerek sahte bütünlük oluşturmaktır. Yaptıkları çalışma da kısıtlı örnekleme yöntemini tanıtarak derlemi genişletmeye çalışmışlardır. Ayrıca, artırma sonrasında daha yüksek kaliteli verileri seçmek için değerlendirme çerçevesi de oluşturmuşlardır. Yani sözdizimsel ve anlamsal hataları azaltmak için ayrıştırıcı alt modelini kullanmışlardır. Deneysel sonuçları, yöntemlerinin, dört külliyattan sekiz dil çiftinde hem küçük hem de büyük ölçekli külliyatlarda diğer yöntemlerden 2,38-4,18 Bleu puanı ile daha iyi performans gösterdiğini belirtmişlerdir.

Longchao ve ark. (2022) [35] düşük kaynaklı makine çevirisi (SMÇ) modellerinin genellikle yalnızca paralel cümle çiftlerine dayandığını ve bu nedenle düşük kaynak

durumlarında performanslarının düştüğünü belirtmişlerdir. Çalışmalarında, sözdizimi gibi tek dilli bilgilerin açıkça dahil edilmesinin SMÇ için etkili olduğunu, ancak mevcut yaklaşımların bu potansiyeli tam olarak kullanamadığını vurgulamışlardır. Bu sorunu çözmek için, kaynak dildeki sözdizimsel bilgiyi çok kafalı öz-dikkat mekanizması ile birleştiren bir sinir ağı modeli olan "sözdizimi-grafik güdümlü öz dikkati" önermişlerdir. Ayrıca, kaynak cümlenin anlamlı temsillerini korumak için, öz-dikkat alt ağlarında rastgele düğüm bırakma stratejisi kullanmışlardır.

Shiyu ve ark. (2022) [36] video destekli makine çevirisi üzerinde çalışmışlardır. Amaçları, video bilgilerini ek uzay-zamansal bağlam olarak kullanarak bir kaynak dil açıklamasını hedef dile çevirmektir. Görsel kavramların, dillerin hizalanması ve çevirisinin iyileştirilmesi açısından faydalı olduğunu ve genellikle ihmal edildiğini belirtmişlerdir. Bu amaçla, video destekli makine çevirisi sorununu çözmek için çift düzeyli geri çeviri yöntemini kullanmışlardır. Önce, kaba taneli semantiği elde etmek için cümle düzeyinde geri çeviriyi uygulamışlardır. Daha sonra, ince taneli semantik tutarlılığı ve indirgene bilirliği keşfetmek için video kavramı düzeyinde geri çeviri modülü önermişlerdir. İki gerçek dünya veri kümesi üzerinde deney yaparak, önerdikleri çözümün etkinliğini göstermişlerdir.

Görgün, O. (2022) [37] yapmış olduğu doktora tez çalışmasında Türkçe-İngilizce dil çifti için istatistiksel sözdizimi ağacı tabanlı makine çevirisi için paralel derlem çalışmaları yapmıştır. 17,000 ve 8300 paralel cümle olmak üzere iki adet derlem oluşturmuştur. Derlemi oluştururken; çevrilmiş ağaçların alt ağaçlarının yeniden sıralamış ve kelime değişimi ile sınırlandırmıştır. İngilizce ağaçları Penn Treebank'tan el ile dönüştürerek; çevrilmiş kelimelerin morfolojik analizi ve hedef ağacın morfolojik olarak zenginleştirilmesi yolunu izlemiştir. Çeviri tutarlılığı amacı ile bir yazılım araçları seti de geliştirmiştir. Çalışmalarında 12.8 ve 26.8 Bleu puanlarını elde etmiştir.

İncelen çalışmalar genel olarak çeviri kalitesini ölçebilmek için Bleu puanını kullanmışlardır. Fakat Bleu puanının hesaplanması için farklı yöntemler kullanılmaktadır. Bu yüzden incelenen çalışmalarda Bleu kıyaslamasına gidilmemiştir. Tez çalışması süresince incelenen çalışmalar, yöntemler ve kullanılan data setleri Çizelge 2.1' de özet olarak verilmiştir.

Çizelge 2.1: Geçmişten günümüze Türkçe – İngilizce makine çeviri sistemleri.

Çalışma	Yöntem	Veri Seti Boyutu
Turhan, C. K. (1997) [38]	Kural Tabanlı Makine Çeviri Sistemi	
Altıntaş, K. (2001). (Yüksek Lisans Tezi) [39]	Sonlu Durum Makinesi, Kural Tabanlı Makine Çeviri Sistemi	
Durgar El-Kahlout, İ., & Oflazer, K. (2006). [40]	İstatistiksel Makine Çeviri Sistemi	20K
Oflazer, K., & Durgar El-Kahlout, İ. (2007). [16]	İstatistiksel Makine Çeviri Sistemi	45K
Alp, N. D., & Turhan, C. (2008) [41]	Örnek Tabanlı Makine Çeviri Sistemi	140
Türe, F. (2008) (Yüksek Lisans Tezi) [42]	Kural tabanlı + istatistiksel (Hibrit)	
Kopru, S. (2009). [43]	İstatistiksel Makine Çeviri Sistemi	20K-IWSLT2009
Bisazza, A., & Federico, M. (2009). [17]	Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K-IWSLT2009
Durgar El-Kahlout, İ. (2009) Doktora Tezi [44]	İstatistiksel Makine Çeviri Sistemi	45K
El-Kahlout, I. D., & Oflazer, K. (2009). [45]	Morfem Düzeyinde Yeniden Sıralama, İstatistiksel Makine Çeviri Sistemi	45K
Matusov, E., & Kopru, S. (2010). [46]	Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K-IWSLT2010
Mermer, C., Kaya, H., & Doğan, M. U. (2010). [18]	Morfolojik Segmentasyon, Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K-IWSLT2010
Yeniterzi, R., & Oflazer, K. (2010). [19]	Syntactic Etiketleme, Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	50K
Mermer, C. (2010). [47]	Morfolojik Segmentasyon, Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K
Wu, C., Xia, F., Deleger, L., & Solti, I. (2011). [48]	İstatistiksel Makine Çeviri Sistemi	5K
Mermer, C., & Saraclar, M. (2011). [49]	Morfolojik Segmentasyon, Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K
Mermer, C., & Saraçlar, M. (2011) [49]	Bayes Çıkarım Yöntemi, İstatistiksel Makine Çeviri Sistemi	20K
Görgün, O., & Yıldız, O. T. (2012). [50]	Morfolojik Düzenleme, İstatistiksel Makine Çeviri Sistemi	20K
Yılmaz, E., El-Kahlout, I. D., Aydın, B., Özil, Z. S., & Mermer, C. (2013). [14]	Hiyerarşik Phrase Tabanlı İstatistiksel Makine Çeviri Sistemi	20K, 173K-IWSLT2013

Çizelge 2.2 (devam): Geçmişten günümüze Türkçe – İngilizce makine çeviri sistemleri.

Çalışma	Yöntem	Veri Seti Boyutu
Mermer, C., Saraçlar, M., & Sarikaya, R. (2013). [51]	Bayes Çıkarım Yöntemi, Gibbs Örneklem, İstatistiksel Makine Çeviri Sistemi	20K
Eyigöz, E., Gildea, D., & Oflazer, K. (2013). [52]	Kelime-Morfem Hizalama, İstatistiksel Makine Çeviri Sistemi	50K
Yıldırım, E., & Tantuğ, A. C. (2013). [53]	Çıktının Yeniden Sıralanması, Google Translate	Test:720
Yıldız, O. T., Solak, E., Görgün, O., & Ehsani, R. (2014). [54]	Paralel Treebank (Ağaç Tabanlı) Veri Seti Oluşturulması	5K
Biçici, E., & Yuret, D. (2014). [55]	Eğitim Seti Özellik Seçimi, İstatistiksel Makine Çevirisi	352K
Yılmaz, E., & El-Kahlout, I. D. (2014). [21]	TSA, Sinirsel Makine Çeviri Sistemi	160K-SETIMES, 160K-TED
Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H. C., & Bengio, Y. (2015) [22]	Dil Modeli (LSTM), TSA, Sinirsel Makine Çeviri Sistemi	160K-SETIMES
Sennrich, R., Haddow, B., & Birch, A. (2015). [23]	Tek Dilli Veri, TSA, Sinirsel Makine Çeviri Sistemi	20K
Zoph, B., Yuret, D., May, J., & Knight, K. (2016) [24]	Transfer Öğrenme, Dikkat Mekanizması + Sinirsel Makine Çevirisi	20K
Li, X., Tracey, J., Grimes, S., & Strassel, S. (2016) [26]	Morfem Hizalamaya Bağlı Paralel Külliyat Oluşturulması	
Ding, S., Duh, K., Khayrallah, H., Koehn, P., & Post, M. (2016). [56]	Phrase, Hiyerarşik Phrase, Söz Dizimi Tabanlı İstatistiksel Makine Çeviri Sistemi	200K-WMT2016
Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., & Bengio, Y. (2017). [57]	Çok Dilli Sinirsel Makine Çeviri Sistemi	785K
García-Martínez, M., Caglayan, O., Aransa, W., Bardet, A., Bougares, F., & Barrault, L. (2017). [58]	BPE, Sinirsel Makine Çeviri Sistemi	200K-WMT2017
Nguyen, T. Q., & Chiang, D. (2017). [27]	Transfer Öğrenme, BPE, Sinirsel Makine Çevirisi	200K
Ataman, D., Negri, M., Turchi, M., & Federico, M. (2017). [59]	Kelime Azaltma, Sinirsel Makine Çevirisi	283K
Currey, A., Miceli-Barone, A. V., & Heafield, K. (2017) [28]	Veri Artırma, Sinirsel Makine Çevirisi	207K
Shen, Y., Tan, X., He, D., Qin, T., & Liu, T. Y. (2018). [60]	Dense Sinir Ağı, Dikkat Mekanizması, Sinirsel Makine Çevirisi	200K
Ataman, D., & Federico, M. (2018).[61]	BPE, LMVR, Dikkat Mekanizması, Sinirsel Makine Çevirisi	135K

Çizelge 2.3 (devam): Geçmişten günümüze Türkçe – İngilizce makine çeviri sistemleri.

Çalışma	Yöntem	Veri Seti Boyutu
Deng, Y., Cheng, S., Lu, J., Song, K., Wang, J., Wu, S., & Chen, B. (2018) [62]	Veri Artırma, Aç Gözlü Arama, Transformer	207K-WMT2018
Marie, B., Wang, R., Fujita, A., Utiyama, M., & Sumita, E. (2018). [63]	Tek Dilli Veri, BPE, Transformer	207K-WMT2018
Schamper, J., Rosendahl, J., Bahar, P., Kim, Y., Nix, A., & Ney, H. (2018) [64]	Veri Artırma, İnce Ayar, Transformer	207K-WMT2018
Dabre, R., & Fujita, A. (2019). [30]	Parametre Paylaşımı, Transformer 208K	208K
Luo, G., Yang, Y., Yuan, Y., Chen, Z., & Ainiwaer, A. (2019) [31]	Transfer Öğrenme + BPE + Transformer	200K
Aji, A. F., Bogoychev, N., Heafield, K., & Sennrich, R. (2020) [65]	Transfer Öğrenme + BPE+ Transformer	207K
Behnke, M., & Heafield, K. (2020). [66]	Dikkat Başlığı Budama + Transformer	200K
Yılmaz, S. (2020). [67]	Paralel Külliyat Oluşturma, İstatistiksel Makine Çevirisi	200K
Escolano, C., Costa-Jussà, M. R., & Fonollosa, J. A. (2021). [68]	Varyasyonel Auto Encoder, Transformer	200K
Liu, Q., Kusner, M., & Blunsom, P. (2021). [69]	Veri Artırma, Dil Modeli , Transformer	200K
Yirmibesoglu, Z. (2021). (Yüksek Lisans Tezi) [70]	Geri Çeviri (Back Translation), BPE, Transformer	656K, 2.2M
Ciora, C., Iren, N., & Alikhani, M. (2021). [71]	Cinsiyet Ön Yargı Araştırması, Google Translate	
Görgün, O., & Yildiz, O. T. (2022). [72]	İstatistiksel Sözdizimi Ağacı Tabanlı Makine Çevirisi	17K
Ata, M., & Debreli, E. (2021). [73]	İngilizce Öğreniminde Online Çeviri sistemlerinin incelenmesi, Google Translate	
Yirmibeşoğlu, Z., & Güngör, T. (2022). [74]	Geri Çeviri (Back Translation), BPE, Transformer	656K, 4.5M
Görgün, O. (2022) (Doktora Tezi) [37]	İstatistiksel Sözdizimi Ağacı Tabanlı Makine Çevirisi	17K

2.3 Türkçe-İngilizce Paralel Külliyatlar

Türkçe-İngilizce arasında çeviri sistemi oluşturabilmek için kullanılabilecek veri setleri sınırlıdır. Bu yönüyle Türkçe düşük kaynak dil olarak tanımlanmaktadır. Tatoeba

paralel veri seti [75], çeşitli dillerde metinlerin çeviri eşlerini içeren bir veri kümesidir. Tatoeba Projesi, gönüllülerin katkılarıyla oluşturulan ve dünya genelinde birçok dilde cümle örneklerini barındıran bir dil veri tabanıdır. Paralel veri seti, bu cümlelerin farklı dillerdeki karşılıklarını içerir ve makine çevirisi, dil modelleme, doğal dil işleme gibi alanlarda kullanılabilir. Bu veri seti, özellikle düşük kaynaklı diller için büyük önem taşır, çünkü birçok dil için yeterli miktarda çeviri verisi bulmak zor olabilir. Tatoeba paralel veri seti, bu tür dillerde yapılan çalışmalara katkı sağlamak amacıyla geniş bir yelpazede dil çiftlerini kapsar.

Wikimedia, WikiMatrix gibi paralel külliyatlar, Wikimedia projeleri (özellikle Wikipedia) tarafından oluşturulan çok dilli metinlerin çevirilerini içeren bir veri kümesidir. Bu veri seti, farklı dillerdeki Wikipedia makalelerinin çevirilerinden oluşur ve makine çevirisi, doğal dil işleme ve dil modelleme gibi alanlarda araştırma yapmak için kullanılır. Wikimedia paralel külliyat, özellikle geniş kapsamlı ve çeşitli dillerde metinler içermesi nedeniyle değerli bir kaynak olarak kabul edilir. Bu veri seti, dil çiftleri arasındaki metinsel uyumu incelemek ve çeviri modellerini eğitmek için kullanılabilir. Ayrıca, dil bilimciler ve araştırmacılar için dilin yapısal ve anlamsal özelliklerini karşılaştırmak için de önemli bir kaynaktır. Türkçe-İngilizce dil çifti için ulaşılabilen paralel külliyatlar Çizelge 2.2'de gösterilmiştir.

Çizelge 2.4: Açık kaynak Türkçe İngilizce paralel külliyatlar.

Paralel Külliyat	Cümle Sayısı	Türkçe Jeton Sayısı	İngilizce Jeton Sayısı
Tatoeba	676,920	4,497,091	3,317,786
wikimedia	668,099	16,897,660	13,158,338
QED	482,964	4,628,021	6,667,098
WikiMatrix	477,736	6,939,641	8,579,371
SETimes	207,678	3,654,669	4,428,278

3. MAKİNE ÇEVİRİSİ

Makine çevirisi Doğal dil işlemenin alt görevlerinden birisidir. Makine çevirisi bilgisayarlar yardımıyla metin veya konuşmayı insan müdahalesi olmadan bir dilden diğerine dönüştüren çalışma alanıdır. Makine çevirisi akademik iletişim ve uluslararası iş müzakereleri gibi birçok pratik uygulamada işçilik maliyetlerini azaltabilmek için büyük bir potansiyele sahiptir. Makine çevirisi son yıllarda çok büyük gelişim gösterse de uzun yıllardır çalışılan bir alandır. Makine çevirisi araştırmaları 1951 yılında bu alanda ilk araştırmacı olan Yehoshua Bar-Hillel'in MIT'de ilk Uluslararası Makine Çevirisi Konferansı'nı düzenlemesiyle başlamıştır [76]. Zaman içerisinde Kural Tabanlı Makine çevirisi [77], İstatistiksel Makine Çevirisi [78,79] ve Sinirsel Makine Çevirisi [80] olmak üzere gelişimini sürdürmüştür. Tezin bu bölümünde makine çevirisinin tarihsel gelişimi ile birlikte güncel makine çevirisi yaklaşımlarından detaylıca bahsedilecektir.

3.1 Kural Tabanlı Makine Çeviri Sistemi

Kural tabanlı makine çevirisi makine çevirisindeki ilk yöntemdir. Bu yöntem tüm farklı dillerin aynı anlamı temsil eden bir sembolü olduğu hipotezine dayanır. Bu yöntemde bir dildeki bir kelime başka bir dilde aynı anlama gelen başka bir kelimeyle eşleştirilmektedir. Yani bu yöntemde çeviri süreci kaynak cümlede kelime değişimi olarak ele alınmaktadır. Farklı diller aynı cümle anlamını farklı kelime sıralarında temsil ettiğinden kelime değiştirme yöntemi her iki dilin söz dizim kurallarını dikkate almak zorundadır. Kural tabanlı yöntem teoride uygulanabilir görünmesine rağmen uygulama da büyük problemler yaşamıştır. Bir cümlenin uyarlanabilir kurallarının belirlenmesi hesaplama açısından yetersiz olabilmektedir. Ayrıca dil bilgisi kurallarının düzenlenmesi de zordur, çünkü dil bilimciler dilbilgisi kurallarını özetlemektedir ve bir dilde çok fazla söz dizim kuralı bulunmaktadır. İki söz dizim kuralının birbiriyle çakışması bile mümkündür [76]. Bunun yanında kural tabanlı çeviri sistemlerinin en ciddi dezavantajlarından biri de çeviri sürecinde bağlam bilgisini göz ardı etmesidir. Bir metinde bulunan herhangi bir cümle çevrilirken metnin tamamının göz önünde bulundurulması gerekmektedir.

“The pen is in the box” - “The box is in the pen” örnek cümlelerinde bulunan “pen” kelimesi kalem ve çit anlamlarında kullanılmaktadır. Kural tabanlı makine çevirisinde bağlamın tamamını ele almadan doğru çeviri yapılması mümkün değildir [76].

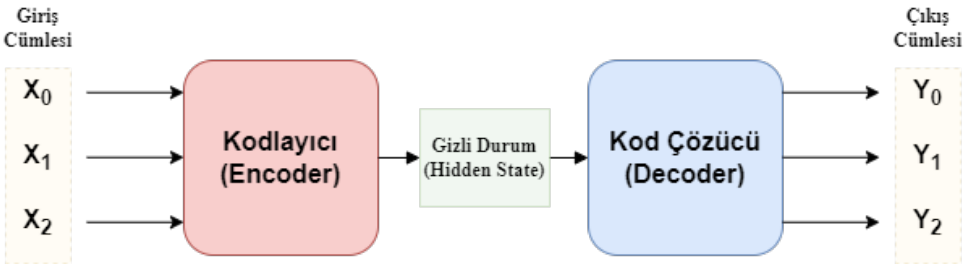
3.2 İstatistiksel Makine Çeviri Sistemi

İstatistiksel makine çevirisi, kural tabanlı makine çevirisinden farklı olarak çeviri görevini istatistiksel bir bakış açısı olarak ele almaktadır. İstatistiksel makine çevirisinde iki dilli derlem aracılığıyla aynı anlama gelen kelimeleri veya kelime öbeklerini istatistik yoluyla bulmaktadır. Bu çeviri yönteminde, bir cümle çevrileceği zaman sistem onu birkaç parçaya böler. Bu parçalara sözcük grubu veya ifade (phrase) adı verilir. Her sözcük grubu hedef tarafta karşılık gelen sözcük grubu ile değiştirilir. Bu yönteme ifade tabanlı istatistiksel makine çevirisi (phrase based statistical machine translation) adı verilir. Bu yöntem ön işleme, cümle hizalama, sözcük hizalama, tümce çıkarma, tümce özellik hazırlama ve dil modeli eğitimini içeren işlem adımlarından oluşur. Bu modelin temel bileşeni, kaynak dildeki kelime gruplarını hedef dildeki kelime gruplarıyla eşleştiren cümle tabanlı bir sözlüktür. Sözlük, iki dilli bir derlem olan eğitim veri setinden oluşturulmuştur. Çeviri modeli, bu çeviride deyimleri kullanarak, deyimlerdeki bağlam bilgisini kullanabilir. Böylelikle model kelimeden kelimeye çeviri yöntemlerinden daha iyi performans gösterebilir.

3.3 Sinirsel Makine Çevirisi

Sinirsel makine çeviri çalışmaları yaklaşık 30 yıl öncesine dayanmaktadır [76]. İlk zamanlardaki düşük performans ve bilgi işleme kaynaklarının sınırlı olması son 10 yıla kadar sinirsel makine çevirisine yeterince ilgi çekememiştir. 2010 yılında Derin öğrenmenin yaygınlaşması nedeniyle doğal dil işleme ile birlikte sinirsel makine çevirisi büyük bir gelişme göstermiştir. Sinirsel makine çevirisi basit mimarisi ve cümlede uzun süreli bağımlılık yakalama becerisi gibi avantajları sayesinde istatistiksel makine çevirisine kıyasla büyük başarılar elde etmiştir. İlk başarılı sinirsel makine çevirisi modeli Kalchbrenner ve Blunsom [81] tarafından önerilmiştir. Sinirsel makine çeviri görevi başlangıçta uçtan uca (End-to-End) bir öğrenme görevi olarak tasarlanmıştır. Bir kaynak diziyi (sequence) doğrudan bir hedef diziye eşler. Bu haliyle iki cümleyi anlamsal uzayda eşleştirmeye çalışan yüksek boyutlu bir sınıflandırma problemi olarak görülmüştür. Güncel sinirsel makine çeviri

modellerinde çeviri süreci iki adıma ayrılmıştır. Bunlar kodlayıcı (Encoder) ve kod çözücü (Decoder) olarak adlandırılır. Şekil 3.1'de kodlayıcı kod çözücü yapısı verilmiştir.



Şekil 3.1: Kodlayıcı Kod Çözücü mimarisi.

Kodlayıcı, kaynak cümleyi anlamsal vektöre temsil etmek için kullanılırken, kod çözücü bu anlamsal vektörden bir hedef cümleye tahmin yapar. Uçtan Uca mimarisi, modelin açıklanabilir ara sonuç olmaksızın kaynak verileri doğrudan hedef verilere dönüştürmesi anlamına gelir.

Olasılıksal olarak SMÇ sistemlerinde $S(s_1, s_2, \dots s_n)$ kaynak cümle olmak üzere $T(t_1, t_2, \dots t_m)$ hedef dizisini üretebilmek için maksimum koşullu olasılık kullanılır. Kaynak cümle uzunluğu n, hedef cümle uzunluğu m ile gösterilir (Denklem 3.1).

$$argmax\ P(T|S) \tag{3.1}$$

Denklemi genişletecek olursak SMÇ sistemi hedef cümlenin her bir kelimesini üretirken hem önceden tahmin ettiği kelimeden hem de kaynak cümleden gelen bilgileri kullanır. Bu bağlamda i kelime sırasını göstermek üzere [76] (Denklem 3.2):

$$argmax \prod_{i=1}^{m} P\ (t_i|t_{j<i}, S) \tag{3.2}$$

İfadesi her bir t_i kelimesinin önceki kelimeler $(t_{j<i})$ ve kaynak cümle S koşulu altında gerçekleşme olasılıklarının çarpımını temsil eder. Bu durum, olayların zincir kuralı kullanılarak modellenmesidir. Yani her bir kelimenin oluşma olasılığı o kelimeden önceki tüm kelimeler ve kaynak cümle altında hesaplanır. $(argmax)$ İfadesi, bu olasılıklar çarpımının en büyük olduğu durumları bulmayı amaçlar. Başka bir deyişle, modelin parametrelerini veya olaylar dizisini, toplam koşullu olasılığı en yüksek yapacak şekilde optimize eder.

İlk derin öğrenme tabanlı sinirsel makine çeviri sistemi Cho ve ark. [82]tarafından istatistiksel makine çeviri sistemine yinelenen sinir ağı tabanlı dil modelinin eklenmesiyle oluşturulmuştur. Yapılan çalışma da çeviri sisteminin ana gövdesi istatistiksel yöntem

olmasına ve hibrit bir yöntem olmasına rağmen şu an oluşturulan derin öğrenme tabanlı makine çeviri sistemlerine yön vermiştir.

3.4 Sinirsel Makine Çevirisi Temel Bileşenleri

Güncel çalışmalar incelendiğinde sinirsel makine çeviri sistemlerinin tasarımında birçok ağ tasarımı görülmektedir. Yinelenen, evrişimli ve dikkat mekanizmaları en yoğun kullanılan sinirsel makine çeviri modelleridir. Tezin bu bölümünde temel bileşenler detaylıca anlatılmıştır.

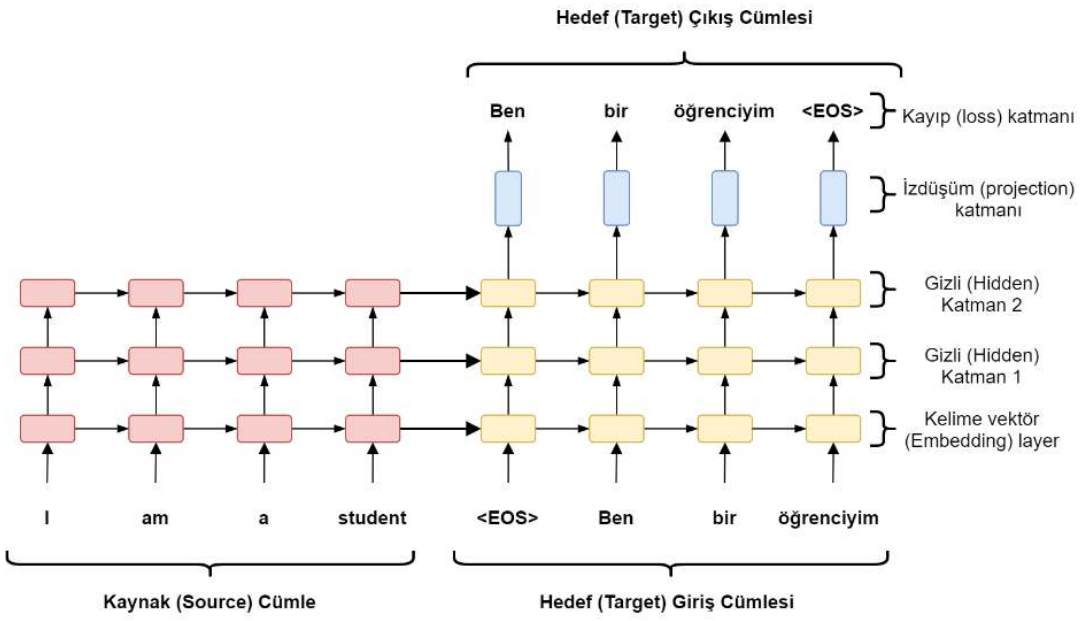
3.4.1 Kodlayıcı -Kod Çözücü mimarisi

Kodlayıcı-kod çözücü (Encoder-Decoder) mimarisi, SMÇ'nin en özgün ve klasik yapısıdır. Doğrudan doğal dil modellerinden (Natural Language Model) ilham alınarak ve Kalchbrenner ve Blunsom [81] tarafından önerilmiştir. Küçük farklılıklarına rağmen neredeyse tüm modern SMÇ modelleri tarafından kullanılmaktadır. Kodlayıcı-kod çözücü yapısının orijinal yapısı kavramsal olarak basittir. Mimarisinde, her biri çeviri sürecinin farklı bir parçası için birbirine bağlı iki ağ (kodlayıcı ve kod çözücü) içerir. Kodlayıcı ağı bir kaynak cümle aldığında, kaynak cümleyi kelime kelime okur ve değişken uzunluklu diziyi her gizli durumda sabit uzunluklu bir vektöre sıkıştırır. Bu işleme kodlama denir. Daha sonra, kodlayıcının son gizli durum vektörü (Context Vector) verildiğinde, kod çözücü, bu vektörü kelime kelime hedef cümleye dönüştürür. Kodlayıcı-Kod çözücü yapısı çeviri görevini kaynak cümleden doğrudan hedef cümleye dönüştürdüğü için yani ara süreçte bir sonuç üretmediği için buna uçtan uca çeviri denilmektedir. SMÇ'nin Kodlayıcı-Kod Çözücü yapısının ilkesi, anlamsal uzayda bir ara vektör aracılığıyla kaynak cümleyi hedef cümle ile eşlemek olarak görülmektedir. Bu ara vektör aslında her iki dilde de aynı semantik anlamı temsil etmektedir.

3.4.2 Yinelenen sinir ağı tabanlı mimariler

Yapay sinir ağlarının (YSA) kullanılmaya başlamasıyla görüntü sınıflandırma gibi sıralı olmayan verilerin işlenmesi problemlerinde büyük başarılar elde edilmiştir. Bir görüntü işleme probleminde YSA için kendisinden önceki veya sonraki örneğin durumu o adımda işlenen veri için önemsiz bir bilgidir yani sonucu etkilemez. Fakat doğal dil işleme uygulamalarında cümle içerisindeki bir kelime işlenirken kendisinden önceki veya sonraki kelimeler sonucu etkiler. Örneğin bir duygu analizi (sentiment analysis) uygulamasında “bu

ürünü çok beğendim" veya "ürün çok kötüydü" gibi iki yorumun olumlu mu yoksa olumsuz mu olduğu tespit edilmeye çalışılsın. Oluşturulan sistem de kelimeler teker teker değerlendirilirse "çok" kelimesi sistem için başarıyı olumsuz yönde etkileyen bir kelime olmaktadır. Doğal dil işleme problemlerinde bu tür problemlerin çözülebilmesi için bir kelime dizisinin tamamı ele alınmalıdır. Klasik YSA ile bu durum çözülemeyeceğinden Yinelen sinir ağları (RNN- Recurrent Neural Network) önerilmiştir [83]. Yinelenen sinir ağlarında ağ giriş (input) olarak bir dizi alır. Ağın çıktısı ise problemin türüne göre tek bir çıkış veya bir dizi olarak oluşturulur. Belirtilen problem türü yinelenen sinir ağının türünü de belirtir. Örneğin resim altyazısı (image captioning) uygulamasında giriş tek bir vektör çıkış ise bir cümle olabilmektedir (one to many). Duygu analizi probleminde ise giriş bir cümle çıkış ise duyguyu belirten tek bir puan (many to one) olabilmektedir. Makine çevirilerinde ise giriş bir cümle ve çıkış olarak ta bir cümle (many to many) üretilir. Şekil 3.2'de yinelenen sinir ağı tabanlı SMÇ'nin eğitim süreci verilmiştir. Şekilde "<EOS>" (End Of Sentence) sembolü dizinin sonu anlamına gelir. Gömme katmanı ön işleme içindir. Diziyi temsil etmek için iki yinelenen sinir ağı katmanı kullanılır.



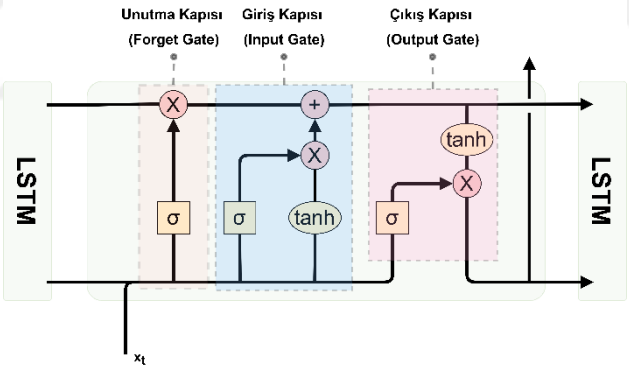
Şekil 3.2: Yinelenen sinir ağı tabanlı SMÇ'nin eğitim süreci.

Yinelenen sinir ağlarında aynı yapay sinir ağlarında olduğu gibi ağın optimizasyonu için gradyan (gradient) hesaplanır ve geri yayılım (backpropagation) yoluyla optimizasyon yapılır. Fakat giriş dizisi çok uzunsa geri yayılım sırasında kaybolan (vanishing) veya patlayan (exploding) gradyan problemi ile karşılaşılır. Geri yayılım hesaplanırken her bir yinelenen sinir ağı hücresi için ağırlık (weight) matrisiyle defalarca çarpım söz konusudur.

Eğer ilk hesaplanan gradyan küçükse bu çarpımlar sırasında giderek küçülerek kaybolacaktır. Aynı şekilde gradyan büyük bir değerse her çarpımda giderek büyüyerek patlayan gradyan problemine yol açacaktır. Bu problemin üstesinden gelebilmek için yinelenen sinir ağlarının türevleri olan farklı mimariler önerilmiştir. Bunlar uzun kısa süreli bellek (Long Short Term Memory (LSTM)) ve geçitli yinelenen birim (Gated recurrent units – (GRU)) modelleridir.

3.4.2.1 Uzun kısa süreli bellek

Uzun kısa süreli bellek (Long Short Term Memory (LSTM)) yinelenen sinir ağlarındaki gradyan problemi için önerilmiştir ve birçok doğal dil uygulamasında yüksek başarı oranları elde etmiştir. LSTM hücrelerinde bulunan kapılar (gate) uzun süreli bağımlılıklarda neyin hatırlanıp neyin hatırlanmayacağına karar verir. Yani girilen bir cümlede sonuca etki etmeyen önemsiz kelimelerin unutulmasını (forget gate) yada sonuca etki eden önemli bir kelimeyse sonraki adımlara aktarılmasını (cell state) sağlar. Bir LSTM hücresi sinir ağının hafızası (cell state), giriş (input), unutma (forget) ve çıkış (output) kapısı olmak üzere 4 birimden oluşur. Şekil 3.3'te LSTM hücresinin yapısı verilmiştir.



Şekil 3.3: LSTM hücresinin yapısı.

1. Hücre Durumu (Cell state): Şekil 3.3'te de gösterildiği gibi en üstten geçen bağlantıdır ve bağlantının amacı bilgiyi LSTM hücreleri arasında taşımaktır. Unutma kapısından gelen sonuç ile bir önceki hücreden gelen sonuç çarpılır çıkan sonuç giriş kapısından gelen veriyle toplanarak bir sonraki hücreye iletilir (Denklem 3.3). Buradaki çarpım işlemi skalar çarpım olarak adlandırılan noktasal çarpımıdır.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{3.3}$$

Denklemde:

C_t : Hücre durumunu belirtir.

f_t : Unutma kapısını belirtir.

C_{t-1}: Bir önceki hücreden gelen durumu belirtir.

i_t: Giriş kapısını belirtir.

$\tilde{C}_t$: Yeni aday hücre durumunu belirtir.

2. Giriş Kapısı (Input gate): Uzun süreli hafızayı güncellemek için kullanılır. Bir önceki hücrenin çıktısıyla o anda ki giriş önce *sigmoid* fonksiyonundan sonrasında *tanh* fonksiyonundan geçirilir. Çıkan iki sonuç çarpılarak cell state de bulunan veriye eklenir (Denklem 3.4 ve 3.5).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] \; + \; b_i) \tag{3.4}$$

$$\tilde{C}_t = tanh\,(W_C \cdot [h_{t-1}, x_t] \; + \; b_C) \tag{3.5}$$

Denklemde:

σ: Sigmoid aktivasyon fonksiyonunu belirtir.

W_i, W_C : Güncellenebilir parametreleri belirtir.

h_{t-1}:t-1 zaman adımında gizli durumu belirtir.

x_t: t zamanın da giriş vektörünü belirtir.

b_i b_C: Bias terimini belirtir.

$tanh$: Tanjant hiperbolik aktivasyon fonksiyonunu belirtir.

3. Unutma Kapısı (Forget gate): Bir önceki hücrenin çıktısıyla birlikte bu adımda girilen veri sigmoid fonksiyonundan geçer. Sigmoid fonksiyonu 0-1 arası bir çıktı oluşturur. Fonksiyon sonucu 0'a ne kadar yakınsa önceki veri o kadar unutulacak, tam tersi olarak 1'e ne kadar yakınsa veri o kadar saklanacaktır (Denklem 3.6).

$$f_t = \; \sigma(W_f \cdot [h_{t-1}, x_t] + \; b_f) \tag{3.6}$$

Denklemde:

f_t : Unutma kapısını belirtir.

W_f: Unutma kapısı güncellenebilir parametreleri belirtir.

h_{t-1}:t-1 zaman adımında gizli durumu belirtir.

x_t: t zamanın da giriş vektörünü belirtir.

b_f: Unutma kapısı bias terimini belirtir.

4. Çıkış Kapısı (Output gate): Bir sonraki hücreye aktarılacak veriye karar verildiği kapıdır. İlk olarak bir önceki hücreden gelen değer ile bu adımdaki değer sigmoid fonksiyonundan geçirilir. Sonrasında Cell state bağlantısından yani ağın tüm hafızasından gelen veri *tanh* fonksiyonundan geçirilir. Bu iki sonuç birbiriyle çarpılarak bir sonraki hücreye, bir önceki hücrenin çıkışı olarak iletilir (Denklem 3.7 ve 3.8).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.7}$$

$$h_t = o_t * tanh\,(C_t) \tag{3.8}$$

Denklemde:

o_t : Çıkış kapısını belirtir.

σ: Sigmoid aktivasyon fonksiyonunu belirtir.

W_o: Çıkış kapısı güncellenebilir parametreleri belirtir.

h_t:t zaman adımında gizli durumu belirtir.

h_{t-1}:t-1 zaman adımında gizli durumu belirtir.

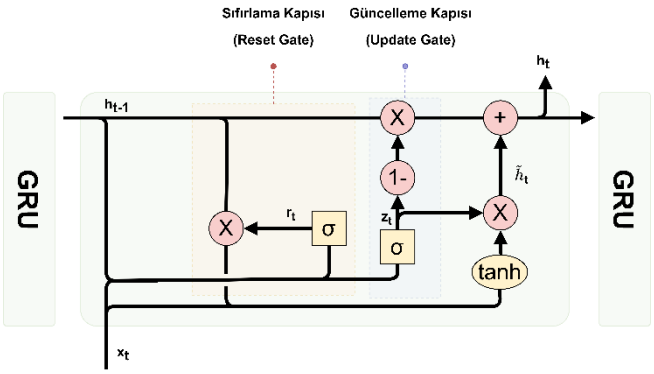
x_t: t zamanın da giriş vektörünü belirtir.

b_o: Çıkış kapısı bias terimini belirtir.

C_t : Hücre durumunu belirtir.

3.4.2.2 Geçitli yinelenen birim

Geçitli yinelenen birim (Gated Recurrent Unit -GRU) LSTM ağlarının bir türevi olarak oluşturulmuştur. LSTM ağlardan farklı olarak cell state kaldırılmıştır yani hücrenin çıktısıyla ağın hafızası birleştirilmiştir. Bir GRU hücresi sıfırlama (reset) ve güncelleme(update) kapısından oluşur. Şekil 3.4'de GRU hücresinin yapısı verilmiştir.



Şekil 3.4: GRU hücresinin yapısı.

1. *Güncelleme Kapısı (Update gate):* Güncelleme kapısı LSTM ağında bulunan hangi bilgilerin saklanacağına veya atılacağına karar verilmesi için kullanılan bağlantıdır (Denklem 3.9).

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{3.9}$$

Denklemde:

z_t : Güncelleme kapısını belirtir.

σ: Sigmoid aktivasyon fonksiyonunu belirtir.

W_z: Güncelleme kapısı güncellenebilir parametreleri belirtir.

h_{t-1}:t-1 zaman adımında gizli durumu belirtir.

x_t: t zamanın da giriş vektörünü belirtir.

2. *Sıfırlama Kapısı (Reset gate):* önceki bilgilerin ne kadarının saklanacağına karar veren bağlantıdır (Denklem 3.10, 3.11 ve 3.12).

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{3.10}$$

$$\tilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{3.11}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t) \tag{3.12}$$

Denklemde:

r_t : Reset kapısını belirtir.

σ: Sigmoid aktivasyon fonksiyonunu belirtir.

W_r: Reset kapısı güncellenebilir parametreleri belirtir.

h_{t-1}:t-1 zaman adımında gizli durumu belirtir.

x_t: t zamanın da giriş vektörünü belirtir.

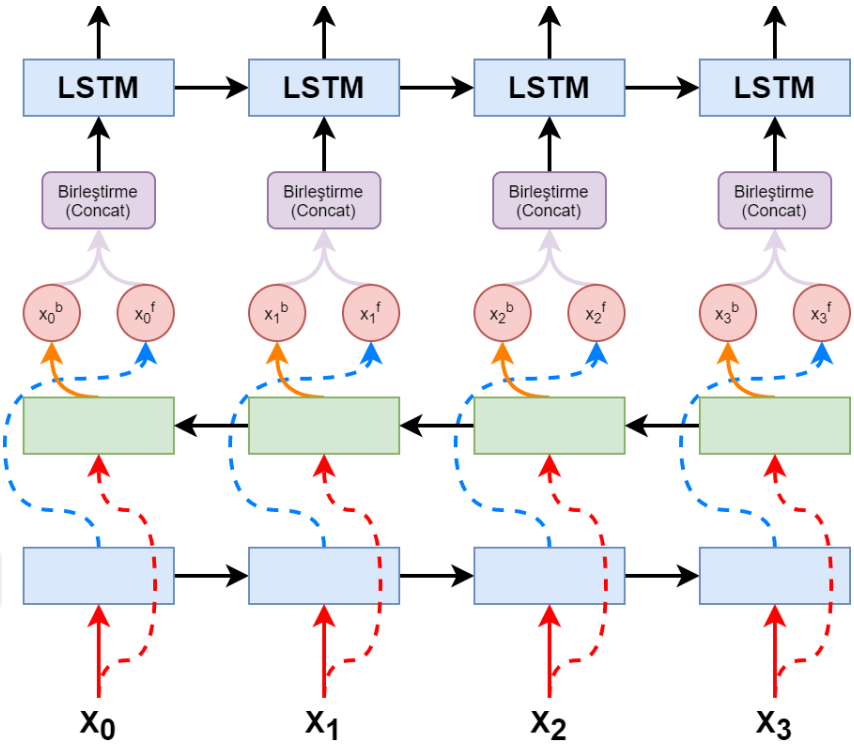
$\tilde{h}_t$: Aday gizli durum vektörünü belirtir.

$tanh$: Tanjant hiperbolik aktivasyon fonksiyonunu belirtir.

z_t : Güncelleme kapısını belirtir.

Yinelenen sinir ağında şu ana kadar anlatılan mimariler tek yönlü hücrelerden oluşmaktadır. Yani veri akışı ilk kelimeden son kelimeye doğru aktarılmaktadır. Fakat makine çevirisi gibi uygulamalarda girilen verinin iki yönlü işlenmesi gerekmektedir. Yani cümledeki son kelimenin ilk kelime üzerindeki etkisi hesaplanmalıdır. Bu tür problemler için iki yönlü yinelenen sinir ağları geliştirilmiştir. İki yönlü yinelenen sinir ağları en basit

şekliyle iki adet yinelenen sinir ağının üst üste eklenmesiyle oluşturulur. Şekil 3.5'te iki yönlü Bi-LSTM ağ yapısı gösterilmiştir.



Şekil 3.5: İki yönlü BiLSTM ağ yapısı.

SMÇ uygulamasında hem Bahdanau ve ark. [25] hem de Wu ve ark. [84] bağlam bilgisini yakalamak için alternatif olarak alt katmanda çift yönlü yinelenen sinir ağı kullanmışlardır. Bu kullanımda birinci katman cümleyi "soldan sağa", ikinci katman cümleyi ters yönde okumaktadır. Daha sonra birleştirilirler ve bir sonraki katmana beslenirler. Uygulanan yöntem sezgisel olsa da genellikle tek yönlü ağlardan daha iyi bir performansa sahip olmuştur.

3.4.3 Evrişimsel sinir ağı tabanlı mimariler

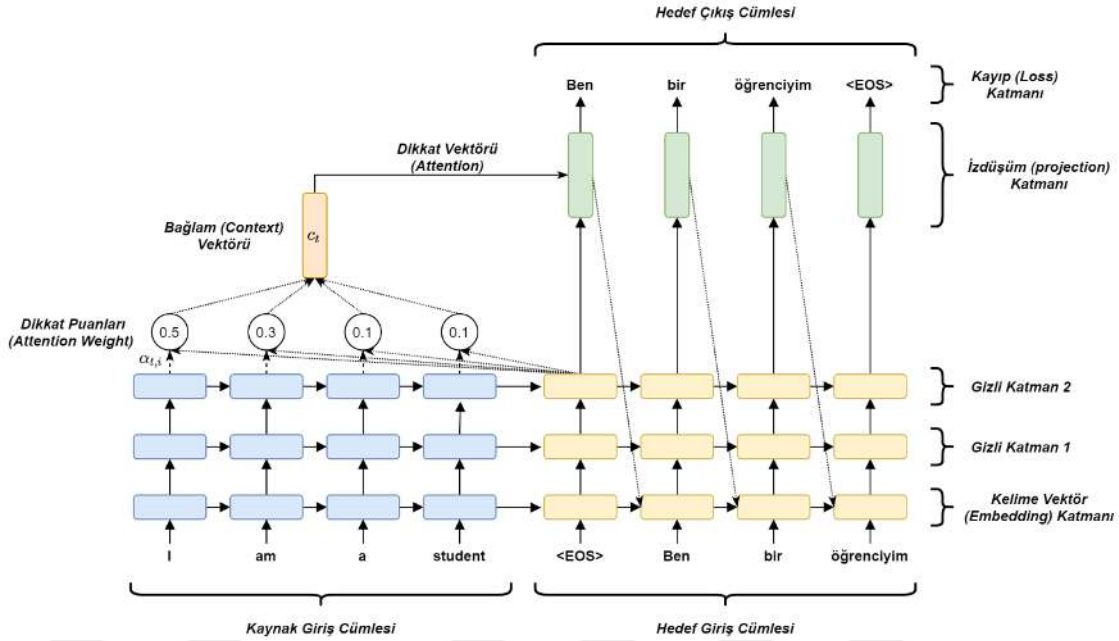
Evrişimsel sinir ağı (Convolutional Neural Network) tabanlı mimariler görüntü işleme uygulamalarında oldukça başarılı sonuçlar elde etmiştir. Doğal dil işleme ve Makine çeviri sistemlerine uyarlanması ise ilk olarak ESA kodlayıcı ve yinelenen sinir ağı kod çözücü ağlarının birlikte kullanılmasıyla başlamıştır [81]. Sonrasında öz yinelemeli ESA kodlayıcı ve yinelenen sinir ağı kod çözücü birleştirilerek yeni bir mimari önerilmiştir [82]. Fakat bu mimari yinelenen sinir ağı kodlayıcıdan daha kötü bir performans göstermiştir. Daha sonrasında Kaiser ve Bengio tarafından [85] , genişletilmiş sinirsel GPU'yu [86]

uygulayan tamamen ESA tabanlı bir SMÇ önerilmiştir. Önerilen yöntemler içerisinde ESA tabanlı makine çeviri sistemlerinde en iyi performansı ESA kodlayıcı + dikkat mekanizması + LSTM kod çözücünün birlikte kullanıldığı Gehring ve ark. [87] tarafından önerilen model elde etmiştir. Gehring ve ark. [88] daha sonra dikkat mekanizması eklenen ESA tabanlı bir SMÇ önererek önceki çalışmalarını geliştirmişlerdir.

Yinelenen sinir ağı tabanlı SMÇ ile karşılaştırıldığında, ESA tabanlı modellerin eğitim hızında avantajı vardır. Bunun nedeni, ESA'nin içsel yapısının, girdi verilerini işlerken farklı filtreler için paralel hesaplamalara izin vermesidir. Ayrıca, model yapısı, ESA tabanlı modellerin gradyan yok olma problemini çözmesini kolaylaştırır. Ancak, çeviri kalitelerini etkileyen önemli bir dezavantajı bulunmaktadır. Orijinal ESA tabanlı model yalnızca filtrelerinin genişliği içindeki sözcük bağımlılıklarını yakalayabilir. Fakat sözcüklerin uzun bağımlılığı yalnızca yüksek düzey evrişim katmanlarında bulunabilir. Bu yönüyle yinelenen sinir ağı tabanlı modeller daha iyi performans göstermektedir.

3.4.4 Dikkat mekanizması ile sinirsel makine çevirisi

İlk Sinirsel makine çeviri sistemleri dizi içindeki bağımlılıkları yakalamada umut verici performans göstermiştir. Fakat kaynak cümle çok uzun olduğunda performans düşüşü görülmektedir. Orijinal SMÇ kodlayıcı bir cümleyi sabit uzunlukta bir vektöre sıkıştırmak zorundadır. Girdi cümlesi uzadıkça performans düşer çünkü kodlayıcı ağın çıktısı sabit uzunlukta bir vektördür. Bu da tüm cümleyi temsil etmede sınırlamalara sahip olmaktadır. Bu durum bir miktar bilgi kaybına neden olmaktadır. Gizli durum vektörünün boyutunu artırmak sezgisel bir çözüm olsa da, yinelenen sinir ağı eğitim hızı doğal olarak yavaş olduğundan, daha büyük bir vektör boyutu daha da kötü bir eğitim süresine neden olmaktadır. Dikkat mekanizması bu problem durumundan dolayı ortaya çıkmıştır. Bahdanau ve ark. [25] başlangıçta bu yöntemi, kod çözme sürecinde ek kelime hizalama bilgisi sağlayabilen ve böylece giriş cümlesi çok uzun olduğunda bilgi kaybını azaltabilen bir ek olarak kullanmıştır Şekil 3.6’da dikkat mekanizmasına sahip sinirsel makine çeviri sistemi verilmiştir.



Şekil 3.6: Dikkat mekanizmasına sahip sinirsel makine çeviri sistemi.

Temelde dikkat mekanizması, kodlayıcı ve kod çözücü arasında, kelime korelasyonunun (kelime hizalama bilgisinin) dinamik olarak belirlenmesine yardımcı olabilecek bir ara bileşen olarak tasarlanmıştır. Dikkat mekanizması özetle giriş cümlesini bir vektör dizisi halinde kodlar ve çevirinin kod çözme adımında bu vektörlerin bir alt kümesini uyarlamalı olarak seçmektedir. Bir makine çevirisinde bir kelime oluşturulduğunda, en ilgili bilginin yoğunlaştığı kaynak cümlede bir konumlar kümesi arar. Daha sonra model, bu kaynak konumları ve önceden oluşturulmuş tüm hedef sözcüklerle ilişkili bağlam vektörlerine dayalı olarak bir hedef sözcük tahmin eder. İlk olarak Bahdanau ve Ark. [25] önerdiği sonrasında ise Luong ve arkadaşları [89] tarafından geliştirilen dikkat mekanizması yapısal olarak benzerdir ve algoritması şu adımlardan oluşur.

1. Kodlayıcı, giriş cümlesinden bir dizi ek açıklama oluşturur. (h_t)
2. Bu ek açıklamalar, bir hizalama modeline ve önceki gizli kod çözücü durumuna beslenir. Hizalama modeli bu bilgiyi dikkat puanları oluşturmak için kullanır. (s_t) (Denklem 3.13)

$$s_t = score(h_t, h_s) \tag{3.13}$$

Denklemde:

s_t : t Zaman adımındaki skor değerini gösterir.

h_t : Hedef dizinin (target sequence) t zaman adımındaki gizli durumunu belirtir.

h_s : Kaynak dizinin (source sequence) s zaman adımındaki gizli durumunu belirtir.

3. Dikkat puanlarına, 0 ile 1 aralığında ağırlık değerlerine normalleştiren bir softmax işlevi uygulanır. ($\alpha_{t,i}$) (Denklem 3.14)

$$\alpha_{t,i} = \frac{exp\ (s_t)}{\sum_{s=0}^{S} exp\ (s_t)} \tag{3.14}$$

Denklemde:

$\alpha_{t,i}$: t zaman adımında i pozisyonundaki dikkat ağırlığını gösterir.

s_t : t Zaman adımındaki skor değerini gösterir.

4. Daha önce hesaplanan açıklamalarla birlikte, bu ağırlıklar, açıklamaların ağırlıklı toplamı yoluyla bir bağlam vektörü oluşturmak için kullanılır. (c_t) (Denklem 3.15)

$$c_t = \sum_s^S \alpha_t h_s \tag{3.15}$$

Denklemde:

c_t : t zaman adımında bağlam vektörünü gösterir.

α_t : t zaman adımında dikkat ağırlığını gösterir.

h_s : Kaynak dizisinin (kodlayıcı) gizli durumunu gösterir.

5. Bağlam vektörü, son çıktıyı hesaplamak için önceki gizli kod çözücü durumu ve önceki çıktı ile birlikte kod çözücüye beslenir. (y_t)
6. 2-6 arasındaki adımlar dizinin sonuna kadar tekrarlanır.

Dikkat puanı hesaplanırken en çok bilinen iki farklı yöntem vardır: (Denklem 3.16)

$$score(h_i, h_t) = \begin{cases} h_i^T W h_t \ \{Luong\} \\ v_a^T tanh(W_1 h_i, h_t) \{Bahdanau\} \end{cases} \tag{3.16}$$

Denklemde:

$score$: Dikkat skor değerini gösterir.

h_i : Kodlayıcının i pozisyonundaki gizli durumunu gösterir.

h_t : Kod çözücünün t zaman adımındaki gizli durumunu gösterir.

W : Öğrenilebilir ağırlık matrisini gösterir.

v_a^T: Öğrenilebilir bir ağırlık vektörünü gösterir.

$tanh$: Tanjant hiperbolik aktivasyon fonksiyonunu gösterir.

Dikkat mekanizmasını SMÇ sistemine uygulamak için ilham, metin verilerini okuma ve tercüme etmedeki insan davranışından gelir. İnsanlar genellikle cümle içindeki bağımlılığı araştırmak için metni tekrar tekrar okurlar, bu da her kelimenin birbiriyle farklı bağımlılık ağırlığına sahip olduğu anlamına gelir. ESA mimarisindeki havuzlama katmanı veya N-gram dil modeli gibi kelime bağımlılık bilgilerinin yakalanmasında diğer modellerle karşılaştırıldığında, dikkat mekanizması global bir kapsama sahiptir. Bir dizideki bağımlılığı bulurken, N-gram modeli arama kapsamını küçük bir aralıkta sabitleyecektir, genellikle N 2 veya 3'e eşittir. Dikkat mekanizması ise kaynak cümlede o anda oluşturulan kelimenin diğer kelimelerle olan ilişkisini hesaplamaktadır.

Vaswani ve ark. 2017 [90] yılında dikkat mekanizmasını geliştirerek Öz dikkat (Self-Attention) ya da bilinen diğer adıyla Transformer mimarisini geliştirmişlerdir. Öz dikkat mimarisi Bölüm 4'te detaylı olarak ele alınacaktır.

3.5 Kelime Temsil Yöntemleri

Çeviri sistemlerinin önemli bileşenlerinden birisi de kelime temsil yöntemleridir. Çeviri sistemlerinde eğitim sırasında görülmeyen bir kelime ile karşılaşıldığında çeviri sisteminin doğru çeviri yapması mümkün değildir. Bu probleme kelime dağarcığı dışı (Out-of-Vocabulary (OOV)) adı verilmektedir. Bu problemin üstesinden gelebilmek için çeviri sistemlerine girdi olarak kelimelerin değil de kelime parçalarının vektörlere dönüştürülerek verilmesi gerekmektedir. Kelimeleri uygun şekilde ayırabilmek için Byte Pair Encoding [29], WordPiece[91] ve UniGram[92] gibi yöntemler önerilmiştir.

3.5.1 Byte Pair Encoding tokenizasyon yöntemi

Byte Pair Encoding (BPE) [29], makine çeviri sistemleri tarafından kullanılan bir kelime temsil yöntemidir. BPE, bir metni kelime parçalarına ayırır ve bu parçaların en sık kullanılan çiftlerini birleştirir. Bu işlem tekrarlandıkça, külliyatta bulunan kelime parçaları azaltılarak daha fazla sıkıştırılabilir hale getirilir. BPE, çeviri sisteminin öğrenmesini kolaylaştırmaktadır. Dil için özel olmayan kelime parçalarının azaltılması daha doğru çeviriler üretilebilmesine olanak sağlar. BPE yöntemi GPT-2[93], RoBERTa [94] gibi dil modellerinde kullanılmıştır. BPE yönteminde örnek olarak "aaabdaaabac" dizisini ele alalım. "aa" en sık görülen bayt çiftidir. Bu nedenle veri içerisinde bulunmayan başka bir karakter D= "aa" ile değiştirilir. Yeni dizi "DabdDabac" şeklinde oluşmuştur. Sonraki adımda E = "ab" ile değiştirilir. Bu adımda yeni dizi "DEdDEac" şeklinde oluşmuştur.

Sonraki adımda ise F=“DE” ile değiştirilir. Bu şekilde dizinin son hali “FdFac” şeklinde temsil edilir.

3.5.2 WordPiece tokenizasyon yöntemi

WordPiece, doğal dil işlemede kullanılan bir alt kelime bölütleme algoritmasıdır. Kelime dağarcığı, dildeki tek tek karakterlerle başlatılır, ardından kelime dağarcığındaki en sık sembol kombinasyonları yinelemeli olarak kelime dağarcığına eklenir. WordPiece, Google'ın BERT[95] modelini önceden eğitmek için geliştirdiği simgeleştirme algoritmasıdır.

3.5.3 Unigram tokenizasyon yöntemi

Unigram[92] tokenizasyon yöntemi bir metni kelime ya da alt kelime parçalarına ayırır. Unigram yönteminde her token, metindeki en küçük anlamlı birim olarak kabul edilir ve bağımsızdır. Bu tür tokenizasyon, özellikle dil modellemede temel bir adımdır ve daha karmaşık tokenizasyon yöntemlerinin de temelini oluşturur. Yöntem de öncelikle metin bölme işlemi gerçekleştirilir. Bu aşamada, girdi kelimelere veya belirli alt kelime birimlerine ayrılır. Daha sonra, tüm bu birimlerin sıklığı hesaplanır ve belirli bir sıklık eşiği üzerinde olan birimler token olarak kabul edilir. Bu şekilde bir token kümesi oluşturulur. Unigram yönteminin avantajları arasında basitliği, uygulanmasının ve anlaşılmasının kolay olması yer alır. Ayrıca, daha karmaşık yöntemlere göre hızlı çalışır ve çoğu dilde kullanılabilir, çünkü yalnızca kelime veya alt kelime birimlerinin sıklığını dikkate alır. Ancak, bu yöntemin bazı dezavantajları da vardır. Unigram ayırdığı tokenları birbirinden bağımsız olduğu için bağlam bilgisini hesaba katmaz. Ayrıca, büyük veri kümelerinde çok sayıda token oluşabilir, bu da bellekte daha fazla yer kaplamasına neden olabilir. Verilen 3 tokenizasyon algoritması için örnek kelime temsilleri Çizelge 3.1 de verilmiştir. Örnek kelime temsilleri için Wikipedia (huggingface.co/datasets/wikimedia/wikipedia) Türkçe veri seti kullanılarak ayrı ayrı tokenizer oluşturulmuştur.

Çizelge 3.1: Örnek kelime temsilleri kodlaması.

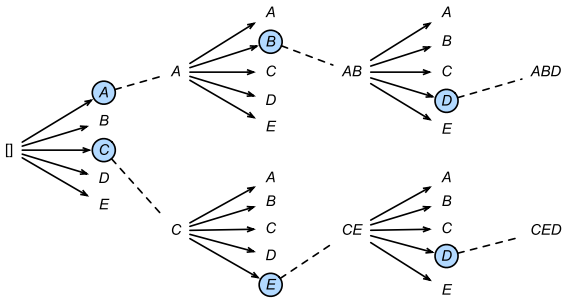
Örnek Cümle	Tokenizasyon Yöntemleri
"Çocuklar oyun bahçesin de oynuyordu."	**BPE:** ['Çocuklar', 'oyun', 'bahç', 'esinde', 'oyn', 'uyordu', '.']
	WordPiece: ['Çocuk', '##lar', 'oyun', 'bahçe', '##sinde', 'oyn', '##uyordu', '.']
	UniGram: ['Ç', 'o', 'cu', 'k', 'lar', 'o', 'y', 'un', 'b', 'ah', 'ç', 'e', 'sinde', 'o', 'y', 'n', 'u', 'y', 'or', 'du', '.']
"Elâzığ Doğu Anadolu Bölgesinde yer alır."	**BPE:** ['El', 'â', 'z', 'ığ', 'Doğu', 'Anadolu', 'Bölg', 'esinde', 'yer', 'alır', '.']
	WordPiece: ['El', '##âz', '##ığ', 'Doğu', 'Anadolu', 'Bölgesi', '##nd', '##e', 'yer', 'alır', '.']
	UniGram: ['E', 'l', 'â', 'z', 'ı', 'ğ', 'D', 'o', 'ğ', 'u', 'A', 'n', 'a', 'd', 'ol', 'u', 'B', 'ö', 'l', 'g', 'e', 'sinde', 'yer', 'a', 'lı', 'r', '.']
"Geometrik şekiller, düz ve eğri çizgilerle oluşturu labilir."	**BPE:** ['Ge', 'ometrik', 'şekil', 'ler', ',', 'düz', 've', 'eğ', 'ri', 'çizg', 'ilerle', 'oluşturul', 'abilir', '.']
	WordPiece: ['Ge', '##ometrik', 'şekil', '##ler', ',', 'düz', 've', 'eğr', '##i', 'çizgi', '##lerle', 'oluşturul', '##abilir', '.']
	UniGram: ['G', 'e', 'o', 'm', 'e', 't', 'r', 'ik', 'ş', 'e', 'k', 'il', 'ler', ',', 'd', 'ü', 'z', 've', 'e', 'ğ', 'r', 'i', 'ç', 'iz', 'g', 'i', 'ler', 'le', 'ol', 'u', 'ş', 'tur', 'u', 'la', 'b', 'il', 'ir', '.']

3.6 Çıkarım yöntemleri

Doğal dil üretme problemlerinde her bir kelimenin üretilmesi kendinden önceki kelimelere bağlı olarak değişmektedir. Her bir kelime üretme adımında sadece o an da olasılığı en yüksek kelimeyi seçmek bazen cümle bazında başarıyı düşürmektedir. Cümle tabanlı olasılığı artırabilmek için farklı arama stratejileri önerilmiştir [96]. Bu arama stratejileri makine çeviri sistemleri için zorunlu değildir. Fakat başarıyı artırdığı gözlemlenmiştir [97]. Çeviri üretiminde iki temel arama stratejisi bulunmaktadır.

3.6.1 Işın arama algoritması

Işın arama (Beam Search), oluşturulacak mümkün olan en yüksek (k) kelime miktarını hafızada tutar. Bir sonraki adım da yeni bir çeviri oluşturmak için her aday kelime yeni kelimeyle birleştirilir. Bu süreç çeviri tamamlanana kadar devam eder. En uygun cümle seçimi için log olasılığı hesaplanır [96,98]. Işın arama algoritmasının k=2 için örnek gösterimi Şekil 3.7'de gösterilmiştir.



Şekil 3.7: Işın arama algoritması (k=2).

3.6.2 Açgözlü arama algoritması

Açgözlü arama (Greedy Search) algoritması, her kelime üretme adımında en iyi seçeneği seçmeyi amaçlar. Şekil 3.8'de aç gözlü arama algoritması gösterilmiştir. Üretilen kelimeden sonraki adımda daha iyi bir seçenek olup olmadığını kontrol etmez. Bu yöntem, çeviri oluşturma sürecini hızlandırabilir, ancak en iyi çözümü bulmayı garanti edemez [99].

A	0.5	0.1	0.2	0.0
B	0.2	0.4	0.2	0.2
C	0.2	0.3	0.4	0.2
<eos>	0.1	0.2	0.2	0.6

Şekil 3.8: Aç gözlü arama algoritması.

3.7 Değerlendirme

Makine çevirilerinde en önemli problemlerden biri de makine çevirisinin kalitesini değerlendirmektir. Makine çevirileri genel olarak manuel insan çevirmenler ve otomatik bilgisayar tabanlı yazılımlar aracılığıyla olmak üzere iki farklı şekilde değerlendirilmektedir. Otomatik değerlendirme ölçütleri sistem çıktısını altın standart bir referans çeviri ile karşılaştırarak puanlamaya çalışmaktadır. Böyle bir karşılaştırmanın nasıl yapılacağı aslında bir problem oluşturmaktadır. Otomatik çeviri sistemlerinde iki temel zorluk bulunmaktadır. Bu zorluklardan birincisi, çeviri görevinin doğasında var olan belirsizliktir. İki profesyonel çevirmen bile neredeyse her zaman aynı cümle için iki farklı çeviri üretebilmektedir. Tek bir insan referans çevirisini eşleştirmek, makine çevirisi için kalitenin tek şartı olması mümkün değildir. İkinci zorluk ise sadece eşleşen kelimelerin değil, aynı zamanda kelime sırasının da çeviri kalitesinde ölçülmesi gerekliliğidir. Bazı kelime sırası farklılıkları anlamsal değişikliklere yol açmayabilir fakat bazıları ise anlamsız bir sonuç doğurabilmektedir. Bu

zorluklar nedeniyle, doğru kelime sayısını sayma (kelime sırasını dikkate almayan) veya kelime hata oranı gibi basit yöntemler iyi sonuç vermeyecektir [100].

3.7.1 Manuel değerlendirme

Makine çevirisi kalitesini değerlendirmek ve ölçmek için en yaygın seçenek insan değerlendirmesidir. MÇ çıktısının kalitesi, çeviri ve dilbilim uzmanları tarafından iki farklı açıdan değerlendirilmektedir. Çizelge 3.2'de makine çevirisi manuel değerlendirme ölçütleri verilmiştir. İlk bakış açısında, çevirinin hedef metne ve hedef dil gramer ve açıklık gibi özelliklere uygunluğu bulunmaktadır. Bu kalite değerlendirme perspektifi akıcılık olarak bilinir[100]. Akıcılık değerlendirilirken kaynak metin göz ardı edilerek değerlendiriciler, yalnızca değerlendirilen çeviriye bakarak bir ölçek dâhilinde puanlama yaparlar. Akıcılık ölçülürken yalnızca hedef dilde akıcı bir uzman yeterlidir. İkinci bakış açısında ölçülen doğruluktur. Hedef metnin, kaynak metnin bilgi içeriğini ne kadar iyi temsil ettiğine bağlı olarak hesaplanır. Değerlendiriciler, değerlendirilen kaynak metne ve çevirilere erişebilir. Sıklıkla, bir cümlenin bağlamı da dikkate alınır. Değerlendiriciler hem kaynak hem de hedef dilde iki dilli olmalıdır. Yeterlilik ve akıcılık genellikle bir 5 puanlık bir ölçekte değerlendirilir. İnsanlar tarafından yapılan değerlendirme zaman alıcı ve maliyeti yüksektir. Aynı zamanda doğası gereği özneldir. Öznellik sorununu hafifletmek için, genellikle daha fazla uzman kullanılarak değerlendirmeleri sağlanır sonrasında ortalama alınarak kalite skor elde edilebilir.

Çizelge 3.2: Makine Çevirisi manuel değerlendirme ölçütleri

Yeterlilik	Akıcılık
5 – Tüm Anlam	5- Kusursuz Dil
4 – Çoğu Anlam	4 – İyi Dil
3 – Yeterli Anlam	3 – Ana Dili olmayan
2 – Yetersiz Anlam	2 – Akıcı olmayan Dil
1 – Anlamsız	1 – Anlaşılmaz Dil

3.7.2 Otomatik değerlendirme

Makine çeviri sistemlerinin çıktısını referans çeviriyle karşılaştırarak değerlendiren bilgisayar tabanlı sistemlerdir. Referans çeviri sayısı ne kadar fazlaysa çeviri için ölçüm o kadar sağlıklı olabilmektedir. Makine çeviri sistemleri ve çeviri değerlendirme sistemleri paralel bir şekilde devamlı geliştirilmektedir. Çeviri değerlendirme sistemlerinde bulunan

temel problemleri giderebilmek için farklı çeviri metrikleri geliştirilmiştir. Otomatik değerlendirme ölçütleri, insan değerlendirmesine ücretsiz alternatiflerdir. En çok bilinen değerlendirme metrikleri f1score, Bleu, Meteor ve Ter olarak bilinen yöntemlerdir.

3.7.2.1 Temel değerlendirme ölçütleri

Makine çevirilerini değerlendirmenin en temel ölçütleri kesinlik (precision), duyarlılık (recall), ve F1Skor puanlarını hesaplamaktır. Bu ölçütler makine öğrenmesi problemlerinde de sıkça kullanılmaktadır. Eşleşen kelime sayılarının oranlarına göre değerlendirilmektedir[101]. Makine çevirisinde kesinlik, modelin ürettiği çevirideki kelimelerin veya n-gramların referans çeviride bulunma oranını ifade eder. Yani, modelin doğru tahmin ettiği kelimelerin, toplam tahmin edilen kelimelere oranıdır (Denklem 3.17).

$$kesinlik = \frac{Eşleşen\ Kelime\ Sayısı}{Makine\ Çevirisi\ Toplam\ Kelime\ Sayısı} \tag{3.17}$$

Duyarlılık, referans çevirideki kelimelerin veya n-gramların model tarafından doğru bir şekilde tahmin edilme oranını ifade eder. Yani, referans çevirideki doğru kelimelerin, toplam referans kelimelerine oranıdır. (Denklem 3.18).

$$duyarlılık = \frac{Eşleşen\ Kelime\ Sayısı}{Referans\ Toplam\ Kelime\ Sayısı} \tag{3.18}$$

F1Skor, duyarlılık ve kesinlik arasındaki dengeyi ölçen bir metriktir. duyarlılık ve kesinlik arasındaki harmonik ortalama olarak hesaplanır ve iki değerin dengeli bir şekilde değerlendirilmesini sağlar. (Denklem 3.19 ve 3.20).

$$f1Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.19}$$

$$f1Score = \frac{Eşleşen\ Kelime\ sayısı}{(MÇ\ Kelime\ Sayısı + Ref\ Kelime\ sayısı)/2} \tag{3.20}$$

Örnek olarak referans çeviri: "the cat is on the mat" ve model çevirisi: "the cat is on mat" olsun. Modelin doğru tahmin ettiği kelimeler: "the", "cat", "is", "on" (5 kelime) ve modelin tahmin edip de referansta olmayan kelime (0) bulunmamaktadır. Referansta olup da modelin tahmin edemediği kelime ise "the" (1) kelimedir. Kesinlik (5/(5+0) = 1.0) olarak duyarlılık ise (5/(5+1)=0.833) hesaplanır. F1Skor (2*(1.0*0.833)/(1.0+0.833) = 0.909) olarak bulunmaktadır.

3.7.2.2 Bleu değerlendirme ölçütü

Bleu skorun arkasındaki temel fikir, çeviride yalnızca referansla eşleşen sözcük sayısını değil, aynı zamanda daha büyük sıralı n-gram eşleşmelerini de sayma işlemidir. Böylece, eşleşen kelime çiftlerinin (bigramlar), hatta üç(trigramlar) veya dört kelimelik (4-gram) dizilerin eşleşme oranına göre puanlama yapar. Bu şekilde doğru kelime sırası da ödüllendirilmiş olur. Bu yöntemle birden fazla referans çeviri kullanılabilir. Ana fikir tek bir insan referans çevirisini eşleştirmek zor olduğundan, birden fazla insan referans çevirisine sahip olarak doğru puanlamayı yapabilmektir. Makine çevirisinin bunlardan herhangi biriyle n-gram eşleşmesi varsa, bu daha yüksek puan demektir (Denklem 3.21). Bir diğer özellik ise Bleu makine çevirisindeki sadece n-gramların referans çeviriyle eşleşme oranı hesaplanır (kesinlik) yani eşleşen kelime sayısının tüm çevrilen kelimelere oranı hesaplanmaz (duyarlılık). Bu nedenle Bleu bunun yerine bir kısalık cezasının kullanılmasıyla nihai puanı hesaplar (brevity penalty) (Denklem 3.22). Makine çevirisindeki kelime sayısı ile referans çevirisi arasındaki orana dayanır ve bu oran 1'in altındaysa (yani makine çevirisi çok kısaysa) devreye girer[102].

$$Bleu = KısalıkCezası \times exp \sum_1^4 log \frac{eşleşen\ i-gram\ sayısı}{M.Ç.toplam\ i-gram\ sayısı} \tag{3.21}$$

$$KısalıkCezası = \min\left(1, \frac{çıktı\ uzunluğu}{ref.\ uzunluğu}\right) \tag{3.22}$$

3.7.2.3 METEOR değerlendirme ölçütü

Meteor (Metric for Evaluation of Translation with Explicit Ordering) metriği Bleu skor hesaplanırken karşılaşılan bazı sorunların giderilmesi için önerilmiştir [103]. Eş anlamlı kelime veya kelime köklerinin eşleşmesi Bleu tarafından puanlanmaz. Meteor puanı hesaplanırken temel fikir aynı anlama gelen farklı kelimelerin puanlama sistemine dahil edilmesini sağlamaktır. Puanlama yapılırken ilk olarak kesinlik ve duyarlılık puanlarının harmonik ortalamasıyla F_{mean} puanı hesaplanır (Denklem 3.23).

$$F_{mean} = \frac{10KD}{D+9K} \tag{3.23}$$

Denklemde:

F_{mean}: Harmonik ortalamayı gösterir.

K: Kesinlik (precision) değerini gösterir.

D: Duyarlılık (recall) değerini gösterir.

Sonrasında çeviri cümle ile referans cümlede bitişik olmayan eşleşmeler ceza olarak eklenir (Denklem 3.24 ve 3.25)

$$c = 0.5\left(\frac{parça\ sayısı}{eşleşen\ unigram\ sayısı}\right)^3 \tag{3.24}$$

$$M = F_{mean}(1 - c) \tag{3.25}$$

Denklemde:

c: Ceza puanı (penalty) değerini gösterir.

F_{mean}: Harmonik ortalamayı gösterir.

M: Meteor puanını gösterir.

Meteor puanının en önemli dezavantajı, hesaplama yönteminin ve formülünün Bleu puanına göre çok daha karmaşık olmasıdır. Bir diğer dezavantajı ise Meteor puanı hesaplanırken morfolojik kökler ve eşanlamlı veri tabanları gibi dilsel kaynakların gerekli olmasıdır.

3.7.2.4 TER (Translation Error Rate)

TER, çevirideki son düzenleme miktarına göre hesaplanır [46]. Yani, aday cümleyi referans cümleyle hizalamak için gereken eylem sayısıdır. Dilden bağımsız olarak kullanılır. Puan ne kadar yüksek olursa çevirideki hata oranı da o kadar yüksek olur. Örneğin 10 kelimelik bir çevirinin referans cümleyle aynı olması için 1 kelime değiştirme ve 1 kelime kaydırma gibi 2 işlem olduğunu varsayalım. Bu durumda 10 kelime içinde 2 hata olduğu varsayılır ve TER puanı %20 olarak hesaplanır. Bleu puanlarının hesaplanması faktör olarak 4 gram kesinliğe sahiptir, ancak bir cümlenin çevirisi 4 gramlık bir eşleşmeye sahip olmayabilir ve bu da Bleu puanının 0 olmasına neden olur. Ter, Bleu kadar yaygın olarak kullanılmasada, çeviri kalitesi hakkında fikir vermektedir. Ter puanını hesaplamanın en büyük dezavantajı hesaplama maliyetidir.

4. ÖZ DİKKAT AĞLARI

Dikkat, insanların çevrelerindeki belirli uyaranlara odaklanabilme yeteneğidir. Günlük yaşamda dikkat, bir konuşmayı dinlerken diğer sesleri arka plana itme, araba kullanırken yola odaklanma veya bir kitabı okurken kelimelere konsantre olma gibi pek çok durumda kendini gösterir. Dikkat, algılama, uzun süreli hafızada saklama ve öğrenme gibi birçok bilişsel işlevin temelini oluşturur. İnsan beyninde dikkat, belirli bir bilgiye odaklanma ve diğer bilgileri filtreleme süreci olarak tanımlanabilir [104]. Beyin, sınırlı kaynaklarını en verimli şekilde kullanmak için dikkat mekanizmalarını kullanır. Bu mekanizma, önemli bilgilerin seçilmesine ve işlenmesine olanak tanırken, önemsiz bilgilerin göz ardı edilmesini sağlar. Bu süreç, dikkat çekici uyarıcıların daha hızlı ve etkili bir şekilde işlenmesini sağlar.

Dikkat, çevremizde çok fazla bilgi olduğu için gereklidir. İnsan beyni, aynı anda sadece belli bir miktar bilgiyi işleyebilir. Bu yüzden, beyin hangi bilgilere öncelik vereceğine karar vermek zorundadır. Dikkat, beynin bu bilgi işlem sürecini nasıl yönettiğini anlatan bir terimdir. Dikkatin etkileri, bilinçli düşünce, beyin dalgaları ve beyin görüntüleme teknikleriyle ölçülebilir.

İlk olarak Bahdanau ve ark. [25] tarafından 2014 yılında önerilen dikkat derin sinir ağlarında yaygın bir şekilde kullanılmaya başlamıştır. Güncel derin öğrenme ve doğal dil işleme uygulamalarında dikkat modelleri yaygın olarak kullanılmaktadır [104]. Sinir ağlarında dikkat mekanizmaları, bilgi akışını ve mevcut kaynakları dinamik olarak yöneterek öğrenmeyi geliştirir. Bu mekanizmalar, gereksiz bilgileri filtreler ve ağın uzun mesafeli ilişkileri kolayca anlamasına yardımcı olur. Pek çok dikkat modeli, basit, ölçeklenebilir ve esnek olup, çeşitli alanlarda başarılı sonuçlar vermektedir. Dikkat mekanizmaları, başlangıçta doğal dil anlama, duygu analizi, makine çevirisi, soru yanıtlama ve transfer öğrenimi gibi zor görevlerde kullanılmıştır. Daha sonra bilgisayarla görme, takviyeli öğrenme ve robotik gibi diğer alanlarda da uygulanmaya başlanmıştır.

Klasik kodlayıcı-kod çözücü mimarisinde, tüm bilgiyi sabit uzunlukta bir vektöre sığdırma zorunluluğu bir darboğaz sorunu yaratır. Bu sorun, dikkat araştırmaları için ilk motivasyon olmuştur. Kodlayıcı, kaynak cümleyi bir vektöre dönüştürür ve kod çözücü bu vektörü kullanarak çeviri yapar. Ancak, kaynak cümle uzadıkça bu yöntem performans

kaybına neden olur. Cho ve ark. [80] , giriş cümlesi uzadıkça performansın hızla düştüğünü göstermişlerdir. Bu darboğazı aşmak için Bahdanau ve ark. [25], birlikte hizalama ve çeviri öğrenen bir model önermiştir. Bu model, her adımda çevrilecek kelimeyi üretirken, kaynak cümlede en alakalı kelimeleri arar ve bu kelimelere bağlı olarak hedef kelimeyi tahmin eder.

Bu yöntemin avantajı, tüm kaynak cümleyi sabit bir vektöre sıkıştırmak yerine, cümleyi bir dizi vektör olarak kodlamasıdır. Çeviri yaparken, bu vektörlerden en uygun olanlarını seçer. Dikkat mekanizması, bilgi darboğazını ortadan kaldırarak daha fazla bilginin ağ üzerinden yayılmasını sağlar. Bu yaklaşım, uzun cümlelerde klasik kodlayıcı-kod çözücü mimarilerinden daha iyi performans göstermiştir ve dikkat mekanizmaları araştırmalarının temelini oluşturmuştur. Dikkat modülü, birçok farklı uygulamada yaygın olarak kullanılmaktadır. Ses tanıma [105] uygulamalarında, dikkat mekanizması, sesin ilgili bölümlerine odaklanarak ek açıklamalar oluşturur. Konuşma modellemesinde [106], modelin konuşmanın son bölümlerine odaklanarak yanıt oluşturmasına yardımcı olur. Görüntü sınıflandırmada [107] ise, dikkat modülü, görüntünün en önemli kısımlarına odaklanarak başarıyı artırır. Metin analizinde, modelin kelimelere odaklanmasına izin vererek analiz ağacı oluşturulmasını sağlar.

Dikkat katmanı, bağlamı sabit bir vektörde özetlemez. Bunun yerine, her adımda dikkat hesaplanır ve önceki katmanların temsilleriyle birlikte bir sonraki katmana aktarılır. Bu şekilde, dikkat her adımda bağımsız olarak işlev görür ve o anki sorgu ve bağlamın bir işlevidir. Bu basitleştirme, dikkat katmanının sorgu ve bağlam arasındaki dikkati öğrenmeye odaklanmasını sağlar ve bu da modelin performansını artırır [104].

Derin dikkat mekanizmaları, hafif dikkat (global attention, soft attention), yoğun dikkat (local attention, hard attention) ve öz dikkat (self-attention, intra-attention) olarak kategorize edilebilir.

Hafif dikkat yönteminde, her giriş öğesine ne kadar dikkat edilmesi gerektiğini belirleyen ağırlıklar atanır. Bu ağırlıklar, 0 ile 1 arasında değerler alır ve dikkat katmanları tarafından hesaplanır. Ancak, dikkat modelinin doğru çalışabilmesi için bu ağırlıkların toplamının 1 olması gerekir. Bu, modelin tüm girdilere ne kadar dikkat edeceğini belirleyen bir çeşit normalizasyon işlemidir. Bu normalizasyon işlemi için softmax fonksiyonu kullanılır. Softmax fonksiyonu, bir dizi değeri alır ve her bir değeri 0 ile 1 arasında bir değere dönüştürürken toplamlarının 1 olmasını sağlar. Böylece, her bir giriş öğesinin ne kadar önemli olduğunu gösteren bir olasılık dağılımı elde edilir. Deterministik olması, modelin

belirli bir girdiye her zaman aynı çıktıyı vermesi anlamına gelir. Bu, modelin tahminlerinin tutarlı ve öngörülebilir olmasını sağlar. Türevlenebilir olması ise, modelin eğitim sürecinde hata geri yayılımı (backpropagation) kullanarak kendini optimize edebilmesini ifade eder. Türevlenebilirlik, modelin ağırlıklarının güncellenebilmesi ve performansının artırılabilmesi için gereklidir. Softmax fonksiyonu, bu iki özelliği sağlar: Hem dikkat ağırlıklarını normalleştirerek deterministik hale getirir, hem de bu ağırlıkların türevlenebilir olmasını sağlar. Bu sayede, dikkat mekanizması modelin öğrenme sürecine katılarak daha iyi performans göstermesini mümkün kılar.

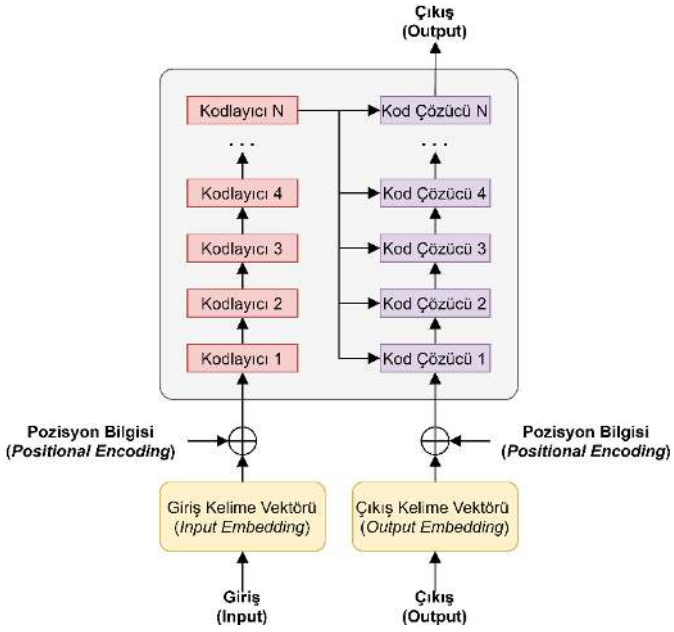
Yoğun dikkat, sinir ağları ve derin öğrenme modellerinde kullanılan bir dikkat mekanizmasıdır. Bu yöntemde, modelin dikkati belirli giriş öğelerine yoğunlaştırılır ve her bir öğe için belirli bir karar verilir. Yoğun dikkat, bir girdi öğesine ya 0 ya da 1 ağırlık atar, yani bir öğe ya tamamen dikkate alınır ya da tamamen göz ardı edilir. Bu mekanizma, sistemin girdisi ile hedef arasındaki karşılıklı bağımlılığı yansıtır. Örneğin, bir çeviri modelinde, yoğun dikkat, çevrilecek bir kelimeyi seçerken sadece en alakalı kelimeye odaklanır. Bu sayede, modelin dikkati belirli bölgelere veya öğelere yoğunlaştırılarak daha hassas ve odaklanmış bir bilgi işleme süreci sağlanır. Ancak, yoğun dikkat yönteminin bazı dezavantajları vardır. Ağırlıklar ya 0 ya da 1 olduğu için, model türevlenemez hale gelir. Bu, modelin eğitim sürecinde hata geri yayılımı (backpropagation) kullanarak kendini optimize edememesi anlamına gelir. Bu nedenle, yoğun dikkatle modelleri eğitmek için pekiştirmeli öğrenme teknikleri kullanılır. Bu teknikler, modelin performansını ödül ve ceza sistemleriyle artırmaya çalışır. Yoğun dikkat modeli, hesaplama açısından daha verimlidir çünkü tüm girdi öğeleri depolanmaz veya işlenmez. Bu da, modelin daha düşük çıkarım süresi ve hesaplama maliyetine sahip olmasını sağlar. Yoğun dikkat, büyük veri setleriyle çalışırken özellikle avantajlıdır çünkü sadece en önemli bilgilere odaklanarak kaynakların daha verimli kullanılmasını sağlar.

Öz dikkat (self-attention) mekanizması, sinir ağlarında giriş öğelerinin birbirleriyle etkileşimini belirleyerek dikkat dağılımını optimize eden bir yöntemdir. Bu mekanizma, her bir giriş öğesinin diğer tüm giriş öğeleriyle olan ilişkisini değerlendirir ve bu ilişkiler doğrultusunda hangi öğelere daha fazla dikkat edilmesi gerektiğini belirler. Model, her bir giriş öğesi için bir dikkat ağırlığı hesaplar; bu ağırlıklar 0 ile 1 arasında değerler alır ve softmax fonksiyonu ile normalize edilir, böylece tüm dikkat ağırlıkları toplamı 1 olur. Bu süreç, modelin uzun mesafeli bağımlılıkları ve karmaşık dil yapılarını anlamasını sağlar. Her bir giriş öğesi, diğer öğelerden topladığı bilgiyi bu ağırlıklar aracılığıyla değerlendirir ve bu

sayede daha anlamlı ve zengin temsiller oluşturur. Öz dikkat mekanizması, tüm giriş öğeleri arasındaki bağımlılıkları aynı anda hesaplayabildiği için paralel hesaplama yeteneğine sahiptir ve bu da modeli daha verimli hale getirir. Ayrıca, giriş verisinin uzunluğundan bağımsız olarak çalışabilmesi, modelin hem kısa hem de uzun giriş dizileriyle etkili bir şekilde başa çıkmasını sağlar. Bu özellik, özellikle doğal dil işleme görevlerinde, dilin karmaşık yapılarını ve uzun mesafeli ilişkilerini anlamada büyük bir avantaj sağlar. Öz dikkat mekanizması, makine çevirisi, metin özetleme, soru yanıtlama gibi doğal dil işleme görevlerinde yaygın olarak kullanılır. Görüntü sınıflandırma ve nesne tanıma gibi görüntü işleme görevlerinde de öz dikkat mekanizması, görüntünün belirli bölgelerine odaklanarak daha doğru ve hızlı sonuçlar elde eder. Öz dikkat, basit ve kolayca paralelleştirilebilir matris hesaplamaları kullanarak tüm giriş öğeleriyle dikkati kontrol eder, bu da modeli hem esnek hem de güçlü kılar ve birçok modern derin öğrenme uygulamasında yaygın olarak kullanılmasını sağlar [104].

4.1 Öz Dikkat Ağları

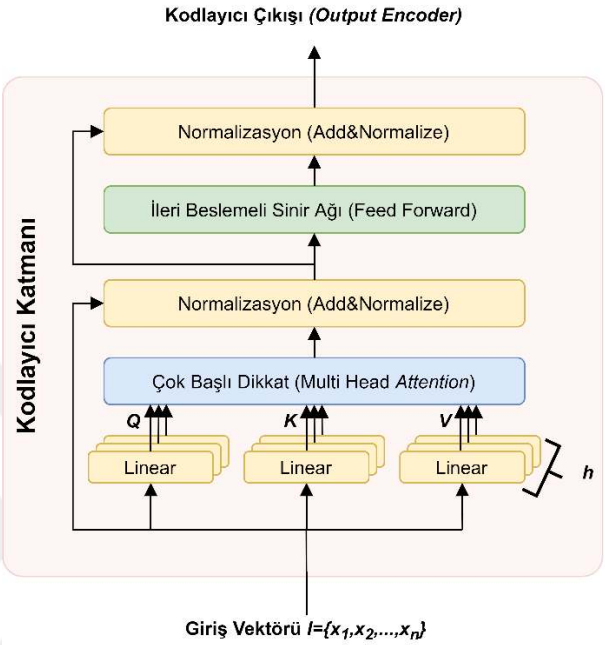
Vaswani ve ark. [90] tarafından önerilen öz dikkat modeli (Self Attention) güncel doğal dil işleme uygulamalarında kullanılan temel modeldir. Transformer olarak bilinen mimari isteğe bağlı (N) sayıda birleştirilmiş kodlayıcı ve kod çözücü katmanlarından oluşur. Şekil 4.1'de öz dikkat mimarisi genel yapısı gösterilmiştir.



Şekil 4.1: Öz dikkat mimarisi genel yapısı.

4.1.1 Kodlayıcı katmanı

Öz dikkat ağlarında her kodlayıcı (encoder), çok başlı dikkat (multi-head attention), ileri beslemeli ağ (feed-forward network), katman normalizasyonu (layer normalization) ve atlamalı bağlantılar (residual connections) bölümlerinden oluşur. Şekil 4.2'de kodlayıcı katmanı gösterilmiştir.



Şekil 4.2: Kodlayıcı katman yapısı

Giriş cümlesi, kelime dizisi olarak alınır. Giriş cümlesi, tokenizasyon adı verilen bir işlemle daha küçük parçalara jetonlara (tokenlere) bölünür. Genellikle, bu parçalar kelime veya alt kelime düzeyinde olur. Her token, yüksek boyutlu bir vektöre dönüştürülür. Bu işleme kelime gömme (Word embedding) adı verilir. Gömme vektörleri, kelimelerin anlamlarını temsil eden sabit boyutlu vektörlerdir. Giriş cümlesi (input) $I = \{x_1, \dots, x_n\}$ olsun. x_i i. Sırada bir jetonu temsil etmektedir. Bu jetonlara karşılık gelen gömme vektörü e_i olarak gösterilsin. Bu vektörlerin boyutu d_{model} ile gösterilir. Gömme matrisi $E = \{e_1, \dots, e_n\}$ olarak gösterilir.

Öz dikkat modeli, kelimelerin sırayla işlendiğini ve sıralarının önemli olduğunu bilmesi için pozisyonel kodlama (positional encoding) kullanır. Pozisyonel kodlama, her kelimenin pozisyon bilgisini gömme vektörüne ekler. Bu kodlama, sinüs ve kosinüs fonksiyonları kullanılarak hesaplanır. Denklem 4.1 ve 4.2' de pozisyon bilgisinin hesaplanması gösterilmiştir.

$$PK_{poz,2i} = \sin\left(\frac{poz}{10000^{2i/d_{model}}}\right) \quad (4.1)$$

$$PK_{poz,2i+1} = \cos\left(\frac{poz}{10000^{2i/d_{model}}}\right) \quad (4.2)$$

Denklemde:

$PK_{poz,2i}$: Çift sırada olan giriş için pozisyon bilgisini gösterir.

$PK_{poz,2i+1}$: Tek sırada olan giriş için pozisyon bilgisini gösterir.

poz : Bir dizi (sequence) içindeki belirli bir pozisyonu gösterir.

i : Modelin gizli boyutunun indeksi gösterir.

d_{model} : Vektör boyutunu gösterir.

Son olarak, kelime gömme vektörleri ve pozisyonel kodlama vektörleri toplanır. Denklem 4.3'de z vektörü kodlayıcı katmanının giriş vektörünü göstermektedir. Bu işlem modelin hem kelimenin anlamını hem de sırasını dikkate almasını sağlar.

$$z_i = e_i + PK_i \quad (4.3)$$

4.1.2 Öz dikkat katmanı

Öz dikkat (self-attention), her bir giriş öğesinin (örneğin bir kelimenin) diğer tüm giriş öğeleriyle ne kadar ilişkili olduğunu belirleyen bir mekanizmadır. Bu yöntem, her bir öğenin diğer öğelerle olan bağımlılıklarını ve ilişkilerini hesaplayarak, hangi öğelere daha fazla dikkat edilmesi gerektiğini belirler. Her kelime için üç farklı vektör hesaplanır. Bunlar Sorgu (query), Anahtar (key) ve Değer (value) vektörleridir. Denklem 4.4'de Z giriş matrisini göstermek üzere, Q sorgu, K anahtar, V değer matrislerini göstermektedir. W ise öğrenilebilir ağırlık matrisidir.

$$Q = ZW^Q, \qquad K = ZW^K, \qquad V = ZW^V \quad (4.4)$$

Her kelimenin diğer kelimelerle olan ilişkisi, sorgu ve anahtar vektörlerinin nokta çarpımı ile hesaplanır. Bu işlem, kelimelerin benzerliklerini ölçmek içindir. Denklem 4.5'de dikkat puanının hesaplanması gösterilmiştir.

$$Dikkat\_puanı(Q, K) = \frac{QK^T}{\sqrt{d_k}} \quad (4.5)$$

Denklem 4.5'de $\sqrt{d_k}$ dikkat puanını normalize etmek için kullanılır. Dikkat puanları, softmax fonksiyonu kullanılarak 0-1 aralığında normalize edilir. Bu, her bir kelimenin diğer

kelimelere ne kadar dikkat etmesi gerektiğini belirler. Denklem 4.6'da dikkat ağırlıklarını hesaplanma yöntemi verilmiştir.

$$dikkat\_ağırlıkları = softmax(\frac{QK^t}{\sqrt{d_k}}) \quad (4.6)$$

Her kelimenin yeni temsili, değer vektörlerinin bu dikkat ağırlıklarıyla çarpılarak hesaplanır. Bu, modelin her kelimenin bağlamını daha iyi anlamasını sağlar. Denklem 4.7'de hesaplama yöntemi gösterilmiştir.

$$dikkat = dikkat\_ağırlıkları \, . V \quad (4.7)$$

Örneğin, "Ali balık tutmayı çok sever." cümlesini ele alalım. Bu cümlede, her kelimenin diğer kelimelerle ilişkisi şöyle hesaplanır. "balık" Kelimesinin "Ali", "tutmayı", "çok", "sever" kelimeleriyle olan ilişkisi sorgu ve anahtar vektörlerinin nokta çarpımı ile belirlenir. Bu ilişkiler softmax fonksiyonu ile normalize edilir ve dikkat ağırlıkları hesaplanır. "balık" Kelimesinin yeni temsili diğer kelimelerin değer vektörlerinin bu dikkat ağırlıkları ile çarpılması sonucu elde edilir.

4.1.3 Çok başlı dikkat

Çok başlı dikkat (Multi-Head Attention), modelin farklı temsil alanlarında daha iyi dikkat dağılımı yapmasını sağlar. Çok başlı dikkat, aynı anda birden fazla dikkat mekanizması (başlık, head) kullanarak giriş vektörlerini paralel olarak işler. Her başlık, kendi dikkat mekanizmasını kullanarak bağımsız olarak çalışır ve sonuçta bu başlıkların çıktıları birleştirilir. Denklem 4.8'de h dikkat başlığı sayısını göstermektedir. W^O Öğrenilebilir ağırlık matrisini göstermektedir.

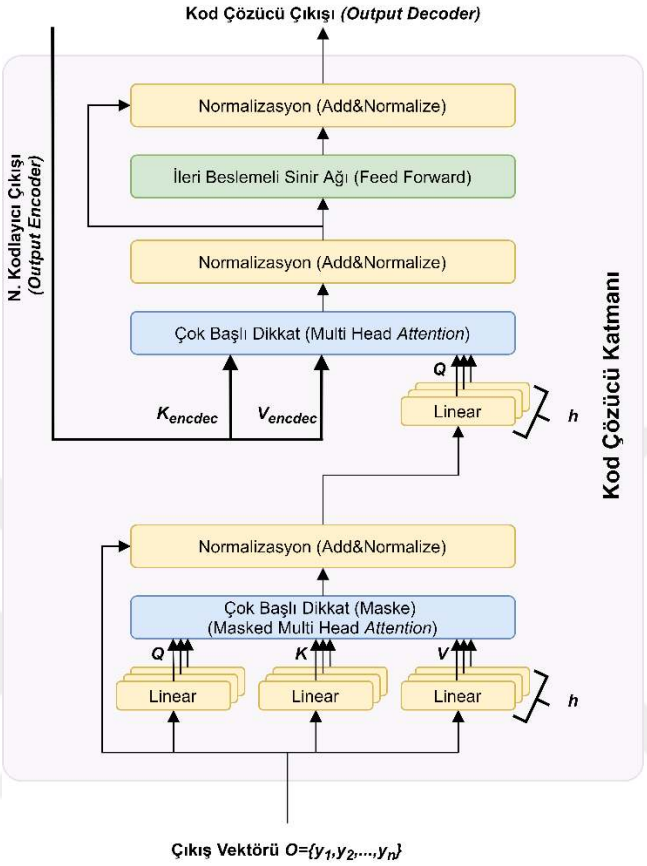
$$Çokbaşlıdikkat(Q, K, V) = Birleştir(dikkat_1, \dots, dikkat_h)W^O \quad (4.8)$$

Birleştirilmiş başlıkların çıktısı, son bir tam bağlantılı katman ile işlenerek nihai çok başlı dikkat çıktısı elde edilir. Bu, modelin daha zengin ve kapsamlı temsiller oluşturmasını sağlar. Ayrıca çok başlı dikkat, farklı başlıkların bağımsız olarak çalışması sayesinde paralel hesaplama yapabilir ve bu modeli daha verimli hale getirir.

4.1.4 Kod çözücü

Öz dikkat modelindeki kod çözücü (decoder), kodlayıcı tarafından üretilen gizli temsilleri (hidden representations) alır ve bunları kullanarak hedef diziyi (örneğin, çeviri cümlesini) üretir. Kod çözücü, kodlayıcı gibi çok katmanlıdır ve her katman üç ana

bileşenden oluşur. Bunlar öz dikkat, kodlayıcı-kod çözücü dikkat (encoder-decoder attention) ve beslemeli ileri ağdır (feed-forward network). Şekil 4.3'de kod çözücü katman yapısı gösterilmiştir.



Şekil 4.3: Kod çözücü katman yapısı

Kodlayıcı da olduğu gibi, kod çözücüde de pozisyonel kodlama kullanılır. Bu, her kelimenin pozisyon bilgisini ekleyerek modelin kelimelerin sırasını anlamasını sağlar. Maskelenmiş kendine dikkat mekanizması, her kelimenin sadece kendisinden önce gelen kelimelere dikkat etmesini sağlar. Bu, modelin gelecekteki kelimeleri görmemesini sağlar.

Modelin her adımda sadece önceki kelimelere bakmasını sağlamak için bir üst üçgen matris kullanılır. Bu matris, modelin gelecekteki kelimelere bakmasını engeller. Matrisin köşegeninin üzerindeki değerler $-\infty$ olarak ayarlanır, diğer değerler 0 olur. Maske matrisi Denklem 4.9'da gösterilmiştir.

$$\begin{bmatrix} 0 & -\infty & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.9}$$

Maskelenmiş dikkat mekanizması, dikkat skorlarına bu maskeyi ekler. Bu sayede, gelecekteki pozisyonların dikkat skorları çok düşük (yaklaşık sıfır) olur ve model bu pozisyonları dikkate almaz. Denklem 4.10'da gösterilmiştir.

$$maskelenmiş\_dikkat\_ağırlıkları = softmax(\frac{QK^t}{\sqrt{d_k}} + maske) \quad (4.10)$$

Kod çözücü, kodlayıcının gizli temsillerine dikkat eder. Bu adımda, kod çözücünün çıktıları ve kodlayıcının çıktıları arasında dikkat hesaplanır. Böylece, kod çözücü, kodlayıcının hangi çıktısına ne kadar dikkat etmesi gerektiğini öğrenir. Kodlayıcı, giriş cümlesini alır ve her kelimeyi bir gizli temsile (hidden representation) dönüştürür. Bu temsiller, kodlayıcının her katmanından çıkan vektörlerdir ve her kelimenin bağlamını yakalar. Kodlayıcı-kod çözücü dikkat mekanizmasında, sorgu vektörleri kod çözücünün çıktılarından, anahtar ve değer vektörleri ise kodlayıcının çıktılarından alınır. Denklem 4.11'de $Z_{kodçözücü}$ kod çözücü giriş vektörlerini, $Z_{kodlayıcı}$ kodlayıcı giriş vektörlerini belirtmektedir. W^Q, W^K, W^V öğrenilebilir ağırlık matrisleridir.

$$Q = Z_{kodçözücü}W^Q, \qquad K = Z_{kodlayıcı}W^K, \qquad V = Z_{kodlayıcı}W^V \quad (4.11)$$

Sonrasında dikkat puanları kodlayıcı katmanında gösterildiği şekilde hesaplanır.

4.1.5 Tam bağlantılı Katman

Öz dikkat modelinin son katmanı, modelin çıktısını alıp kelime olasılıklarına dönüştüren bir tam bağlantılı katman (fully connected layer) ve ardından softmax fonksiyonundan oluşur. Bu katman, modelin her adımda bir kelime tahmin etmesini sağlar. Kod çözücünün son katmanından gelen gizli temsiller (hidden representations), modelin her pozisyondaki kelime için ürettiği vektörlerdir. Bu vektörler, kelimenin bağlamını ve anlamını yakalar. Kod çözücünün son çıktıları, bir tam bağlantılı katmana (linear layer) aktarılır. Bu katman, her bir gizli temsili modelin kelime dağarcığındaki (vocabulary) her kelime için bir lojit (logit) değere dönüştürür. Denklem 4.12'de tam bağlantılı katmanın ağırlıkları (W) ve biasları (b) öğrenilebilir parametrelerdir.

$$lojit = gizli\_temsiller.W + b \quad (4.12)$$

Burada, $gizli\_temsiller$ kod çözücünün son katmanından gelen temsillerdir. Lojitler, softmax fonksiyonu aracılığıyla normalize edilerek kelime olasılıklarına

dönüştürülür. Softmax fonksiyonu, her kelimenin olasılığını hesaplar ve tüm olasılıkların toplamı 1 olur. Denklem 4.13'de gösterilmiştir. $P(kelime_i)$ Kelimenin olasılığıdır.

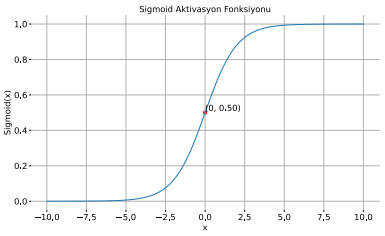
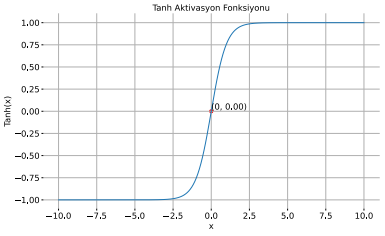
$$P(kelime_i) = \frac{e^{lojit_i}}{\sum_j e^{lojit_j}} \tag{4.13}$$

Tam bağlantılı katman ve softmax fonksiyonu, modelin her pozisyondaki kelime için olasılık dağılımını hesaplamasını sağlar. En yüksek olasılığa sahip kelime seçilerek modelin tahmini olarak çıkış verilir. Genellikle, ışın arama veya aç gözlü arama gibi arama stratejileri kullanılarak en olası kelime dizisi bulunur [90].

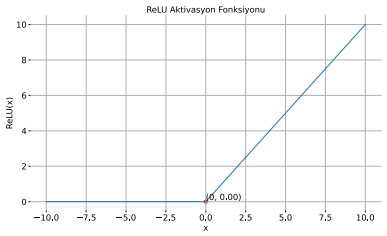
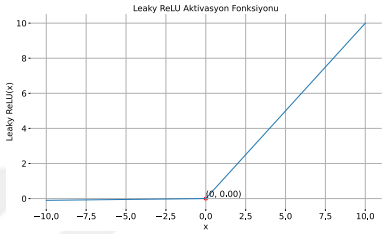
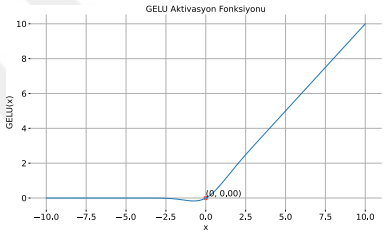
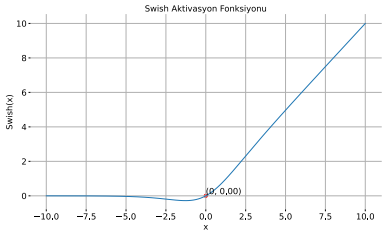
4.2 İleri Beslemeli Sinir Ağları için Aktivasyon Fonksiyonları

Sinir ağları, son yıllarda çok sayıda sorunu çözmek için aktif olarak kullanılmıştır. Farklı türdeki problemlerle başa çıkmak için çeşitli sinir ağları tanıtılmıştır. Aktivasyon fonksiyonları, sinir ağlarının sadece doğrusal hesaplamalar yapmasını engelleyip daha karmaşık hesaplamalar yapabilmelerini sağlamaktır. Bu fonksiyonlar, her bir nöronun çıkışını hesaplayarak ağın öğrenmesini ve veri üzerinde daha iyi tahminler yapmasını sağlar. En yaygın aktivasyon fonksiyonları Sigmoid, Tanh, ReLU, GELU, Swish gibi aktivasyon fonksiyonlarıdır[108]. Ayrıca eğitilebilir parametreye sahip GLU gibi aktivasyon fonksiyonları da bulunmaktadır Çizelge 4.1'de Aktivasyon fonksiyonları verilmiştir.

Çizelge 4.1: Aktivasyon fonksiyonları

Aktivasyon Fonksiyonu	Denklem	Grafik
Sigmoid	$sigmoid(x) = \frac{1}{1+e^{-x}}$	
Tanh	$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	

Çizelge 4.2 (devamı): Aktivasyon fonksiyonları

Aktivasyon Fonksiyonu	Denklem	Grafik
ReLU	$relu(x) = \max(x, 0)$	
Leaky Relu	$L\_relu(x) = \max(x, 0.1x)$	
GELU	$GELU(x) = x * \frac{1}{2}[1 + \text{erf}(x/\sqrt{2})]$ $\approx 0.5x(1 + \tanh[\sqrt{\frac{2}{\pi}}(x + 0.0044715x^3)])$ $\approx \frac{x}{1 + e^{-1.702x}}$ $\approx x * sigmoid(1.702x)$ *erf= Gauss hata fonksiyonu*	
Swish (SiLU)	$Swish(x) = \frac{x}{1 + e^{-x}}$ $= x * sigmoid(x)$	

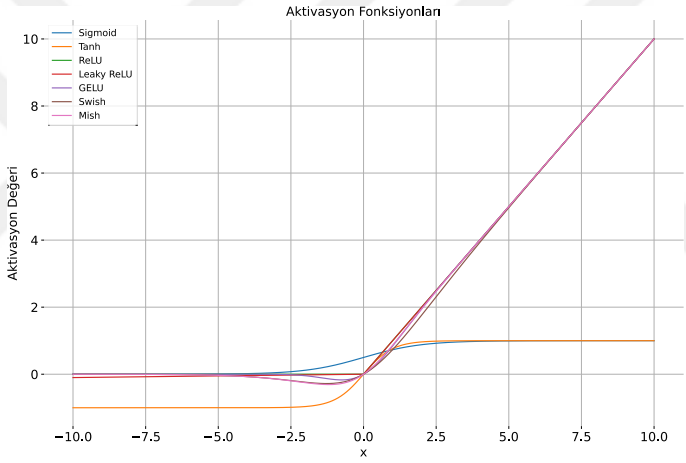
Öz dikkat [90] mimarisi, iki ileri beslemeli sinir ağı katmanı arasındaki doğrusal olmama durumu için ReLU aktivasyon fonksiyonunu kullanmıştır.

Zamanla, birkaç çalışma ReLU dışındaki farklı aktivasyon fonksiyonlarının kullanımını araştırmışlardır. Ramachandran ve ark. [109], öz dikkat modelindeki ReLU'yu Swish fonksiyonu ile değiştirmişlerdir. Bunun WMT 2014 İngilizce-Almanca veri kümesindeki performansı iyileştirdiğini göstermişlerdir.

GPT [93] dil modelinin ön eğitiminde ReLU'yu Gauss Hatası Doğrusal Birimi (GELU) [110] ile değiştirilmiştir. GELU birçok önceden eğitilmiş dil modeli için varsayılan uygulama haline gelmiştir [95,111].

Ayrıca [112]-[113], Geçitli Doğrusal Birimleri (GLU), ve varyantlarını ReLU'nun yerine kullanmayı araştırmıştır. Ön eğitim deneyleri, GLU varyantlarının ReLU aktivasyonu ile klasik öz dikkat mimarisini sürekli olarak iyileştirdiğini göstermiştir. GLU fazladan parametreler getirmiştir ve yapılan çalışma da parametre sayısını diğer modellerle eşleştirebilmek için ileri beslemeli sinir ağının parametre sayısını azalttığı görülmüştür [114].

Öz dikkat mimarilerinde kullanılan tüm aktivasyon fonksiyonlarının gösterimi Şekil 4.4'de verilmiştir.



Şekil 4.4: Transformer mimarisinde kullanılan aktivasyon fonksiyonları.

4.3 Ön Eğitimli Dil Modelleri

Ön eğitimli dil modelleri (Pre-Trained Language Model), büyük miktarda metin üzerinde eğitilmiş yapay zekâ modelleridir. Bu modeller, metinleri anlamak, yorumlamak ve yeni metinler oluşturmak için kullanılır. Öz dikkat (self-attention) mimarisi, bu modellerin temel bileşenidir. Öz dikkat, bir metindeki kelimelerin birbirleriyle olan ilişkilerini öğrenir ve bu ilişkiler doğrultusunda hangi kelimenin diğerlerine ne kadar önemli olduğunu belirler. Bu mekanizma, modelin metin içerisindeki bağlamı daha iyi anlamasına yardımcı olur. Bu sayede, model cümlelerin anlamını daha derinlemesine kavrar ve daha tutarlı ve anlamlı metinler üretebilir. Öz dikkat mimarisine dayalı modeller, dil işleme

görevlerinde önemli başarılar elde etmiştir. Örneğin, BERT ve GPT gibi modeller, bu mimariyi kullanarak metinleri işler ve anlam çıkarır.

BERT (Bidirectional Encoder Representations from Transformers)[95], öz dikkat mimarisini kullanarak metnin her iki yönündeki bağlamı öğrenir. Bu model, iki ana strateji kullanarak eğitilir: maskelenmiş dil modelleme ve ikili cümle tahmin görevi. Maskelenmiş dil modelleme stratejisinde, cümledeki bazı kelimeler gizlenir (maskelenir) ve modelin bu kelimeleri tahmin etmesi beklenir. Bu sayede BERT, kelimelerin bağlamını daha iyi anlayarak metinleri daha doğru bir şekilde işler. İkili cümle tahmin görevinde ise, modelin iki cümlenin birbirini takip edip etmediğini belirlemesi istenir. Bu, modelin cümleler arası ilişkileri daha iyi kavramasını sağlar. BERT, bu stratejilerle eğitildiğinde, doğal dil işleme görevlerinde üstün performans gösterir, çünkü metin içindeki karmaşık ilişkileri ve bağlamları öğrenmiştir.

GPT (Generative Pre-trained Transformer) [93] ise metin üretimi için özelleşmiş bir modeldir. GPT, otoregresif dil modelleme stratejisi kullanarak eğitilir, yani bir cümledeki sonraki kelimeyi tahmin etmek için eğitilir. Bu süreçte, model, cümlenin başından sonuna kadar kelimeleri sırayla tahmin eder. Her adımda, bir sonraki kelimeyi tahmin etmek için önceki kelimelerin bilgisini kullanır. Bu strateji, modelin doğal ve akıcı metinler oluşturabilmesini sağlar. GPT, büyük miktarda metin üzerinde ön eğitimden geçer ve bu süreçte dilin yapısını ve kalıplarını öğrenir. Eğitim süreci tamamlandıktan sonra, GPT çeşitli dil işleme görevlerine uygulanabilir ve yüksek kaliteli metin üretimi, özetleme ve çeviri gibi görevlerde etkili performans sergiler. Hem BERT hem de GPT, ön eğitim sırasında büyük miktarda metin kullanarak dilin inceliklerini öğrenir ve bu bilgiyi farklı dil işleme görevlerinde başarıyla kullanabilirler.

mBART (multilingual BART) [32], çok dilli metin işleme ve çeviri için özelleşmiş bir modeldir. mBART, BART (Bidirectional and Auto-Regressive Transformers) mimarisinin çok dilli versiyonudur. BART, hem kodlayıcı (encoder) hem de kod çözücü (decoder) mimarisine sahip bir modeldir ve sıralama düzeltme, metin tamamlama gibi görevlerde başarılıdır. mBART, çeşitli dillerdeki büyük miktarda metin üzerinde ön eğitim alarak, çok dilli dil modellerinin gücünü artırır. Model, öncelikle bir dildeki metni bozarak, ardından bu metni yeniden oluşturarak eğitilir. Bu süreç, modelin dilin yapısını ve bağlamını öğrenmesine yardımcı olur. mBART, özellikle dil çiftleri arasında çeviri yaparken yüksek performans gösterir. Bu, onu çok dilli metin işleme ve çeviri görevlerinde değerli bir araç

haline getirir. mBART'in eğitimi sırasında, metinlerin hem kaynak hem de hedef dildeki bağlamını öğrenmesi sağlanır, bu da modelin çeviri kalitesini artırır.

Özetle, BERT, GPT ve mBART gibi ön eğitimli dil modelleri, öz dikkat mimarisi kullanarak metinleri işler ve çeşitli dil işleme görevlerinde üstün performans sergiler. Herbir modelin kendine özgü ön eğitim stratejileri ve kullanım alanları vardır, bu da onları belirli görevlerde güçlü kılar. BERT, bağlamı derinlemesine anlama; GPT, doğal ve akıcı metin üretimi; mBART ise çok dilli metin işleme ve çeviri konularında öne çıkar.

5. GELİŞTİRİLEN YÖNTEMLER

Bu bölümde tez sürecinde yapılan dört çalışma detaylıca tanıtılmıştır. Bu çalışmalardan ilki Türkçe-İngilizce sinirsel makine çevirilerinin düşük kaynak problemini giderebilmek için oluşturulan paralel külliyattır. Türkçe-İngilizce dil çifti arasında literatürde sınırlı sayıda ve miktarda paralel veri bulunmaktadır. Düşük kaynak probleminin giderilmesi için yapılan ilk çalışmada akademik metinlerden oluşturulmuş paralel külliyat sunulmuştur. Hazırlanan ikinci çalışmada ise Bert, DistilBert, Electra gibi Türkçe ön eğitimli dil modellerinin dil anlama görevlerindeki başarısı ölçülmüştür. Bu çalışmanın amacı Türkçe dilini en iyi anlayabilecek dil modelinin belirlenerek çeviri sistemlerinde kullanmanın yollarını araştırmaktır. Hazırlanan üçüncü çalışmada ise çok dilli (multilingual) ön eğitimli dil modellerinin çeviri sistemlerine uyarlanması çalışmasıdır. Bu çalışmanın temel amacı birden fazla dilde ön eğitime tutulmuş dil modelini düşük kaynaklı Türkçe-İngilizce dil çiftine uyarlamaktır. Bu görev için parametre etkili yöntemler kullanılarak modelin ön eğitim sırasında öğrendiği bilgileri unutmadan yeni bir göreve uyum sağlamasını sağlamaktır. Son olarak Türkçe-İngilizce sinirsel makine çevirilerinde akıcılığı artırmak için ön eğitimli dil modellerinin kullanımı araştırılmıştır. Bu çalışmada ilk çalışma da oluşturulan paralel külliyat kullanılarak akademik çevirilerde kalitenin artırılması sağlanmıştır. Tamamen akademik metinlerle ön eğitimden geçen SciBert dil modeli öz dikkat mimarisiyle oluşturulmuş sinirsel makine çeviri sistemine entegre edilmiştir. Yapılan tüm çalışmalar Türkçe-İngilizce dil çifti arasındaki düşük kaynak problemini ortadan kaldırarak çeviri kalitesini artırmayı amaçlamıştır.

5.1 Uygulama 1: Türkçe İngilizce Paralel Külliyat oluşturulması

Çeviri sistemlerinde paralel külliyat, yüksek kaliteli ve doğru çeviri sonuçları elde edebilmek için kritik bir öneme sahiptir. Paralel külliyat, iki dilde karşılıklı olarak eşleşmiş cümle çiftlerinden oluşur. Yüksek kaliteli ve büyük ölçekli paralel külliyatlar, makine çeviri modellerinin daha iyi performans göstermesini sağlar. Daha fazla veri, modelin dildeki çeşitli kalıpları, deyimleri ve özel ifadeleri öğrenmesini mümkün kılmaktadır. Türkçe-İngilizce makine çevirilerinin düşük kaynaklı dil çifti olarak gösterilmesinin en önemli sebebi yeterli büyüklükte ve ölçekte paralel külliyatın olmamasıdır. Tezin bu uygulamasında

ilk olarak bu sorun üzerinde durulmuştur. Literatürde 676K cümleye sahip Tatoeba [75] en yüksek boyutlu paralel külliyat olarak tespit edilmiştir. Yüksek kaynaklı dillerde bu sayı 50M paralel cümleye çıkabilmektedir. Yapılan bu çalışma ile TR-EN dilleri arasında yaklaşık 1.2M paralel cümleye sahip külliyat oluşturulmuştur. Paralel külliyat oluşturulurken lisansüstü tezlerin özet kısımları kullanılmıştır. Veri web kazıma yöntemleriyle toplanmıştır. Toplanan veri önce vektörel düzeyde sinir modelleri yardımıyla hizalanmıştır. Ek olarak kelime tabanlı sözlük kullanılarak da hizalama yapılmıştır. Bu iki yöntemin kesişim kümesi belirlenerek TR-EN için paralel külliyat oluşturulmuştur. Oluşturulan külliyatın kalitesini ölçebilmek için Bi-LSTM tabanlı sinirsel makine çeviri sistemi geliştirilmiştir. Geliştirilen sistemin temel amacı paralel külliyatın kalitesini ölçmektir. Sonuç olarak 15.8 Bleu puanı elde edilmiştir.

5.1.1 Materyal

Külliyatın oluşturulması için lisansüstü tezlerin özet bölümleri kullanılmıştır. Yöktez (https://tez.yok.gov.tr/UlusalTezMerkezi/) sisteminde açık erişim olarak yayınlanan lisansüstü tezlerin özet bölümleri İngilizce ve Türkçe olarak kullanıma açıktır. Web kazıma yöntemleri ile 2015-2020 yılları arasında 178 farklı alanda yayınlanmış 238,233 tez bu çalışma için kullanılmıştır. Çalışma da kullanılan tezler yayınlandığı çalışma alanına göre taranmıştır. Kullanılan tezlerin alanlara göre dağılımı Çizelge 5.1'de verilmiştir.

Çizelge 5.1: Kullanılan tezlerin alanlara göre dağılımı

Alanlar	Oran (%)
Eğitim ve Öğretim	6.1
İşletme	3.1
Psikoloji	2.9
Ziraat	2.77
Makine Mühendisliği	2.73
Kimya	2.45
Tarih	2.44
Türk Dili ve Edebiyatı	2.29
Matematik	2.23
Bilgisayar Bilimleri	2.18
Diğerleri	68.5

5.1.1.1 Metin önişleme adımları

Taranan tezler konu alanlarına göre web kazıma yapılarak içerikleri metin dosyalarına kayıt edilmiştir. Sonrasında sırasıyla aşağıdaki önişlem adımları uygulanmıştır.

- Her dosya teker teker okunarak ve farklı alan adlarıyla kayıt edilen aynı tezler filtrelenerek tekrar eden tezler çıkarılmıştır.
- Metin verileri cümlelere bölünmüştür ve bu aşamada cümle sayısı yaklaşık 3M Türkçe ve İngilizce cümlelerden oluşmuştur.
- Gereksiz boşluklar, sekmeler ve yeni satır karakterleri temizlenmiştir.
- HTML etiketleri, emoji, URL gibi özel karakterler metinden çıkarılmıştır.
- Metin küçük harflere çevrilmiştir.
- Cümle boyutları 20-400 karakter aralığında seçilerek çok uzun veya çok kısa cümleler filtrelenmiştir.

5.1.1.2 Cümle hizalama

Kullanılan tezlerin özet bölümleri, birden fazla cümleden oluşan Türkçe ve İngilizce cümlelere sahiptir. Çeviri sistemlerinde kullanılabilmesi için külliyatta her bir cümleye karşılık gelen bir cümle olmak zorundadır. Paralel külliyat oluştururken karşılaşılan en büyük zorluklardan biri cümlelerin hizalanmasıdır. Literatürde cümle hizalaması ile ilgili iki temel yaklaşım bulunmaktadır: HunAlign [115] ve Vecalign [116] hizalama algoritmaları.

HunAlign, istatistiksel yöntemlere dayalı bir hizalama algoritmasıdır ve kelime çiftlerinin birlikte görünme olasılıklarına göre cümleleri hizalar. Bu yöntem, büyük miktarda veri ile eğitildiğinde yüksek doğruluk sağlayabilir ancak dil bilgisel farklılıklar ve düşük frekanslı kelimelerle başa çıkmada zorluk yaşayabilir.

Vecalign ise, cümlelerin vektör temsillerini kullanarak hizalama yapar. Bu yöntem, kelime gömme vektörleri (word embeddings) ve sinir ağları (neural networks) kullanarak cümlelerin anlamlarını sayısal olarak temsil eder ve bu temsiller üzerinden hizalama yapar. Vecalign, anlam temelli hizalama yapabildiği için dil bilgisel farklılıkları ve düşük frekanslı kelimeleri daha iyi yönetebilir.

Bu çalışmada, cümlelerin hizalanması için hem HunAlign hem de Vecalign algoritmaları kullanılarak ayrı ayrı hizalama yapılmıştır. Her iki algoritmanın da kendi avantajları ve dezavantajları vardır, bu nedenle her iki yöntemin sonuçları karşılaştırılarak kesişen hizalamalar paralel külliyata eklenmiştir. Bu yaklaşım, her iki algoritmanın güçlü yönlerinden yararlanarak daha doğru ve güvenilir bir paralel külliyat oluşturmayı sağlamıştır.

HunAlign cümle hizalama algoritması

HunAlign [115], metin çiftlerini (kaynak ve hedef dildeki metinler) cümle düzeyinde hizalamak için kullanılan bir araçtır. İstatistiksel yöntemlere dayanır ve kelime çiftlerinin birlikte görünme olasılıklarını kullanarak cümle hizalama işlemini gerçekleştirir. HunAlign, kaynak ve hedef dildeki metinlerin cümlelere bölünmüş ve belirli bir formatta hazırlanmış olmasını gerektirir. Genellikle her cümle bir satıra gelecek şekilde düzenlenir. HunAlign, cümle hizalaması yapabilmek için kelime hizalama modellerine ihtiyaç duyar. Bu modeller, iki dil arasındaki kelime çiftlerinin birlikte görünme olasılıklarını belirler. Model eğitimi, genellikle önceden hizalanmış küçük bir paralel metin kümesi (sözlük) kullanılarak yapılır. Bu çalışma da Pavlick ve Ark. [117] tarafından hazırlanan sözlük kullanılmıştır. Bu metinler üzerinden kelime eşleşme olasılıkları hesaplanır ve bir kelime hizalama modeli oluşturulur. HunAlign, cümlelerin uzunlukları arasındaki ilişkiyi de göz önünde bulundurur. Kaynak ve hedef dildeki cümle uzunlukları arasındaki olasılık dağılımları hesaplanır. Örneğin, uzun bir cümlenin uzun bir cümle ile hizalanma olasılığı daha yüksektir. HunAlign, nispeten hızlı ve basit bir algoritmadır. Büyük veri kümelerinde de hızlı bir şekilde çalışabilir. Dezavantaj olarak dil bilgisel farklılıklar ve düşük frekanslı kelimeler cümle hizalama sürecinde problem oluşturabilir.

Vecalign cümle hizalama algoritması

Vecalign [116], cümle hizalama işlemi için cümlelerin anlam temsillerini kullanır. Cümleleri vektör temsilleriyle ifade eder ve bu temsiller üzerinden hizalama yapar. Bu sayede, dil bilgisel farklılıkları ve anlam benzerliklerini dikkate alarak daha doğru hizalamalar yapabilir. Kaynak ve hedef dildeki her cümle, kelime gömme teknikleri (örneğin, Word2Vec, GloVe) veya cümle düzeyinde gömme modelleri (örneğin, BERT, Sentence-BERT) kullanılarak vektörlere dönüştürülür. Bu çalışma da Facebook araştırmacıları tarafından önerilen [118] LASER çok dilli cümle yerleştirme (sentence embedding) vektörlerini kullanılarak cümleler hizalanmaya çalışılmıştır.

Örneğin Türkçe (T_i) ve İngilizce (E_j) cümleler ile gösteriliyor olsun. Bu cümlelerin vektör gösterimleri ise $v(T_i)$, $v(E_j)$ olarak gösterilir. İki cümle vektörü arasındaki benzerlik, kosinüs benzerliği kullanılarak hesaplanır [119,120]. Kosinüs benzerliği, iki vektör arasındaki açıyı ölçer ve Denklem 5.1'deki gibi hesaplanır.

$$benzerliği(v(T_i), v(E_j)\ ) = \frac{v(T_i).v(E_j)}{\|v(T_i)\| \times \|v(E_j)\|} \tag{5.1}$$

Denklem de:

$benzerlik\big(v(T_i), v(E_j)\big)$: Cümle vektörlerinin benzerlik oranını gösterir.

$v(T_i).v(E_j)$: Cümle vektörlerinin Skaler (nokta) çarpımını gösterir.

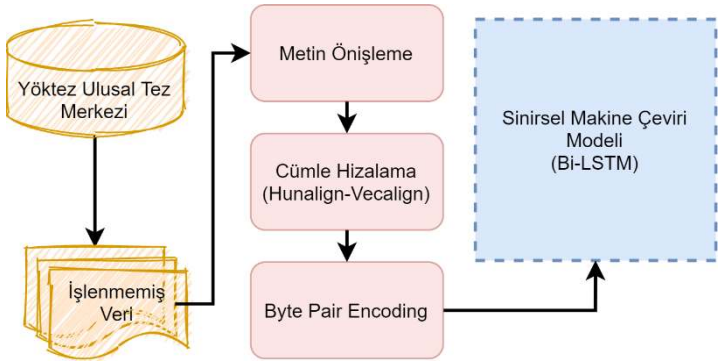
$\|v(T_i)\| \times \|v(E_j)\|$: Cümle vektörlerinin büyüklüklerinin çarpımını gösterir.

Vecalign algoritmasın da her iki dildeki cümle çiftleri arasındaki benzerlik matrisi oluşturulur ve bu matrise dayanarak optimal hizalamalar belirlenir. Eğer bir Türkçe cümle birden fazla İngilizce cümle ile yüksek benzerlik gösteriyorsa (örneğin, ikinci en yüksek benzerlik skoru 0.8'den büyükse), bu cümleler birleştirilerek hizalanır.

Vecalign algoritması, cümlelerin anlam temsillerini kullanarak daha esnek ve doğru hizalamalar yapmayı sağlar. Özellikle, bir cümlenin birden fazla cümle ile hizalanması durumunda, bu yaklaşım yüksek benzerlik skorlarına dayalı olarak en uygun hizalamayı belirler.

5.1.2 Bi-LSTM tabanlı sinirsel makine çeviri

Oluşturulan paralel külliyatın kalitesini test edebilmek için Bi-LSTM tabanlı sinirsel makine çeviri sistemi [76] geliştirilmiştir. Geliştirilen sistem dikkat mekanizmasına sahip Kodlayıcı-Kod Çözücü mimarisini kullanır. Kelime giriş yöntemi olarak Bayt Çifti Kodlamasını kullanır. Şekil 5.1'de önerilen yöntem gösterilmiştir.

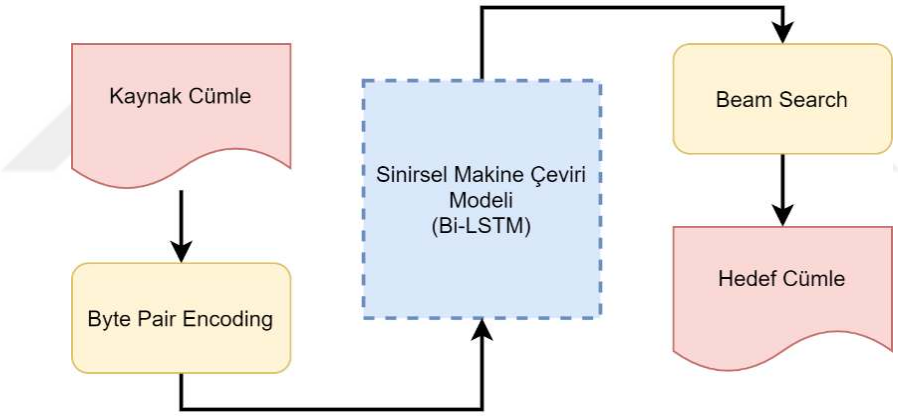


Şekil 5.1: Önerilen yöntem.

Önerilen yöntem özet olarak şu adımlardan oluşur:

- Yöktez sisteminden ham-işlenmemiş veri (Raw Data) web kazıma yöntemleri ile metin dosyalarına kaydedilir.
- Tüm dosyalar taranarak tekrarlayan tezler silinir. Sonrasında tüm metinler küçük harfe çevrilir

- İlk olarak HunAlign yöntemi kullanılarak cümleler hizalanır. Sonrasında Vecalign algoritmasıyla başka bir hizalanmış veri elde edilir. Sonrasında iki veri birleştirilerek kesişim kümeleri nihai paralel külliyatı oluşturur.
- Nadir bulunan kelime sorununun üstesinden gelebilmek için kelimeler bayt çifti kodlaması yöntemiyle tokenize edilmiştir.
- Bi-LSTM mimarisine sahip (N=2, FFNN=1024) sinirsel makine çeviri sistemi oluşturulmuştur. Parametre sayısı yaklaşık 31M civarındadır.
- Mimari Google Colab bulut hesaplama üzerinde 16 GB GPU üzerinde test edilmiştir.
- Tüm işlemlerde Python Programlama dili kullanılmıştır. Web kazıma için "Selenium", metin önişleme adımları için "NLTK" ve çeviri sistemi için "Facebook-Fairseq" kütüphaneleri kullanılmıştır.
- Model eğitildikten sonra kelime çıkarım için Işın Arama [96] algoritması kullanılmıştır. Şekil 5.2'de cümle çeviri yöntemi gösterilmiştir.



Şekil 5.2: Cümle çeviri yöntemi.

- Modelin test aşamasında TED test seti [121] kullanılmıştır. Başarı puanlarının hesaplanmasında Bleu [122] metriği kullanılmıştır.
- Önerilen yöntemin test aşamasında Ted veri seti kullanılmıştır. Ted veri setini oluşturan eğitim ve test datalarından test seti kullanılarak modelin başarımı hesaplanmıştır. Bu yöntem Sıfır Vuruşlu Eğitim (Zero Shot Learning) [84] olarak adlandırılmaktadır.
- Oluşturulan paralel külliyat ve önerilen Bi-LSTM tabanlı mimari Ted veri setinde Türkçe-İngilizce çeviriler için 15.8 Bleu İngilizce-Türkçe çeviriler için 14.9 Bleu puanı elde etmiştir.

5.1.3 Sonuç

Yapılan bu çalışma ile Türkçe İngilizce dilleri arasında makine çeviri kalitesinin artırılması amacıyla paralel külliyat oluşturulmuştur. Oluşturulan paralel külliyat yaklaşık 1.2M cümleye sahiptir ve literatürde Türkçe için bulunan en yüksek boyutlu ve bilimsel veri seti olma özelliği göstermektedir. Oluşturulan paralel külliyatı test edebilmek için önerilen Bi-LSTM tabanlı sinirsel makine çeviri sistemi farklı bir veri setinde yaklaşık 15 Bleu puanı elde etmiştir. Bu puanlar dikkate alındığında veri setinin TR-EN üzerinde düşük kaynak problemini azaltabileceği tespit edilmiştir. Gelecekte yapılacak olan çalışmalarda araştırmacıların rahatlıkla kullanabilmeleri için oluşturulan veri seti açık kaynak olarak yayınlanmıştır. (https://github.com/ilhamisel/tr-en_translate)

5.2 Uygulama 2: Ön Eğitimli Dil Modellerinin Türkçe Dili Üzerinde Başarısının Ölçülmesi

İngilizce için oluşturulan ön eğitimli dil modelleri, dil anlama görevlerinde başarılı sonuçlar elde etmişlerdir [94,95]. Bu modeller arasında BERT (Bidirectional Encoder Representations from Transformers), DistilBERT ve Electra, en yaygın ve etkili olanlardandır. Bu modeller, metin sınıflandırma, duygu analizi, soru yanıtlama, metin tamamlama ve makine çevirisi gibi çeşitli doğal dil işleme (NLP) görevlerinde üstün performans göstermektedir. Örneğin, BERT[95], doğal dil anlama (GLUE) benchmark'larında ve SQuAD (Stanford Question Answering Dataset) gibi veri setlerinde yüksek doğruluk oranları elde etmiştir. DistilBERT[123] , BERT'in daha hafif ve hızlı bir versiyonu olarak benzer başarılar elde ederken, daha az hesaplama kaynağı gerektirir. Electra[124] ise, ön eğitim sürecinde hem kelime düzeyinde hem de cümle düzeyinde ikili sınıflandırma görevlerini kullanarak daha verimli ve hızlı sonuçlar üretir.

Tez çalışmasının bu bölümünde, Türkçe ön eğitim ile oluşturulmuş sınırlı sayıdaki dil modellerinden transfer öğrenme ile dil anlama görevlerinde deneyler yapılmıştır. Dil anlama görevi olarak, yazar profili oluşturma (Author Profiling) uygulamalarından biri olan cinsiyet tespiti seçilmiştir. Yazar profili oluşturma, bir metnin içeriği ve üslubuna bakarak yazarın yaş, cinsiyet ve kişilik özellikleri gibi unsurları tahmin etme görevi olarak tanımlanabilir. Bu görev, çeşitli doğal dil işleme uygulamalarında önemli bir yere sahiptir. Çevrimiçi iletişimde, özellikle sosyal medya, forumlar, bloglar ve e-posta gibi platformlarda, kişinin cinsiyetini belirlemek, sahtekarlık ve dolandırıcılık gibi siber suçların tespit edilmesinde büyük önem taşır. Kişinin yazı stiline ve dil kullanımına dayalı olarak yapılan cinsiyet tespiti, sahte profillerin ve kötü niyetli kullanıcıların belirlenmesinde yardımcı olabilir.

Ayrıca, yazar profili oluşturma, sahte haber yayma gibi adli olayların incelenmesinde de kritik bir rol oynar. Sahte haberlerin yayılması, kamuoyunu yanıltma ve toplumsal huzursuzluk yaratma potansiyeline sahiptir. Bu nedenle, sahte haberlerin yazarlarını tespit etmek ve sorumlularını belirlemek için yazar profili oluşturma teknikleri kullanılabilir. Bu teknikler, adli soruşturmalarda delil toplama ve suçluların izini sürme süreçlerinde etkili bir şekilde kullanılabilir.

Reklam ve pazarlama alanında da yazar profili oluşturmanın önemi büyüktür. Markalar ve şirketler, hedef kitlelerini daha iyi anlamak ve onlara daha etkili bir şekilde ulaşmak için yazar profili oluşturma tekniklerinden faydalanabilir. Kişilerin demografik

bilgilerini ve kişilik özelliklerini tahmin ederek, daha özelleştirilmiş ve etkili reklam kampanyaları oluşturabilirler. Bu, müşteri memnuniyetini artırabilir ve satışları yükseltebilir.

Psikolojik ve sosyolojik araştırmalarda da yazar profili oluşturma teknikleri kullanılmaktadır. Araştırmacılar, insanların yazı stillerine dayanarak psikolojik durumlarını, kişilik özelliklerini ve sosyokültürel arka planlarını analiz edebilirler. Bu, bireylerin ve toplumların davranışlarını daha iyi anlamaya ve analiz etmeye yardımcı olabilir. Örneğin, depresyon belirtilerini gösteren yazıları tespit ederek, erken müdahale ve tedavi olanakları sunulabilir.

Kısa metinlerden cinsiyet tespiti yapmak, diğer dillerde olduğu gibi Türkçe için de oldukça zor bir görevdir. Özellikle Twitter gönderileri, dil kurallarına uymayan kelimeler, kısaltmalar ve farklı cümle yapıları içerebilir. Bu tür kısa ve genellikle kuraldışı metinlerde cinsiyet tespiti yapmak, dil modelleri ve doğal dil işleme teknikleri açısından ciddi bir zorluk teşkil eder. Ancak, bu zorluklara rağmen, Twitter gönderileri ve benzeri kısa metinler, cinsiyet belirleme görevlerinde sıkça kullanılmaktadır. Bu, modellerin gerçek dünya uygulamalarında ne kadar etkili olduğunu gösterir.

Bu çalışma, cinsiyet belirleme görevini ikili sınıflandırma problemi olarak ele almıştır. Sınıflandırma problemini çözmek için üç farklı yöntem izlenmiştir. İlk olarak makine öğrenmesi metodu (TF-IDF + SVM) kullanılmıştır. Sonrasında derin öğrenme yöntemleri (LSTM, CNN) mimarileri geliştirilmiştir. Son olarak Türkçe için oluşturulmuş ön eğitimli dil modellerini (BERT, DistilBERT, Electra) ince ayar yaparak problem çözülmeye çalışılmıştır. Yapılan deneysel çalışmalarda kelime boyutu 128K olan BERT ön eğitimli dil modeli en yüksek başarıyı (%80.1) elde etmiştir.

Bu çalışmanın temel amacı, çeviri sistemlerinde kullanılması planlanan Türkçe için hazırlanan ön eğitimli dil modellerinin dil anlama görevlerindeki başarısını ölçmektir. Sonuçlar, Türkçe için geliştirilen dil modellerinin, yazar profili oluşturma gibi karmaşık dil anlama görevlerinde etkili olduğunu göstermektedir. Bu, Türkçe dil modellerinin gelecekteki çeviri sistemlerinde ve diğer doğal dil işleme uygulamalarında daha geniş bir yelpazede kullanılabileceğini göstermektedir.

5.2.1 Arka plan ve ilgili çalışmalar

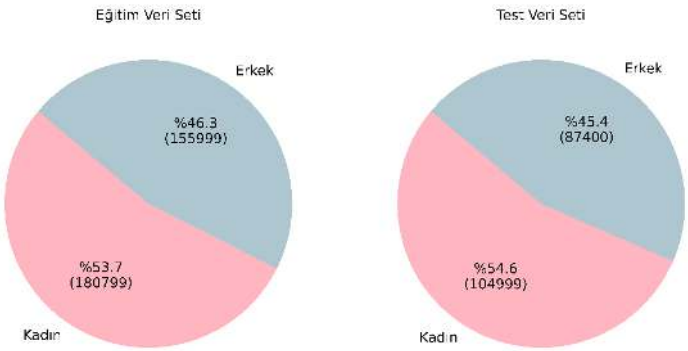
Her yıl uluslararası düzenlenen CLEF (Conference and Labs of the Evaluation Forum) konferanslarında PAN (yazar profili oluşturma) görevleri bulunmaktadır. Yaş, cinsiyet, eğitim gibi temel özelliklerin tespiti görevlerinin yanında, herhangi bir yazarın sahte haber yaymaya istekli olup olmadığının belirlenmesi gibi özel görevler de bulunmaktadır [125]. 2013 yılından itibaren [124, 125] bu görevler içerisine giren cinsiyet belirleme görevi özellikle siber suçların tespitinde önemli bir yer tutmaktadır. İngilizce gibi yüksek kaynaklı dillerde çok fazla sayıda çalışma olmasına rağmen Türkçe için yeterli sayıda çalışma yapılmamıştır. Klasik duygu analizi gibi sınıflandırma problemlerinden farklı olarak sosyal medya paylaşımlarıyla sınıflandırma yapmak oldukça karmaşık ve başarısı düşük olabilmektedir. Bu durumun genel sebepleri şunlardır:

- Metinlerin boyutu bazen bir veya iki kelime gibi oldukça kısa olması,
- Kelimelerin sözlükte kullanıldığı şekilde yazılmayıp bazı harflerin eksik veya bazılarının fazla yazılması,
- Metin içinde değinme (Mention), etiket (Hashtag), bağlantı (url) veya sındamga (emoji) gibi web ifadelerin oldukça fazla olması

Türkçe için sınırlı sayıda çalışma ve veri seti bulunmaktadır. Sezerer ve Ark. tarafından [128] hazırlanan ve Twitter üzerinden toplanan cinsiyetlerin etiketlendiği veri seti bu çalışma için ulaşabildiğimiz en kapsamlı veri setidir. Yazarlar bu veri setiyle yaptıkları çalışmada (Bag-Of-Words + SVM) %72.32'lik başarıya ulaşmışlardır. Aynı veri seti bu çalışma için de kullanılmıştır.

5.2.2 Materyal

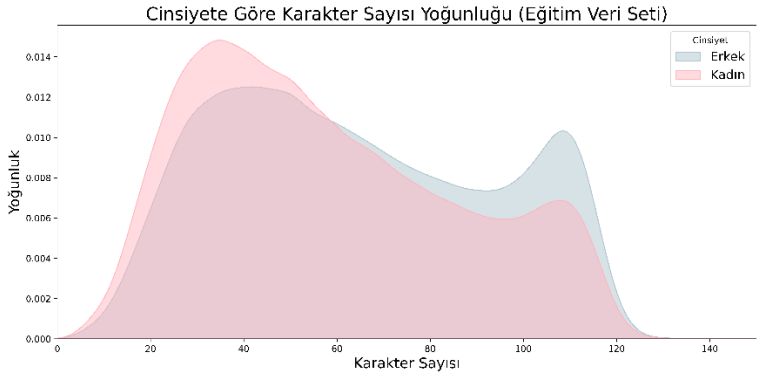
Bu çalışma da Sezerer ve Ark. [128] tarafından hazırlanan veri seti kullanılmıştır. Veri seti her biri 100 adet olmak üzere 5292 farklı kişi tarafından gönderilen mesajlardan oluşmaktadır. Yazarlar veri setini 3368 eğitim ve 1924 test olmak üzere ayırmışlardır. Veri setine ait dağılım bilgileri Şekil 5.3'te verilmiştir. Kadın olarak etiketlenen gönderi sayısı tüm veri setinin %53,68'ini oluşturmaktadır. Bu yönüyle veri seti için dengeli olduğu söylenebilir.



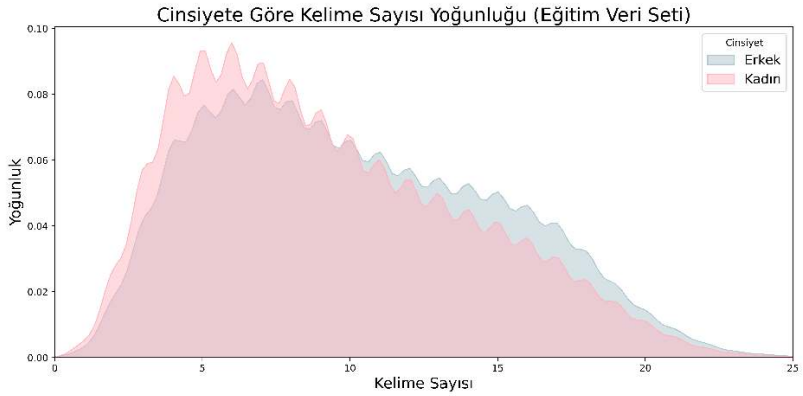
Şekil 5.3: Eğitim ve test veri setleri cinsiyete göre dağılımı.

Tüm veri seti incelendiğinde Şekil 5.4 ve 5.5'te de görüldüğü gibi veri seti hakkında şu bilgiler gözlenmiştir;

- 25-45 karakter aralığındaki gönderilerin kadın kullanıcılar tarafından daha fazla tercih edildiği, 100 karakter üzeri gönderilerde ise durumun tam tersi olduğu,
- 40 karakterlik gönderilerde cinsiyetin kadın olma ihtimalinin erkek olma ihtimalinden daha fazla olduğu
- 110 karakterlik gönderilerde cinsiyetin erkek olma ihtimalinin kadın olma ihtimalinden daha fazla olduğu
- 100 karakter bulunduran gönderilerin her iki cinsiyet içinde beklenmeyen bir şekilde az tercih edildiği,
- 6 kelimelik gönderilerde cinsiyetin kadın olma ihtimalinin erkek olma ihtimalinden daha fazla olduğu 15 kelimelik gönderilerde durumun tam tersi olduğu tespit edilmiştir.



Şekil 5.4: Gönderilerin karakter sayısının cinsiyete göre oranı.



Şekil 5.5: Gönderilerin kelime sayısı yoğunluğu.

5.2.3 Metin önişleme

Sezerer ve ark. hazırlamış [128] olduğu veri seti her bir kullanıcı için xml türü dosyalardan oluşmuştur. Bu dosyaların her birinde bir kişiye ait 100 adet sosyal medya gönderisi bulunmaktadır. Araştırmacılar için açık erişim olarak sunulan veri seti üzerinde uygulanan metin ön işleme adımları şu şekildedir:

- Xml dosyaları sırasıyla işlenerek verilen etiketler sayesinde veri seti tek bir dosya halinde csv formatına çevrilmiştir. Kadınlar için "0" erkekler için "1" etiketi kullanılmıştır.
- Tüm gönderiler küçük harfe dönüştürülmüştür.
- Cinsiyet tespitinde yazım tarzı veya üslup gönderinin içeriğinden daha belirleyici bir unsur olduğundan "Stop Words" olarak belirtilen gereksiz kelimeler, noktalama işaretleri, emoji gibi ifadeler çıkarılmamıştır.
- Gönderilerde bulunan sosyal medyaya özgü etiket (#), değinme (@) ve web sayfası bağlantıları (http) şeklinde etiketlenerek bu ifadelerin çeşitliliği azaltılmıştır.
- Bu işlemler sonucunda toplam token sayısı 420k olarak hesaplanmıştır.
- Gönderileri sadece 1 gönderi üzerinden sınıflandırmaya çalışmak yeterli sonuçları vermeyebilir. Bunun yerine veri setinde aynı yazar tarafından gönderilen metinler n boyutunda bir çerçeve kullanılarak birleştirilmiştir. n=[1,2,4,5,10]
- Gönderi uzunlukları veri setindeki örneklerin %99'unu kapsayacak şekilde belirlenmiştir. Sistem de kullanılan veri seti özellikleri Çizelge 5.2'de gösterilmiştir.

Çizelge 5.2: Gönderi sayısına göre (n) oluşturulan veri setleri.

n	Eğitim Örnek sayısı	Test Örnek Sayısı	Örnek Uzunluğu
1	336800	192400	25
2	168400	96200	50
4	84200	48100	80
5	67360	38480	100
10	33680	19240	175

5.2.4 Yöntem

Çalışma kapsamında Makine Öğrenmesi, Derin Öğrenme ve Ön eğitimli Dil modellerinin kullanımı gibi 3 farklı yönteme bağlı olarak 8 farklı model oluşturulmuştur.

5.2.4.1 Makine öğrenmesi yöntemleri

Klasik makine öğrenmesi modellerinde, kelimeleri tokenleştirmek yani vektör haline getirmek için sınıflandırıcıdan ayrı yöntemler kullanılmaktadır. Bu yöntemler, kelimelerin metin içerisindeki frekanslarına ve birlikte kullanıldığı diğer kelimelere dayalı olarak oluşturulur. En yaygın kullanılan tokenizasyon yöntemleri arasında N-Gram, TF-IDF (Term Frequency-Inverse Document Frequency) ve BOW (Bag of Words) yer alır [129] .

- N-Gram: N-Gram yöntemi, kelimelerin veya karakterlerin n-uzunluklu dizilerini oluşturarak metni temsil eder. Örneğin, “Elma çok güzeldir” cümlesi için 2-gram (bi-gram) jetonları (token) “Elma çok”, “çok güzeldir” gibi diziler oluşturur. Bu yöntem, kelimeler arasındaki lokal bağımlılıkları yakalamada etkilidir.
- TF-IDF: TF-IDF, kelimenin bir belge içindeki frekansını (Term Frequency) ve kelimenin tüm belgeler arasındaki yaygınlığını (Inverse Document Frequency) kullanarak kelimenin önemini belirler. Bu yöntem, sık kullanılan kelimelerin ağırlığını azaltırken nadir kelimelerin ağırlığını artırır, böylece daha anlamlı temsiller elde edilir.
- Bag of Words (BOW): BOW yöntemi, metindeki kelimeleri sıklıklarına göre vektörleştirir. Bu yöntemde, kelimelerin sırası ve gramer yapısı göz ardı edilir. Metin, sadece kelimelerin frekanslarına dayalı bir vektör olarak temsil edilir.

Çalışmanın bu bölümünde, PAN 2019 görevinde kullanılan ve yüksek başarı sağlamış yöntem oluşturduğumuz veri setleri üzerinde kullanılmıştır [130]. Bu çalışmada, kelime tokenizasyonu ve sınıflandırma işlemi için N-Gram ve destek vektör makineleri (SVM) kullanılmıştır. N-Gram (1-3) yöntemi, kelime dizilerini 1-gram, 2-gram ve 3-gram olarak temsil eder. Bu sayede hem tek tek kelimeler hem de kelime çiftleri ve üçlüleri dikkate alınarak daha zengin ve anlamlı temsiller oluşturulur. Destek vektör makineleri (SVM), sınıflandırma işlemi için kullanılan güçlü bir denetimli öğrenme algoritmasıdır. SVM, veriyi yüksek boyutlu bir uzaya projeler ve sınıflar arasındaki en iyi ayrımı sağlayan bir hiper düzlem bulur. Bu çalışmada, N-Gram ile elde edilen kelime temsilleri, SVM modeline giriş olarak verilmiş ve sınıflandırma işlemi gerçekleştirilmiştir.

5.2.4.2 Derin öğrenme yöntemleri

Yinelenen sinir ağları ve evrişimli sinir ağları, metin sınıflandırmada sıklıkla kullanılan derin öğrenme modelleri olmuştur. Bu modeller, yüksek parametre sayısı ve yoğun hesaplama gerektirmelerine rağmen, sınıflandırma problemlerinde klasik makine öğrenmesi modellerine göre başarıyı önemli ölçüde artırmışlardır[129] . Bu çalışmada, metin verisinin jetonlaştırılması için kelime çantası (Bag of Words (BOW)) yöntemi kullanılmıştır. BOW yöntemi, metindeki kelimeleri sıklıklarına göre vektörize eder ve bu vektörler, yapay sinir ağına gömme katmanı olarak aktarılır. Gömme katmanı, kelime vektörlerini düşük boyutlu bir uzaya projelendirir. Bu çalışmada, gömme boyutu (embedding size) 50 olarak belirlenmiştir. Model eğitim sırasında bu vektörleri güncelleyerek öğrenmeyi sağlar ve daha anlamlı temsiller oluşturur.

Derin öğrenme yöntemlerinden ilk olarak yinelenen sinir ağı modeli kullanılmıştır. yinelenen sinir ağı mimarilerinden iki yönlü uzun kısa süreli bellek [131] modeli kullanılmıştır. İki yönlü uzun kısa süreli bellek hem ileri hem de geri yönde veri işleyerek uzun mesafeli bağımlılıkları yakalamada etkili olur. Model, 128 adet gizli katmana sahip iki yönlü uzun kısa süreli bellek katmanı ve çıkış katmanı olarak tek bir çıkışa sahip yapay sinir ağı eklenerek oluşturulmuştur. Model sıralı verilerdeki bağlam bilgilerini daha iyi öğrenir ve bu sayede metin sınıflandırma görevlerinde yüksek performans sağlar.

Evrişimsel sinir ağı modeli olarak, 128, 64 ve 32 boyutunda üç evrişim katmanı kullanılmıştır. ESA, metindeki n-gram benzeri özellikleri yakalayarak metin sınıflandırmada etkili olur. Her ESA katmanı, farklı boyutlarda filtreler kullanarak metin verisindeki

özellikleri çıkarır ve bu özellikleri birleştirir. Modelin sonuna, tek bir çıkışa sahip yapay sinir ağı eklenmiştir.

Her iki modelde de aşırı öğrenme (overfitting) problemini önlemek için dropout katmanları eklenmiştir. Dropout katmanları, belirli nöronların eğitim sırasında rastgele kapatılmasıyla modelin genelleme yeteneğini artırır ve aşırı uyumu önler. Çıkış fonksiyonu olarak sigmoid fonksiyonu seçilmiştir. Sigmoid fonksiyonu, her bir sınıf için olasılık tahminleri sağlar ve özellikle ikili sınıflandırma problemlerinde yaygın olarak kullanılır.

5.2.4.3 Türkçe ön eğitimli dil modelleri

Dil modelleri özetle verilen bir metinde maskelenmiş kelimeleri tahmin etmeye çalışan yapılardır. Bu maskelenen veri, tek bir kelime olabileceği gibi bir cümle de olabilir. 2017 yılında Vaswani ve ark.[90] tarafından önerilen öz dikkat mekanizmasına sahip derin öğrenme mimarisi, makine çevirisi alanında çığır açmıştır. Transformer mimarisinin başarısından sonra, bu mimari kullanılarak büyük miktarda metin verileriyle ön eğitimli dil modelleri oluşturulmuştur. Örneğin, BERT 16 GB, GPT-3 ise 40 GB metin verisi ile eğitilmiştir. Bu büyük dil modelleri, transfer öğrenmesi yöntemleri ile diğer doğal dil işleme görevlerinde (downstream tasks) yaygın olarak kullanılmaktadır.

Transfer öğrenmesi, önceden eğitilmiş bir modelin başka bir göreve uyarlanmasıdır. Bu yöntem, büyük miktarda veriyi ve hesaplama gücünü gerektiren eğitim sürecini daha verimli hale getirir. Ön eğitimli dil modelleri, metin sınıflandırma, duygu analizi, soru yanıtlama, özetleme ve daha birçok doğal dil işleme görevinde kullanılabilir.

Bu çalışma süresince, BERT, DistilBERT ve Electra olmak üzere üç farklı dil modeli ve versiyonları kullanılarak sınıflandırma başarıları test edilmiştir. Bu modeller, büyük Türkçe metin veri kümeleriyle eğitilmiştir. Bu modeller, HuggingFace [137] kütüphanesinde bulunmaktadır ve araştırmacılar ve geliştiriciler tarafından kolayca erişilebilir. HuggingFace, çeşitli dil modelleri ve araçları sunan popüler bir kütüphanedir ve doğal dil işleme projelerinde geniş bir yelpazede kullanılmaktadır.

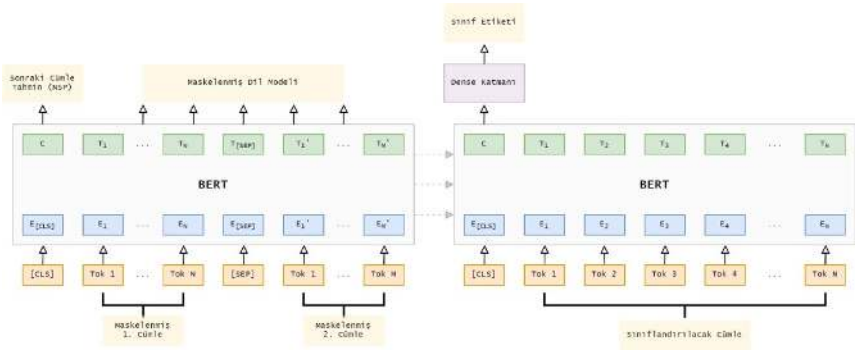
Çalışmada, kelimelerin vektörle temsil edilebilmesi için her modelin önceden eğitimli embedding vektörleri kullanılmıştır. Bu vektörler, kelimelerin anlamını ve bağlamını yakalayarak metin işlemede daha doğru ve anlamlı sonuçlar elde edilmesini sağlar. Örneğin, BERT modeli, çift yönlü dikkat mekanizması kullanarak kelimenin hem önceki hem de sonraki bağlamını dikkate alır, bu da daha zengin ve bağlamsal temsiller oluşturur.

BERT dil modeli

BERT [95]dil modeli, öz dikkat mimarisinin encoder ağı kullanılarak Google Brain ekibi tarafından Wikipedia ve BookCorpus verileriyle eğitilmiştir. BERTBase ve BERTLarge olmak üzere iki farklı versiyonu bulunmaktadır. BERTBase modeli 12 katman, 768 gizli boyut ve 12 dikkat başlığına sahip olup toplamda 110 milyon parametre içerir. BERTLarge modeli ise 24 katman, 1024 gizli boyut ve 16 dikkat başlığına sahip olup toplamda 340 milyon parametre içerir. BERT modelinde, giriş verisinin bir kısmı mask etiketiyle maskelenir. Model, eğitildikçe bu maskelenen kelime veya cümle verisini tahmin etmeye çalışır. Bu yöntem, modelin metin içindeki bağlamı öğrenmesini sağlar ve böylelikle soru cevap sistemleri, varlık tanıma gibi görevlerde başarısını artırır. Maskeli dil modelleme (Masked Language Modeling, MLM) olarak bilinen bu teknik, modelin hem ileri hem de geri yönde bağlamı dikkate alarak öğrenmesini sağlar. BERT modeli, kelimelerin konum bilgilerini de dikkate alarak farklı vektör temsilleri oluşturur. Bu, pozisyonel kodlama (positional encoding) sayesinde gerçekleştirilir. Dolayısıyla, bir kelimenin cümle başında veya ortasında olması, modelin o kelimeye uygulayacağı dikkat miktarını etkiler. BERT'in bu çift yönlü dikkat (bidirectional attention) mekanizması, kelimeler arasındaki karmaşık bağımlılıkları daha iyi öğrenmesine olanak tanır.

Ön eğitimli BERT dil modelinin son katmanından sonra ek katmanlar eklenerek farklı görevler için ince ayar yapılabilir. Bu, transfer öğrenmesi yöntemiyle gerçekleştirilir ve modelin belirli bir göreve uyarlanmasını sağlar. Örneğin, ikili sınıflandırma yapmak için BERT modeline tek hücreli bir sinir ağı eklenebilir. Benzer şekilde, soru cevap sistemi geliştirmek için cümle boyutunda bir katman eklenmesi yeterlidir. Bu ek katmanlar, modelin belirli görevler için optimize edilmesini sağlar ve performansını artırır.

BERT'in ön eğitim ve ince ayar stratejileri, çeşitli doğal dil işleme görevlerinde kullanılmaktadır. Ön eğitim süreci, modelin geniş bir metin veri kümesi üzerinde genel dil bilgisi edinmesini sağlar. İnce ayar süreci ise, modelin belirli bir görev için özelleştirilmesini ve optimize edilmesini sağlar. Bu iki aşamalı eğitim stratejisi, BERT'in esnekliğini ve etkinliğini artırır. Şekil 5.4'te, BERT modelinin nasıl eğitildiğini ve farklı NLP görevlerine nasıl adapte edildiğini gösteren ön eğitim ve ince ayar stratejileri verilmiştir.



Şekil 5.4: Bert ön eğitim ve sınıflandırma ince ayar stratejileri.

Bu çalışmada Bert temel mimarisi ile geliştirilmiş 2 farklı model kullanılmıştır. Eğitim verisi 35Gb metin dosyaları ve 44M token'dan oluşmaktadır. Kelime temsilleri birinci modelde 32k ikinci modelde 128k boyutunda vektörlerden oluşmaktadır.

DistilBert dil modeli

DistilBERT (Distillation BERT) [123], damıtılmış BERT olarak adlandırılır. Damıtma terimi, daha büyük bir modelin davranışını yeniden üretmek için daha küçük bir modelin eğitildiği sıkıştırma tekniğini ifade eder. Bu süreçte, daha büyük ve güçlü bir model (öğretmen model) kullanılarak daha küçük bir model (öğrenci model) eğitilir. DistilBERT, BERT'in küçük ve verimli bir versiyonudur ve bu teknikle eğitilmiştir. Parametre sayısı BERT dil modelinin neredeyse yarısı (66 milyon) kadardır. DistilBERT, BERT'e göre daha kısa eğitim ve transfer öğrenme süresi sunmaktadır. Daha az hesaplama gücü ve bellek gerektirir, bu da onu daha hızlı ve daha ekonomik hale getirir. DistilBERT, BERT'in performansını neredeyse korurken, önemli ölçüde daha verimli çalışır. Doğal dil anlama görevlerinde, BERT'e kıyasla yaklaşık %2'lik bir başarı kaybı yaşamaktadır. Bu küçük performans kaybı, daha hızlı işlem süreleri ve daha düşük kaynak gereksinimleri ile telafi edilebilir.

DistilBERT'in eğitim süreci, bilgi damıtma (knowledge distillation) adı verilen özel bir yöntemle gerçekleştirilir. Bu yöntemde, öğretmen modelin ürettiği yumuşak hedefler (soft targets) kullanılarak öğrenci model eğitilir. Yumuşak hedefler, öğretmen modelin çıktılarıdır ve bu çıktılar, öğrenci modelin daha doğru ve etkili bir şekilde öğrenmesini sağlar. DistilBERT, bu yumuşak hedefleri kullanarak, BERT'in öğrendiği dil bilgisini ve bağlamsal ilişkileri yeniden üretir.

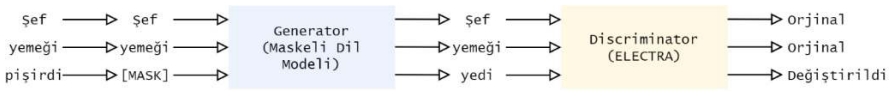
Electra dil modeli

BERT gibi maskeli dil modelleme (MLM) ön eğitim yöntemleri, bazı simgeleri [MASK] etiketiyle değiştirerek girdiyi bozar ve ardından orijinal simgeleri yeniden yapılandırmak için bir model eğitir. Bu yöntem, modelin metin içerisindeki bağlamı öğrenmesine yardımcı olur ve doğal dil işleme görevlerinde yüksek başarı elde edilmesini sağlar. Ancak, bu yöntem genellikle büyük miktarda hesaplama gerektirir ve eğitim süreci oldukça yoğundur. Electra mimarisi [124], bu duruma alternatif olarak, örnek açısından daha verimli bir ön eğitim yöntemi olan "değiştirilen belirteç tespiti" (replaced token detection) yaklaşımını önermektedir. Electra'da, girdiyi maskelemek yerine, bazı simgeler küçük bir üretici (generator) ağından örneklenen benzer değerlerle değiştirilir. Bu, metni bozar ve jeneratör, orijinal metni yeniden yapılandırmak yerine, değiştirilmiş simgeleri üretir. Ardından, bu bozuk simgelerin orijinal değerlerini tahmin eden bir model eğitmek yerine, bozuk girdideki her bir simgenin jeneratör tarafından değiştirilip değiştirilmediğini öngören bir ayırıcı (discriminator) model eğitilir.

Electra'nın eğitim süreci iki aşamadan oluşur. Küçük ve hafif bir model olan üretici, girdideki bazı simgeleri değiştirir. Örneğin, " kahverengi hızlı tilki" ifadesinde "hızlı" kelimesi üretici tarafından "çabuk" ile değiştirilebilir. Jeneratör, BERT gibi maskeli dil modellerine benzer şekilde çalışır ve belirli simgeleri değiştirme amacı güder. Ayırıcı model, değiştirilmiş metni alır ve her bir simgenin jeneratör tarafından değiştirilip değiştirilmediğini tahmin eder. Bu, modelin bağlamsal ilişkileri ve dil bilgilerini öğrenmesini sağlar. Ayırıcı, bozuk simgeleri tespit ederek orijinal simgeler ile değiştirilmiş simgeler arasındaki farkı anlamaya çalışır.

Bu yaklaşım, örnek açısından daha verimlidir çünkü model, her adımda sadece maskelenmiş simgelerle değil, tüm simgelerle çalışır. Bu da eğitim sürecini hızlandırır ve daha az hesaplama kaynağı gerektirir.

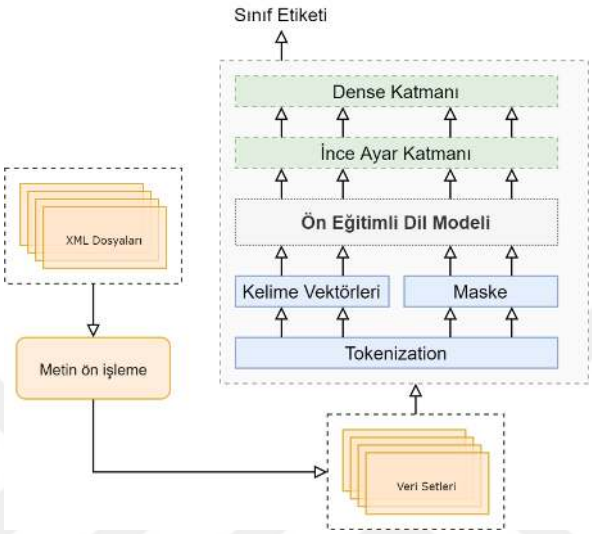
Şekil 5.5'te Electra dil modelinin ön eğitim yöntemi görsel olarak gösterilmiştir. Electra'nın bu yenilikçi yaklaşımı, aynı model boyutunda BERT ile karşılaştırıldığında daha iyi performans göstermektedir. Bunun nedeni, Electra'nın eğitim sürecinde daha fazla veri noktası kullanarak daha etkili bir şekilde öğrenmesidir.



Şekil 5.5: Electra dil modeli ön eğitim yöntemi.

5.2.5 Uygulama

Bu uygulama makine öğrenme yöntemleri, derin öğrenme mimarileri ve ön eğitimli dil modelleri kullanılarak Twitter gönderilerinden oluşturulmuş veri seti üzerinde cinsiyet tespiti yapılmaya çalışılmıştır. Önerilen mimarinin genel çalışma prensibi Şekil 5.6'da verilmiştir.



Şekil 5.6: Önerilen Model.

Çalışma kapsamında toplamda 8 adet farklı model oluşturulmuştur. Her bir model 5 farklı veri boyutuyla eğitilmiştir. Derin öğrenme yöntemleriyle ve dil modelleri kullanılarak oluşturulan modellerin parametre sayıları Çizelge 5.3'de verilmiştir. Dil modellerinde önceden oluşturulmuş parametreler güncellenmemiş sadece son katman güncellenmiştir.

Çizelge 5.3: Oluşturulan modellerin parametre sayıları

Modeller	P. Sayısı	Güncellenen P. Sayısı
Bi-LSTM	45M	
CNN	50M	
dbmdz/bert-base-turkish-128k-cased	185M	
dbmdz/bert-base-turkish-cased	110M	104k
dbmdz/distilbert-base-turkish-cased	67M	
dbmdz/electra-base-turkish-cased-discriminator	110M	
dbmdz/electra-base-turkish-cased-generator	35M	37k

Hiperparametreler ve uygulama ortamı

Eğitim süresince öğrenme oranı için 0.00001(1e-5) ile 0.005(5e-3) arası değerler test edilmiş en önemli yakınsama '0.001' değerinde yakalanmıştır. Tüm modeller için aynı öğrenme oranı ve 'adam' optimizer kullanılmıştır. Uygulama 2'li sınıflandırma problemi

olarak ele alınmış ve başarım değerlendirmesi için ‘binary accuracy’ , kayıp fonksiyonu için ‘binary cross entropy’ seçilmiştir.

Tüm uygulamalarda python programlama dili kullanılmıştır. Ön işlem adımların da ‘XML’, ‘NLTK’ kütüphaneleri makine öğrenmesi metodunda ‘SckitLearn’ kütüphanesi, Derin öğrenme ve dil modellerinin eğitimi sırasında ‘Keras’ kütüphaneleri kullanılmıştır. Ayrıca dil modellerinin önceden eğitilerek kullanılmasını sağlayan ‘transformers’ kütüphanesi kullanılmıştır. Çalışmalarda kullanılan bilgisayar 2080Ti 11Gb ekran kartı, 9. Nesil i7 işlemci ve 16 Gb ram donanımına sahiptir. Çalışmanın kodlarına ve sonuçlarına Github repomuzdan ulaşabilmektedir.

5.2.6 Değerlendirme ve sonuçlar

Oluşturulan modellerin başarım ölçümleri için doğruluk (accuracy) metriği kullanılmıştır. Erkek cinsiyeti için pozitif, kadın cinsiyeti için negatif değişkenleri atanmıştır. Her bir kayıt için modelin doğru tahmin ettiği erkekler doğru pozitif (True Positive DP), doğru tahmin ettiği kadın cinsiyeti için doğru negatif (True Negative DN) olarak belirlenmiştir. Tam tersi olarak modelin yanlış tahmin ettiği erkek cinsiyeti için yanlış negatif (False Negative YN), kadın cinsiyeti için yanlış pozitif (False Positive YP) olarak belirlenmiştir. Doğruluk metriğinin hesaplanması Denklem 5.3’de verilmiştir [23].

$$Başarı = \frac{(DP+DN)}{(DP+YN+DN+YP)} \tag{5.2}$$

Denklem de:

$Accuracy$: Başarı puanını gösterir.

DP : Gerçekte pozitif olan ve model tarafından da pozitif olarak tahmin edilen örneklerdir.

DN : Gerçekte negatif olan ve model tarafından da negatif olarak tahmin edilen örneklerdir.

YP : Gerçekte negatif olan ama model tarafından pozitif olarak tahmin edilen örneklerdir.

YN : Gerçekte pozitif olan ama model tarafından negatif olarak tahmin edilen örneklerdir.

Ayrıca en başarılı olan modelin F1-score (Denklem 5.4), duyarlılık (Sensitivity) (Denklem 5.5), özgüllük (Specificity) (Denklem 5.6) başarı puanları ölçülmüştür (Tablo 4).

$$F1Score = \frac{2DP}{2DP+YP+YN} \tag{5.3}$$

$$\text{duyarlılık} = \frac{DP}{DP+YN} \tag{5.4}$$

$$\text{özgüllük} = \frac{DN}{YP+DN} \tag{5.5}$$

Yapılan uygulamalar sonucunda ulaşılan doğruluk puanları Çizelge 5.4'te verilmiştir.

Çizelge 5.4: Geliştirilen modeller ve başarı puanları

	n=1	***n=2***	***n=4***	***n=5***	***n=10***
TF-IDF + SVM	0.6812	0.7116	0.7540	0.7680	0.7910
Bi-LSTM	0.6823	0.7153	0.7471	0.7524	0.7660
ESA	0.6781	0.6830	0.6981	0.7158	0.7459
Bert	0.6967	0.7163	0.7524	0.7651	0.7931
Bert-128k	**0.7017**	**0.7304**	**0.7630**	**0.7708**	**0.8012**
Distilbert	0.6452	0.6525	0.6824	0.7028	0.7457
Electra Discriminator	0.6825	0.7157	0.7383	0.7451	0.7747
Electra Generator	0.6910	0.7238	0.7538	0.7657	0.7952

En başarılı olan Bert-128k modelinin farklı başarım metrikleri Çizelge 5.5'te verilmiştir.

Çizelge 5.5: Bert-128k modelinin farklı başarım metrikleri

	DP	***DN***	***YN***	***YP***	***Doğruluk***	***F1-Score***	***Duyarlılık***	***Özgüllük***
n=1	61198	73809	26202	31191	0.7017	0.6808	0.7002	0.7029
n=2	31868	38396	11832	14104	0.7304	0.7108	0.7292	0.7314
n=4	16642	20059	5208	6191	0.7630	0.7449	0.7616	0.7642
n=5	13483	16177	3997	4823	0.7708	0.7535	0.7713	0.7703
n=10	7002	8413	1738	2087	0.8012	0.7855	0.8011	0.8012

5.2.7 Sonuç

Bu çalışma kapsamında Türkçe Twitter gönderilerinden cinsiyet tespiti yapılmıştır. Problem, sınıflandırma görevi olarak ele alınmış ve klasik makine öğrenmesi yöntemleri, derin öğrenme modelleri ve ön eğitimli dil modelleri kullanılmıştır. Yapılan deneyler sonucunda en başarılı model, BERT mimarisi ile hazırlanan ve kelime boyutunu 128K vektörle temsil eden model olmuştur. Bu model, derin öğrenme mimarilerine kıyasla çok daha az parametre eğitimi ile çok daha yüksek puan almıştır. Electra Generator mimarisine sahip dil modeli ise en az parametreye sahip olmasına rağmen yüksek oranda başarı elde etmiştir. Dolayısıyla bu çalışma, Türkçe dil modellerinin az sayıda parametre eğitimi ile sınırlı sayıdaki eğitim seti ve kısa, düzensiz kelime grupları üzerinde doğal dili anlama becerilerini ölçmüştür. Önerilen model ve yöntemlerin, farklı metin sınıflandırma problemlerinde kullanılabilir olduğu gösterilmeye çalışılmıştır. Veri seti içeriği, düzensiz gönderiler, kısa metinler, farklı yazım yanlışları ve çok sayıda kelime türü gibi

sınıflandırmayı etkileyen olumsuz öğelerden oluşmaktadır. Başarımı artırmak için farklı yöntemlere ihtiyaç duyulmaktadır. Bu sorunu giderebilmek için bu çalışmada gönderileri tek başına değerlendirmenin yanında, belirli bir çerçeve boyutuyla ek çalışmalar yapılmıştır. Sonuç olarak, tek bir tweet için cinsiyet tespiti zor olmasına rağmen, 5-10 arası tweet gönderileri değerlendirmeye alındığında %80'e kadar doğru tahmin edilebilmektedir. Çalışma, Türkçe gönderilerden cinsiyet tespiti için hazırlanan en kapsamlı çalışmalardan biri olmaktadır. Ayrıca kısa metin sınıflandırma görevi olarak ele alındığında, ön eğitimli dil modellerinin Türkçe için başarımları kıyaslanarak sonraki çalışmalar için bir rehber olmaktadır.

Aynı veri seti ile ulaşılan tek yayın olan diğer çalışmada [128] (Bag-Of-Words + SVM) %72.32 oranında başarıya ulaşıldığı belirtilmektedir. Ancak yazarların veri setini ham bir şekilde paylaşmasından ve metin ön işleme adımlarını detaylıca belirtmemesinden dolayı, bu çalışmada kullanılan SVM sınıflandırma yöntemi ile aynı sonuçlara ulaşılamamıştır. Oluşturulan en başarılı BERT modeli ile n=1 için %70, n=2 için %73 başarım elde edilmiştir. Ön eğitimli dil modellerinin eğitimi için genellikle yüksek miktarda veri kullanılmaktadır. Bu veriler çoğunlukla düzenli, anlamlı ve standart cümle yapılarından oluşmaktadır. Dil modellerinde kelimelerin cümle içindeki anlamı, diğer kelimelerle ilişkisi ve hatta konumu bile kodlanmaktadır (encoding). Bu yüzden metin sınıflandırma, duygu analizi gibi uygulamalarda klasik makine öğrenme modellerine oranla yüksek başarımlar elde edebilmektedirler. Ancak, Twitter gibi düzensiz kelime öbeklerinden oluşan, sohbet dili kullanılmış metin kümeleri üzerindeki uygulamalarda aynı oranda fark oluşmamaktadır. SVM gibi sınıflandırıcılar, dil modellerinin aksine, metin içindeki cümlelerin yapısıyla veya kelimelerin anlamları ile ilgilenmezler. Bu sebeple klasik makine öğrenme algoritmaları, cinsiyet belirleme görevi için yüksek parametreli dil modellerine yakın sonuçlar elde etmiştir.

Sonuç olarak, bu çalışma, Türkçe gönderilerden cinsiyet tespiti konusunda önemli bir katkı sunmuş ve kısa metin sınıflandırma görevlerinde ön eğitimli dil modellerinin etkinliğini göstermiştir. Ayrıca, bu modellerin yüksek başarımları, diğer metin sınıflandırma problemlerinde de kullanılabilir olduğunu kanıtlamıştır. Bu bulgular, gelecekte yapılacak çalışmalar için değerli bir rehber niteliğindedir.

5.3 Uygulama 3: Düşük Kaynaklı Dil Çevirisi için Çok Dilli Modellerin Etkili Adaptasyonu

Bu çalışmada, düşük kaynaklı dil çevirilerinde çok dilli (multilingual) ön eğitimli dil modellerinin transfer öğrenme yöntemleri araştırılmıştır. Dil modelleri genellikle metin içindeki kelimelerin veya cümlelerin anlamını tahmin eder, metin üretir, cümleleri tamamlar ve metinler arasındaki ilişkileri analiz eder. Bu süreç, dil modelinin eğitim sürecinde öğrendiği dil bilgisi ve bağlam bilgisine dayanır. Tek dilli (monolingual) dil modelleri, belirli bir dilde metin işleme görevlerini yerine getirmek üzere eğitilir ve o dilde yüksek performans sağlar. Tek dilli dil modellerinden farklı olarak, çok dilli dil modelleri ise birden fazla dilde eğitilir ve bu sayede çeviri modellerinde kullanılabilir. Çok dilli modeller, birden fazla dili aynı anda işleyebilir ve farklı diller arasında çeviri yapma yeteneğine sahiptir. Çok dilli dil modelleri, düşük kaynaklı dillerde bile iyi performans gösterebilir, çünkü yüksek kaynaklı dillerden öğrenilen bilgiyi transfer ederek bu dillerdeki çevirileri oluşturabilirler. Çok dilli ön eğitimli dil modelini çeviri sistemine uyarlayabilmek için farklı transfer öğrenme yöntemleri kullanılabilir. Tam ince ayar en sık kullanılan transfer öğrenme yöntemlerinden biridir. Parametre verimli ince ayar stratejileri büyük dil modellerinin tam ince ayar yapılmasının maliyetini azaltabilmek için kullanılan transfer öğrenme yöntemlerinden biridir. Bu çalışmada transfer öğrenme için parametre verimli bir yöntem önerilmiştir.

Yapılan çalışmada parametre verimli yöntemler olan LoRA (Low-Rank Adaptation) [132] ve dar boğaz adaptörlerinin (Bottleneck Adapters) [133] dil modellerine eklenmesi üzerine yoğunlaşılmıştır. Parametre verimli yöntemler de genellikle ön eğitimli dil modeline küçük ek katmanlar veya parametreler eklenir. Bu katmanlar, modelin performansını artırırken, mevcut model parametrelerinin büyük kısmını sabit tutar ve böylece ince ayar işlemi daha verimli hale gelir.

Bu çalışmada hem çok başlı dikkat katmanına hem de ileri beslemeli sinir ağına ek katmanlar eklenerek transfer öğrenme gerçekleştirilmiştir. Çok başlı dikkat katmanına, matrislerin boyutunu düşürebilmek için düşük ranklı matris temsilleri eklenmiştir. Bu, LoRA adaptör mimarisinde olduğu gibi rank (derece) kullanılarak yapılmıştır. İleri beslemeli sinir ağına ise darboğaz adaptörleri eklenmiştir. Bu adaptörler, sinir ağının içindeki bilgi akışını kontrol ederek, modelin parametre verimliliğini artırır ve hesaplama maliyetlerini düşürür. Yapılan deneyler, LoRA ve darboğaz adaptörlerinin birlikte kullanılmasının, diğer yöntemlere kıyasla daha iyi performans gösterdiğini ortaya koymuştur. Bu kombinasyon,

önceden eğitilmiş model parametrelerinin yalnızca %5'inin ince ayarı ile elde edilmiştir. Bu yöntem, parametre verimliliğini artırmış ve hesaplama maliyetlerini düşürmüştür. Çok dilli önceden eğitilmiş dil modelinin tam ince ayarına kıyasla, Bleu skorunda yalnızca %3'lük bir fark göstermiştir. Bu, önemli ölçüde daha düşük maliyetle neredeyse aynı performansın elde edildiği anlamına gelir. Ayrıca, yalnızca darboğaz adaptörlerini kullanan modeller, daha yüksek bir parametre sayısına sahip olmalarına rağmen daha kötü performans göstermiştir. Bu, darboğaz adaptörlerinin tek başına yeterli olmadığını ve diğer yöntemlerle birlikte kullanılmasının gerekliliğini göstermektedir. Yalnızca dikkat mekanizmalarına adaptör eklemek yeterli performans göstermemiş olsa da hem dikkat mekanizmalarına hemde ileri beslemeli sinir ağlarına adaptör eklemek makine çevirisi performansını önemli ölçüde geliştirmiştir. Önerilen yöntem, çeviri kalitesini korurken daha az bellek ve hesaplama yükü gerektirmiştir. Bu, düşük kaynaklı dil çevirisi için daha uygun ve verimli bir çözüm sunar. Bu çalışma, özellikle sınırlı kaynaklara sahip donanımlarda, yüksek performanslı makine çevirisi sistemleri geliştirmek için değerli bir yaklaşım sağlamaktadır.

5.3.1 Arka plan ve ilgili çalışmalar

2017 yılında Vaswani ve arkadaşları tarafından tanıtılan öz dikkat modeli [90], doğal dil işleme alanında bir dönüm noktası olmuştur. Öz dikkat modeli, yinelenen sinir ağlarından farklı olarak, sıralı veri işleme yerine, "kendi kendine dikkat" olarak bilinen bir mekanizmayı kullanarak girdinin tüm kısımlarını aynı anda işler. Bu durum, modelin cümledeki kelimelerin bağlamını daha etkili bir şekilde yakalamasını sağlamıştır. Daha iyi bağlam sayesinde öz dikkat modelleri çeviri ve metin üretimi gibi görevlerin kalitesinde önemli bir artışa yol açmıştır. Öz dikkat hızla BERT[95], GPT[93] ve türevleri gibi yeni nesil büyük dil modellerinin temeli haline gelerek, geniş bir alanda DDİ görevlerinde başarıları en üst seviyeye çıkarmıştır. Bu modeller sadece daha doğru değil, aynı zamanda daha esnek olup, görev özelinde ince ayar ile çeşitli dil işleme görevlerini yerine getirebilirler.

Öz dikkat mimarileri insan dilinin karmaşıklıklarını ele almada ve doğal dil anlama görevlerinde kritik bir rol oynamıştır. Ancak, bu mimariyi düşük kaynaklı dillere uygulamak önemli bir zorluk olmaya devam etmektedir. Düşük kaynaklı diller, eğitim için mevcut verinin sınırlı olması ile tanımlanmıştır. Bu durum yüksek performans elde etmek için büyük miktarda paralel külliyat gerektiren makine çevirisi görevlerinde önemli bir zorluk oluşturur.

Bu zorluğa yanıt olarak, araştırmamız, düşük kaynaklı makine çevirisi için büyük çok dilli dil modellerinin ince ayarına yönelik yenilikçi bir yaklaşım önermektedir. Önerilen yöntem parametre verimli adaptör yöntemleri kullanarak, önceden eğitilmiş dil modellerinin mevcut yapısına küçük, eğitilebilir modüller eklemeyi içermektedir. Bu adaptörler, hedef düşük kaynaklı dil için özel olarak tasarlanmış olup, modelin davranışını genel yapıyı minimal düzeyde ayarlayarak özelleştirme olanağı sağlamıştır. Bu yaklaşım, modelin genel yetenekleri ile her dilin özel incelikleri arasında bir denge sunarak, tam model yeniden eğitiminin hesaplama ve veri taleplerini ortadan kaldırmayı amaçlamıştır. Son araştırmalar, büyük modelleri minimum eğitimle yeni görevlere uyarlamak için parametre verimli yöntemlere odaklanmıştır [134]. Bu, önceden eğitilmiş bir modele eklenen küçük, eğitilebilir modüller olan adaptör katmanları gibi teknikleri içerir [132,135,136]. Karşılaştırmalı analizler, bu yöntemlerin önemli ölçüde daha az kaynak gerektirirken neredeyse güncel yüksek performanslara ulaşabileceğini göstermektedir [137].

5.3.2 Materyal ve yöntem

Makine çeviri sistemleri, kaynak ve hedef diller için geliştirilmiş paralel külliyatları kullanır. İngilizce, Fransızca ve Almanca gibi diller için çeviri sistemleri, çeviri kalitesini olumlu yönde etkileyen zengin bir veri deposuna sahiptir. Buna bağlı olarak, bu diller arasındaki makine çevirisinin kalitesi dikkate değer derecede yüksektir. Ancak, Türkçe gibi düşük kaynaklı dillerde çeviri kalitesini artırma çabaları devam etmektedir. Bu çalışma özellikle TR-EN dil çiftine odaklanmaktadır. Bu araştırma kapsamında, WMT17[3] benchmark testinden alınan TR-EN paralel külliyat kullanılmaktadır. Bu veri seti yaklaşık 200 bin paralel cümle içermektedir. Çizelge 5.6'da WMT17 (TR-EN) paralel külliyatı gösterilmiştir.

Çizelge 5.6: WMT17 (TR-EN) paralel külliyatı.

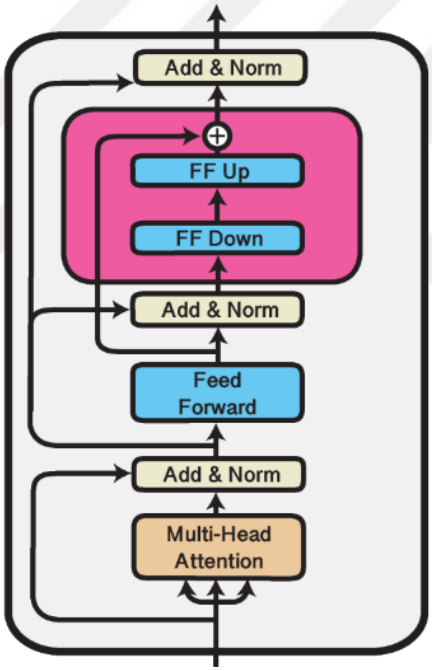
Eğitim	Test	Doğrulama
205,756	3007	3000

mBART gibi çok dilli olarak eğitilen büyük dil modelleri makine çeviri sistemlerinin ince ayarında yaygın olarak kullanılmaktadır. Bu araştırmada da mBART dil modeli kullanılmıştır. MBART[138], birçok dilde büyük ölçekli tek dilli kütüphaneler üzerinde önceden eğitilmiş bir dizi sıralı otomatik kodlayıcıdır (AutoEncoder). mBART, BART[139] modelinin (Bidirectional and Auto-Regressive Transformers) çok dilli bir uyarlamasını

temsil eder ve birden fazla dilde metin anlama ve üretme görevlerini yerine getirebilir. mBART modeli hem kodlayıcı hem de kod çözücü bileşenlerinde 12 katman içerir. Her katman 16 dikkat başlığı kullanır ve 1024 boyutunda gizli bir vektör kullanır. Hem kodlayıcı hem de kod çözücü bloklarındaki katmanların boyutları 4096'dır. MBART dil modeli yaklaşık 610 milyon parametreye sahiptir. Bu çalışmada, çeşitli ince ayar ve adaptör metodolojileri kullanılarak mBART modeli ile TR-EN çeviri sistemleri geliştirilmiştir.

5.3.2.1 Darboğaz adaptörleri

Dar boğaz adaptörleri (Bottleneck adapters), NLP ve diğer makine öğrenimi uygulamalarında kullanılan bir transfer öğrenme tekniğidir. Önceden eğitilmiş modellerin daha verimli bir şekilde ince ayarını yapmak için tasarlanmışlardır. Şekil 5.7'de darboğaz adaptör tasarımı gösterilmiştir.



Şekil 5.7: Darboğaz adaptör tasarımı.

Bir modelin her katmanına bir dar boğaz adaptörü eklenir. Bu adaptör genellikle iki ayrı doğrusal (veya tam bağlantılı) katmandan oluşur: bir daralma katmanı ve bir genişleme katmanı. Orijinal katmanın boyutunun d olduğunu varsayalım, daralma katmanı $d \times r$ boyutuna, genişleme katmanı ise $r \times d$ boyutuna sahiptir. Burada r dar boğaz boyutudur ve genellikle r≪d olacak şekilde seçilir. Daralma ve genişleme katmanları arasına tipik olarak bir Düzeltilmiş Doğrusal Birim ($ReLU$) veya benzer bir doğrusal olmayan aktivasyon fonksiyonu yerleştirilir. Adaptör, modelin orijinal katmanlarına eklenir. Orijinal katmanın

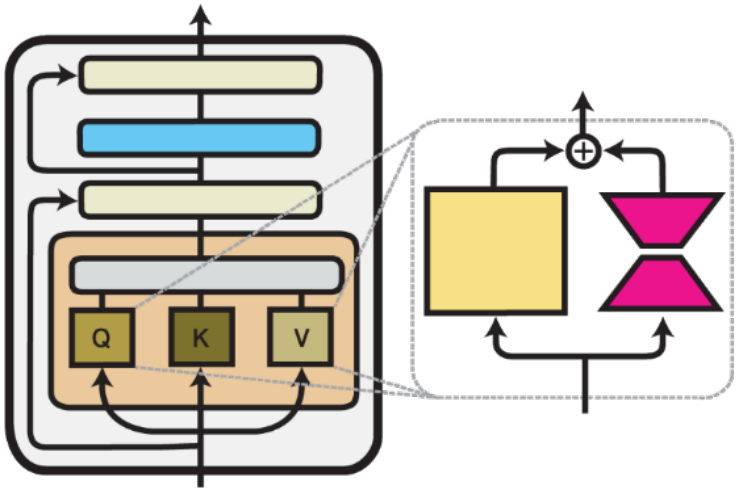
çıktısının x olduğunu varsayalım, adaptörün çıktısı x' Denklem 5.7'de gösterildiği gibi hesaplanır (Denklem 5.7):

$$x' = W_2\big(\sigma(W_1 x + b_1)\big) + b_2 \tag{5.7}$$

Bu adaptörlerde daralma ve genişleme katmanlarının ağırlıkları sırasıyla $W1$ ve $W2$, önyargı terimleri ise $b1$ ve $b2$ olarak adlandırılır. Aktivasyon fonksiyonu ise σ (sigmoid) olarak gösterilir. İnce ayar sırasında yalnızca adaptör katmanlarının (yani, $W1, W2, b1, b2$) parametreleri güncellenir. Modelin diğer tüm parametreleri sabit kalır. Bu yapı, modelin ana parametrelerini değiştirmeden belirli görevler için gerekli uyarlamaların yapılmasına olanak tanıyacak şekilde tasarlanmıştır. Dar boğaz adaptörlerinin kullanımı, modelin orijinal yapısını ve öğrenilmiş bilgilerini korurken görev odaklı ince ayar yapılmasını sağlar. Bu, özellikle büyük ve karmaşık modeller için hesaplama maliyetlerini ve eğitim süresini önemli ölçüde azaltır.

5.3.2.2 LoRA adaptörleri

LoRA (Low-Rank Adaptation of Large Language Models), önceden eğitilmiş derin öğrenme modellerinin ağırlıklarını düşük-ranklı matrisler aracılığıyla ayarlayarak çalışır. Şekil 5.8'de LoRA adaptör tasarımı gösterilmiştir.



Şekil 5.8: LoRA adaptör tasarımı.

Kavramsal olarak, modelin orijinal ağırlıklarını W olarak düşündüğümüzde, ana değişiklik, bu ağırlıklara küçük, düşük dereceli matrisler ΔW eklenerek yapılır. Bu yaklaşım, modelin yapısının bütünlüğünü korurken onu yeni görevlere uyarlamaktadır. LoRA, d

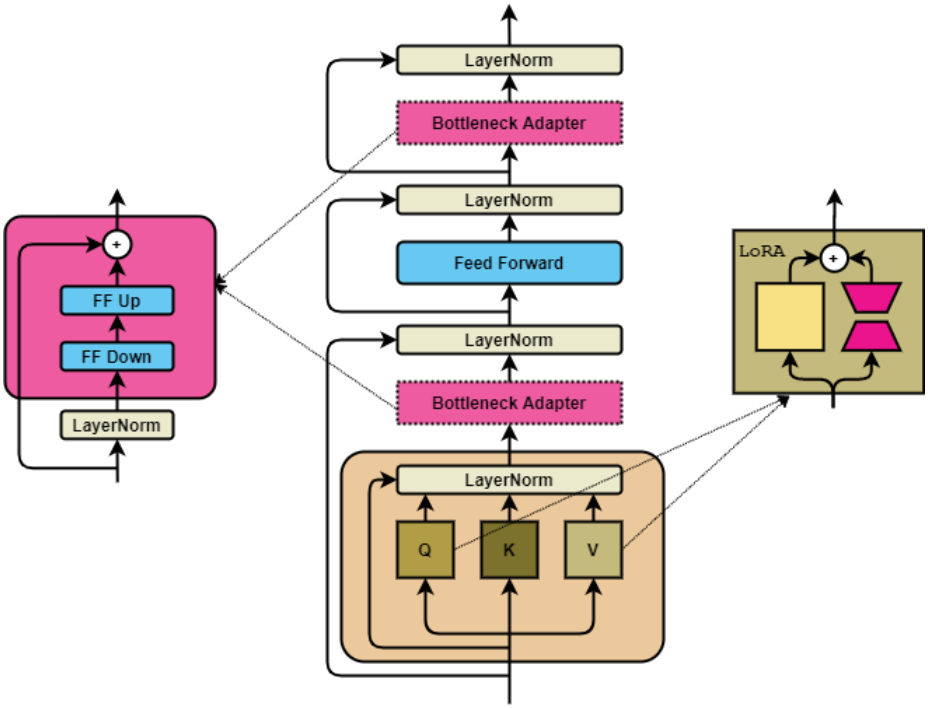
orijinal ağırlık matrisinin boyutu ve r seçilen rank değeri olmak üzere, iki düşük boyutlu matris $A \in R^{dxr}$ ve $B \in R^{rxd}$ kullanır. Genellikle, r değeri d'den çok daha küçük olacak şekilde seçilir, yani $r \ll d$. Düşük dereceli değişiklik matrisi ΔW olmak üzere $\Delta W = AB$ olarak hesaplanır. AB çarpımı düşük dereceli bir matris verir. Bu düşük dereceli değişiklik matrisi daha sonra orijinal ağırlık matrisine eklenir: $h = W + \Delta W$, burada h ince ayar yapılmış yeni ağırlık vektörü Denklem 5.8'de gösterildiği gibi hesaplanır. (Denklem 5.8).

$$h = W_0 x + \frac{d}{r} BAx \tag{5.8}$$

Bu yaklaşımın avantajı, tüm ağırlık matrisini güncellemek yerine, yalnızca iki küçük matris A ve B'nin eğitilmesidir. Bu, özellikle büyük modellerde, hesaplama ve bellek gereksinimlerini önemli ölçüde azaltır. LoRA, modelin çekirdek yapısını ve öğrenilmiş temsillerini koruyarak, modeli yeni görevlere uyarlamak için minimal fakat gerekli ayarlamaları yapar.

5.3.2.3 Önerilen model

Önerilen model, ileri beslemeli sinir ağına darboğaz adaptörleri ile çok başlı dikkat mekanizmasına LoRA adaptörlerin eklenmesiyle oluşturulmuştur. Şekil 5.9'da önerilen model gösterilmiştir.



Şekil 5.9: Önerilen model.

Dar boğaz adaptörleri, modellere eklenen modüllerdir ve parametre sayısını azaltarak verimliliği artırır. Bu adaptörler, giriş ve çıkış boyutlarını düşürerek hesaplama yükünü azaltır ve öğrenme sürecini hızlandırır. Dar boğaz adaptörlerinin kullanımı, özellikle büyük veri setleri ile çalışırken model performansını artırmak için önemlidir.

LoRA, modellerin belirli katmanlarındaki ağırlık matrislerinin düşük ranklı temsillerini kullanarak parametre sayısını azaltan bir tekniktir. Bu yaklaşım, ince ayar sürecinde güncellenmesi gereken parametre sayısını önemli ölçüde azaltırken modelin kapasitesini korur. LoRA, modelin öğrenme kapasitesini ve performansını koruyarak, daha az parametre güncellemesi ile etkili ince ayar yapılmasına olanak tanır. Bu da hesaplama maliyetlerini düşürür ve eğitim süreçlerini hızlandırır. LoRA, büyük modellerde parametre verimliliğini artırmada özellikle etkili bulunmuştur. LoRA'nın düşük ranklı temsiller kullanması, bellek ve hesaplama gereksinimlerini azaltırken model performansını korur. Bu, özellikle kaynak kısıtlı ortamlarda veya sınırlı veri mevcudiyeti durumlarında önemli bir avantajdır. Bu teknik, doğal dil işleme gibi büyük ölçekli modellerin daha verimli ince ayarını sağlar. Özetle, dar boğaz ve LoRA adaptörlerinin birlikte kullanımı, parametre verimliliğini artırır ve hesaplama yükünü azaltır, bu da daha hızlı ve verimli öğrenme süreçlerine yol açar.

5.3.3 Sonuç ve öneriler

Çalışmada, mBART önceden eğitilmiş dil modeli ve WMT17 (TR-EN) değerlendirme veri seti kullanılmıştır. Kullanılan veri seti ve dil modeli, Huggingface[140] kütüphanesinde açık kaynak olarak sağlanmaktadır. Adaptör yöntemleri, AdapterHub[133] kütüphanesi kullanılarak uygulanmıştır. Çevirilerin kalitesini ölçebilmek için Bleu [122] puan kullanılmıştır. Bleu (Bilingual Evaluation Understudy) skoru, makine çevirisinin kalitesini ölçmek için kullanılan otomatik bir metriktir. Bu skor, çeviri sonucunun insan çevirmenler tarafından üretilen referans çevirilere ne kadar benzer olduğunu matematiksel olarak değerlendirir. Esasen, makine tarafından üretilen çevirinin referans çeviri ile kelimeler ve ifadeler açısından ne kadar iyi eşleştiğine bakar. Bleu skoru, makine çevirisi çıktısını bir veya daha fazla referans çeviri ile karşılaştırarak hesaplanır. Çevirideki kelimeler ile referanslardaki kelimeler arasındaki eşleşmeler sayılır ve bu eşleşmelerin sayısı, çevirideki ve referanslardaki toplam kelime sayısına bölünerek normalize edilir. Bleu puanı tipik olarak 0 ile 100 arasında değişir. Daha yüksek bir skor, makine çevirisinin insan çevirisine daha yakın olduğunu gösterir.

Çalışma boyunca oluşturulan modellerin parametre sayıları, kayıp ve Bleu puanları Çizelge 5.7'de verilmiştir.

Çizelge 5.7: Oluşturulan modeller.

Adaptör	Konfigürasyon	Eğitilen Parametre Sayısı	Kayıp	Bleu Puanı
Darboğaz	mh_adapter=True Output Adapter = True r= 16	6.343.680 (1.04%)	1.46	12.82
	mh_adapter=False Output Adapter = True r= 16	3.171.840 (0.52%)	1.49	13.15
	mh_adapter=True Output Adapter = False r= 16	3.171.840 (0.52%)	1.67	11.23
	mh_adapter=True Output Adapter = True r= 8	12.638.208 (2.07%)	1.31	14.3
	mh_adapter=True Output Adapter = True r= 4	25.276.416 (4.14%)	1.21	16.3
	mh_adapter=True Output Adapter = True r= 2	50.552.832 (8.28%)	1.10	16.6
LoRA	r=8	1.179.648 (0.19%)	3.62	8.5
	r=4	589.824 (0.1%)	4.35	8.1
	r=16	2.359.296 (0.39%)	3.40	8.9
Dar Boğaz+LoRA	**mh_adapter=True, (r=16) Output Adapter = True, (r=2) LoRA (r= 16)**	**30.733.824 (5%)**	**1.05**	**21.3**
Tam İnce Ayar		610.851.840 (100%)	0.59	21.95

Bu çalışmada, düşük kaynaklı dil çiftlerinde dil modellerinin ince ayarı için parametre verimli adaptör yöntemlerinin performansını değerlendirilmiştir. Deneyler, bu adaptör tekniklerinin sınırlı veri koşullarında çeviri kalitesini nasıl etkilediğini anlamayı amaçlamıştır.

Test edilen çeşitli parametre verimli yöntemleri arasından, çok başlı dikkat ile ileri beslemeli sinir ağına eklenerek oluşturulan model, diğer yaklaşımlara kıyasla en yüksek

performansı göstermiştir. Özellikle, daha yüksek parametre sayısına sahip yalnızca dar boğaz adaptörlerini kullanan model daha düşük skorlar elde etmiştir. Ancak, LoRA entegre edildiğinde çeviri kalitesi önemli ölçüde iyileşmiştir. Bu birleşim sadece parametre verimliliğini artırmakla kalmayıp, aynı zamanda hesaplama gereksinimlerini de azaltarak düşük kaynaklı dil çiftlerinde çeviri kalitesinde önemli iyileşmeler sağlamıştır. Ayrıca, bu model toplam parametrelerin sadece %5'ini güncelleyerek verimli ince ayar ve optimize edilmiş kaynak kullanımı sağlamıştır. Tam ince ayar ile yapılan deneylerde ise performansta sadece %3'lük bir iyileşme gözlenmiştir. Bu sonuç, yüksek hesaplama maliyetlerine rağmen tam ince ayarın sınırlı performans artışı sağladığını göstermektedir. IA3 ve önek ayarlama gibi diğer parametre verimli yöntemlerle yapılan deneyler de gerçekleştirilmiş, ancak bu yöntemler yeterli sonuçlar vermemiştir.

Deneysel sonuçlarımız, LoRA ve dar boğaz adaptörlerinin birleşiminin, sinirsel makine çeviri modellerinin performansını optimize etmek için etkili bir yaklaşım olduğunu göstermektedir. Bu yöntem, model üzerinde daha az bellek ve hesaplama yükü oluşturarak çeviri kalitesini korur veya iyileştirir. Bu, özellikle sınırlı kaynaklara sahip uygulamalar için faydalıdır. Ayrıca, sadece %5 parametrenin güncellenmesi hesaplama maliyetlerini önemli ölçüde azaltarak genel verimliliği artırmıştır. Bu çalışmanın bulguları, dil modellerinin ince ayarında parametre verimli yöntemlerin önemini vurgulamaktadır. Bu yaklaşım, ön eğitimli modelin güncelleme sürecini daha hızlı ve düşük maliyetle tamamlamıştır.

Gün geçtikçe duyurulan ön eğitimli dil modellerinin parametre sayıları da devamlı artmaktadır. Bu dil modellerinin göreve özgü uyarlamaları için tam ince ayar yapmak kişisel bilgisayarda neredeyse imkânsız hale gelmektedir. Buna bağlı olarak parametre verimli adaptör yöntemler ise giderek önem kazanmaktadırlar. Gelecek çalışmalarda önerilen adaptör yöntemler farklı dillerde ve farklı görevlerde kullanılmaya çalışılması gerekmektedir. Ayrıca farklı görevlerde diğer parametre verimli tekniklerle karşılaştırmalı çalışmalar da yapılması gerekmektedir.

5.4 Uygulama 4: Sinirsel Makine Çeviri Sistemlerine Ön Eğitimli Dil Modellerinin Dâhil Edilmesi Stratejileri

İngilizce dünyada akademik dil olarak kabul edilmektedir. Bu, diğer dilleri konuşanlar için akademik çalışmalarında İngilizce kullanımını zorunlu kılmaktadır. Bu araştırmacılar İngilizce dilinin kullanımında yetkin olsalar bile akademik bir makale yazarken bazı hatalar yapabilmektedir. Akademisyenler bu sorunu çözmek için otomatik çeviri programlarını kullanma ya da ileri düzeyde İngilizce bilen kişilerden yardım alma eğilimindedir. Bu çalışma, araştırmacılara akademik makale yazma sürecinde yardımcı olacak bir uzman sistem sunmaktadır. Bu çalışmada kaynak dil olarak düşük kaynak dilleri arasında sayılan Türkçe kullanılmıştır. Önerilen model, öz dikkat modelini sığ füzyon yöntemiyle önceden eğitilmiş Sci-BERT dil modeliyle birleştirir. Model sığ füzyon yönteminde tam dikkat katmanı (fully attentional network) kullanır. Bu sayede kelime düzeyinde dikkat arttırılarak daha yüksek bir başarı oranı elde edilebilir. Oluşturulan modelin değerlendirilmesinde farklı metrikler kullanılmıştır. Deneyler sonucunda oluşturulan model 45.1 Bleu ve 73.2 Meteor puanlarına ulaşmıştır. Ayrıca önerilen model, Dünya makine çevirisi (World Machine Translation) (2017–2018) test veri kümelerinde sıfır vuruşlu çeviri yöntemiyle sırasıyla 20,12 ve 20,56 puan almıştır. Önerilen yöntem, dil modelini çeviri sistemine dâhil etmek için diğer düşük kaynaklı dillere ilham verebilir.

5.4.1 Arka plan ve ilgili çalışmalar

Google Translate (GT) ve DeepL çeviri sistemleri şu anda Türkçeden İngilizceye en yaygın kullanılan çeviri sistemleridir. Her iki çeviri sistemi de öz dikkat mimarisini kullanır ve çok başarılı çeviriler oluşturur. Ancak akademik terminoloji çevirileri günlük konuşma dilindeki çevirilerden farklıdır. Bu bağlamda, çeviri sistemleri bazen akademik terimleri hatalı çevirir. Çizelge 5.8'de Google Translate (Sürüm: 6.36, Tarih: 11 Mayıs 2022) sisteminin farklı çevirdiği bazı akademik terimler verilmiştir.

Özellikle son yıllarda büyük şirketler tarafından geliştirilen çeviri hizmetleri (Google, Facebook, DeepL) büyük ilerlemeler kaydetmiştir. Günlük konuşma dili çevirilerinde sisteme girilen kaynak cümle düzgün ve dilbilgisi açısından doğru sıralı bir cümle olduğunda buna paralel olarak hedef cümle yapısı net ve anlaşılır olacaktır. Akademik çalışmalarda bilimsel terimlerin bolluğu ve çeşitliliği bu çeviri sistemlerinin etkinliğini azaltabilir. Aspec [141] ve Scielo Corpus [142] gibi diğer diller için akademik çeviri sistemlerinin örnekleri varken, Türkçe dili üzerine böyle bir araştırma bulunmamaktadır.

Akademik bir çeviri sistemine olan ihtiyaç ile doğal dil işleme ve makine çevirisindeki gelişmeler, bu çalışmanın ortaya çıkmasında motive eden faktörlerdir. Bu çalışmada Türkçe-İngilizce akademik çeviriler için sinirsel makine çeviri modeli önerilmiştir. Bu çalışma, araştırmacıların ve lisansüstü öğrencilerin İngilizce akademik makale yazarken daha iyi çeviriler elde etmeleri için bir çeviri sistemi sunmaktadır. Bu sayede yazım ve dilbilgisi hataları en aza indirilebilir. Bu çalışmada temsil edilen çeviri mimarilerinin temeline eklenen sığ füzyon yönteminin kullanılması, farklı dil çiftleri üzerine yapılacak daha sonraki çalışmalara ilham verebilir. Önerilen model makine çevirisi alanında çalışan araştırmacılar için açık kaynak kodlu olarak sunulmuştur.

Çizelge 5.8: Google Translate sisteminin farklı çevirdiği bazı akademik terimler.

Turkish	Google Translate	In the literature
evrişimsel sinir ağı girişleri	convolutional neural network entries	Convolutional neural network inputs
sözel olmayan yakınlık becerileri	non-verbal intimacy skills	Non-verbal immediacy skills
nitel araştırmalarda çeşitleme	diversification in qualitative research	Triangulation in qualitative research
sınıf içi öğretmen davranışları	classroom teacher behavior	Teacher behavior within classroom
belirsizlik hoşgörüsü seviyesi	level of uncertainty tolerance	Ambiguity tolerance level

5.4.2 Materyal

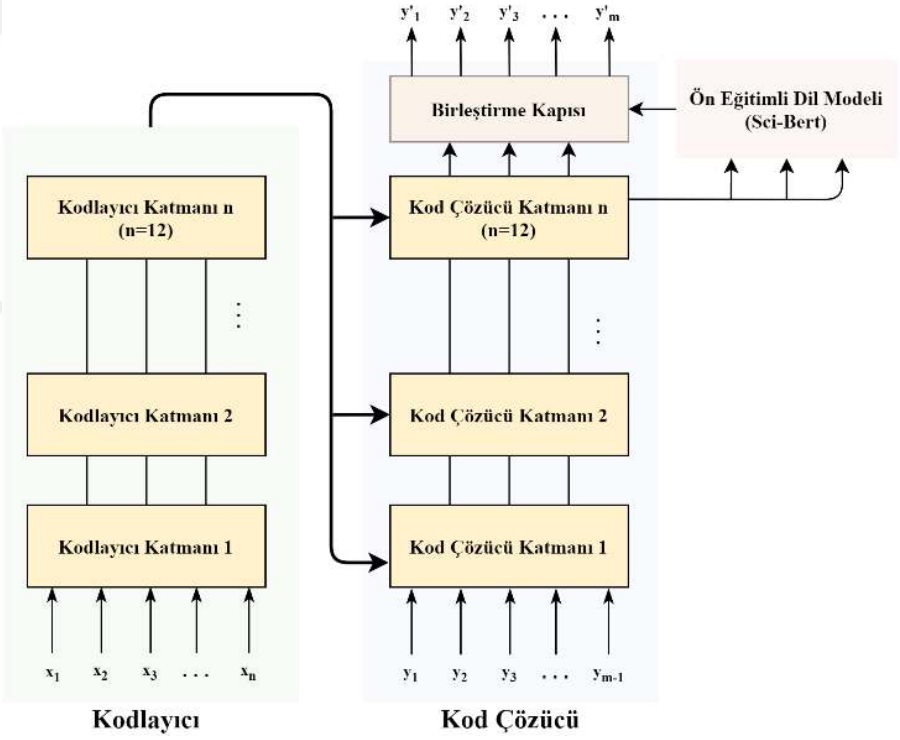
Bu çalışmada daha önce oluşturulan paralel külliyat genişletilerek kullanılmıştır [143]. Önceki çalışmada 2015-2020 yılları arasındaki tezler kullanılarak oluşturulan külliyat 2013-2022 yılları da dahil edilerek veri sayısı artırılmıştır. Ön işlemler için önceki çalışmada belirtilen adımlar izlenmiştir. Önceki çalışma dan farklı olarak külliyat 'tan rastgele seçilen 1000 cümle bir uzman tarafından incelenmiştir. Çevirilerin birbirini karşılayıp karşılamadığı göz önünde bulundurulmuştur. Örnek cümlelerin kalitesi incelenmemiştir. Rastgele seçilen 1000 cümle içerisinde sadece 2 cümle çiftinin yeterince birbirini karşılamadığı tespit edilmiştir. Bu oranın ise göz ardı edilebilir bir oran olduğu öngörülmüştür. Tüm işlemlerden sonra 1.217.300 cümle elde edilmiştir. 2.3M cümle göz ardı edilmiştir. Oluşturulan paralel külliyat yaklaşık 30K farklı token'den (belirteç) oluşmuştur. Oluşturulan veri seti Çizelge 5.9'da gösterilmiştir.

Çizelge 5.9: Kullanılan paralel külliyat.

Veri	Cümle Sayısı
Eğitim	1.197.300
Doğrulama	10.000
Test	10.000
Toplam	1.217.300

5.4.3 Yöntem

Çalışma da Sinirsel makine çevirisi için öz dikkat tabanlı bir model önerilmiştir. Çevirilerin kalitesini artırmak için, ön eğitimli dil modelleri oluşturulan modele sığ füzyon yöntemiyle eklenmiştir. Ayrıca bayt çifti kodlama, ışın arama gibi yöntemler de eklenmiştir Şekil 5.10'da önerilen model gösterilmiştir. Bu bölümde modeli oluşturan yöntemler ve değerlendirme kriterleri detaylı olarak anlatılmaktadır.



Şekil 5.10: Önerilen model.

Öz dikkat modeli

Çeviri sisteminin temel mimarisi öz dikkat [90] tabanlı kodlayıcı kod çözücüdür. Öz dikkat mimarisi için katman olarak iki farklı model için 6 ve 12 seçilmiştir. İleri beslemeli sinir ağı (FFNN) için boyut 2048 olarak belirlenmiştir. Kelime vektörleri için 512 boyut belirlenmiştir. Ayrıca Query(sorgu), Key(Anahtar) ve Value(Değer) vektörleri de 64 boyutlu

olarak belirlenmiştir. Çok başlı dikkat mekanizması için h=8 olarak belirlenmiştir. Öz dikkat mimarisinin detayları Bölüm 4'te detaylıca anlatılmıştır.

Ön eğitimli dil modelinin çeviri sistemine dahil edilmesi birleştirme katmanı

Çeviri sisteminin bu katmanında hedef çeviriyi daha akıcı hale getirmek için dil bilgisi eklenmeye çalışılmıştır [135,136]. Bu görev için hedef dil olan İngilizce için önceden eğitilmiş bir dil modeli sisteme eklenmiştir. Çalışma da akademik bir çeviri sistemi oluşturulması hedeflenmektedir ve bu sebeple SciBERT [146] ön eğitimli dil modeli kullanılmıştır. SciBERT dil modeli 1.14 milyon akademik makaleyi içermekte ve toplamda 3.1 milyar jetondan oluşmaktadır. Çeviri sistemlerinde, yaygın olarak derin (deep) ve sığ (shallow) füzyon yöntemleri kullanılarak önceden eğitilmiş dil modelleri eklenmektedir. Derin füzyon yönteminde, dil modeli kod çözücü ağına entegre edilir. Bir sonraki kelimenin çıkış olasılığını hesaplarken, model hem kod çözücü ağının hem de dil modelinin [144] gizli durumlarını kullanmak için ince ayarlıdır. Sığ füzyon yönteminde, kod çözücü ağının çıktı katmanı ile dil modelinin çıktı katmanı birleştirilir [147]. Bu çalışmada sığ füzyon yöntemi kullanılmıştır. Birleştirme kapısı adı verilen bir kapı, kod çözücü ağ çıktısını dil modeli çıktısıyla birleştirmek için kullanılır. Bilinen sığ füzyon yönteminde [147] bu kapı için ileri beslemeli ağ yapısı önerilmektedir. Bu çalışmada kapı tasarımında tam dikkat katmanı (Fully Attentional Network - FAN) önerilmiştir.

Sığ füzyon- ileri beslemeli sinir ağı

Çeviri sistemlerinde her bir belirtecin ortaya çıkma olasılığı hesaplanırken önceden oluşturulmuş belirteçler dikkate alınır. Bu aynı zamanda kod çözücü ağı ve ön eğitimli dil modeli için de aynıdır (Denklem 5.9).

$$y' = \sigma_{dec}^{t} \cdot y'_{dec} + \sigma_{lm}^{t} \cdot y'_{lm} \tag{5.9}$$

Denklem de:

t : o anki çıkış tokeni gösterir.

$y' = \log p\ (y_t = k|x, y_{<t})$: Nihai çıktıyı gösterir.

$y'_{dec} = \log p_{dec}\ (y_t = k|x, y_{<t})$: Kod çözücü ağın çıktısını gösterir.

$y'_{lm} = \log p_{lm}\ (y_t = k|x, y_{<t})$: Dil modelinin çıktısını gösterir.

$\sigma_{dec}, \sigma_{lm} \epsilon (0,1)$: çeviri sisteminin dil modeli ile kod çözücü ağı arasında seçim yapmasına izin veren kapılardır.

FFN : Bir gizli katmanı ve iki çıkışı olan bir ileri beslemeli ağdır (Denklem 5.10).

$$\left[\sigma_{dec}^{(t)}, \sigma_{lm}^{(t)}\right]^T = FFN(s_{trans}^t) \tag{5.10}$$

Sığ füzyon- tam dikkat sinir ağı (Fully Attentional Network)

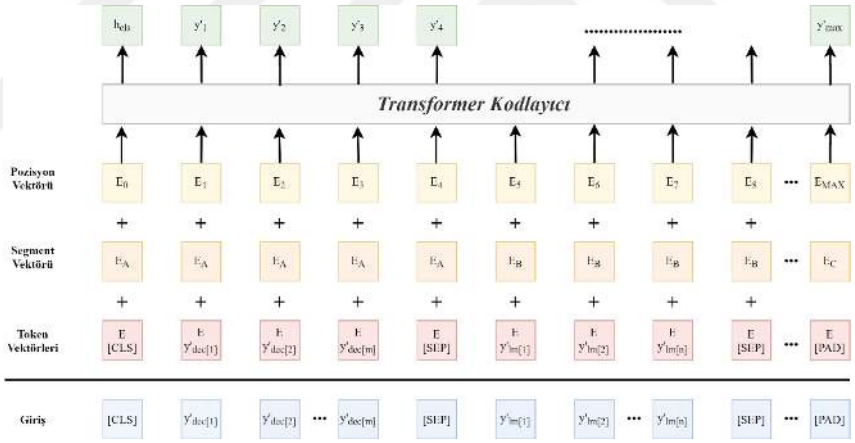
Çalışma da önerilen yöntemde birleştirme katmanı için tam dikkat ağı kullanılmaktadır (Şekil 5.11). Kod Çözücü ağ çıkışı ve önceden eğitilmiş dil modeli çıkışı düz birleştirilmiştir (Denklem 5.11).

$$y'' = concat([< cls >, y'_{dec}, < sep >, y'_{lm}]) \tag{5.11}$$

Oluşturulan cümle çiftleri (y'') belirteç yerleştirme (tokenizasyon), segment yerleştirme ve konum yerleştirme katmanlarında işlenir. Bu işlemlerden sonra öz dikkat kodlayıcı ağına girdi olarak verilir (Denklem 5.12).

$$y' = TransformerKodlayıcı(E_{token}(y'') + E_{seg}(y'') + E_{pos}(y'')) \tag{5.6}$$

Bu konfigürasyonda, eğitim sırasında dil modeli ağırlıklarını dondurmak önemlidir ve çeviri modeli, dil modeline ne zaman ve ne kadar güvenebileceğine karar verebilmelidir.



Şekil 5.11: Önerilen birleştirme kapısı.

Çeviri sisteminde kelime giriş yöntemi olarak Bayt çifti kodlaması (BPE) seçilmiştir. Çevirilerin üretilmesi aşamasında ise Işın Arama (Beam Search) yöntemi kullanılmıştır. Çevirilerin puanlanmasında Bleu, Meteor ve Ter puanları kullanılmıştır.

Uygulama ayarları ve hiper parametreler

Önerilen çeviri sisteminde temel Transformer yöntemine ek olarak, önceden eğitilmiş bir dil modeli kullanılmıştır. Nadir kelime problemini çözmek için Bayt Çifti Kodlaması kullanılmıştır. Çeviri kalitesini artırmak için model eğitildikten sonra Işın arama algoritması kullanılmıştır. Oluşturulan modellerin hiper parametreleri aşağıda verilmiştir

Çizelge 5.10'da oluşturulan modellerin hiper parametreleri verilmiştir. Çalışma boyunca Python programlama dili kullanılmıştır. Web kazıma ve metin önişleme adımlarında Selenium, NLTK ve SpaCy kütüphaneleri kullanılmıştır. Modellerin eğitimi için Pytorch derin öğrenme çerçevesi kullanılmıştır. Önceden eğitilmiş SciBert modeli, Huggingface kitaplığı aracılığıyla kullanılmıştır. Modellerin eğitimi sırasında aktivasyon fonksiyonu olarak "adam" fonksiyonu kullanılmıştır. Ayrıca "smoothing" etiketi seçilmiş ve yumuşatma oranı (0.1) olarak belirlenmiştir. Eğitim süresini azaltmak için yarı duyarlıklı (FP16) 16 bit kayan nokta kullanılmıştır. Eğitim, GPU tabanlı bulut bilişim sistemleri kullanılarak tamamlanmış olup, 4×Nvidia RTX 3090 ekran kartları kullanılmıştır.

Çizelge 5.10: Oluşturulan modellerin hiper parametreleri.

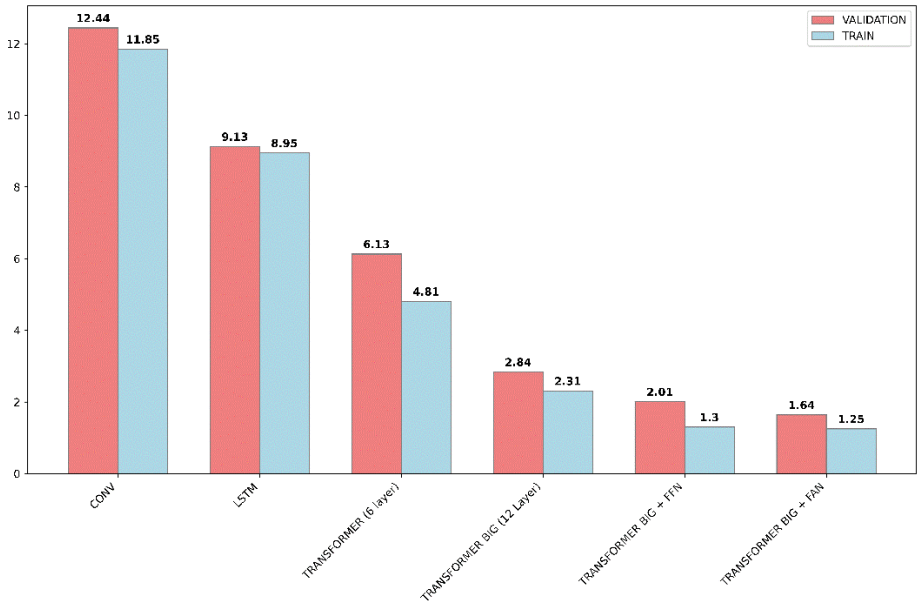
	Katman	FFNN	d_{model}	dQ, dK, dV
Conv	20	1024	-	-
LSTM	2	1024	-	-
Transformer	6	2048	512	64
Transformer Big	12	2048	512	64
FFN-Birleştirme Kapısı	1	1024	-	-
FAN-Birleştirme Kapısı	4	2048	512	64

Ayrıca oluşturulan modellerin parametre sayıları Çizelge 5.11'de verilmiştir. Eğitimi durdurmak için şaşkınlık (perplexity) puanı kullanılmıştır.

Çizelge 5.11: Oluşturulan modellerin parametre sayıları.

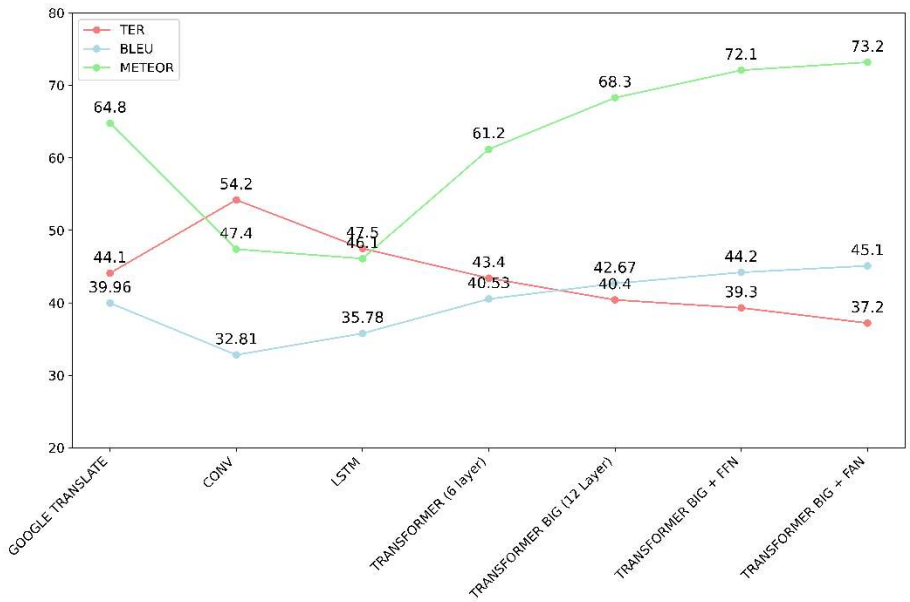
Oluşturulan Modeller	Parametre Boyutu
Conv	96.841.408
LSTM	31.066.208
Transformer (6 Katman)	73.033.074
Transformer Big (12 Katman)	98.278.770
Transformer Big + FFN	136.303.045
Transformer Big + FAN	158.523.484

Doğrulama testi setindeki şaşkınlık puanı 10 dönem boyunca düşmezse eğitim durdurulmuştur. Doğrulama ve eğitim veri setlerindeki modellerin son şaşkınlık (perplexity) puanları Şekil 5.12 de verilmiştir.



Şekil 5.12: Oluşturulan modellerin şaşkınlık (perplexity) puanları.

Oluşturulan tüm modellerin Ter, Bleu ve Meteor puanları ayrı ayrı hesaplanmıştır. Ayrıca test seti Google Translate kullanılarak çevrilmiştir. Bu çevirinin değerlendirme puanları da hesaplanmıştır. Tüm sonuçlar Şekil 5.13 da gösterilmektedir.



Şekil 5.13: Oluşturulan modellerin Ter, Bleu ve Meteor puanları.

5.4.4 Tartışma ve gelecek çalışmalar

Bu çalışmada, düşük kaynaklı dil çiftlerinde çeviri kalitesini artırmak için önceden eğitilmiş dil modelleri kullanılmıştır. Çeviri sistemi için Türkçe-İngilizce dil çifti seçilmiştir.

Çeviri sistemi için öncelikle akademik çevirilerden oluşan paralel bir külliyat oluşturulmuştur. Paralel derlem oluşturulurken Türkiye'de hazırlanmış tezlerin özet bölümlerinden yararlanılmıştır. Çeviri sistemi için öz dikkat mimarisine sahip bir kodlayıcı-kod çözücü ağı, önceden eğitilmiş bir dil modeli ve bir tam dikkat mimarisinden oluşan yeni bir model önerilmiştir. Önerilen modelde, temel öz dikkat kod çözücü ağının son katmanı, önceden eğitilmiş dil modelinin son katmanı ile tam dikkat katmanı kullanılarak birleştirilir. Bu çalışmada sığ füzyon yöntemi olarak tam dikkat ağı kullanılmıştır. Ek olarak, Eğitim sırasında nadir görülen kelime probleminin üstesinden gelmek için BPE kullanılmıştır. Model eğitildikten sonra Işın Arama algoritması ile alternatif çeviriler arasından en iyi çeviri seçilmeye çalışılmıştır. Önerilen modelin test aşamasında Ter, Bleu ve Meteor puanları hesaplanmıştır. Önerilen model, temel öz dikkat mimarisine göre (+2 Bleu, +4 Meteor) daha yüksek puanlar elde etmiştir. 40 ve üzeri Bleu puanı kaliteli çeviri olarak bilinmektedir. https://cloud.google.com/translate/automl/docs/evaluate. Ayrıca önerilen model, Google Translate çeviri sisteminden daha yüksek puanlar da almıştır. Bu durumun temel nedeni akademik cümle yapısı ve test setindeki akademik terimlerin fazlalığı olabilir. Önerilen model kullanılarak WMT17 ve WMT18 test setleri üzerinde sıfır atış yöntemi kullanılarak çeviri yapılmıştır. Önerilen model, bu test setlerinde sırasıyla 20.12 ve 20.56'lık bir Bleu puanı elde etmiştir. En son çalışmalara bakıldığında Behnke ve ark. [66], WMT17 test seti için 21,9 Bleu puanı elde etmiştir. Ayrıca Pan ve ark. [148], WMT18 test seti için 23,1 Bleu puanı elde etmiştir. Bahsi geçen çalışmalarda çeviri türüne özel paralel derlem ve tek dilli verilerin kullanıldığı görülmektedir. Önerilen çeviri modeli, rekabetçi Bleu puanlarına ulaşmıştır. Ayrıca çeviri sistemlerinde kullanılan SciBert modeli yerine BERT, DistilBert, GPT gibi dil modelleri ile WMT test setlerinin çevirisi yapılmaya çalışılmıştır. Ancak sonuçlar beklenenden çok daha yüksek çıkmıştır. Bunun nedeni, önceden eğitilmiş dil modellerini eğitirken etiketlenmemiş verilerin kullanılması olabilir. Önceden eğitilmiş dil modellerinin eğitimi sırasında WMT test setlerinin kullanıldığı tahmin edilmektedir. Model, İngilizce makale hazırlayan araştırmacılar için yararlı bir kaynak olarak kullanılabilir.

Yüksek kaynaklı dillerle karşılaştırıldığında bu puanların temel nedeni yetersiz paralel külliyattır. Diğer yüksek kaynaklı dil çiftleri şu anda 50M üstü paralel külliyata sahiptir [149]. Henüz TR-EN dil çifti arasında bu büyüklükte bir külliyat bulunmamaktadır. Bundan sonraki çalışmalarda sadece akademik çeviriler için değil genel çeviriler için de kullanılabilecek bir derlem ve çeviri sistemi oluşturulmaya çalışılacaktır.

Önerilen yöntem ile çeviri kalitesini artırmak için önceden eğitilmiş dil modeli kullanılmıştır. Bu yöntem, temel öz dikkat mimarisinden daha iyi çeviriler üretmiştir. Doğal dil işleme görevlerinde güncel puanların elde edilmesinde önceden eğitilmiş dil modellerinin kullanımı yaygındır. Ancak önceden eğitilmiş dil modellerinin doğal dil oluşturma görevlerinde en etkin şekilde nasıl kullanılacağı halen önemli bir araştırma konusudur. Özellikle düşük kaynaklı çeviri sistemlerinde önceden eğitilmiş dil modellerinin etkisinin artırılması gerekmektedir. Ayrıca, gelecekteki çalışmalar bu soruna odaklanacaktır.

6. SONUÇLAR VE ÖNERİLER

Derin öğrenmenin hızla gelişmesi, doğal dil işleme uygulamalarında büyük başarılar elde edilmesine olanak sağlamıştır. Özellikle makine çevirisi alanında kaydedilen ilerlemeler, diller arası iletişimi kolaylaştırarak küresel bilgi paylaşımını artırmıştır. Derin öğrenme modellerinin artan kapasitesi ve karmaşıklığı, farklı diller arasındaki çevirilerin doğruluğunu ve doğallığını önemli ölçüde artırmıştır. Ancak, bu alanda kaydedilen tüm ilerlemelere rağmen, düşük kaynaklı diller için hâlâ çözülmesi gereken birçok zorluk bulunmaktadır. Bu zorluklar, düşük kaynaklı dillerdeki veri eksikliğinden ve bu dillerin yapısal farklılıklarından kaynaklanmaktadır.

Türkçe-İngilizce, çeviri sistemleri için yeterli veri kaynağına sahip olmaması nedeniyle düşük kaynaklı bir dil çifti olarak kabul edilmektedir. Ayrıca Türkçe'nin yapısal özellikleri, makine çevirisi çalışmalarında önemli zorluklar teşkil etmektedir. Türkçe, eklemeli bir dil olmasının yanı sıra karmaşık dilbilgisi kuralları ve esnek kelime sıralaması ile dikkat çekmektedir. Bu durum, çeviri sistemlerinin yüksek doğrulukla çalışmasını zorlaştırmakta ve Türkçe İngilizce dil çiftinde yüksek kaliteli çeviriler elde edilmesini engellemektedir. Özellikle, yeterli ve kaliteli veri setlerinin eksikliği, bu alanda çalışan araştırmacılar için önemli bir engel oluşturmaktadır. Mevcut veri setleri genellikle sınırlı kapsam ve kalitededir, bu da çeviri modellerinin öğrenme süreçlerini olumsuz etkilemektedir.

Veri eksikliği, çeviri modelinin o dile ait tüm varyasyonlarını ve nüanslarını öğrenmesini zorlaştırmaktadır. Bu da çevirilerin doğruluğunu ve doğallığını olumsuz etkilemektedir. Bu nedenle, düşük kaynaklı dillerde çeviri sistemlerinin performansını artırmak için yeni ve kapsamlı veri setlerine ihtiyaç duyulmaktadır. Bu tez çalışması, düşük kaynaklı diller için veri seti oluşturma ve bu veri setlerini kullanarak çeviri sistemlerinin performansını optimize etme konularında yeni yaklaşımlar sunmayı amaçlamıştır.

Bu tez çalışmasının birinci uygulaması, Türkçe-İngilizce çeviri sisteminin kalitesini artırmak amacıyla yeni bir paralel külliyat oluşturmaktır. Bu amaç doğrultusunda, Yükseköğretim Kurulu Ulusal Tez Merkezi (Yök Tez) veri sisteminden yüksek lisans ve doktora tezlerinin özet kısımları kullanılmıştır. Çalışma kapsamında, eğitim öğretim, psikoloji, mühendislik, işletme gibi 178 farklı alandan 2013-2022 yılları arasında yayınlanmış tez çalışmaları kullanılmıştır.

Yüksek kaliteli ve doğru çeviriler elde edebilmek için paralel külliyatın oluşturulması sürecinde, cümle hizalama işlemi büyük önem taşımaktadır. Bu doğrultuda, Hunalign ve

Vecalign gibi gelişmiş cümle hizalama algoritmaları kullanılarak, tez özetlerinin cümleleri hizalanmıştır. Hunalign algoritması, çift dilli metinlerin hizalanmasında etkin bir yöntem sunarken, Vecalign algoritması ise vektör temelli yaklaşımları ile hizalama doğruluğunu artırmaktadır. Bu algoritmaların entegrasyonu sayesinde, cümle hizalama işlemi daha yüksek doğrulukla gerçekleştirilmiş ve veri setinin kalitesi önemli ölçüde artırılmıştır.

Oluşturulan paralel külliyat, 1.2 milyondan fazla paralel cümlelere sahip olup, Türkçe-İngilizce çeviri sistemlerinin eğitiminde kullanılmak üzere hazırlanmıştır. Bu külliyat, mevcut çeviri modellerinin performansını artırmada kritik bir kaynak sağlamakta ve çeviri kalitesinin yükseltilmesine katkıda bulunmaktadır. Yeterli ve kaliteli veri setlerinin eksikliği, düşük kaynaklı dillerde çeviri sistemlerinin geliştirilmesinde önemli bir engel teşkil etmektedir. Bu çalışmada oluşturulan külliyat, bu engelin aşılmasına yönelik önemli bir adım olarak değerlendirilmektedir.

Bu tez çalışmasının birinci uygulaması, düşük kaynaklı dillerde çeviri sistemlerinin kalitesini artırmak için yeni külliyatlar oluşturmanın önemini vurgulamaktadır. Elde edilen külliyat, Türkçe-İngilizce dil çiftinde yapılan çeviri çalışmalarında kullanılmakta olup, modelin performansını optimize etmeye yönelik önemli katkılar sağlamaktadır. Bu bağlamda, çalışma, yalnızca Türkçe-İngilizce çeviri sistemlerinin geliştirilmesine değil, aynı zamanda diğer düşük kaynaklı dillerde de çeviri kalitesinin artırılmasına yönelik önemli bir referans noktası oluşturmaktadır.

Bu tez çalışmasının ikinci uygulaması, Türkçe için ön eğitimden geçen dil modellerinin performansını değerlendirmektir. Bu dil modellerinin özellikle Türkçe dilini anlama görevlerindeki başarısının ölçülmesi amaçlanmıştır. Bu doğrultuda, yazar profili oluşturma (Author Profiling) görevlerinden biri olan cinsiyet belirleme uygulaması geliştirilmiştir. Bu uygulama, Türkçe Twitter gönderilerinden cinsiyet tespiti yapmaya odaklanmaktadır. Problem, bir sınıflandırma görevi olarak ele alınmış ve çeşitli makine öğrenmesi ve derin öğrenme yöntemleri kullanılarak çözülmeye çalışılmıştır.

Çalışmada kullanılan yöntemler arasında, geleneksel makine öğrenmesi metotları olan TF-IDF (Term Frequency-Inverse Document Frequency) ve SVM (Support Vector Machine) ile derin öğrenme yöntemleri olan LSTM ve ESA bulunmaktadır. Ayrıca, Türkçe ile oluşturulmuş BERT, DistilBERT ve Electra dil modellerinin de performans değerlendirmesi için kullanılmıştır. Bu modeller, dilin anlamını ve bağlamını daha iyi kavrayarak cinsiyet tespiti görevinde yüksek başarımlar elde etmeyi hedeflemiştir.

Yapılan deneyler sonucunda, en yüksek başarıma (%80.1) kelime boyutunun 128k olduğu BERT modeli ile ulaşılmıştır. BERT modeli, çift yönlü dil modeli olması ve derin bağlamlı temsil yeteneği sayesinde, Türkçe dilini anlama görevinde üstün performans sergilemiştir. Bu sonuç, BERT modelinin Türkçe için etkili bir dil modeli olduğunu ve diğer modellerle kıyaslandığında daha yüksek doğrulukla çalıştığını göstermektedir.

Bu çalışmanın esas amacı, Türkçe dilini en iyi anlayan dil modellerini tespit etmek ve bu modelleri çeviri sistemlerine entegre ederek çeviri kalitesini artırmaktır. Türkçe-İngilizce dil çiftinde çeviri sistemlerinin performansını optimize etmek için, dilin yapısal ve anlamsal özelliklerini en iyi şekilde kavrayabilen modellerin kullanılması büyük önem taşımaktadır. Bu bağlamda, yapılan deneyler ve elde edilen sonuçlar, çeviri sistemlerinin geliştirilmesinde hangi dil modellerinin daha etkili olacağına dair değerli bilgiler sunmaktadır. Bu çalışma, düşük kaynaklı dillerde çeviri kalitesinin artırılmasına yönelik önemli bir adım olarak değerlendirilmektedir ve gelecekte yapılacak araştırmalar için sağlam bir temel oluşturmaktadır.

Bu tez çalışmasının üçüncü uygulaması, çok dilli ön eğitimli dil modellerinin çeviri sistemlerine uyarlanması için yeni bir model geliştirilmesini kapsamaktadır. Bu yöntemde, parametre verimli yöntemler kullanılarak, dil modellerinin çeviri performansı optimize edilmeye çalışılmıştır. Çalışmalar, çok başlı dikkat mekanizmasına ve ileri beslemeli sinir ağına adaptör eklenmesi üzerine yoğunlaşmıştır.

Dikkat başlığı ve ileri beslemeli sinir ağına tek tek veya birlikte adaptör eklenerek oluşturulan modeller, Türkçe-İngilizce makine çevirisi için ince ayar yapılmasına yardımcı olmuştur. Deneyler sonucunda, her iki katmana birlikte eklenen adaptörlerin kombinasyonu, diğer yöntemlere kıyasla daha iyi performans göstermiştir. Bu kombinasyon, önceden eğitilmiş dil model parametrelerinin yalnızca %5'inin ince ayarını gerektirmiş ve bu sayede parametre verimliliğini artırmıştır. Oluşturulan modelin en büyük katkısı hesaplama maliyetlerini düşürerek, çeviri sistemlerinin daha ekonomik bir şekilde optimize edilmesini sağlamasıdır. Önerilen yöntemin performansı, tam ince ayar yapılan çok dilli önceden eğitilmiş dil modeline kıyasla yalnızca %3'lük bir Bleu skoru farkı göstermiştir. Bu, neredeyse aynı performansın önemli ölçüde daha düşük maliyetle elde edilebildiğini göstermektedir.

Bu çalışma, düşük kaynaklı dillerde makine çevirisi için parametre verimliliğini artırma ve hesaplama maliyetlerini düşürme konularında önemli katkılar sağlamaktadır. Önerilen yöntem, düşük kaynaklı dillerde çeviri sistemlerinin geliştirilmesine yönelik yeni

yaklaşımlar sunarak, dil modellerinin entegrasyonu ve çeviri kalitesinin yükseltilmesi için değerli bilgiler sunmaktadır. Bu bağlamda, çalışma, yalnızca Türkçe-İngilizce dil çiftinde değil, diğer düşük kaynaklı dillerde de çeviri sistemlerinin performansını artırmaya yönelik önemli bir referans noktası oluşturmaktadır.

Bu tez çalışmasının dördüncü uygulaması, yeni bir çeviri sistemi ve sonradan düzenleme (post-editing) için yeni bir model oluşturulmasıdır. Önerilen çeviri sistemi, öz dikkat ağına sahip kodlayıcı kod çözücü mimarisini sığ füzyon yöntemiyle önceden eğitilmiş Sci-BERT dil modeliyle birleştirmektedir. Bu model, sığ füzyon katmanında tam dikkat ağ katmanı (Full Attention Layer) kullanır. Bu sayede, kelime düzeyinde dikkat artırılarak daha yüksek bir başarı oranı elde edilebilir.

Modelin değerlendirilmesi için farklı metrikler kullanılmıştır. Deneyler sonucunda oluşturulan model, 45.1 Bleu ve 73.2 Meteor puanlarına ulaşmıştır. Bu yüksek puanlar, modelin çeviri kalitesini ve doğruluğunu yansıtmaktadır. Ayrıca, önerilen model, World Machine Translation (2017–2018) test veri kümelerinde sıfır vuruşlu çeviri (zero-shot translation) yöntemiyle sırasıyla 20,12 ve 20,56 puan almıştır. Bu sonuçlar, modelin düşük kaynaklı dillerde bile etkili bir şekilde çalışabileceğini göstermektedir.

Önerilen yöntem, dil modelini çeviri sistemine dâhil etmek için diğer düşük kaynaklı dillere ilham vermektedir. Öz dikkat mimarisinin güçlü özelliklerini ve önceden eğitilmiş Sci-BERT dil modelinin bilgi zenginliğini birleştirerek, daha hassas ve anlamlı çeviriler üretmek mümkün olmuştur. Bu yöntem, düşük kaynaklı dillerdeki veri eksikliğini telafi etmekte ve çeviri sistemlerinin kalitesini artırmakta önemli bir rol oynamıştır.

Bu uygulama, düşük kaynaklı diller için çeviri sistemlerinin geliştirilmesine yönelik yeni yaklaşımlar sunarak, dil modellerinin entegrasyonu ve çeviri kalitesinin yükseltilmesi konularında önemli katkılar sağlamaktadır. Bu çalışma, gelecekte yapılacak araştırmalar için sağlam bir temel oluşturmakta ve doğal dil işleme alanında yeni gelişmelere kapı aralamaktadır.

Özet olarak bu tez çalışması, dört ana uygulama aracılığıyla düşük kaynaklı dillerde çeviri sistemlerinin kalitesini artırmaya yönelik yenilikçi yaklaşımlar sunmaktadır. Bu dört uygulama, düşük kaynaklı dillerde çeviri kalitesini artırmak için farklı ve etkili yaklaşımlar sunmaktadır. Geliştirilen veri setleri ve modeller, Türkçe-İngilizce dil çiftinde çeviri sistemlerinin performansını optimize ederken, diğer düşük kaynaklı diller için de örnek teşkil etmektedir. Özellikle parametre etkili yöntemler kullanılarak hesaplama maliyetlerinin düşürülmesi, çeviri sistemlerinin daha geniş çapta ve ekonomik bir şekilde

uygulanabilirliğini artırmaktadır. Bu tez, düşük kaynaklı dillerde çeviri kalitesinin yükseltilmesi, maliyetlerin azaltılması ve çeviri sistemlerinin daha erişilebilir hale getirilmesi konularında önemli katkılar sunarak, doğal dil işleme alanında yeni araştırmalara ve uygulamalara ilham kaynağı olmayı hedeflemektedir.

Gelecek çalışmalarda, düşük kaynaklı dil çiftleri üzerinde daha geniş deneyler yapılarak, Türkçe-İngilizce dil çiftinde elde edilen başarılı sonuçların diğer düşük kaynaklı dillerde de doğrulanması hedeflenmektedir. Ayrıca, düşük kaynaklı diller için daha büyük ve çeşitli veri setlerinin oluşturulması, çeviri modellerinin eğitimini ve test edilmesini kolaylaştıracaktır. Bu veri setlerinin, farklı konularda ve dil yapılarına sahip metinleri içermesi amaçlanmaktadır. Transfer öğrenme yöntemlerinin incelenmesi, yüksek kaynaklı dillerden elde edilen bilgi ve modellerin düşük kaynaklı dillere uyarlanmasını sağlayarak çeviri sistemlerinin performansını artıracaktır.

Geliştirilen çeviri sistemlerinin gerçek dünya uygulamaları üzerinde test edilmesi ve kullanıcı deneyimlerinin incelenmesi, sistemlerin pratikte ne kadar etkili olduğunu ve kullanıcılar tarafından nasıl değerlendirildiğini gösterecektir. Bunun yanı sıra, çeviri sistemlerinin adil ve eşit şekilde çalışmasını sağlamak için dil çiftleri arasındaki kültürel ve dilsel farklılıkları dikkate alan yeni yaklaşımlar geliştirilecektir. Bu sayede, çevirilerin doğru ve anlamlı olması sağlanacaktır. Son olarak, tezde sunulan yöntemler, doğal dil işleme ve yapay zekâ alanında yeni araştırmalara ve uygulamalara ilham kaynağı olabilir. Bu doğrultuda, mevcut yöntemlerin farklı alanlarda ve uygulamalarda nasıl kullanılabileceği üzerine çalışmalar yapılacaktır. Bu çalışmalar, düşük kaynaklı dillerde çeviri sistemlerinin kalitesini daha da artırarak, bu alandaki yenilikçi yaklaşımların geliştirilmesine katkı sağlayacaktır.

7. KAYNAKLAR

[1] **Haddow, B., Bawden, R., Valerio, A., Barone, M., Helcl, J. and Birch, A. (2022).** Survey of Low-Resource Machine Translation. https://doi.org/10.1162/COLI

[2] **Haddow, B., Huck, M., Federmann, C., Yepes, A.J., Neves, M., Negri, M. et al. (2016).** Findings of the 2016 Conference on Machine Translation (WMT16). **2**, 131–98.

[3] **Haddow, B., Post, M., Rubino, R., Federmann, C., Huck, M., Specia, L. et al. (2017).** Findings of the 2017 Conference on Machine Translation (WMT17). **2**, 169–214.

[4] **Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M. et al. (2018).** Findings of the 2018 Conference on Machine Translation (WMT18). *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, Association for Computational Linguistics (ACL). p. 272–303. https://doi.org/10.18653/v1/W18-64028

[5] **Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y. et al. (2019).** Findings of the 2019 Conference on Machine Translation (WMT19). **2**, 1–61. https://doi.org/10.18653/v1/w19-5301

[6] **Bawden, R., Di Nunzio, G.M., Grozea, C., Jauregi Unanue, I., Jimeno Yepes, A., Mah, N. et al. (2020).** Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages. *Proceedings of the Fifth Conference on Machine Translation*, 660–87.

[7] **Ranathunga, S., Lee, E.-S.A., Skenduli, M.P., Shekhar, R., Alam, M. and Kaur, R. (2021).** Neural Machine Translation for Low-Resource Languages: A Survey. ACM Computing Surveys, 55(11), 1-37.

[8] **Tayir, T. and Li, L. (2024).** Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*, ACM New York, NY. **23**, 1–22.

[9] **Koehn, P. (2020).** Neural machine translation. Cambridge University Press

[10] **Koehn, P. (2005).** Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit x: Papers*, p. 79–86.

[11] **Tiedemann, J. (2016).** OPUS-Parallel Corpora for Everyone. *Baltic Journal of Modern Computing*, **4**.

[12] **Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M. et al. (2020).** ParaCrawl: Web-scale acquisition of parallel corpora. Association for Computational Linguistics (ACL).

[13] **Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M. et al. (2018).** Findings of the 2018 Conference on Machine Translation (WMT18). *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, Association for Computational Linguistics (ACL). p. 272–303. https://doi.org/10.18653/v1/W18-64028

[14] **Ylmaz, E., El-Kahlout, I.D., Aydn, B., Özil, Z.S. and Mermer, C. (2013).** TÜBİTAK Turkish-English submissions for IWSLT 2013. *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*,.

[15] **Ma, Q., Bojar, O. and Graham, Y. (2018).** Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 671–88.

[16] **Oflazer, K. and Durgar El-Kahlout, Ilknur**. **(2007).** Exploring different representational units in English-to-Turkish statistical machine translation.

[17] **Bisazza, A. and Federico, M. (2009).** Morphological pre-processing for Turkish to English statistical machine translation. *Workshop on Spoken Language Translation*, 129–35.

[18] **Mermer, C., Kaya, H. and Doğan, M.U. (2010).** The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2010. *International Workshop on Spoken Language Translation (IWSLT) 2010*, 183–8.

[19] **Yeniterzi, R. and Oflazer, K. (2010).** Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 454–64.

[20] **Bakay, Ö., Avar, B. and Yildiz, O.T. (2019).** A tree-based approach for English-to-Turkish translation. *Turkish Journal of Electrical Engineering and Computer Sciences*, Turkiye Klinikleri Journal of Medical Sciences. **27**, 437–52. https://doi.org/10.3906/elk-1807-341

[21] **Ylmaz, E. and El-Kahlout, I.D. (2014).** The use of recurrent neural networks language model in Turkish-English machine translation. *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, p. 1247–50.

[22] **Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C. et al. (2015).** On using monolingual corpora in neural machine translation. *ArXiv Preprint ArXiv:150303535*,.

[23] **Sennrich, R., Haddow, B. and Birch, A. (2015).** Improving neural machine translation models with monolingual data. *ArXiv Preprint ArXiv:151106709*,.

[24] **Zoph, B., Yuret, D., May, J. and Knight, K. (2016).** Transfer learning for low-resource neural machine translation. *ArXiv Preprint ArXiv:160402201*,.

[25] **Bahdanau, D., Cho, K.H. and Bengio, Y. (2015).** Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

[26] **Li, X., Tracey, J., Grimes, S. and Strassel, S. (2016).** Uzbek-English and Turkish-English morpheme alignment corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2925–30.

[27] **Nguyen, T.Q. and Chiang, D. (2017).** Transfer learning across low-resource, related languages for neural machine translation. *ArXiv Preprint ArXiv:170809803*,.

[28] **Currey, A., Miceli-Barone, A.V. and Heafield, K. (2017).** Copied monolingual data improves low-resource neural machine translation. *Proceedings of the Second Conference on Machine Translation*, p. 148–56.

[29] **Sennrich, R., Haddow, B. and Birch, A. (2016).** Neural machine translation of rare words with subword units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3, 1715–25.

[30] **Dabre, R. and Fujita, A. (2019).** Recurrent stacking of layers for compact neural machine translation models. *Proceedings of the AAAI Conference on Artificial Intelligence*, p. 6292–9.

[31] **Luo, G., Yang, Y., Yuan, Y. and Chen, Z. (2019).** Hierarchical Transfer Learning Architecture for Low-Resource Neural Machine Translation. *IEEE Access*, IEEE. **7**, 154157–66. https://doi.org/10.1109/ACCESS.2019.2936002

[32] **Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M. et al. (2020).** Multilingual Denoising Pre-training for Neural Machine Translation.

[33] **Sun, M., Wang, H., Pasquine, M. and A. Hameed, I. (2021).** Machine translation in low-resource languages by an adversarial neural network. *Applied Sciences*, MDPI. 11, 10860.

[34] **Maimaiti, M., Liu, Y., Luan, H. and Sun, M. (2022).** Data augmentation for low-resource languages NMT guided by constrained sampling. *International Journal of Intelligent Systems*, Wiley Online Library. 37, 30–51.

[35] **Gong, L., Li, Y., Guo, J., Yu, Z. and Gao, S. (2022).** Enhancing low-resource neural machine translation with syntax-graph guided self-attention. *Knowledge-Based Systems*, Elsevier. 246, 108615.

[36] **Chen, S., Zeng, Y., Cao, D. and Lu, S. (2022).** Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems*, Elsevier. 245, 108598.

[37] **Görgün, O. (2022).** English to Turkish machine translation using synchronous grammars. Işık Üniversitesi.

[38] **Turhan, C.K. (1997).** An English to Turkish machine translation system using structural mapping. *Fifth Conference on Applied Natural Language Processing*, p. 320–3.

[39] **Altıntaş, K. (2001).** Turkish to Crimean Tatar machine translation system. Bilkent Universitesi (Turkey).

[40] **Durgar El-Kahlout, I. and Oflazer, K. (2006).** Initial explorations in English to Turkish statistical machine translation.

[41] **Alp, N.D. and Turhan, C. (2008).** English to Turkish Example-Based Machine Translation with Synchronous SSTC. *Fifth International Conference on Information Technology: New Generations (Itng 2008)*, p. 674–9.

[42] **Türe, F. (2008).** A Hybrid Machine Translation System from Turkish to English. Sabancı Üniversitesi, Doktora Tezi.

[43] **Kopru, S. (2009).** AppTek turkish-english machine translation system description for IWSLT 2009. *Proceedings of the 6th International Workshop on Spoken Language Translation: Evaluation Campaign*,.

[44] **Durgar El-Kahlout, Ilknur**. **(2009).** A prototype English-Turkish statistical machine translation system.

[45] **El-Kahlout, I.D. and Oflazer, K. (2009).** Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE. 18, 1313–22.

[46] **Matusov, E. and Kopru, S. (2010).** AppTek's APT Machine Translation System for IWSLT 2010. *Proceedings of the 7th International Workshop on Spoken Language Translation: Evaluation Campaign*,.

[47] **Mermer, C.C. (2010).** Unsupervised Search for The Optimal Segmentation for Statistical Machine Translation, Proceedings of the ACL 2010 Student Research Workshop.

[48] **Wu, C., Xia, F., Deleger, L. and Solti, I. (2011).** Statistical machine translation for biomedical text: are we there yet? *AMIA Annual Symposium Proceedings*, p. 1290.

[49] **Mermer, C. and Saraclar, M. (2011).** Unsupervised Turkish morphological segmentation for statistical machine translation. *Workshop of MT and Morphologically-Rich Languages*,.

[50] **Görgün, O. and Yldz, O.T. (2012).** Using morphology in English-Turkish statistical machine translation. *2012 20th Signal Processing and Communications Applications Conference (SIU)*, p. 1–4.

[51] **Mermer, C., Saraçlar, M. and Sarikaya, R. (2013).** Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE. 21, 1090–101.

[52] **Eyigöz, E., Gildea, D. and Oflazer, K. (2013).** Simultaneous word-morpheme alignment for statistical machine translation. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 32–40.

[53] **Yldrm, E. and Tantuğ, A.C. (2013).** The feasibility analysis of re-ranking for N-best lists on English-Turkish machine translation. *2013 IEEE INISTA*, p. 1–5.

[54] **Yildiz, O.T., Solak, E., Görgün, O. and Ehsani, R. (2014).** Constructing a Turkish-English parallel treebank. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 2, 112–7. https://doi.org/10.3115/v1/p14-2019

[55] **Biçici, E. and Yuret, D. (2014).** Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE. 23, 339–50.

[56] **Ding, S., Duh, K., Khayrallah, H., Koehn, P. and Post, M. (2016).** The JHU machine translation systems for WMT 2016. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, p. 272–80.

[57] **Firat, O., Cho, K., Sankaran, B., Vural, F.T.Y. and Bengio, Y. (2017).** Multi-way, multilingual neural machine translation. *Computer Speech & Language*, Elsevier. 45, 236–52.

[58] **García-martínez, M., Caglayan, O., Aransa, W., Bardet, A., Bougares, F. and Barrault, L. (2017).** LIUM Machine Translation Systems for WMT17 News Translation Task.

[59] **Ataman, D., Negri, M., Turchi, M. and Federico, M. (2017).** Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 331–42. https://doi.org/10.1515/pralin-2017-0031

[60] **Shen, Y., Tan, X., He, D., Qin, T. and Liu, T.-Y. (2018).** Dense information flow for neural machine translation. *ArXiv Preprint ArXiv:180600722*,.

[61] **Ataman, D. and Federico, M. (2018).** An evaluation of two vocabulary reduction methods for neural machine translation. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, p. 97–110.

[62] **Deng, Y., Cheng, S., Lu, J., Song, K., Wang, J., Wu, S. et al. (2018).** Alibaba's neural machine translation systems for wmt18. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 368–76.

[63] **Marie, B., Wang, R., Fujita, A., Utiyama, M. and Sumita, E. (2018).** NICT's neural and statistical machine translation systems for the WMT18 news translation task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 449–55.

[64] **Schamper, J., Rosendahl, J., Bahar, P., Kim, Y., Nix, A. and Ney, H. (2018).** The RWTH Aachen University supervised machine translation systems for WMT 2018. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 496–503.

[65] **Aji, A.F., Bogoychev, N., Heafield, K. and Sennrich, R. (2020).** In neural machine translation, what does transfer learning transfer?

[66] **Behnke, M. and Heafield, K. (2020).** Losing heads in the lottery: Pruning transformer attention in neural machine translation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2664–74.

[67] **Yılmaz, S. and Uğuz, İ. (2020).** Turkish-English Machine Translation System, Boğaziçi Üniversitesi.

[68] **Escolano, C., Costa-Jussà, M.R. and Fonollosa, J.A.R. (2021).** From bilingual to multilingual neural-based machine translation by incremental training. *Journal of the Association for Information Science and Technology*, Wiley Online Library. 72, 190–203.

[69] **Liu, Q., Kusner, M. and Blunsom, P. (2021).** Counterfactual data augmentation for neural machine translation. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 187–97.

[70] **Yirmibeşoğlu, Z. (2021).** Morphologically Motivated Input Variations in Turkish-English Neural Machine Translation. Boğaziçi Üniversitesi. Yüksek Lisans Tezi.

[71] **Ciora, C., Iren, N. and Alikhani, M. (2021).** Examining covert gender bias: A case study in Turkish and English machine translation models. *ArXiv Preprint ArXiv:210810379*,.

[72] **Görgün, O. and Yildiz, O.T. (2022).** Evaluating the English-Turkish parallel treebank for machine translation. *Turkish Journal of Electrical Engineering and Computer Sciences*, **30**, 184–99.

[73] **Ata, M. and Debreli, E. (2021).** Machine Translation in the Language Classroom: Turkish EFL Learners' and Instructors' Perceptions and Use. *IAFOR Journal of Education*, ERIC. 9, 103–22.

[74] **Yirmibeşoğlu, Z. and Güngör, T. (2022).** Morphologically Motivated Input Variations and Data Augmentation in Turkish-English Neural Machine Translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, ACM New York, NY.

[75] **Tiedemann, J. (2020).** The Tatoeba Translation Challenge -- Realistic Data Sets for Low Resource and Multilingual MT.

[76] **Yang, S., Wang, Y. and Chu, X. (2020).** A Survey of Deep Learning Techniques for Neural Machine Translation.

[77] **Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A. et al. (2011).** Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25, 127–44. https://doi.org/10.1007/s10590-011-9090-0

[78] **Koehn, P., Och, F.J. and Marcu, D. (2003).** Statistical Phrase-Based Translation. In 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003) (pp. 48-54). Association for Computational Linguistics.

[79] **Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N. et al. (2007).** Moses: Open source toolkit for statistical machine translation. *Prague: Proceedings of 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–80.

[80] **Cho, K., van Merrienboer, B., Bahdanau, D. and Bengio, Y. (2015).** On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. 103–11. https://doi.org/10.3115/v1/w14-4012

[81] **Kalchbrenner, N. and Blunsom, P. (2013).** Recurrent continuous translation models. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1700–9.

[82] **Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. et al. (2014).** Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–34. https://doi.org/10.3115/v1/d14-1179

[83] **Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K.J. (1989).** Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 328–39. https://doi.org/10.1109/29.21701

[84] **Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z. et al. (2017).** Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, **5**, 339–51. https://doi.org/10.1162/tacl_a_00065

[85] **Kaiser, Ł. and Bengio, S. (2016).** Can active memory replace attention? *Advances in Neural Information Processing Systems*, 3781–9.

[86] **Kaiser, Ł. and Sutskever, I. (2016).** Neural GPUs learn algorithms. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 1–9.

[87] **Gehring, J., Auli, M., Grangier, D. and Dauphin, Y.N. (2016).** A Convolutional Encoder Model for Neural Machine Translation.

[88] **Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N. (2017).** Convolutional Sequence to Sequence Learning. In International conference on machine learning (pp. 1243-1252). PMLR.

[89] **Luong, M.T., Pham, H. and Manning, C.D. (2015).** Effective approaches to attention-based neural machine translation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1412–21. https://doi.org/10.18653/v1/d15-1166

[90] **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. (2017).** Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5999–6009.

[91] **Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. et al. (2016).** Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 1–23.

[92] **Kudo, T. (2018).** Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.

[93] **Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2018).** Language Models are Unsupervised Multitask Learners. OpenAI blog, 1(8), 9.

[94] **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. et al. (2019)**. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[95] **Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, K. (2018).** Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:181004805*,.

[96] **Britz, D., Goldie, A., Luong, M.T. and Le, Q. V. (2017).** Massive exploration of neural machine translation architectures. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1442–51. https://doi.org/10.18653/v1/d17-1151

[97] **Sel, İ. and Hanbay, D. (2022).** Fully Attentional Network for Low-Resource Academic Machine Translation and Post Editing. *Applied Sciences*, MDPI AG. 12, 11456. https://doi.org/10.3390/app122211456

[98] **Freitag, M. and Al-Onaizan, Y. (2017).** Beam Search Strategies for Neural Machine Translation. arXiv preprint arXiv:1702.01806.

[99] **Shao, C., Feng, Y. and Chen, X. (2018).** Greedy Search with Probabilistic N-gram Matching for Neural Machine Translation.

[100] **Maučec, M.S. and Donaj, G. (2019).** Machine Translation and the Evaluation of Its Quality. Recent trends in computational intelligence, 143.

[101] **Rivera-Trigueros, I. (2022).** Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, Springer Science and Business Media B.V. 56, 593–619. https://doi.org/10.1007/s10579-021-09537-5

[102] **Chatzikoumi, E. (2020).** How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, Cambridge University Press. 26, 137–61. https://doi.org/10.1017/S1351324919000469

[103] **Lavie, A. and Agarwal, A. (2007).** Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments.

[104] **Correia, A. de S. and Colombini, E.L. (2021).** Attention, please! A survey of Neural Attention Models in Deep Learning. https://doi.org/10.1007/s10462-022-10148-x

[105] **Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O. (2015).** Listen, Attend and Spell. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4960-4964). IEEE.

[106] **Vinyals, O. and Le, Q. (2015)**. A Neural Conversational Model. arXiv preprint arXiv:1506.05869.

[107] **Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M. (2022).** Transformers in Vision: A Survey. *ACM Computing Surveys*, Association for Computing Machinery (ACM). **54**, 1–41. https://doi.org/10.1145/3505244

[108] **Dubey, S.R., Singh, S.K. and Chaudhuri, B.B. (2021).** Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. Neurocomputing, 503, 92-108.

[109] **Ramachandran, P., Zoph, B. and Le, Q. V. (2017).** Searching for Activation Functions. arXiv preprint arXiv:1710.05941.

[110] **Hendrycks, D. and Gimpel, K. (2016).** Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415.

[111] **He, P., Liu, X., Gao, J. and Chen, W. (2020).** DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint arXiv:2006.03654.

[112] **Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G. et al. (2017).** Outrageously Large Neural Networks: The Sparsely-Gated Mixture-Of-Experts Layer. arXiv preprint arXiv:1701.06538.

[113] **Dauphin, Y.N., Fan, A., Auli, M. and Grangier, D. (2017).** Language Modeling with Gated Convolutional Networks. In International conference on machine learning (pp. 933-941). PMLR.

[114] **Lin, T., Wang, Y., Liu, X. and Qiu, X. (2022).** A survey of transformers. *AI Open*, Elsevier BV. 3, 111–32. https://doi.org/10.1016/j.aiopen.2022.10.001

[115] **Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. and Trón, V. (2005).** Parallel corpora for medium density languages. *International Conference Recent Advances in Natural Language Processing, RANLP*, 2005-Janua, 590–6. https://doi.org/10.1075/cilt.292.32var

[116] **Thompson, B. and Koehn, P. (2020).** Vecalign: Improved sentence alignment in linear time and space. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on*

Natural Language Processing, Proceedings of the Conference, 1342–8. https://doi.org/10.18653/v1/d19-1136

[117] **Pavlick, E., Post, M., Irvine, A., Kachaev, D. and Callison-Burch, C. (2014).** The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2, 79–92. https://doi.org/10.1162/tacl_a_00167

[118] **Artetxe, M. and Schwenk, H. (2019).** Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, **7**, 597–610. https://doi.org/10.1162/tacl_a_00288

[119] **Sel, I., Karci, A. and Hanbay, D. (2019).** Feature Selection for Text Classification Using Mutual Information. *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*,. https://doi.org/10.1109/IDAP.2019.8875927

[120] **Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2020).** Deep Learning Based Text Classification: A Comprehensive Review. *ArXiv*, 54.

[121] **Qi, Y., Sachan, D.S., Felix, M., Padmanabhan, S.J. and Neubig, G. (2018).** When and why are pre-trainedword embeddings useful for neural machine translation? *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, **2**, 529–35. https://doi.org/10.18653/v1/n18-2084

[122] **Post, M. (2018).** A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Stroudsburg, PA, USA. 186–91. https://doi.org/10.18653/v1/W18-6319

[123] **Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019).** DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*, 2–6.

[124] **Clark, K., Luong, M.T., Le, Q. V. and Manning, C.D. (2020).** Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, 1–18.

[125] **Pardo, F.M.R., Giachanou, A., Ghanem, B. and Rosso, P. (2020).** Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. *CLEF 2020 Labs and Workshops, Notebook Papers*, 22–5.

[126] **Álvarez-Carmona, M.A., Pellegrin, L., Montes-Y-Gómez, M., Sánchez-Vega, F., Escalante, H.J., López-Monroy, A.P. et al. (2018).** A visual approach for age and gender identification on Twitter. *Journal of Intelligent and Fuzzy Systems*, 34, 3133–45. https://doi.org/10.3233/JIFS-169497

[127] **Rangel, F., Rosso, P., Montes-Y-Gómez, M., Potthast, M. and Stein, B. (2018).** Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. *CEUR Workshop Proceedings*, 2380.

[128] **Sezerer, E., Polatbilek, O. and Tekir, S. (2019).** A Turkish Dataset for Gender Identification of Twitter Users. 203–7. https://doi.org/10.18653/v1/w19-4023

[129] **Qian, L.I., Peng, H., Jianxin, L.I., Congyin, X.I.A., Lifang, H.E., Lichao, S.U.N. et al. (2020).** A Survey on Text Classification: From Shallow to Deep Learning. *ArXiv*, 31, 1–21.

[130] **Pizarro, J. (2019).** Using N-grams to detect Bots on Twitter Notebook for PAN at CLEF 2019. *CEUR Workshop Proceedings*, 2380, 9–12.

[131] **ElSayed, S. and Farouk, M. (2020).** Gender identification for Egyptian Arabic dialect in twitter using deep learning models. *Egyptian Informatics Journal*, Faculty of Computers and Artificial Intelligence, Cairo University. 21, 159–67. https://doi.org/10.1016/j.eij.2020.04.001

[132] **Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. et al. (2021).** LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[133] **Poth, C., Sterz, H., Paul, I., Purkayastha, S., Engländer, L., Imhof, T. et al. (2023).** Adapters: A unified library for parameter-efficient and modular transfer learning. *ArXiv Preprint ArXiv:231111077*,.

[134] **Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A. et al. (2019).** Parameter-Efficient Transfer Learning for NLP. In International conference on machine learning (pp. 2790-2799). PMLR.

[135] **Li, X.L. and Liang, P. (2021).** Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv preprint arXiv:2101.00190.

[136] **He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T. and Neubig, G. (2021).** Towards a Unified View of Parameter-Efficient Transfer Learning.

[137] **Mao, Y., Mathias, L., Hou, R., Almahairi, A., Ma, H., Han, J. et al. (2021).** UNIPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. arXiv preprint arXiv:2110.07577.

[138] **Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V. et al. (2020).** Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv Preprint ArXiv:200800401*,.

[139] **Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O. et al. (2019).** Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv Preprint ArXiv:191013461*,.

[140] **Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A. et al. (2019).** HuggingFace's Transformers: State-of-the-art Natural Language Processing.

[141] **Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. et al. (2016).** ASPEC: Asian Scientific Paper Excerpt Corpus. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2204–8.

[142] **Neves, M., Yepes, A.J. and Névéol, A. (2016).** The scielo corpus: A parallel corpus of scientific publications for biomedicine. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2942–8.

[143] **Sel, İ., Üzen, H. and Hanbay, D. (2021).** Creating a Parallel Corpora for Turkish-English Academic Translations. *Bilgisayar Bilimleri*, Anatolian Science - Bilgisayar Bilimleri Dergisi. 335–40. https://doi.org/10.53070/BBD.990959

[144] **Gulcehre, C., Firat, O., Xu, K., Cho, K. and Bengio, Y. (2017).** On integrating a language model into neural machine translation. *Computer Speech and Language*, Elsevier Ltd. 45, 137–48. https://doi.org/10.1016/j.csl.2017.01.014

[145] **Yan, R., Li, J., Su, X., Wang, X. and Gao, G. (2022).** Boosting the Transformer with the BERT Supervision in Low-Resource Machine Translation. *Applied Sciences (Switzerland)*, MDPI. 12. https://doi.org/10.3390/app12147195

[146] **Beltagy, I., Lo, K. and Cohan, A. (2019).** SCIBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676.

[147] **Skorokhodov, I., Rykachevskiy, A., Slotin, S. and Ponkratov, A. (2018).** Semi-Supervised Neural Machine Translation with Language Models. In Proceedings of the AMTA 2018 workshop on technologies for MT of low resource languages (LoResMT 2018) (pp. 37-44).

[148] **Pan, Y., Li, X., Yang, Y. and Dong, R. (2020).** Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation.

[149] **Ranathunga, S., Lee, E.-S.A., Skenduli, M.P., Shekhar, R., Alam, M. and Kaur, R. (2021).** Neural Machine Translation for Low-Resource Languages: A Survey. ACM Computing Surveys, 55(11), 1-37.

8. ÖZGEÇMİŞ

Ad Soyad **:** İlhami SEL

ÖĞRENİM DURUMU:

- **Lisans** **:**2007, Fırat Üniversitesi, Teknik Eğitim Fakültesi, Elektronik Bilgisayar Eğitimi Bölümü, Bilgisayar Öğretmenliği
- **Lisans** **:**2018, İnönü Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü
- **Yüksek Lisans** **:**2013, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik Bilgisayar Eğitimi Bölümü
- **Doktora** **:**2024, İnönü Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Bölümü

MESLEKİ DENEYİM:

- **2007- halen** **:** Millî Eğitim Bakanlığı, Teknik Öğretmen

DOKTORA TEZİNDEN TÜRETİLEN ÇALIŞMALAR

1. **Uluslararası hakemli dergilerde yapılan çalışmalar (SCI & SSCI):**
   - Sel, İ., & Hanbay, D. (2022). Fully Attentional Network for Low-Resource Academic Machine Translation and Post Editing. Applied Sciences, 12(22), 11456. https://doi.org/10.3390/app122211456
2. **Ulusal Ulakbim'de ve Dergipark'ta taranan dergilerde yapılan çalışmalar:**
   - Sel, İ. & Hanbay, D. (2021). Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 33 (2) , 675-684 . DOI: 10.35234/fumbd.929133.
   - Sel, İ., Üzen, H. & Hanbay, D. (2021). Creating a Parallel Corpora for Turkish-English Academic Translations. Computer Science, 5th International Artificial Intelligence and Data Processing symposium, 335-340. DOI: 10.53070/bbd.990959
3. **Uluslararası Bilimsel Toplantılarda Sunulan ve Bildiri Kitabında Basılan Bildiriler:**
   - Sel, İ., & Hanbay, D. (2019). E-mail classification using natural language processing. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE. DOI: 10.1109/SIU.2019.8806593.

- Sel, İ., Karci, A., & Hanbay, D. (2019). Feature Selection for Text Classification Using Mutual Information. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-4). IEEE. DOI:10.1109/IDAP.2019.8875927.
- Sel, İ., Yeroğlu, C., & Hanbay, D. (2019). Feature Selection by Using Heuristic Methods for Text Classification. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-6). IEEE. DOI:10.1109/IDAP.2019.8875892.