T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNGİLİZCEDEN TÜRKÇE'YE İSTATİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE PARALEL DERLEM BOYUTU ve KALİTESİNİN ETKİLERİ

ERAY YILDIZ

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ. DR. BANU DİRİ

EŞ DANIŞMAN

YRD. DOÇ. DR. A. CÜNEYD TANTUĞ

İSTANBUL, 2014

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

İNGİLİZCEDEN TÜRKÇE'YE İSTATİKSEL MAKİNE ÇEVİRİSİ SİSTEMLERİNDE PARALEL DERLEM BOYUTU ve KALİTESİNİN ETKİLERİ

Eray Yıldız tarafından hazırlanan tez çalışması 08.07.2014 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Eş Danışman

Yrd. Doç. Dr. A. Cüneyd TANTUĞ

İstanbul Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. A. Cüneyd TANTUĞ

İstanbul Teknik Üniversitesi _____

Yrd. Doç. Dr. Gülşen Eryiğit

İstanbul Teknik Üniversitesi _____

Yrd. Doç. Dr. Arzucan ÖZGÜR

Boğaziçi Üniversitesi _____

ÖNSÖZ

Bu tez çalışmasının gerçekleşmesinde değerli bilgilerinden ve yardımlarından yararlandığım, bana her daim yol gösteren ve yardımcı olan danışman hocalarım Yrd. Doç. Dr. A. Cüneyd Tantuğ'a ve Doç. Dr. Banu Diri'ye, çalışmalarımı inceleme ve takip etme nezaketini gösteren Prof. Dr. Eşref Adalı'ya sonsuz teşekkürlerimi bir borç bilirim.

Desteklerini benden esirgemeyen Ezgi ve İsmail başta olmak üzere tüm laboratuvar arkadaşlarıma, yaşadığım her türlü sıkıntıyı, zorluğu ve sevinci benimle yaşayan değerli arkadaşım Duygu'ya ve tüm eğitim hayatım boyunca her zaman yanımda olan ve bana destek veren aileme çok teşekkür ederim.

Temmuz, 2014

Eray Yıldız

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ

ALPAC	Automatic Language Processing Advisory Committee
AMD	Advanced Micro Devices
BÇ	Bilgisayarlı Çeviri
B	Bulma
BLEU	BiLingual Evaluation Understudy
BSA	Bilingual Sentence Aligner
DARPA	Defense Advanced Research Projects Agency
DDİ	Doğal Dil İşleme
DVM	Destek Vektör Makineleri
ENIAC	Electronical Numerical Integrator and Calculator
F1	F1 Ölçütü
FAHQT	Fully Automatic - High Quality output -unrestricted Text
GB	Gigabyte
IBM	International Business Machines
İBÇ	İstatiksel Bilgisayarlı Çeviri
LDC	Linguistic Data Consortium
MAT	Machine Asisted Translation
METEOR	Method for Evaluation of Translation with Explicit Reordering
NB	Naive Bayes
ÖTBÇ	Örnek Tabanlı Bilgisayarlı Çeviri
RKO	Rastsal Karar Ormanı
RTF	Radyal Tabanlı Fonksiyon
SETimes	South-East European Times
T	Tutturma
WEKA	Waikato Environment for Knowledge Analysis
YSA	Yapay Sinir Ağı

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

İNGİLİZCE'DEN TÜRKÇE'YE İSTATİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE PARALEL DERLEM BOYUTU ve KALİTESİNİN ETKİLERİ

Eray YILDIZ

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Banu DİRİ

Eş Danışman: Yard. Doç. Dr. A. Cüneyd TANTUĞ

Bilgisayalı Çeviri (BÇ) bir dilde yazılmış bir ifadenin başka bir dile bilgisayar tarafından otomatik olarak çevrilmesi işlemidir. BÇ konusunda yapılan çalışmalar 1950'lerin ilk yıllarında başlamıştır. İkinci Dünya Savaşı sonrası önemi arttığı düşünülen bu alana siyasal, sosyal ve ticari sebeplerden oldukça fazla yatırım yapılmış; birçok araştırmacı bu konuda çalışmıştır. Takip eden yıllarda ise yine birçok akademik ve ticari çevrelerde önemli çalışmalar yapılmasına, büyük bütçeler ayrılmasına rağmen beklentileri karşılayan sonuçlar alınamamış ve 1960'lı yılların ortalarından itibaren bu alana yapılan yatırımlar ve bu konuda çalışmalar azalmaya başlamıştır. BÇ ile ilgili olarak kalite, maliyet, öngörüler, beklentiler ve ihtiyaçlar konusunda çalışmalar yapan Automatic Language Processing Advisory Committee (ALPAC) kuruluşunun 1964 yılında yayınladığı olumsuz rapor sonrasında bu alanda motivasyon ve yatırım kaybı oluşmuştur. BÇ'nin ilk dönemi olarak görülen bu dönemde sistemler daha çok çeşitli dilbilgisel düzeylerde (biçimbilimsel, sözdizimsel, anlamsal) çalışan kural tabanlı sistemler olarak gerçekleştirilmiştir. 1990'lı yıllardan itibaren gelişen internet teknolojisinin etkisiyle öne çıkan istatiksel yöntemler, ses işleme, doğal dil işleme konularında da değerlendirilmeye başlanmıştır. IBM'in öncülüğünde yapılan İBÇ (İstatiksel Bilgisayarlı Çeviri) çalışmaları BÇ alanındaki duraksamayı ortadan kaldırmış; birçok araştırmacı gelişen bu yeni alanda çalışmalara başlamıştır. Yine 90'lı yıllardan sonra ortaya çıkan veriye dayalı diğer bir yöntem de örnek tabanlı BÇ yöntemidir.

Günümüzde çeşitli kaynaklardan BÇ için veri elde etme nisbeten daha kolay olduğu için istatiksel yöntemlerin de katkısıyla BÇ çalışmaları belirli bir başarıya ulaşmış ve çeşitli alanlardaki uygulamaları giderek artmıştır. Fakat bir BÇ sisteminden beklenen özelliklerin hepsini birlikte başaran sistemler üzerine araştırma-geliştirme faaliyetleri hızla devam etmektedir. BÇ sisteminden beklenen bu özellikler: anlaşılır ve aslına uygun çeviri yapabilmesi, insan etkisi olmadan otomatik çeviri yapabilmesi ve belirli bir konuya bağlı olmadan genel amaçlı çeviri yapabilmesi olarak sıralanabilir.

Örnek tabanlı ve istatiksel yöntemlerin eğitim için kullandığı verilerden en önemlisi paralel derlemlerdir. Birbirinin çevirisi olan metinlerden oluşan ve cümle seviyesinde hizalanmış olan paralel derlemler BÇ'nin yanı sıra sözcük belirsizliği giderme, bilgi erişimi gibi diğer doğal dil işleme alanlarında da kullanılmaktadır.

Bu çalışmada BÇ tarihi, yöntemleri hakkında genel bilgiler toparlanılmış; İBÇ yöntemlerinin günümüzde geldiği nokta araştırılmıştır. Ayrıca, erişilebilir Türkçe-İngilizce paralel derlemler incelenmiş ve çeşitli kaynaklardan yeni paralel derlemler oluşturularak Türkçe-İngilizce paralel derlem sayısının artmasına katkıda bulunulmuştur. İngilizce'den Türkçe'ye istatistiksel BÇ sistemleri üzerinde paralel derlemin büyüklüğünün ve kalitesinin etkisi araştırılmıştır. Paralel cümle çiftlerinin kalitesininin otomatik ölçülebilmesi için cümle çiftlerinden çeşitli özellikler çıkaran makine öğrenmesi yöntemleri kullanılarak cümle çiftlerini kaliteli ve kalitesiz olarak sınıflandıran bir sınıflandırıcı geliştirilmiştir. Yapılan deneylerde elimizdeki paralel derlemlerden oluşturulan farklı boyutlarda paralel derlemlerle İBÇ sistemleri eğitilerek paralel derlemin büyüklüğünün etkisini araştırmak amacıyla başarıları karşılaştırılmıştır. Daha sonra paralel derlemin kalitesinin etkilerini gözlemleyebilmek için farklı boyutlardaki her bir derlemin sadece sınıflandırıcının kaliteli olarak işaretlediği örnekleri kullanarak İBÇ sistemleri eğitilmiştir. Paralel derlemin boyutu arttıkça daha yüksek başarılara ulaşıldığı gösterilirken; içerisinde hatalı veya kalitesiz örnekleri temizlenmiş daha az sayıda örnek içeren paralel derlemler ile aynı veya daha yüksek başarılara ulaşıldığı gösterilmiştir.

Anahtar Kelimeler: Makine Öğrenmesi, Yapay Zeka, Doğal Dil İşleme, Makine Çevirisi, Bilgisayarlı Çeviri, İstatiksel Bilgisayarlı Çeviri, Paralel Derlem, Paralel Derlem Filtreleme, Örnek Seçimi

ABSTRACT

THE EFFECT OF PARALLEL CORPUS QUALITY VS SIZE IN ENGLISH-TO-TURKISH STATISTICAL MACHINE TRANSLATION

Eray Yıldız

Department of Computer Engineering

MSc. Thesis

Adviser: Assoc.Prof. Banu DİRİ

Co-Adviser: Assist.Prof. A. Cüneyd TANTUĞ

Machine Translation (MT) is the process of translating an expression to another language automatically with the aid of computers. MT has been studied since the early 1950s. MT, which is thought to increase in importance after World War 2, has been invested due to political, social and economic facts. Although, many important studies have been conducted in the following years, the results couldn't meet expectations. The investments and studies in this field began to decline from the middle of 1960. The Automatic Language Processing Advisory Committee (ALPAC) which studies about costs, projections, expectations and requirements about MT, has issued a negative report about MT and caused loss of motivation and investment in MT field. During this first period of MT studies, MT was primarily performed using rule based transfers of some representation levels like morphological, syntactical or semantic representations. The statistical approaches which are developep under the fluence of internet and big data technologies have started to be utilized in signal processing and natural language processing. The hesitancy in MT has eliminated by Statistical Machine Translation (SMT) studies pioneered by IBM and many researchers has started to work in developing this new field. Another MT approach that based on training data is example based machine translation (EBMT).

Nowadays, MT systems have reached a certain success and its applications in various fields have steadily increased because of the convenience of data acquisition. But, the

research and development activities on the systems that are able to combine all of the features expected, is proceeding rapidly. The featetures that expected from a successful MT system are as follws: ability to process understandable and literal translations, ability to process automatic translations without any human intervention and ability to process general-purpose texts without any domain restriction.

The most important training data for example based MT models and statistical MT models are parallel corpus. Parallel Corpus are consist of texts that translation of each other and aligned at sentence level. In addition to MT, parallel corpus are widely utilized in word disambiguation, information retrieval and some of other natural language processing fields.

In this study, general information about history of MT and methods are presented, the point reached by SMT is investigated. Furthermore, publicly avaible parallel corpus between Turkish and English languages are studied and severalTurkish - English parallel corpus are constructed from various sources. The aim of this study is to figure out the effects of parallel corpus size and quality in statistical machine translation between Turkish and English languages. In this study, a machine learning based classifier is developed to classify parallel sentence pairs in a parallel corpus as high-quality or poor quality. This calassifier has been applied to a parallel corpus contains 1 million parallel English – Turkish sentence pairs and 600K high-quality parallel sentence pairs were obtained. The multiple SMT systems with various sizes of entire raw parallel corpus and filtered high quality corpus, their performances are evaluated in our experiments. As expected, the experiments show that the size of parallel corpus is a major factor in translation performance. However, instead of extended corpus with all available “so-called” parallel data, a better translation performance and reduced time-complexity can be achieved with a smaller high-quality corpus using a quality filter.

Keywords: Machine Learning, Artificial Intelligence, Natural Language Processing, Machine Translation, Statistical Machine Translation, Parallel Corpus, Parallel Corpus Filtering, Data Selection

BÖLÜM 1

GİRİŞ

Küreselleşme ve internet çağında iletişim birçok alanda çok önemli bir konuma sahip olmuştur. Günümüzde teknoloji ve bilgi sistemleri siyasal, sosyal ve gündelik hayatın her alanına nüfus ederek insanların organizasyonunda çok önemli bir rol oynamaya başlamıştır. Özellikle iletişim teknolojileri sayesinde birbirinden çok uzakta olan insanların ticari, eğitimsel ve sosyal birçok konuda organize olabilmesi giderek kolaylaşmaktadır. Küreselleşen ve günümüz dünyasında iletişim ve etkileşimin daha gelişmesi ve pratikleşmesi yolunda aşılmaya çalışılan bir engel de farklı coğrafyalarda çok değişik biçimlerde kullanılan diller arası farklılıklardır. Diller arasındaki bu farklılıklar, dilin bilgisayarlar tarafından işlenmesi konusunda her dil için özel bir çaba gereksimine yol açmıştır [1]. İngilizce, Fransızca, Almanca gibi diller üzerinde oldukça fazla çalışma varken son yıllarda Çince, Arapça, Japonca gibi diller üzerinde de birçok çalışma yapılmıştır. Bu konuda Türkçe üzerinde yapılan çalışmaların son on yılda hız kazandığı söylenebilir. Türkçe kendine özgü yapısal ve dilbilgisel özellikleri sebebiyle bilgisayar tarafından işlenebilmesi için birçok zorluk içermekte ve dolayısıyla özel bir ilgi ve çaba gerektirmektedir.

Doğal dilleri otomatik olarak çözümlemeyi, anlamayı, yorumlamayı ve üretmeyi amaçlayan Doğal Dil İşleme konusu, yapay zeka ve dilbilimi alanlarının bir alt dalı olarak görülmektedir[1]. Bilgisayarlar ile Doğal Dil İşleme çok değişik alanlarda uygulama bulmaktadır. Hatalı yazılmış sözcüklerin bulunması ve düzeltilmesi, doğal dilde cümle ve metin üretmek, diller arası metin çevirisi bunlardan bazılarıdır. Doğal Dil İşlemenin bir diğer önemli özelliği ise Dilbilimine deney ortamı sunmasıdır. Bu sebeple Doğal Dil

İşleme Bilgisayar bilimcilerinin ve Dilbilimcilerinin ortak çalışması gereken bir alandır. Doğal Dil İşleme alanındaki bütün uygulamaların en büyük sıkıntısı dillerdeki karmaşıklık ve belirsizliktir. İnsanlar arasında dahi iletişim güçlüklerine ve yanlış anlaşılmalara yol açan dildeki bu belirsizlikler ve karışıklıklar, bilgisayar ortamında doğal dillerin modellenmesinin önündeki en büyük engeli oluşturmaktadır [1].

BÇ doğal dil işlemenin en popüler ve güncel konularından biridir. Küresel iletişimin önündeki dil engelini ortadan kaldıracak BÇ uygulamaları günümüz dünyasında çok önemli bir ihtiyacı giderecektir. BÇ alanındaki gelişmelere baktığımızda oldukça yol kat edildiği görülmekle beraber genel amaçlı ve yüksek başarımlı BÇ sistemlerine henüz ulaşılamamıştır. BÇ çalışmaları genellikle yaygın dil çiftleri arasında yapıldığı için birçok dil için yeterli çalışma bulunmamaktadır. Son yıllarda ise yeni dil çiftleri üzerinde yapılan çalışmaların yoğunlaştığı söylenebilir.

1.1 Literatür Özeti

BÇ tarihine bakıldığında; ilk dönemi temsil eden 1950-1970 arası yıllarda kural tabanlı sistemler gerçeklenmiştir. 1954'te 6 dilbilgisi kuralı ve 250 sözcük için sözlük girdisi içeren; Georgetown Üniversitesi ve IBM ortaklığıyla gerçekleştirilen 'Georgetown Deneyi' altmıştan fazla Rusça cümlenin İngilizce'ye tam otomatik çevirisini kapsıyordu. Bu deney büyük bir başarıydı ve yarattığı iyimser havayla bilgisayarlı çeviri araştırmasına büyük kaynakların aktarılmasının sebep olmuştur [2]. Sözlük yardımıyla insanlar tarafından oluşturulan çeviri ve dilbilgisi kuralları üzerine inşa edilen bu sistemler ilk çalışmalarda umut vadeden sonuçlar vermiş olsa da 1960'lı yılların sonuna doğru yapılan yatırımların ve çalışmaların büyüklüğü göz önüne alındığında bu yıllarda yapılan çalışmaların sonuçları umut vermekten çok uzaktadır. On yıllık araştırmanın hayalleri gerçekleştirmede başarısız olduğunu ortaya koyan 1966'daki ALPAC (Automatic Language Processing Advisory Committee) Raporu'ndan sonra bu alana yapılan yatırım belirgin ölçüde azaltıldı [3]. Yine de 1980'li yılların sonuna kadar kural tabanlı sistemler geliştirilmeye devam etmiştir. 1990'lı yıllardan itibaren elektronik ortamdaki metinlerin ve kullanılabilirliğinin artmasıyla beraber istatiksel ve örnek tabanlı yöntemler yaygınlaşmıştır [1].

İstatiksel ve örnek tabanlı yöntemler insan çevirilerinden oluşan, birbirinin çevirisi olan metinlerden modeller üretirler. Bu iki dilli metin derlemlerine paralel derlem denir ve istatiksel ve örnek tabanlı BÇ'nin eğitim verisini oluştururlar [4]. Paralel derlemler BÇ'nin yanı sıra sözcük belirsizliği giderme, bilgi erişimi gibi diğer doğal dil işleme alanlarında da kullanılmaktadır. Paralel derlem elde etmek için birbirinin çevirisi olan metinleri elde etmek ve daha sonra cümle seviyesinde hizalamak gereklidir. İnsan eliyle paralel derlem oluşturma işlemi pratik olarak mümkün olmadığından otomatik paralel metin elde etme işlemi özellikle son yıllarda başlı başına bir çalışma alanı olarak görülmektedir. Web'den paralel metin toplama, cümle hizalama, gürültülü paralel kaynaklardan filtreleme ile paralel olanları ayıklama gibi alt çalışma alanlarından oluşmaktadır.

İBÇ paralel derlemden yola çıkarak kaynak dildeki bir metinden hedef dildeki bir metine aktarımı için 'çeviri modeli' denilen olasılık dağılımını üretir [4]. Buna ek olarak İBÇ tek dilli derlemlerden yararlanarak çevirisi yapılan metnin hedef dilde karşılaşılabilen bir metin olma olasılığının dağılımını üretir. Bu model ise dil modeli olarak adlandırılır [4]. Kısaca İBÇ bütün Türkçe cümlelerin, bütün İngilizce cümlelerin çevirisi olduğunu kabul eder ve en yüksek olasılıklı çeviriyi bulmaya çalışır. Fikir olarak 1950'li yıllarda temelleri atılmış olsa da, İBÇ özellikle IBM'in katkılarıyla 90'lı yıllardan itibaren gelişmeye başlamıştır. Son yıllarda İstatiksel Bilgisayarlı Çeviri (İBÇ) olan ilgi hızla artmış ve en fazla uygulanan yöntemlerden biri haline gelmiştir. Tüm bu gelişmelere ve çalışmalara karşın arzulanan, otomatik, aslına uygun ve genel amaçlı olma özelliklerinin tümüne sahip olan başarımı yüksek bir sisteme halen ulaşılamamıştır [1]. Geliştirilen sistemler içerisinde, yapısal olarak birbirine benzeyen dil çiftleri arasında çeviri yapan sistemlerin sonuçlarının, yapısal farklılıklar içeren diğer dil çiftleri arası çevirilerden daha kaliteli olduğu görülmektedir. Örneğin, aynı dil ailesinde sınıflandırılan İngilizce ve Almanca arasında gerçeklenen çeviri sistemlerinin başarısı, farklı dil ailelerinde olan İngilizce ve Japonca arasındaki bilgisayarlı çeviri sistemlerinin başarısından daha yüksektir [1].

Türkçe Ural-Altay dil ailesine ait sondan eklemeli bir dildir. Sözcüğün anlamı İngilizce gibi dillere göre oldukça farklıdır. Sözcükler birçok çekim ve yapım eklerinin kök sözcüğe eklenmesi ile oluşur. Her biçimbirim farklı bir bilgi taşımaktadır. Kök sözcüklere

biçimbirimler eklenerek binlerce yeni sözcük türetilebilir [5]. Türkçe biçimsel olarak İngilizceden oldukça farklı bir yapıdadır. İBÇ sistemleri paralel metinler dışında ekstra bir dil bilgisine başvurmadan etkili sonuçlar üretmektedir ancak, Türkçe-İngilizce, Japonca-İngilizce, Çince-İngilizce gibi birbirine uzak ailelerden olan diller arasındaki başarıları birbirine yakın dillere göre daha düşüktür. Bu sebepten Türkçe-İngilizce İBÇ sistemlerine Türkçe'nin biçimsel özelliklerinin de ilave edilmeyi amaçlayan çalışmalar mevcuttur [5], [6], [7].

İngilizce-Türkçe dil çifti için yapılan çalışmalar 1981 tarihine dayanmaktadır [8]. 1997 yılında Turhan tarafından [9] İngilizce-Türkçe yapısal eşleştirme yapan bir BÇ sistemi geliştirilmiştir. Hakkani vd. [10] 1998 yılında geliştirdikleri kural tabanlı BÇ sisteminde İnterlingua denilen diller arası aktarımda ara bir uluslararası gösterim kullanımını denemişlerdir. Bu yaklaşıma göre her kaynak dildeki ifade biçimbirimsel, sözdizimsel ve anlamsal analizler sonucunda aynı ifadeyi aktaran İnterlingua ifadesine dönüştürülür ve hedef dildeki ifadenin elde edilebilmesi için anlamsal, sözdizimsel ve biçimbirimsel sentezleyiciden geçirilir. İngilizce'den Türkçe'ye BÇ için yapılan bu ilk çalışmalar genellikle kural tabanlı yaklaşımlarda yoğunlaşmışlardır.

2006 yılında El-Kahlout ve Oflazer İngilizce-Türkçe İBÇ sisteminin başarısını artırmak için eğitim verisinin Türkçe tarafında biçimbirimsel analiz yaparak bazı ekleri ayrı yazılmasını denemişler ve yeterince tatmin edici olmasa da iyileştirme sağlamışlardır [5]. Yenitrezi ve Oflazer [7] 2010 yılında İBÇ sistemlerinde yeni bir yaklaşım olan eğitim versindeki sözcüklerin yüzeysel biçimlerinin yanısıra cümle biçimbirimsel analizi ile elde edilen sözcük türü, kökü, aldığı ekler vb. bilgileri de kullanan bir model üzerine çalışmışlar ve göreceli olarak %38 iyileşme sağlamayı başarmışlardır.

Akraba veya yakın diler arasında çeviri amaçlı geliştirilen sistemler, farklılıkların büyük olduğu, Türkçe-İngilizce gibi diller arasında BÇ için gerek duyulan karmaşık yöntem-lere göre, daha basit ve kolay gerçeklenebilir yöntemler kullanmaktadırlar. Birçok yönden benzerlikler gösteren Türk Dil Ailesi için de BÇ çalışmaları yapılmıştır [11], [12], [13], [14]. Hamzaoğlu [11] 1993 yılında Türkçe'den Azerice'ye; 2000 yılında Altıntaş [14] Türkçe'den Kırım Tatarcası'na sözcük bazında işlem yapan kural tabanlı BÇ sistemleri geliştirmişlerdir. Tantuğ vd. tarafından [12] 2008 yılında geliştirilen Türkçe'den

Türkmence'ye BÇ sisteminde Tantuğ [1] tarafından önerilen kural tabanlı ve istatiksel yöntemleri birleştiren "Akraba ve Bitişken Diller Arasında Bilgisayarlı Çeviri İçin Karma Bir Model" kullanılmıştır.

Türkçe dili için yapılan bu BÇ çalışmaları dikkate alındığında başarılı BÇ sistemleri tasarlamak için BÇ yöntemlerinin gelişimi incelenmeli ve ilerleme kaydeden yeni yöntemler kullanılmalıdır. Ancak, Türkçe'nin kendine has özelliklerini de dikkate alan ve güncel İBÇ yöntemlerini Türkçe'ye özgü doğal dil işleme tekniklerinin de kullanımıyla zenginleştiren yöntemlerin daha yüksek başarılara ulaşabileceği görülmektedir.

Bir İBÇ sisteminin başarısını modelin etkisi haricinde, eğitim verisi de oldukça etkilemektedir. İngilizce, Almanca, Fransızca vb. diller arasında yeterli miktarlarda paralel veriler kullanılabilir durumdayken, birçok dil çifti için paralel veri miktarlı kısıtlıdır. İnsan emeği ile paralel derlem oluşturmak çok fazla zaman ve kaynak isteyen güç bir işlem olduğu için otomatik paralel metin toplayan sistemler, paralel metinlerde cümle bazında otomatik hizalama yapan yöntemler üzerine çalışmalar yoğunlaşmıştır. Ayrıca, eğitim verisinin temiz ve kaliteli olması da İBÇ sisteminin performansını oldukça etkilemektedir. Gürültülü paralel derlemlerden gürültüyü ayıklamak veya paralel olmayan çift dilli kaynaklardan paralel cümleler çıkarmak gibi paralel derlem filtreleme çalışmalarına da son dönemlerde ihtiyaç giderek artmaktadır. Yüksek başarılı bir Türkçe'den İngilizce'ye İBÇ sistemi için temiz, kaliteli ve yeterli miktarda paralel derlem kullanılması son derece önemlidir. Bu çalışmanın odaklandığı nokta Türkçe-İngilizce dil çifti için daha başarılı İBÇ sistemlerine ulaşabilmek amacıyla mevcut kullanılabilir paralel derlemleri filtreleyerek kaliteli derlemler elde etmek ve çeşitli kaynaklardan yeni paralel derlemler üretmektir.

Çeşitli dil çiftleri için paralel derlem elde etmek için çeşitli kaynaklar ve yöntemler kullanılmaktadır. Resmi kurumların yayınları [15], dini kitaplar[16], kullanma kılavuzları [17], film alt yazıları [18], farklı dillerde yayımlanmış kitaplar [19], farklı dillerde yayın yapan haber siteleri [20] ve web sayfaları [21] gibi kaynaklardan paralel derlemler elde edilmiştir. Toplanılan paralel metinler doküman olarak ayrılmışsa; paralel derlem oluşturmak için önce doküman seviyesinde hizalama daha sonra da cümle bazında hizalama yapılması gereklidir. Resnik ve Noah'ın çalışmasındaki [21] gibi internet

üzerinde çoklu dil ile yayın yapan siteleri elde etme için arama motorlarında çoklu dil ile yayın yapan siteleri döndürmesi muhtemel olan sorguların sonuçlarındaki sitelerin kontrolü yapılır. Sonraki işlemler ise; web sayfasını html, javascript gibi kodlardan temizlemek, sayfa hizalamak ve son olarak cümle seviyesinde hizalama yapılması gelir.

Paralel derlem filtreleme çalışmaları paralel metin madenciliğinin son işleme adımı olarak düşünülebilir. Gale ve Church [22] 1993 yılında cümle hizalama işlemi için geliştirdikleri programda paralel cümlelerin uzunluklarının oranını ölçmüşlerdir. Uzunluk tabanlı yaklaşımlar Fransızca-İngilizce gibi cümle uzunlukları korelasyonu yüksek dil çiftlerinde oldukça iyi çalışabilirlerken; İngilizce-Çince gibi korelasyonu düşük dil çiftlerinde performans oldukça düşmektedir [23]. Chen ve Nie [24] 2000 yılında İngilizce-Çince paralel metinleri toplamak için bir sistem geliştirmişler ve topladıkları veriyi temizlemek için cümle uzunlukları ve dil belirleme işlemiyle elde ettikleri özelliklerden yararlanmışlardır. Resnik ve Smith [21] da 2003 yılında webden paralel metin toplayan bir sistem üzerine çalışmışlar ve çeviri benzerliği puanı ile topladıkları veriyi temizlemişlerdir. Bu çeviri benzerliği kaynak cümledeki sözcüklerin ne kadarının sözlük karşılıklarının hedef cümlede yer aldığına bakarak elde edilmektedir.

Otomatik olarak üretilmiş derlemlerdeki gürültü, kaynak ve hedef dokümandaki farklılıklardan, aslına uygun olmayan çevirilerden veya cümle hizalama hatalarından kaynaklanabilir. Büyük bir derlemde bu hataları elle gidermek oldukça güçtür. Bu sebeple paralel cümle çiftlerini değerlendirecek otomatik sistemlere ihtiyaç duyulmuştur. Bazı çalışmalar paralel derlemleri temizlemek, daha kaliteli hale getirmek üzerine yoğunlaşırken [25], [26] bazı çalışmalar ise paralel olmayan çok dilli kaynaklardan paralel cümleleri çıkarma amacındadır [27], [28]. Bu çalışmalarda genellikle paralel cümle çiftine ait uzunluk oranları, farkı, sözcük eşleşme oranları, benzerlik hesapları vb. özellikler çıkarılarak makine öğrenmesi tabanlı filtreler geliştirilmiştir.

1.2 Tezin Amacı

İBÇ sistemlerinin başarısını etkileyen en önemli etkenlerden biri paralel derlemdir. Çünkü İBÇ çeviri sırasında kullanacağı parametreleri eğitim sırasında paralel derlemden çıkarmaktadır. Paralel derlemin BÇ açısından kalitesi için şu özellikler sıralanabilir:

- Çeviri denkiliği: İki paralel cümlenin birbirinin çeviri olması ve doğru ve eksiksiz bir şekilde aktarılmış olması
- Cümlelerin dilbilgisi kurallarına uygun olarak kurulmuş olması ve yazım hataları içermemesi
- Dilde görülme olasılığı yüksek akıcı cümleler olması
- Farklı alanlara ait örnekler barındırması; kapsayıcılık ve genel amaçlılık

Bu çalışmada paralel derlem büyüklüğünün İBÇ sisteminin başarısına etkisini araştırmak için farklı kaynaklardan oluşan 1M cümlelik karma bir paralel derlemin farklı boyutlarda alt kümeleri ile İBÇ sistemleri eğitilmiştir. Ve paralel derlemin kalitesinin ve temiz olmasının İBÇ sisteminin başarısına nasıl bir etki yapacağını gözlemlemek amacıyla Türkçe-İngilizce paralel çiftlerinin kalitesinin ölçülmesi konusunda deneyler yapılmış ve paralel derlemden kalitesiz çiftlerin elenmesiyle kaliteli bir paralel derlemin elde edilmesini sağlayan bir sistem üzerine çalışılmıştır. Daha sonra aynı deneyler bu filtrelenmiş temiz derlem üzerinde yapılmış ve paralel derlemin kalitesinin de İBÇ sisteminin başarısına olumlu yöndeki katkıları gözlemlenmiştir. Tüm deneylerden elde edilen sonuçlar Bölüm 4'te ayrıntılarıyla gösterilmiş ve yorumlanmıştır.

Türkçe-İngilizce için kullanıma açık paralel kaynaklar sınırlı da olsa mevcuttur. Ayrıca bu çalışma kapsamında Türkçe için yapılacak çalışmalara kaynak olabilmesi için Türkçe-İngilizce paralel metinler toplanılarak yeni derlemler üretilmiştir ve paralel metin toplama, cümle hizalama gibi konularda var olan yöntemler incelenmiş; Türkçe için en uygun yöntemler deneyler sonucunda belirlenmiş; Türkçe'nin biçimbilimsel yapısı göz önüne alınarak yöntemlerde Türkçeye uygun değişiklikler yapılmıştır. Bölüm 3'de paralel derlem oluşturma, kalitesini ölçme, cümle hizalama gibi konularda bilgi verilmiş; kullanıma açık kaynaklar tanıtılmış ve yeni oluşturulan derlemler tanıtılmıştır.

Yapılan her çalışmanın temel noktası Türkçe-İngilizce dilleri arasında çalışan İBÇ sistemlerinin başarılı sonuçlar üretebilmesi için kaynak oluşturmak ve dolayısıyla Türkçe için BÇ ve doğal dil işleme dünyasına katkı sunmaktır.

1.3 Hipotez

Günümüzde BÇ çalışmaları istatiksel yöntemlerde yoğunlaşmaktadır. Dilden bağımsız, veriye dayalı model olan istatistiksel modellerin başarısı ise eğitim verisinin büyüklüğü ve kalitesiyle doğrudan ilişkilidir.

Bu çalışmada İngilizce'den Türkçe'ye İBÇ sistemi, farklı büyüklüklerde eğitim verisi ile çalışılmış ve paralel derlemin büyüklüğü artıkça aynı oranda sistemin başarısının da arttığı gösterilmiştir. 100 bin paralel cümlelik eğitim verisinden 1 milyon paralel cümleye kadar yapılan testlerde başarıdaki artış aynı hızla devam etmektedir. Bu durum başarılı bir İBÇ sistemi için eğitim verisinin büyüklüğünün mümkün olduğunca fazla olması gerektiğini göstermektedir.

Eğitim verisi yüksek boyutlara ulaştıkça fiziksel kaynakların kısıtlılığı ve çok uzun süren eğitim süreleri ortaya çıkmaktadır. Bu çalışma kapsamında geliştirilen paralel derlem filtresi ile gürültülü bir paralel derlemden hatalı olan paralel cümle çiftlerinin ve dilbilgisel olarak bozukluklar içeren çiftlerin elenmesiyle elde edilen güvenilir bir eğitim verisiyle; daha az kaynak ve zamanla neredeyse aynı sonuçlar elde edilebileceği gösterilmiştir. Paralel derlem filtresi kullanarak paralel derlemdeki gürültü ve hata oranına göre başarıyı yükseltebileceği de görülmektedir. Türkçe-İngilizce dil çifti için paralel metin kaynaklarının kısıtlılığı düşünüldüğünde; film altyazıları, Wikipedia sayfaları; web sayfaları gibi güvenilir olmayan kaynakların İBÇ sistemlerindeki eğitim verisi kaynağı olarak kullanılabilmesi ihtiyacı söz konusudur. Geliştirilen bu paralel cümle çiftlerinin doğruluğunu ve kalitesini ölçen filtre ile bu kaynaklardan Türkçe-İngilizce dil çifti için güvenilir, kaliteli paralel derlemler elde etmek mümkün olacaktır.

Çalışmalar göstermiştir ki; bütün bir derlemin eğitim verisi olarak kullanılmasındansa ayıklayıcı tarafından seçilmiş örneklerin kullanılması daha az veriyle; daha az kaynak ve zamanla daha yüksek başarılar elde etmek mümkündür.

BÖLÜM 2

BİLGİSAYARLI ÇEVİRİ

Bir dilin (kaynak dil) diğer bir dile (hedef dil) otomatik olarak çevrilmesi diğer adıyla Bilgisayarlı Çeviri (BÇ) bilgisayar bilimlerinin ve doğal dil işlemenin çok eskiden bu yana ilgilendiği konulardan biridir. Bu tür bir çalışmanın yapılabilmesi için bilgisayarın her iki dili, dillerdeki eşanlamlı sözcükleri, sözcük öbeklerini ve gramerlerini bilmesi gerekir [5].

Ancak günümüz teknolojisi ve teknikleri ile yetkin bir çeviri sisteminin gerçeklenmesi çok zordur. Yetkin bir bilgisayarlı çeviri sisteminin temelde şu üç özelliği barındırması beklenir:

1. Otomatiklik: İnsan müdahalesine gerek kalmadan sonuç üretebilmeli

2. Kaliteli Çeviri Yapabilme: Sistemin ürettiği çıktılar anlaşılabilir ve asıllarına uygun olmalı

3. Geniş Kapsamlılık: Çeviri sistemi her türlü konuyu içeren genel metinler (makale, haber, hikâye, mektup vs.) üzerinde işlem görebilmeli

Bu üç özellik İngilizcede FAHQT (Fully Automatic - High Quality output -unrestricted Text) olarak geçmektedir [1].

Her ne kadar bilgisayarlı çeviri ile istenilen noktalara ulaşılamamışsa da çeşitli hata oranlarıyla çeviri yapan ve yaygın kullanılan sistemler mevcuttur. Bazı sistemler kapsamı daraltılarak belirli konularda çalışması sağlanmıştır. Bu sistemlere örnek olarak İngilizce-Fransızca arasında hava tahminlerini çeviren Météo sistemi örnek gösterilebilir [1]. Bazı sistemlerde ise otomatiklikten ödün vermişlerdir. Bu sistemler

insan eliyle yapılan çeviri faaliyetinin kolaylaştırılması için tasarlanmış ve sistem çıktıları çevirmenler tarafından düzenlenen sistemlerdir. İnternet ortamından bilgi toplama gibi çeviri kalitesinden ödün verilebilecek uygulama alanları olması sebebiyle bazı sistemler de kaliteden ödün vererek yüzeysel bir çeviri yapmaktadır.

2.1 Doğal Dil İşleme

BÇ Doğal Dil İşleme (DDİ) alanının bir alt dalı olarak görülmektedir. Liddy'e göre [29] DDİ'nin motivasyonu uygulamaların insanlar gibi dil işleyebilmesi amacıyla hesaplama teknikleri ile doğal dillerde yazılmış metinleri bir veya daha fazla dilsel çözümleme seviyelerinde çözümlemesi veya üretmesidir.

DDİ kapsamında aşağıdaki konular üzerine çalışmalar yürütülmektedir

- Yazım yardımcı araçlarının geliştirilmesi
- Yazım yanlışlarının düzeltilmesi
- Bul ve değiştir
- Basılı bir metni okuma (optik olarak metin okuma) ve okuma yanlışlarını düzeltme
- Bir metnin özetini çıkarma
- Metnin içerdiği bilgiyi çıkarma
- Bilgiye erişim
- Metni anlama
- Bilgisayarla sesli etkileşim
- Bilgisayarın konuşması (metni seslendirme)
- Konuşmayı anlama (konuşmayı metne dönüştürme)
- Soru yanıt dizgeleri
- Yabancı dil okuma yardımcı araçları
- Yabancı dilde yazma yardımcı araçları
- Doğal diller arası çeviri

2.1.1 Doğal Dil İşleme Bilgi Seviyeleri

DDİ'nin her seviyesi farklı seviyelerde dilsel olarak anlamlı öğeler üretmekten ve çözümlemekten sorumludurlar. Tüm seviyeleri uygulamak zorunlu değildir fakat daha başarılı aktarımlar, daha derinlemesine çözümlemelerle mümkün olmaktadır [29].

2.1.1.1 Sesbilim

Sesbilim (Phonology) sözcüklerin içerisindeki ve arasındaki sesleri yorumlamaktan sorumlu olan seviyedir. Fonetik (phonetic), fonemik (phonemic) ve prosodik (prosodic) kurallar olmak üzere 3 tipte kuralı yönetir. Fonetik kurallar sözcüklerin içerisinde bir araya gelen sabitleri tanımlayan kurallarken, fonemik kurallar sözcükler bir araya gelince oluşan telaffuz varyasyonlarını tanımlar. Prosodik kurallar ise sözcüklerin yükseltme, alçaltma veya vurgulama biçimlerini tanımlayan kurallardır [29].

2.1.1.2 Biçimbilimsel Seviye

Biçimbilimsel seviye (morphology) sözcükler üzerine yapılan çalışmalardan sorumlu olan seviyedir ve bu seviyede yapılan çözümleme sonucu bir sözcüğün anlamlı en küçük birimleri (morphems) bulunur [29]. Şekil 2.1'de "aklımdan" sözcüğünün biçimbirimsel gösterimi yer almaktadır.

Yapısal Biçim: akıl + AD + TEKİL + 1. TEKİL ŞAHIS İYELİK + YÖNELME HALİ
Yüzeysel Biçim: aklımdan

Şekil 2.1 Türkçe bir sözcüğün biçimbilimsel gösterimi

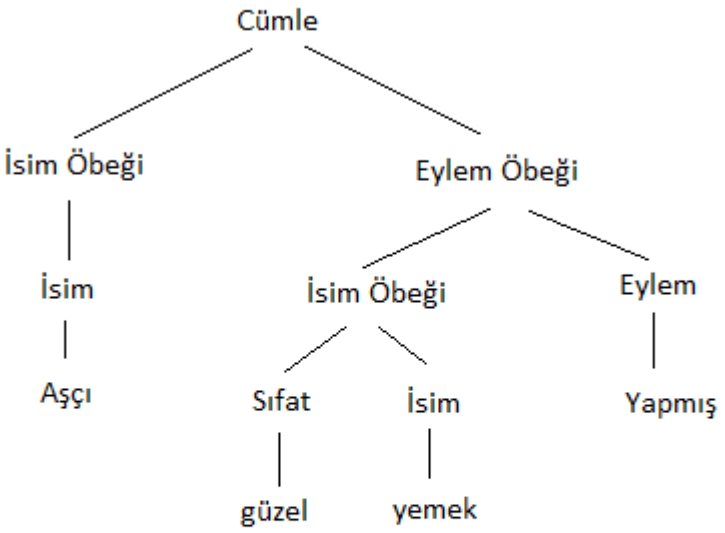
2.1.1.3 Sözlüksel Seviye

Sözcüksel seviye (lexical level) sözcüklerin anlamlarının yorumlanmasından sorumlu olan seviyedir. Sözcükleri en olası anlamını ve cümle içerisindeki görevini (part-of-

speech tags) bularak haritalandırır. Sözlüksel seviyedeki işlemler sözlük gerektirebilmektedir [29].

2.1.1.4 Sözdizimsel Seviye

Sözdizimsel seviye (syntactic) bir cümle içerisindeki sözcüklerin dilbilgisel yapıyı ortaya çıkarmak için nasıl bir araya geldiklerini çalışan seviyedir. Sözcük dizileri dilbilgisel kurallar ve doğal dil sabitleri kullanılarak sözdizimsel ağaçlara aktarılır [29]. Şekil 2.2'de *"Aşçı güzel yemek pişirmiş." Cümlesinin sözdizimsel ağacı gösterilmiştir.*



Şekil 2.2 Türkçe bir cümlenin sözdizim ağacı

2.1.1.5 Anlamsal Seviye

Anlamsal (semantic) çözümleme cümlelerin anlamları üzerine çalışır ve cümle yapısına, diğer bir ifadeyle sözdizimsel ağaca anlamsal görevlerin yüklenmesi işlemini yerine getirir. Birden fazla anlama sahip olan sözcüklerin sebep olduğu belirsizliği gidermek üzerine çalışan anlamsal belirsizlik gidericiler de bu seviyenin parçalarındandır [29].

2.1.1.6 Söylemsel Seviye

Bu seviyede cümleler arasındaki ilişkilere odaklanarak; bir cümlenin anlamının içerisinde yer aldığı metnin içindeki diğer cümleler tarafından etkileşimleri incelenmektedir [29].

2.2 Bilgisayarlı Çeviri Tarihçesi

1930'lu yıllarda başlayan çeviri denemeleri, 1946'da ENIAC (Electronical Numerical Integrator and Calculator) adı verilen elektronik çeviri cihazının İkinci Dünya Savaşı sonrasındaki yeniden yapılanma sürecinde önemli bir rol oynaması, çeviri alanına elektroniğin ve bilgisayarın girmesine öncülük etmiştir. MAT (Machine – Assisted Translation) adı verilen bu sistemler kendi başına çeviri yapamasalar da, bitmiş bir çeviriyi belleğinde bulunan sözcük yapıları ile karşılaştırarak "hata ayıklayabilmektedirler". Yine de çevirmenlerin işini biraz da olsa hafifletmişlerdir [30].

1954'te yapılan 'Georgetown Deneyi' (Georgetown Üniversitesi ve IBM ortaklığıyla gerçekleştirilmiş bu çeviride IBM 701 bilgisayarı kullanılmıştır ve sistemin hafızasında sadece 6 gramer kuralı ve 250 sözcük vardı) altmıştan fazla Rusça cümlenin İngilizce'ye tam otomatik çevirisini kapsıyordu. Bu deney büyük bir başarıydı ve bilgisayarlı çeviri araştırmasına yatırım dönemini başlattı [31]. Yazarlar, üç veya beş yıl içerisinde bilgisayarlı çeviri sorununun çözülebileceğini iddia ediyordu. Ne var ki, gerçek anlamda ilerleme çok daha yavaştı ve on yıllık araştırmanın hayalleri gerçekleştirmede başarısız olduğunu ortaya koyan 1966'daki 'ALPAC (Automatic Language Processing Advisory Committee) Raporu'ndan sonra bu alana yapılan yatırım belirgin ölçüde azaltıldı. Bu rapor, o dönemde, Akademi tarafından oluşturulan, Automatic Language Processing Advisory Committee (ALPAC) kuruluşuna aitti. Bu kuruluşun amacı, BÇ ile ilgili olarak kalite, maliyet, öngörüler, beklentiler ve ihtiyaçlar konusunda çalışmalar yapmaktı. ALPAC raporu, insan çevirmenler konusunda bir yetersizlik olmadığını, genel bilimsel metinlerin çevirisinde kullanılabilecek bir BÇ'nin öngörülmediğini belirtmişti. Bu rapor, ABD hükümetinin BÇ için sağladığı fonun sanal olarak sonu oldu. Daha da kötüsü, sektörde genel bir moral kaybına neden oldu [32].

Yine de, Kanada ve Avrupa'da bu alanda çalışmalara devam edildi. Hava durumu raporlarını İngilizce'den Fransızca'ya çeviren ilk başarılı sistemlerden Meteo 1990'lı yıllara kadar kullanılmıştır. Aynı zamanlarda en popüler ve başarılı kural tabanlı sistemlerden SYSTRAN geliştirilmeye başlanmıştır. SYSTRAN 20 dil arasında doğrudan aktarım modeliyle çalışmaktadır ve Google gibi arama motorlarında; Avrupa Birliği, NATO gibi kurumlarda kullanılmaktadır [33].

Kural tabanlı sistemler özel konularda iyi sonuçlar üretebilmelerine karşın geniş kapsamlı bir çalışma için çok sayıda insan emeğiyle yazılan kurallara ve sözlük kayıtlarına ihtiyaç duymaktadır. Bu da oldukça fazla zaman ve maliyet gerektiren bir durum oluşturmaktadır. Bir dil çifti için yazılan bu kurallar başka dil çiftleri için geçerli olmamakta; genel amaçlı çeviri sistemleri ihtiyacını gideremektedir.

Genel kabul, Bilgi Çağı'nın başlangıcının internetin yaygın olarak kullanılmaya başlandığı 1991 yılı olduğu şeklindedir. Silikon çiplerin çok yüksek miktarda sayısal veriyi muhafaza edebilmeleri, yapay zeka çalışmalarının hızlı gelişimi BÇ'ye de yeni olanaklar sunar [34]. Ses işleme ve tanıma gibi alanlarda başarısı kanıtlanmış olan istatistiksel yöntemlerin BÇ alanında da denenmesi yoluna gidilmiştir. IBM'in önderliğini yaptığı bu çalışmalar sonucunda elde edilen başarılar, kural tabanlı sistemlerde tıkanan ve ileri gidemeyen çalışmaların büyük bir bölümünü derlem tabanlı sistemlere yöneltmiştir. Paralel derlemleri eğitim verisi olarak kullanan ilk yöntem olan örnek tabanlı BÇ yöntemi 1980'li yılların ortalarında öne sürülmüştür. Paralel derlemdeki paralel cümle çiftlerini çeviri örnekleri olarak kullanan örnek tabanlı BÇ yöntemi sözcük ve sözcük öbeği çevirilerini kendisine en çok benzeyen örneklerden çalışma süresi içerisinde öğrenmektedir. En büyük dezavantajı ise çok büyük ve hatasız eğitim verisine ihtiyaç duymaktadır. Son 20 yılda BÇ konusunda en yaygın kullanılan yöntem ise İBÇ (İstatiksel Bilgisayarlı Çeviri) yöntemidir. IBM'in yeni ufuklar açan çalışmaları BÇ'ye ilgi duyan araştırmacıları etkilemiş ve istatiksel yöntemlere yöneltmiştir. Daha az emek ile daha yüksek başarı getirdiği söylenen İBÇ yönteminin ilk yaklaşımı IBM'in sade sözcük tabanlı yöntemidir. SYSTRAN ve IBM'in BÇ sistemi (CANDIDE) arasında yapılan deneylerde istatiksel yöntemlerin kural tabanlı yöntemleri aştığı görülmüştür. Üstelik yeni bir yöntem olan İBÇ'nin diğer konulara ve dil çiftlerine aktarımının oldukça kolay olması BÇ alanındaki ilginin büyük bir kısmını İBÇ'ye

yöneltmiştir. İBÇ sistemlerinin ihtiyaç duyduğu eğitim verisi dilbilgisel olarak iyi ve çeviri açısından da doğru örneklerden oluşan paralel metinlerdir.

2.3 Bilgisayarlı Çeviri Uygulamaları

Giderek yaygınlaşan BÇ sistemleri düz metinleri; dokümanları, elektronik postaları, web sayfalarını, yazılım ara yüzlerini, anlık çeviri için sesleri çevirmek için, çevirmenlere yardımcı olmak için kullanılmaktadır.

BÇ'nin en büyük kurumsal kullanıcısı kurum içi kullanım için belgelerin ilk taslaklarının büyük bölümünün otomatik çevirisini yapmak için ticari amaçlı bilgisayarlı çeviri sistemi olan SYSTRAN'nın üst düzey sürümünü kullanan 'Avrupa Komisyonu'dur [35]. Kural tabanlı olarak geliştirilen ticari bir BÇ sistemi olan SYSTRAN[1] yazılımına akıcılığı ve esnekliği artırmak için istatiksel yöntemler de eklenmiş ve 52 dil arasında çeviri yapabilen bir sistem haline getirilmiştir.

Bir Danimarka çeviri ajansı olan 'Lingtech A/S', 'Trados' ticari CAT araç sistemine dayalı çeviri hafızasıyla birlikte çalışarak 'PaTrans' adlı tecilli, kurallara dayalı bilgisayarlı çeviri sistemini kullanarak 1993'ten beri İngilizce'den Danimarkaca'ya patent uygulamalarını çevirmektedir.

İspanya'nın günlük gazetesi 'Periodico de Catalunya', bir bilgisayarlı çeviri sistemiyle İspanyolca'dan İngilizce'ye çevrilmektedir. 'Google', tecilli istatiksel bilgisayarlı çeviri motoru kullanarak umut verici sonuçların elde edildiğini bildirmiştir. Bu motor halen, yakında daha fazla dil çiftleri SYSTRAN motorundan Google motoruna alınmak üzere, Arapça - İngilizce ve Çince - İngilizce çevirileri için halen Google Çeviri araçlarında kullanılmaktadır. 'Uwe Muegge', İngilizce, Almanca ve Fransızca web sayfalarının tam otomatik, yüksek kalitede bilgisayarlı çeviri yapmak amacıyla Google motoruyla birlikte 'kontrollü bir dil' kullanan bir demo websayfası uygulamaya koymuştur. Son zamanlarda terörizme yoğunlaşarak, Amerika'daki askeri kaynaklar doğal dil mühendisliğine büyük miktarda para yatırmaktadır. 'In-Q-Tel' (özel sektör girişimcileri vasıtasıyla yeni teknolojileri teşvik etmek amacıyla Amerikan İstihbarat Topluluğu

[1] http://www.systransoft.com/

tarafından büyük ölçüde finansmanı sağlanan bir risk sermayesi fonu), 'Language Weaver' gibi şirketler oluşturmuştur. Şu an Amerika'daki askeri kesim Arapça, Paştu ve Dari gibi dillerin çevirisine ve işlemine ilgi duymaktadır. DARPA'daki (Defense Advanced Research Projects Agency) Bilgi İşleme Teknoloji Ofisi, 'TIDES' ve 'Babylon Çeviri' gibi programlara ev sahipliği yapmaktadır. Amerika Hava Kuvvetleri bir dil çeviri teknolojisi geliştirmek için 1 milyon dolarlık bir sözleşme yapmıştır [35].

Google Translate[1] istatiksel bilgisayarlı çevirinin popüler, ücretsiz ve çevrimiçi bir uygulamasıdır. Diğer bir ücretsiz, çevrimiçi istatiksel bilgisayarlı çeviri uygulaması ise Microsoft tarafından geliştirilen Bing[2] uygulamasıdır.

Koehn ve diğerleri [36] kullanıcıların istatiksel bilgisayarlı çeviri uygulamaları geliştirebilmesi için tasarlanmış MOSES[3] isimli için ücretsiz bir araç seti geliştirmişlerdir. Bu çalışmada yapılan deneylerde bu araç setinden faydalanılmıştır.

2.4 Bilgisayarlı Çeviri Yöntemleri

BÇ amacıyla kullanılan yöntemler kural tabanlı, örnek tabanlı ve istatistiksel yöntemler olmak üzere 3 grupta incelenmektedir. Hangi yöntemin kullanılacağı seçilirken üzerine çalışılan dil çifti; BÇ siteminin kapsamı gibi faktörler göz önüne alınmalıdır.

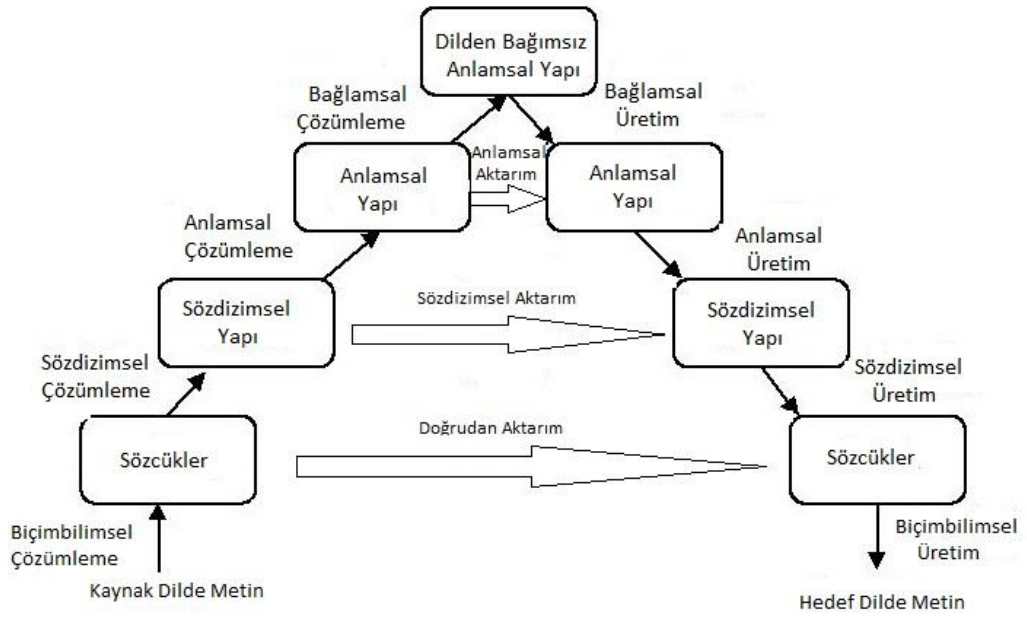
2.4.1 Kural Tabanlı Sistemler

Kural tabanlı çeviri yöntemlerinde, kaynak tümcesinin çeşitli bilgi seviyelerinde gösterimlerini oluşturduktan sonra bu bilgi seviyesinde aktarım yapılmasını öngören bir dizi yöntem kullanır. Bu yöntemleri görselleştirmek için Vauquois Üçgeni yaygın olarak kullanılır (Şekil 2.1) [37].

[1] http://translate.google.com

[2] www.bing.com/translator

[3] www.statmt.org/moses/



Şekil 2.3 Bilgi Tabanlı Yöntemlerin Sınıflandırılması-Vauquois Üçgeni

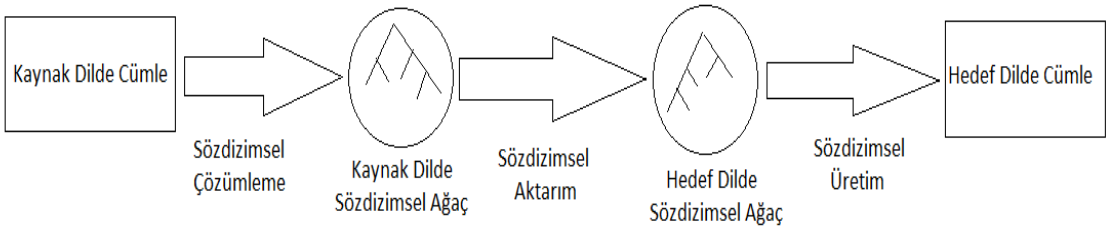
2.4.1.1 Doğrudan Aktarım

Vauquois Üçgeninin en altındaki aktarım en temel çeviri türüdür. Kaynak dildeki sözcüklerin karşılıklarının bulunarak hedef dile çevrilmesidir. Bu basit çeviri türünde bile tam sözcük karşılığının bulunmaması, farklı anlamlar içeren sözcüğün hangi anlamda kullanıldığı gibi sıkıntılar ortaya çıkmaktadır. Dolayısıyla sözcüklerdeki bu belirsizliklerin giderilmesi gerekmektedir.

Her ne kadar doğrudan aktarım yönteminde tümce üzerinde çözümleme yapılması gerekmese de birçok uygulamada biçimbilimsel çözümleme yapılır [1].

2.4.1.2 Sözdizimsel Aktarım

Sözdizim aşamasında biçimbilimsel çözümleyicide ayrıştırılan sözcükler kullanılarak cümledeki öğelerin (isim, sıfat, zarf, ...) dizimsel formülleri oluşturulur [38]. Sözdizimsel çözümlemesi yapılan kaynak dildeki ifadenin sözcükler arası bağlantıları gösteren ağaç verisi elde edilir ve hedef dile bu ağaç aktarılır. Sözdizimsel yapı aktarıldıktan sonra, doğrudan aktarım yönteminde olduğu gibi sözcükler de aktarılır (Şekil 2.4).



Şekil 2.4 Sözdizimsel Aktarım

2.4.1.3 Anlamsal Aktarım

Anlambilim aşamasında, sözdizimsel çözümleme sonucu üretilen sözdizim ağacına anlamsal görevler de yüklenerek anlamsal gösterim oluşturulur. Anlamsal gösterim kaynak dilden hedef dile aktarılarak; anlamsal üretim; sözdizimsel ve biçimbilimsel üretim gerçekleştirilerek hedef dilde cümle elde edilir.

2.4.1.4 Dilden Bağımsız Anlamsal Gösterimin Aktarımı

Anlambilim aşamasında, doğal dillerde kullanılan cümleler "İnterlingua" adı da verilen, cümlenin anlamını dilden bağımsız bir yapıda ifade eden "diller arası" bir formata çevrilir. Bu sayede cümlelerin diğer dillere ya da makine diline çevrilebilmesi mümkün olur. En önemli özelliği; her dil için geliştirilen DDİ araçları ile o dilden bağımsız anlamsal gösterimi oluşturmak ve bu gösterimden ifadeyi üretme işlemi gerçekleştirildiğinde BÇ için ayrı bir çabaya gerek duyulmayacak olmasıdır. Yani, her dil çifti için ayrı ayrı çalışmak yerine; dilleri bilgi seviyelerinde çözümlemesini ve üretimini gerçekleştirecek araçlara sahip olmak yeterli olacaktır. Fakat diller arası var olan büyük farklılıklar dilden bağımsız gösterimin nasıl sağlanabileceği konusu henüz üzerinde anlaşmaya varılmış bir problem değildir.

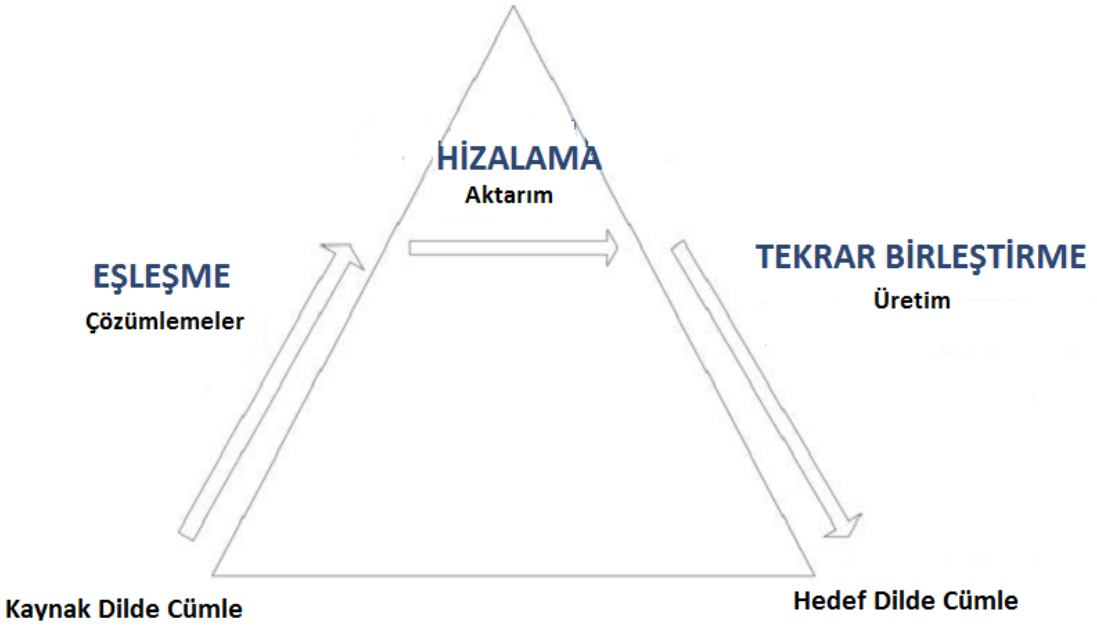
2.4.2 Derlem Tabanlı Yöntemler

90'lı yılların başlarında, bilgisayar teknolojilerindeki gelişme ve büyük miktardaki çevrimiçi metinlere ulaşmanın kolaylaşması sonucu derlem tabanlı yöntemler

gelişmiştir. Derlem tabanlı yaklaşımlar iki alt alandaki çalışmalara yoğunlaşmıştır: Örnek tabanlı Bilgisayarlı Çeviri (ÖT BÇ) ve istatiksel bilgisayarlı çeviri (İBÇ).

2.4.2.1 Örnek Tabanlı Yöntemler

Örnek Tabanlı Bilgisayarlı Çeviri (ÖTBÇ) veya 'analojiyle çeviri' öbeklerin çıkarılması ve birleştirilmesi temeline dayanmaktadır. Nagao'ya göre [4] insanlar bir cümleyi çevirmek için derinlemesine dilsel analizler yapmazlar; cümleyi alt ifade parçalarına ayrıştırır ve bu parçaları çevirip hedef dilde bir cümlede birleştirmeye çalışır. Her bir parçanın çevirisi uygun örneklere başvurarak yapılmaktadır. Nagao ÖTBÇ'nin 3 görevi olduğunu düşünmektedir: eşleştirme (matching), hizalama (aligning) ve tekrar birleştirme (recombination) [39]. Şekil 2.5'te bu görevler, geleneksel BÇ'de (kural tabanlı yöntemler) denk geldiği görevlerle görselleştirilmiştir.



Şekil 2.5 ÖTBÇ'ye uyarlanmış Vauquois üçgeni

Eşleşme görevinde basit karakter eşleştirme algoritmaları veya daha gelişmiş dilsel özellikler kullanılarak giriş cümlesindeki parçalara en çok benzeyen parçaları içeren cümleler bulunmaktadır. İlk önce bütün cümle ele alınarak benzeyen cümle bulunmaya çalışılırken; eşleştirilemeyen parçalar için daha küçük parçalara yönelinir. En küçük parçalar ise sözcüklerdir.

Hizalama görevinde ise eşleşen parçaların hedef cümledeki çevirileri elde edilir. Bu hizalama benzer cümlelerin hepsine bakılarak elde edildiği gibi sözlükten de faydalanılabilmektedir [40].

Yeniden birleştirme görevi ise hizalama sonrası hedef dile çevrilen parçaların birbirine bağlanması işlemini içerir. Bu aşamanın zorluğu hedef dilin dilbilgisel yapısının karmaşıklığıyla ilişkilidir [41].

Chunyu ve diğerlerine göre [42] ÖTBÇ'de 4 basamak vardır: örnek edinme (example acquisition), örnek tabanı yönetimi (example base management), örnek uygulama (example application) ve hedef cümle üretimi (target sentence synthesis).

Örnek edinme, paralel çok dilli derlemden örneklerin elde edilmesi işlemidir. Çok dilli metinlerde çeşitli seviyelerde hizalama yapmak gerekli bir aşamadır. İnsan emeği ile yapılan hizalama her ne kadar güvenilir örnekler elde etmek için bir çözüm gibi görülse de, maliyet ve zaman kaybı göz önüne alındığında hizalama işlemleri otomatik metin hizalama teknolojileri öncülüğünde yapılmaktadır [42].

Örnek tabanı yönetimi ise örnekleri depolama, yönetme (ekleme, silme, değiştirme vb.) ve hızlı erişim işlemlerinin yürütüldüğü basamaktır. Çok sayıda örneği hızlı işleyebilme görevini yürütür [42].

Örnek uygulama basamağı ise çeviri işlemini gerçekleştirebilmek için örneklerin kullanımı işlemlerini kapsar. Kaynak cümlenin parçalara ayrılması, örnekler arasında görülen parçaların hedef dildeki çevirilerinin elde edilmesi işlemlerini kapsar [42].

Hedef cümle üretimi hedef dile çevrilen parçaların birleştirilmesi işlemidir. Dilbilgisi kurallarına uygun, kolay okunabilen, akıcı cümleler üretmeye çalışır [42].

ÖTBÇ'nin verimliliğini artırmak amacıyla örneklerde genellemelere gidilerek İngilizce – Fransızca ve İngilizce – İspanyolca dilleri arasında çeviri sistemleri oluşturulmuştur [43]. Örneklerin içerisindeki bazı sözcükler (özel isim, yer ismi, tarih vb.) etiketlenmiş ve böylelikle benzer cümleleri bulma aşamasında daha genel örneklerin kullanımı ile ÖTBÇ performansı iyileştirilmeye çalışılmıştır.

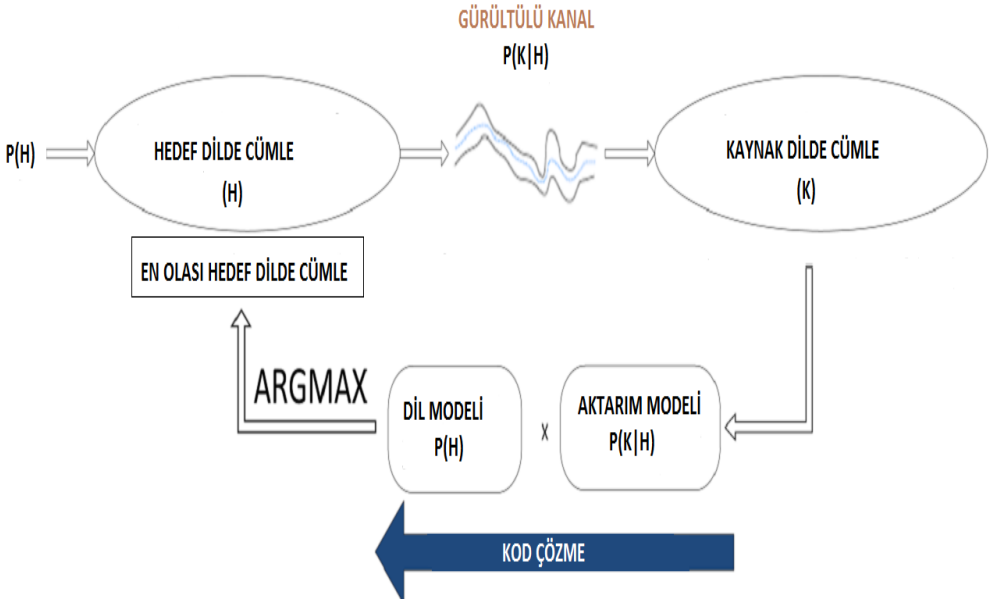
ÖTBÇ'yi kullanan bir diğer çeviri sistemi ise Mayor ve diğerleri tarafından 2006 yılında Bask Dilleri için geliştirilmiştir [44].

2.4.2.2 İstatiksel Yöntemler

BÇ konusunda geçtiğimiz son 20 yılın en popüler yöntemi IBM'in çığır açan çalışmaları ile başlayan İstatiksel Bilgisayarlı Çeviri (İBÇ) dir [45]. Bu yöntemin verimliliği BÇ ile çalışan araştırmacılar üzerinde büyük bir etki yaratmış; yoğun emek gerektiren aktarım ve üretim kuralları; otomatik, hızlı ve kolay eklemlenebilen istatiksel yöntemlerle yer değiştirmeye başlamıştır. İBÇ yöntemleri daha az çabayla kendinden önceki yaklaşımlara göre daha başarılı sonuçlar üretmiştir [33]. Bilgisayarlı çeviri (BÇ) alanındaki çalışmaların istatistiksel yaklaşıma dönüşümü IBM' in CANDIDE sisteminin temel kural-tabanlı yaklaşıma olan üstünlüğünün kanıtlanması ile başlamaktadır. Hesaplama gücünün ve buna bağlı olarak paralel dil verisine ulaşımın kolaylaşması araştırmacıların bu alana olan eğilimlerine destek vermiştir [6].

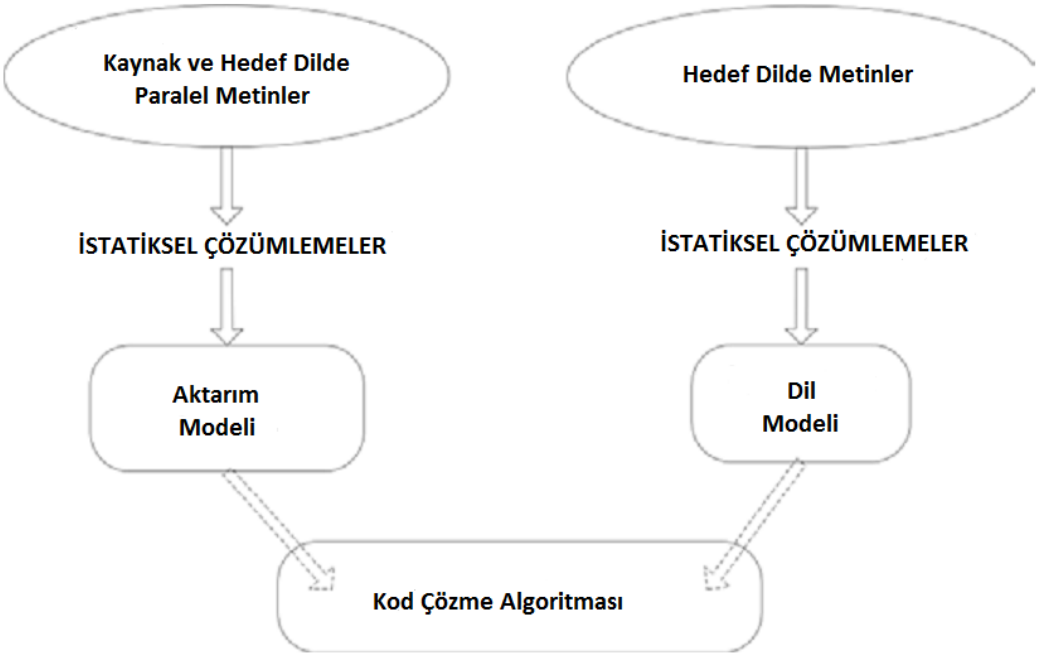
İstatiksel Bilgisayarlı Çeviride (İBÇ) birbirinin çevirisi olan metinlerden oluşan büyük verilerden istatiksel bilgilerin çıkarımından faydalanılmaktadır. İstatistiksel çeviri yöntemlerinin fikirleri 1950'li yıllarda ortaya atılmış olsa da gerçek anlamda ilk olarak IBM tarafından konuşma işleme, doğal dil işleme gibi alanlardaki istatiksel yöntemlerin başarılı uygulamalarından esinlenerek yayınlanmıştır [4].

Shannon tarafından [46] tanıtılan gürültü kanalı modeli İBÇ'yi tanımlamak için oldukça yaygın kullanılmaktadır [47]. Gürültülü kanal modelinde kaynak dildeki K cümlesi hedef dildeki H cümlesinin gürültülü iletişim kanalından geçerken bozulmuş hali olarak düşünülmektedir. İBÇ sisteminin amacı istatiksel yöntemleri kullanarak kaynak dilde verilen K cümlesinin hedef dildeki eşdeğer H cümlesini üretmektir. Kaynak dil ve hedef dil arasındaki İBÇ için gürültülü kanal modeli Şekil 2.6'da gösterilmektedir.



Şekil 2.6 İBÇ için gürültülü kanal modeli

İBÇ iki adet hesaplamalı görevi kapsar: dil modeli olasılıkları ve aktarım modeli olasılıkları. Birbirinin çevirisi olan paralel metinlerden oluşan eğitim verisinden arama ve kod çözme algoritmaları yardımıyla oluşturulan bu olasılıksal modeller İBÇ'nin temel bileşenleridir. Şekil 2.7'de bu bileşenlerin daha iyi anlaşılabilmesi için açıklamalar verilmiştir.



Şekil 2.7 İBÇ'nin bileşenleri

2.4.2.2.1 Dil Modeli

İstatistiksel dil modelleri, metin içinde bir cümlenin veya bir cümle içinde sözcüklerin yer alma olasılıklarını hesaplamada sıkça kullanılır [48]. Cümlelerin sözcük dizilerinden oluştuğu varsayılırsa bu olasılık şöyle yazılabilir:

$$P(C) = P(s_1, s_2, ..., s_n) = \prod_{i=1}^{n} P(s_i | s_1, ..., s_{i-1}) \quad \textbf{(2.1)}$$

Denklem (2.1)'de ***C*** cümleyi, s_i ise sözcükleri göstermektedir. N-gram modelleri denklemin sağındaki olasılıkları sadece geçmişteki N-1 terimi kullanacak şekilde yakınsar.

$$P(C) \approx \prod_{i=1}^{n} P(s_i | s_{i-N+1}, ..., s_{i-1}) \quad \textbf{(2.2)}$$

Yani N-gram modelleri denklem (2.1)'deki dil modeli denklemini denklem (2.2) haline getirir. Her bir terimin hesaplanabilmesi için en iyi olabilirlik (maximum likelihood) kuralı ile eğitim için toplanmış metin verisi kullanılarak her bir N-gram'ların görülme olasılıkları hesaplanır. Eğitim verisinde görülmeyen N-gram'ların sıfır olasılık almamaları için ise bir takım yumuşatma algoritmaları geliştirilmiştir [49].

2.4.2.2.2 Aktarım Modeli

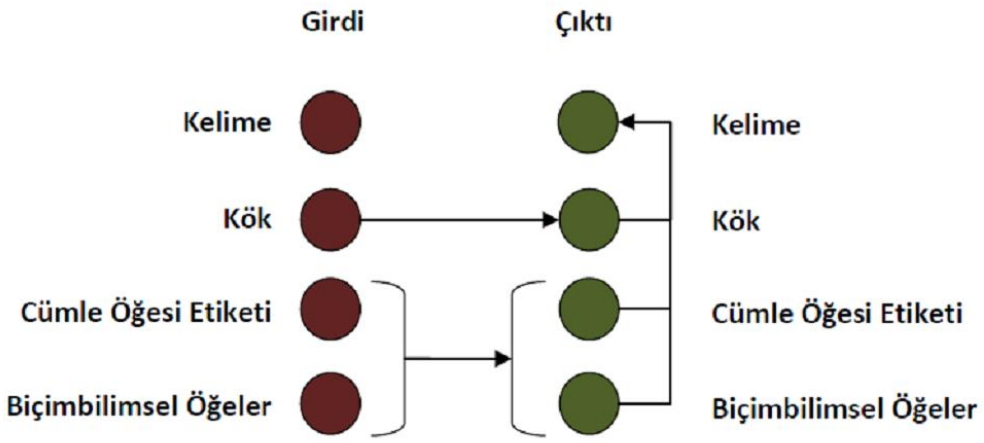
Şekil 2.6'da İBÇ sistemine uyarlanan gürültülü kanal modelinde P(k|h) ifadesiyle gösterilen model aktarım modelidir ve çeviride doğruluğu (faithfullness) temsil eder. Eğitim verisindeki paralel cümleleri kullanarak kaynak cümledeki sözcük veya sözcük öbeklerinin hedef dildeki hangi sözcük veya sözcük öbeklerinin hangi olasılıkla karşılığı olduğu bilgisini üretir. Bu olasılıkların elde edilmesi için paralel derlem içerisinde ilk önce cümle hizalama daha sonra sözcük ve sözcük öbeği hizalama işlemleri yapılır.

İBÇ aktarım modelleri sözcük tabanlı aktarım modeli, sözcük öbeği tabanlı aktarım modeli ve sözdizimi tabanlı aktarım modeli olmak üzere üç farklı grupta incelenir.

Sözcük tabanlı aktarım modelleri İBÇ için geliştirilen orijinal modellerdir. Bu modellerde çeviri işlemi her bir sözcüğün ayrı çevrilmesine bağlanmıştır. Sözcük tabanlı ilk aktarım modelleri farklı hizalama ve parametreler kullanan 5 ayrı sürümü bulunan IBM modelleridir. IBM Model 1'de tüm hizalamalar aynı olasılıkla başlatılırken; IBM Model 2 'de sıfırıncı mertebe (zero-order) hizalama modeli; IBM Model 3'te ters sıfırıncı mertebe hizalama modeli ve kullanılan sözcüğün hizalandığı sözcük sayısını gösteren doğurganlık (fertility) modeli kullanılmaktadır. IBM Model 4'te ise ters birinci mertebe (first-order) hizalama modeli ve doğurganlık modeli kullanılırken; IBM Model 5'te ise IBM Model 4'ün bazı eksiklerini gideren yeniden düzenlenmesi ile oluşturulan model kullanılmaktadır [50]. GIZA++ IBM modellerinin İBÇ sistemlerinin eğitimleri sırasında yaygın bir şekilde kullanılan uygulamasıdır.

Sözcük öbeği tabanlı aktarım modellerinin temel çeviri birimleri sözcükler yerine sözcük öbekleri, yani devamlılık gösteren sözcük dizileridir. Kaynak cümle sözcük öbeklerine ayrılır ve çevrilebilen her bir sözcük öbeği hedef dile aktarılır. Çevrilen öbeklerin sıralaması üzerinde işlemler yapılarak son çıktı üretilir. Eğitim sonrasında üretilen tüm sözcük öbekleri ve çevirileri bir tabloda olasılık değerleri ile birlikte saklanır [51].

Sözdizimi tabanlı İBÇ modelleri kaynak ve hedef dildeki cümlelerin dilbilgisel biçimlerinin de eşleşmesi üzerine kurulmuştur. Diller arasındaki sözcük sıralaması gibi dilbilgisel farklılıkları da içeren modellerdir [52], [53]. Literatürde faktörlü aktarım modelleri olarak geçen bu modeller hem biçimbilimsel ve sözdizimsel öğelerin hem de sözcük köklerinin ayrı olarak eşleştirilmesini gerektirmektedir. Şekil 2.8'de faktörlü İBÇ yaklaşımı gösterilmektedir.



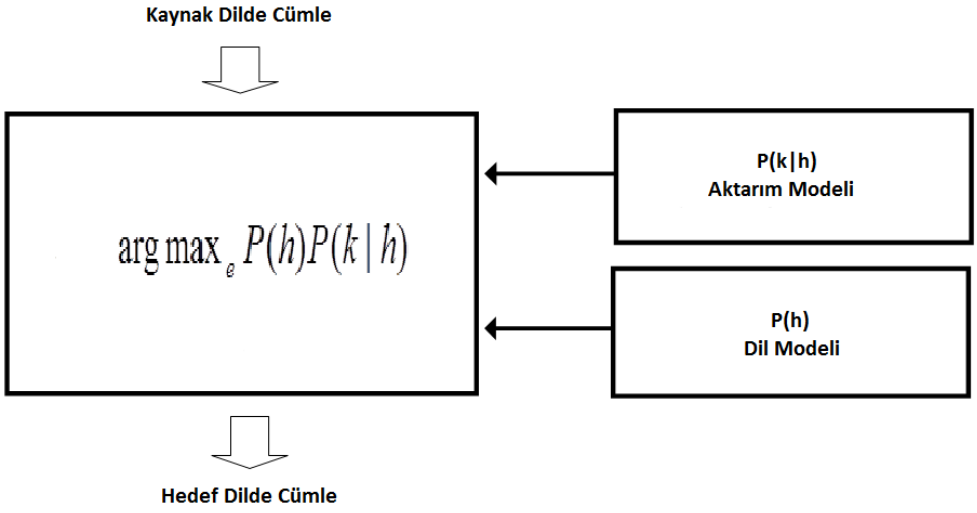
Şekil 2.8 Faktörlü İBÇ yaklaşımı

2.4.2.2.3 Arama veya Kod Çözme Algoritması

Arama veya kod çözme algoritması bütün bu olasılık değerlerini kullanarak kaynak cümlenin en yüksek olasılıklı çevirisi olan hedef cümleyi bulmak için kullanılan algoritmadır. En yüksek olasılıklı cümle Bayes teoremini kullanan denklem (2.3)'teki eşitlik ile bulunmaktadır.

$$e^* = \arg\max_e P(h)P(k \mid h) \qquad \textbf{(2.3)}$$

En olası h cümlesinin nasıl hesaplanacağı bilinse de pratik olarak imkânsızlığından dolayı demetli arama, A* gibi arama yöntemleri kullanılabilmektedir. Şekil 2.9'da İBÇ modellerinin genel mimarisi verilmiştir. P(k|h) kaynak ve hedef dilleri arasında aktarım modelini, P(h) ise hedef dilde dil modelini göstermektedir.



Şekil 2.9 İBÇ modellerinin mimarisi

2.4.3 Melez Yöntemler

Farklı BÇ yaklaşımlarını bir araya getiren bu sistemlerin amacı kural tabanlı sistemlerin çıktılarını istatiksel yaklaşımlarla güçlendirmek veya derlem tabanlı çeviri sistemlerine kılavuzluk etmesi amacıyla kurallar tanımlamaktır. Kural tabanlı bir BÇ sistemi olan ve sonradan akıcılık ve esnekliği artırmak amacıyla istatiksel yöntemler de eklemlenen SYSTRAN yazılımı melez yöntemlere örnek olarak gösterilebilir.

Tantuğ'un geliştirdiği [1] Türkçe ve akraba dilleri arasında çalışan BÇ sistemi de yine sözcükleri istatiksel modellere göre aktaran; dilbilimsel aktarımı ise kural tabanlı yapan melez yöntemlere örnek olarak verilebilir.

2.5 Akraba Diller Arasında Bilgisayarlı Çeviri

Akraba diller arasındaki yapısal benzerlikler yardımı ile bu diller arasında bilgisayarlı çevirinin gerçeklenmesi, farklı dil aileleri arasında çeviri yapmaktan, en azından sezgisel olarak, daha kolay görünmektedir. Çoğu zaman anlamsal çözümlemelere gerek duyulmamaktadır. Aktarım kuralları birbirine uzak olan dillerdeki kurallara göre çok daha az sayıda olabilmektedir. Birbirine akraba olan dil çiftlerinde genellikle biçimbilimsel çözümleme ve sonrasında doğrudan aktarım iyi sonuçlar vermektedir.

Türkçe ve Kırım Tatarcası arasında [14], Çekçe ve Polonyaca [54], İspanyolca ve Bask dilleri [44] gibi birbirine yakın olan diller arasında BÇ çalışmaları yapılmıştır. Bu çalışmaların ortak özelliği birbirine yakın olan iki dil arasındaki dilbilgisel farklılıkların ortaya çıkartılarak aktarım kurallarının yazılması ve istatistiksel yöntemlerden faydalanılmasıdır.

Tantuğ'un 2007 yılında yapmış olduğu çalışmada [1] önerilen yaklaşım, akraba diller örneğin Türk dilleri, arasında çeviri söz konusu olduğunda, olasılık dağılımı esasına göre çalışan "çeviri bileşeninin" kural tabanlı çalışan "aktarım fonksiyonu" ile değiştirilerek istatistiksel dil modeli ile beraber kullanılması yönündedir.

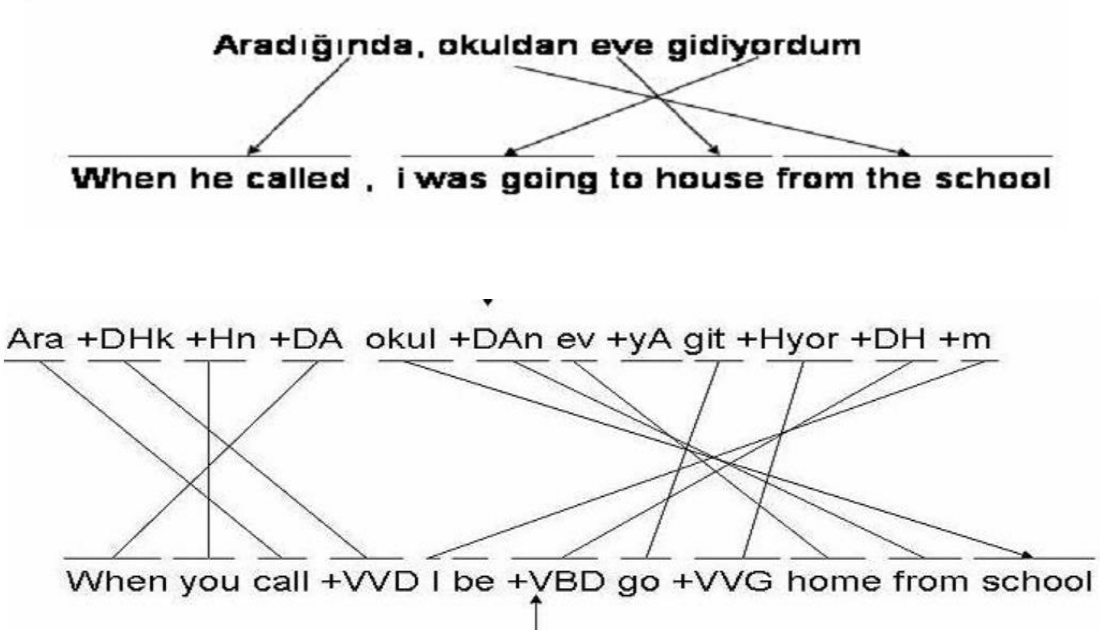
2.6 İngilizceden Türkçeye Bilgisayarlı Çeviri

Türkçe Ural-Altay dil ailesine ait sondan eklemeli bir dildir. Sözcüğün anlamı İngilizce gibi dillere göre oldukça farklıdır. Sözcükler birçok çekim ve yapım eklerinin kök sözcüğe eklenmesi ile oluşur. Her biçimbirim farklı bir bilgi taşımaktadır. Kök sözcüklere biçimbirimler eklenerek binlerce yeni sözcük türetilebilir [5]. Örneğin, *"yapmak"* eylem kökünden gelen *"yaptıklarımınzdaki"* sözcüğü ayrıştırıldığında *"yap+tık+lar+ımız+da+ki"* olarak ayrıştırılır. Türkçe ses olaylarını da hesaba kattığımızda *"yaptıklarımınzdaki"* sözcüğünün sözcüksel biçimi *"yap +DHk +lAr +HmHz +dA +ki"* olarak gösterilir. Bu sözcüğü İngilizce ifade etmek istediğimizde tek bir sözcükle ifade etmek mümkün değildir.

İngilizce-Türkçe dil çifti için yapılan çalışmalar 1981 tarihine dayanmaktadır [8]. 1997 yılında Turhan tarafından [9] İngilizce-Türkçe yapısal eşleştirme yapan bir BÇ sistemi geliştirilmiştir. Hakkani vd. [10] 1998 yılında geliştirdikleri kural tabanlı BÇ sisteminde İnterlingua denilen diller arası aktarımda ara bir uluslararası gösterimi kullanımını denemişlerdir. Bu yaklaşıma göre her kaynak dildeki ifade biçimbirimsel, sözdizimsel ve anlamsal analizler sonucunda aynı ifadeyi aktaran İnterlingua ifadesine dönüştürülür ve hedef dildeki ifadenin elde edilebilmesi için anlamsal, sözdizimsel ve biçimbirimsel sentezleyiciden geçirilir. İngilizce'den Türkçe'ye BÇ için yapılan bu ilk çalışmalar genellikle kural tabanlı yaklaşımlarda yoğunlaşmışlardır.

İBÇ sistemleri paralel metinler dışında ekstra bir dil bilgisine başvurmadan etkili sonuçlar üretmektedir ancak Türkçe-İngilizce, Japonca-İngilizce, Çince-İngilizce gibi birbirine uzak ailelerden olan diller arasındaki başarıları birbirine yakın dillere göre daha düşüktür. Bu sebepten Türkçe-İngilizce İBÇ sistemlerine Türkçenin biçimsel özelliklerini de ilave etmeyi amaçlayan çalışmalar mevcuttur [5], [7], [6].

2006 yılında El-Kahlout ve Oflazer İngilizce-Türkçe İBÇ sisteminin başarısını artırmak için eğitim verisinin Türkçe tarafında biçimbirimsel analiz yaparak bazı ekleri ayrı yazılmasını denemişler ve yeterince tatmin edici olmasa da iyileştirme sağlamışlardır [5]. Şekil 2.10'da "Aradığında okuldan eve gidiyordum" Türkçe cümlesi ile "When he called, I was going to house from school" İngilizce cümlesinin yüzeysel biçimdeki hizalaması ve El-Kahlout ve Oflazer 'in çalışmasındaki gibi ayrıştırılmış ekler ile hizalama gösterilmektedir.



Şekil 2.10 Yüzeysel biçimde hizalama ve ayrıştırılmış ekler ile hizalama [5]

Yenitrezi ve Oflazer [7] 2010 yılında İBÇ sistemlerinde yeni bir yaklaşım olan eğitim verisindeki sözcükleri biçimbirimsel çözümleme ve sözdizimsel çözümleme ile elde edilen sözcük türü, kökü, aldığı ekler vb. bilgileri de kullanan faktörlü İBÇ modellerini

[55] kullanarak Türkçe – İngilizce dilleri arasında çeviri üzerine çalışmışlar ve göreceli olarak %38 iyileşme sağlamayı başarmışlardır. Faktörlü aktarım modelleri hem biçimbilimsel ve sözdizimsel öğelerin hem de sözcük köklerinin ayrı olarak eşleştirilmesini gerektirmektedir. Görgün ve Yıldız [6] da 2012 yılında Türkçe – İngilizce dil çifti için faktörlü modeller üzerine çalışmalar yapmışlardır.

2.7 Çeviri Kalitesinin Değerlendirilmesi

Bilgisayarlı çeviri için geliştirilen yöntemlerin, yöntemler üzerinde yapılan değişikliklerin sonuçlarının değerlendirilebilmesi için BÇ sistemlerinin çıktılarının başarısını ölçmek gerekmektedir.

2.7.1 İnsanlar Tarafından Puanlama

Çeviri kalitesinin değerlendirilmesinde en etkili yol olan her iki dile hâkim olan insanlar tarafından BÇ sisteminin yapmış olduğu çevirilerin doğruluk, akıcılık gibi kriterlere göre oylanmasıdır. Bazı değerlendirme sistemleri ise BÇ sistemlerinin çıktılarının çevirmenler tarafından düzenlenirken gereken çabayı ölçüt olarak almaktadır [1].

2.7.2 Çeviri Kalitesini Otomatik Değerlendiren Yöntemler

İnsanlar tarafından puanlama yöntemleri ile çeviri kalitesinin değerlendirilmesi BÇ sistemlerinin sürekli geliştirildiği ve yeni yaklaşımların denendiği göz önüne alındığında oldukça maliyetli ve yavaş olmaktadır. Bu sebepten çeviri kalitesini değerlendiren otomatik yöntemler geliştirilmiştir.

2.7.2.1 BLEU / NIST

BLEU IBM tarafından 2002 yılında geliştirilen aynı kaynak cümlelerin BÇ sistemi çıktıları ile çevirmenler tarafından çevrilen referans çevrileri arasındaki benzerliği ölçerek çeviri sisteminin kalitesini ölçen bir yöntemdir [56]. Sistem çıktısındaki sözcükler ve ikili, üçlü ve dörtlü sözcük gruplarının eşleştirilmesiyle benzerlik ölçülür. Dörtten uzun sözcük dizilerinin eşleştirilmesine ise gerek duyulmadığı gözlemlenmiştir.

BLEU metodunda her bir cümle için kesinliği (precision) ifade eden değiştirilmiş tutturma (modified precision) değeri hesaplanır. Değiştirilmiş tutturma değeri (p_n) her

n-gram mertebesi için denklem (2.4)'teki gibi hesaplanır. Değiştirilmiş tutturma değeri *C* derlemindeki her *S* aday cümlesinde ve referans cümlede geçen her sözcük veya sözcük grubunun sayısının toplamının; toplam sözcük ve sözcük grubuna oranı ile bulunmaktadır.

$$p(n) = \frac{\sum_{S \in C} \sum_{(n-gram) \in S} A\det_{eşleşe}(n-gram)}{\sum_{S \in C} \sum_{(n-gram) \in S} A\det(n-gram)} \quad \textbf{(2.4)}$$

Her n-gram için (1'den 4'e kadar) hesaplanan değiştirilmiş tutturma değerlerini birleştirmek için geometrik ortalamaya denk gelen tek biçimli (uniform) ağırlıklar ile ortalama logaritma kullanılır.

BLEU bulma (recall) değerinin hesaplanmasındaki zorluklar (uzunlukla alakalı zorluklar, birden fazla referans metin olması vb.) sebebiyle tutturma (precision) ölçütüne göre çalışan bir metottur. Bu nedenle, referans çevirilerden çok daha kısa bir aday çevirinin, yüksek tutturma değeri sayesinde yüksek BLEU puanları almasını engellemek amacıyla denklem (2.5)'teki gibi hesaplanan bir kısalık cezası (Brevity Penalty) tanımlanmıştır.

$$KC = \begin{cases} 1 & eger \quad c > r \\ e^{1-r/c} & eger \quad c \le r \end{cases} \quad \textbf{(2.5)}$$

Bu denklemde *KC* kısalık cezasını, *c* derlemdeki aday çevirilerin tamamının toplam uzunluğunu, *r* ise etkin referans uzunluğunu göstermektedir. Etkin referans uzunluğu, referans cümleler derlemi içerisinde, kendi aday cümlesinin uzunluğuna en yakın olan referansların uzunlukları toplamıdır.

Değiştirilmiş tutturma değeri ve kısalık cezası hesaplandıktan sonra son olarak denklem (2.6)'daki gibi BLEU puanı hesaplanır.

$$BLEU = KC \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad \textbf{(2.6)}$$

BLEU puanı 1 ile 0 aralığındadır. BLEU puanı yükseldikçe BÇ sisteminin başarısının arttığı düşünülmektedir. BLEU yöntemi çeviri kalitesinin ölçümünde hatasız bir çözüm olmasa da günümüzde otomatik ve hızlı olması sebebiyle yaygın olarak kullanılmaktadır.

NIST metodu BLUE metoduna oldukça benzemektedir. BLEU her n-gram için ayrı ayrı hesaplanan değiştirilmiş tutturma değerlerini birleştirmek için geometrik ortalama kullanırken, NIST aritmetik ortalamayı kullanmaktadır. Bir diğer farklılık da NIST yönteminde değiştirilmiş tutturma değerlerinin hesabında n-gram'ların sıklıklarına bakarak daha az sıklığa sahip olan n-gram'lara daha çok önem vermesidir.

Tantuğ ve diğerleri tarafından 2008 yılında geliştirilen BLEU+ aracı ile [57] BLEU yönteminin sondan eklemeli bir dil olan Türkçe için özelleştirilmiştir. Bu araçta aday ve referans metindeki sözcüklerin biçimbilimsel çözümlemeleri yapılarak eşleşmeler için sözcüklerin köklerine ve biçimbilimsel özelliklerine bakılmaktadır.

2.7.2.2 F Ölçütü

Tutturma (precision) ve bulma (recall) ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli olmadığı için her iki ölçütü beraber değerlendiren F-ölçütü tanımlanmıştır. BÇ sistemlerinin başarısının ölçümünde bu yöntem aday cümle ile referans cümle arasında en uzun sözcük dizisi eşleşmeyi bulur. En uzun eşleşmeyi aday cümlesindeki sözcük sayısına bölerek tutturma değerini; referans cümledeki sözcük sayısına bölerek de bulma değerini bulur. F ölçütü tutturma ve bulma değerlerinin harmonik ortalaması olarak tanımlanmıştır.

2.7.2.3 Meteor

Meteor ölçütü F-ölçütünü tutturma ve bulma değerlerine bulma değeri üzerinde ağırlaştırma yapan katsayılar vererek kullanan bir yöntemdir. F ölçütünde tutturma ve bulma değerlerinin sözcük bazında olması dolayısıyla daha değerli olan uzun eşleşmeler hesaba katılmamaktadır. Meteor ölçütü, F ölçütünün yanı sıra uzun eşleşmeleri de hesaba katan bir ceza katsayısı hesaplar ve F ölçütü ile çarpımı sonucu Meteor puanını hesaplar. Ayrıca, Meteor yönteminde bazı dilbilimsel süreçlerde yer aldığından sözcük eşleşmelerinin yanı sıra sözcük köklerinin de eşleşmesine olanak tanınmıştır.

BÖLÜM 3

PARALEL DERLEM OLUŞTURMA ve FİLTRELEME

Farklı dillerde birbirinin çevirisi olan metinlere paralel metinler denir. Paralel metinler bilgisayarlı çeviri ve çok dilli doğal dil işleme uygulamalarında önemli bir rol oynamaktadır [4]. Paralel metinler otomatik sözcük edinimi (automatic lexical acquisition) için kaynak sunarlar [17], istatiksel çeviri modelleri için vazgeçilmez eğitim verisi görevini üstlenirler [4] ve çapraz dilli bilgiye erişim sistemlerinde (cross language information retrieval) sözcükler arası bağlantıları sağlarlar [58]. Ayrıca, son yıllarda tek dilli kaynaklar ve araçlar geliştirmek İngilizce gibi doğal dil işleme çalışmaları açısından zengin dillerdeki geliştirilen kaynak ve araçlardan kelime eşleme yöntemleri (word alignment) ile bağlantılar kurarak faydalanan çalışmalar yapılmaktadır [21].

Görüldüğü gibi birçok doğal dil işleme alanında önemli bir kaynak olarak kullanılan paralel metinler son derece kritik kaynaklardır. Fakat ne yazık ki; henüz yeterli miktarlara ulaşabilmiş değiller. Yakın zamanlara kadar bilgisayarlı çevirideki istatiksel çalışmalar genellikle Fransızca – İngilizce dilleri arasında yoğunlaşmaktaydı, çünkü Kanada Parlamentosu tutanaklarını bu iki dilde yayınlıyordu ve bu diller arasındaki paralel metinler yeterli miktarlara ulaşmıştı. Bugün hala az sayıda dil çifti için yeterli miktarlarda paralel metinlerden söz edilebilmektedir. Bunların da genellikle resmi tutanaklar, haber metinleri ve kullanım kılavuzları gibi özel biçimlerde olduğu ve düzensizlikler içerdiği söylenebilir. Ayrıca, diğer kaynaklarda olduğu gibi paralel kaynakların da ücretli, lisans kısıtlamaları olduğu da bir gerçektir. Tüm bu sebeplerden Church ve Mercer'in "daha çok veri daha iyi başarı" [59] tavsiyesine uymak güçtür. Daha çok veri için tutarlılıktan, kapsayıcılıktan ve dengeden taviz verilmek zorunda kalınmaktadır [21].

Paralel derlem elde etmek için birbirinin çevirisi olan metinleri elde etmek ve daha sonra cümle seviyesinde hizalamak gereklidir. İnsan eliyle paralel derlem oluşturma işlemi pratik olarak mümkün olmadığından otomatik paralel metin elde etme işlemi özellikle son yıllarda başlı başına bir çalışma alanı olarak görülmektedir. Web'den paralel metin toplama, cümle hizalama, gürültülü paralel kaynaklardan filtreleme ile paralel olanları ayıklama gibi alt çalışma alanlarından oluşmaktadır.

Yaygın kullanılan paralel derlemlere örnek olarak Avrupa Birliği yayınlarından oluşan 11 dili kapsayan Europarl derlemi, İngilizce – Fransızca dillerini kapsayan Kanada Parlamentosu yayınlarından oluşan Hansard derlemi ve LDC (Linguistic Data Consortium) derlemi verilebilir. Ne yazık ki, bu derlemlerin hiç biri Türkçe dilini içermemektedir.

3.1 Paralel Derlem Oluşturma

Paralel derlem oluşturma önişlemeler haricinde iki aşamadan oluşmaktadır. Paralel metin elde edimi ve cümle hizalama aşamaları doğal dil işleme uygulamalarına kaynak olabilecek çok dilli verileri oluşturmak için gereken aşamalardır. Toplanılan paralel metinler doküman olarak ayrılmışsa; paralel derlem oluşturmak için önce dokuman seviyesinde hizalama sonra cümle hizalama yapılması gerekmektedir.

3.1.1 Paralel Metin Toplama

Çeşitli dil çiftleri için paralel derlem elde etmek için çeşitli kaynaklar ve yöntemler kullanılmaktadır. Resmi kurumların yayınları [15], dini kitaplar[16], kullanma kılavuzları [17], film alt yazıları [18], farklı dillerde yayımlanmış kitaplar [19], farklı dillerde yayın yapan haber siteleri [20] ve web sayfaları [21] gibi kaynaklardan paralel derlemler elde edilmiştir. Bu kaynaklar genellikle dokümanlar şeklinde olup; cümle hizalama adımına geçmeden önce doküman seviyesinde hizalama yapılması daha makuldür.

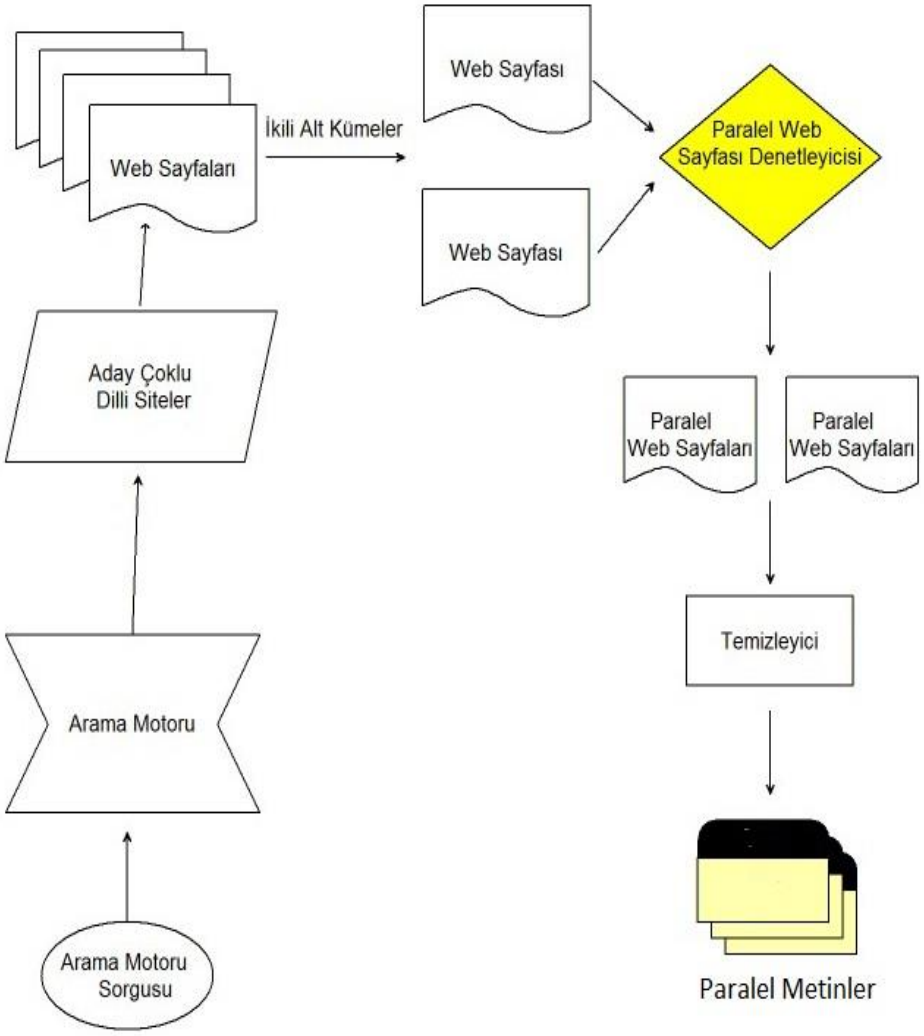
Son 15 yıllık dönemde özellikle paralel kaynakları kısıtlı olan dil çiftleri için Web'den otomatik paralel metinler toplayan sistemlerin geliştirilmesi üzerine çalışmalar yoğunlaşmıştır. Resnik ve diğerleri [21] (Fransızca – İngilizce, Arapça – İngilizce, Çince - İngilizce), Chen ve diğerleri [24] (Çince - İngilizce) Web kaynaklarından paralel metin

toplayan sistemler geliştiren ilk çalışmalardandır ve sonraki çalışmalarda da bu çalışmalar referans alınmıştır [60], [61], [62], [63]. Resnik ve diğerlerinin geliştirdiği [21] STRAND isimli sistemde arama motorlarına bağlantı içeren metinlerde diller ile ilgili ifade arayan sorgular göndererek çoklu dilli siteler elde edilmeye çalışılmış; bu sitelerdeki sayfalar ikili ikili HTML yapısı, uzunluğu, içeriği vb. konularda incelenerek paralel olup olmadıkları denetlenmiştir. Elde edilen paralel olduğu düşünülen sayfalarda ise HTML ve diğer meta bilgileri temizlenerek sayfadaki paralel metinler elde edilmiştir. Bu çalışmayı takip eden sonraki çalışmalarda farklı dil çiftleri için benzer yöntemlerle (sayfa karşılaştırma gibi adımları geliştirerek) çoklu dilli yayın yapan siteleri bulma; birbirinin çevirisi olan paralel sayfaları bulma işlemlerini yapan sistemler geliştirilmiştir. Espla ve Gomis'in çalışmasında [64] Bitextor[1] isimli giriş olarak verilen çoklu dilli sitelerden paralel metin toplayan dilden bağımsız otomatik bir sistem geliştirilmiştir. Arama motoruna çoklu dilli siteleri bulması için gönderilen sorgular sayfa URL'lerinde geçen "language=en, english" gibi dil ifade eden sözcükler veya bağlantı içeren sözcüklerde "in English, English version" gibi sözcükler olabilir. Google[2], Bing[3] gibi arama motorları URL'ler ve bağlantı metinleri (anchor) içerisinde arama yapmaya izin vermektedir. Şekil 3.1'de Web'den paralel metin toplama işlemini gösterilmektedir.

[1] http://bitextor.sourceforge.net/

[2] http://www.google.com

[3] http://www.bing.com



Şekil 3.1 Web'den paralel metin toplayan sistem

Bu çalışma kapsamında Türkçe dili için paralel metin kaynaklarını artırmak amacıyla STRAND'a benzer Web'ten Türkçe – İngilizce paralel metin toplayan bir sistem geliştirilmiş ve Bölüm 3.3.2.3 'te Web Derlemi başlığı altında bahsedilmiştir. Türkçe – İngilizce dil çifti için geliştirilen bu sistemde STRAND'dan farklı olarak arama motorunun döndürdüğü sitelerin gerçekten çoklu dilli olup olmadığını sayfa içerisindeki bağlantıları takip eden ve birbirine bağlantı veren sayfaların bir dil belirleyici yardımıyla birinin Türkçe diğerinin İngilizce olduğu belirlenirse siteyi çoklu dilli olarak etiketleyen bir modül eklenmiştir. Çoklu dilli olarak işaretlenen site içerisindeki paralel sayfaları bulmak için Bitextor yazılımından faydalanılmış ve elde edilen aday paralel metinler 88 824 İngilizce sözcük ve Türkçe karşılığını içeren elektronik bir sözlükten yararlanılarak

içerik karşılaştırılması ile kontrol edilmiştir. Yine bu çalışma kapsamında toplanılan diğer derlemler bölüm 3.3.2'de tanıtılmıştır.

3.1.2 Cümle Bölüştürme

Toplanılan paralel metinlerde cümle hizalaması adımına geçmeden önce cümlelerin bölüştürülmesi gerekmektedir. Bu işlem için cümlenin bittiğini ifade eden noktalama işaretleri, büyük-küçük harf bilgisi gibi özellikleri kullanan ve kısaltmalardaki noktalama işaretleri, büyük harfle başlayan özel isimler gibi istisnai durumları tespit eden algoritmalar kullanılmaktadır.

3.1.3 Cümle Hizalama

İBÇ ve diğer DDİ uygulamaları için kullanılabilir olması için paralel derlem içerisindeki cümlelerin hangilerinin birbirine paralel olması gerektiği belirlenmelidir. Bu işleme cümle hizalama denir. Örneğin, İBÇ yöntemleri sözcük hizalama işlemini bu cümleler arasında yaparken, ÖTBÇ yöntemleri girilen cümleye en çok benzeyen cümleyi ve karşılığını bularak çalışmaktadır.

Cümle hizalama işlemi; çevirmenlerin bazı cümleleri öteki dilde birden fazla cümleyle ifade etmesi; çevrilmeyen cümlelerin bulunması gibi sebeplerden zorluklar içermektedir. Bire bir eşleşen paralel metinler bulmak kolay olmadığı gibi; kullanıldığı uygulamaların başarısına doğrudan etkileyen cümle hizalama işlemini insan emeği ile yapmak zaman ve maliyet açısından mümkün görünmemektedir. Bu sebepten otomatik cümle hizalama yöntemleri geliştirilmiştir.

Cümle hizalama işlemi için farklı yaklaşımlar mevcuttur. Bazı yaklaşımlar dilden bağımsız bir yöntem geliştirmeye çalışmışlardır. Brown ve diğerleri [65] birbirinin çevirisi olan cümlelerin sözcük sayılarının da yakın olacağı düşüncesiyle sözcük sayılarından yola çıkarak cümle hizalama yöntemi geliştirmişlerdir. Gale ve Church [22] de benzer bir yaklaşımla karakter sayılarından yola çıkmışlardır ve geliştirdikleri yöntem hala yaygın bir şekilde kullanılmaktadır. Europarl derlemi Gale ve Church'un algoritması ile cümle seviyesinde hizalanmıştır [67]. Cümle uzunluğu tabanlı bu yaklaşımlar cümle uzunlukları arasında yüksek korelasyonu olan İngilizce – Fransızca dil çifti gibi çiftlerde

oldukça verimli çalışırken, Çince – İngilizce gibi cümle uzunluğu korelasyonu düşük dil çiftlerinde kullanıldığında performans oldukça düşmektedir. Li ve diğerleri [68], Melamed [69] çalışmalarında cümlelerin uzunluğuyla birlikte yer bilgilerini kullanan geometrik tabanlı çalışmalar yapmışlardır. Wu [70] ise noktalama işaretleri, özel isimler ve bir çeviri sözlüğünden yararlanan cümle hizalama yöntemi üzerine çalışmıştır. Ma'nın [71] sözlük tabanlı çalışmasında ise daha az geçen sözcüklere çok ağırlık veren sözlük tabanlı bir yaklaşımla çalışan Champollion[1] isminde bir araç geliştirilmiştir. Sözlük tabanlı bu çalışmaların dezavantajı belirli dil çiftlerinde çalışıyor olması iken, aralarındaki cümle uzunluğu korelasyonu düşük olan dil çiftlerinde oldukça verimli çalıştığı görülmektedir. Senrich ve Volk tarafından [67] geliştirilen bir diğer yaklaşım ise BÇ tabanlı cümle hizalama yöntemidir. Bu yaklaşımda ise hedef dildeki cümlelerle kaynak dildeki cümlelerin BÇ çıktıları arasında BLEU ölçüsü kullanılarak benzerlik hesaplanmaktadır.

Taşçı ve diğerleri [72] Türkçe – İngilizce metinler üzerinde çalışan, cümle uzunlukları ve yer bilgilerini kullanan bir cümle hizalama yöntemi geliştirmişlerdir. E-kitaplar, haber metinleri, akademik çalışmalar ve çeviri şirketlerinden elde edilen dokümanlar bu çalışmada derlenmiş ve testlerde kullanılmıştır.

3.1.3.1 Türkçe – İngilizce Paralel Metinler için Cümle Hizalama Yöntemlerinin Karşılaştırılması

Türkçe – İngilizce cümlelerin hizalamada yukarıda anlatılan yaklaşımların başarısını ölçmek için bu çalışma kapsamında deneyler yapılmıştır. Deneylerde uzunluk tabanlı yaklaşımları temsil etmesi için Moore'un [73] Bilingual Sentece Aligner aracı[2], sözlük tabanlı yaklaşımların temsili için Ma'nın Champollion araç takımı [71], BÇ tabanlı yöntem olan Senrich ve Volk tarafından [67] geliştirilen BLEUAlign[3] aracı kullanılmıştır. Çok dilli haber sitelerinden, çok dilli internet sitelerinin 'hakkımızda' sayfalarından oluşan deney kümesindeki dokümanlar toplam 1035 İngilizce ve 1055 Türkçe cümle

[1] http://champollion.sourceforge.net/

[2] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[3] https://github.com/rsennrich/bleualign

içermektedir. İnsan emeğiyle yapılan hizalamada bire bir eşleşen cümleler (1-1) ve birden fazla sayıda cümleyle olan eşleşmeler (M-N) elde edilmiş ve otomatik cümle hizalama araçlarının çıktıları ile karşılaştırılarak sonuçlar elde edilmiştir.

Sonuçlar araçların ürettiği sonuçların ne kadarının doğru olduğunu gösteren tutturma (precision) ve elle yapılan gerçek hizalamaların ne kadarını bulabildiğini gösteren bulma (recall) değerleri ve iki değerin geometrik ortalaması olan F1 ölçütü ile Tablo 3.1'de gösterilmiştir. T tutturma değerini, B bulma değerini göstermektedir.

Çizelge 3.1 Türkçe – İngilizce Dilleri için Cümle Hizalama Yöntemlerinin Karşılaştırılması

Hizalama Yöntemi		**1-1**			**N-M**			**Hepsi**		
		T	**B**	**F1**	**T**	**B**	**F1**	**T**	**B**	**F1**
Uzunluk Tabanlı Hizalama		0.81	0.82	0,81	-	-	-	0.81	0.82	0,81
BÇ Tabanlı Hizalama		0.94	0.9	0,92	0.49	0.56	0,53	0.90	0.88	0,89
Sözlük Tabanlı Hizalama(yüzeysel biçimde sözcükler)		0.91	0.79	0,84	0.24	0.45	0,33	0.80	0.76	0,78
Sözcük Tabanlı Hizalama (biçimbilimsel çözümleme ile sözcük kökleri)		0.96	0.92	0,94	0.49	0.68	0,58	0.91	0.90	0,91
Sözcük Tabanlı Hizalama (basit kök eşleştirme)	k=2	0.96	0.88	0,92	0.36	0.64	0,48	0.87	0.86	0,87
	k=3	0.97	0.90	0,93	0.46	0.72	0,58	0.90	0.90	0,90
	k=4	0.97	0.94	0,95	0.51	0.72	0,60	0.92	0.92	0,92
	k=5	**0.97**	**0.94**	**0,95**	0.56	**0.76**	**0,65**	**0.92**	**0.93**	**0,92**
	k=6	0.97	0,94	0,95	**0,56**	0,75	0,65	0,92	0,92	0,92

Champollion aracı İngilizce – Arapça ve İngilizce – Çince dilleri için geliştirilmiştir ve sözlüksel eşleştirmeleri gerçekleştirebilmek için İngilizce için sözcükleri sözlüksel biçimine çeviren bir kök bulucu içermektedir. Çalışma kapsamında bu araç için İngilizce – Türkçe dilleri arasında çalışabilmesi için bazı değişiklikler gerçekleştirilmiştir. Türkçe için biçimbilimsel çözümleme İngilizce diline göre çok daha zor ve masraflı olması sebebiyle sözcüksel eşleşmeleri gerçekleştirmek için sözcüklerin ilk *k* harfini sözcüğün kökü olarak ele alan bir yöntem kullanılmış olup deneylerde hem bu basit kök eşleştirme yöntemi hem de Oflazer'in [74] biçimbilimsel çözümleyicisinin yanı sıra Yüret ve Türe'nin [75] sözcük belirsizlik giderici aracı kullanılarak elde edilen sözcük kökleri kullanılmıştır. En uygun k değerini bulabilmek için *k = 2, 3, 4, 5, 6* değerlerinde ayrı ayrı testler gerçeklenmiştir. Sözlük tabanlı olan bu yöntem için 88.824 İngilizce sözcük ve Türkçe karşılıklarını içeren Google Translate[1] 'den elde edilen elektronik bir sözlük kullanılmıştır. BÇ tabanlı hizalama yöntemi için BÇ sistemi olarak ise Google Translate kullanılmıştır.

Sonuçlardan yola çıkarak Türkçe – İngilizce cümleleri hizalamak için en uygun yöntemin Türkçe – İngilizce sözlük yardımıyla kullanılan sözlük tabanlı yöntem olduğu, kök bulmak için maliyetli olan biçimbilimsel çözümleme kullanmak yerine eşleşmeler için yukarıda anlatılan, sözcüğün ilk *k* harfini kök olarak alan basit kök eşleştirici yöntemin kullanılmasının uygun olduğu görülmektedir. En uygun *k* değerinin ise 5 olduğu sonuçlardan görülmektedir. BÇ tabanlı hizalama yöntemi de oldukça makul sonuçlar üretirken, çift dilli bir sözlüğe veya bir BÇ sistemine ulaşmanın zor olduğu durumlarda uzunluk tabanlı yöntemin de kabul edilebilir sonuçlar verdiği yine sonuçlara bakarak gözlemlenebilir.

Bu çalışma kapsamında derlenen paralel derlemlerdeki cümle hizalama işlemi için sözlük tabanlı yöntem olan Champollion aracı Türkçe – İngilizce sözlük kullanarak ve yukarıda anlatılan basit kök eşleştirme yöntemiyle birlikte kullanılmıştır.

[1] http://translate.google.com

3.2 Paralel Derlem Filtreleme

Paralel olmayan cümle çiftlerinin filtrelenmesi işlemi paralel metin madenciliğinin son işleme adımlarından biri olarak düşünülmektedir. Gale ve Church [22] 1993 yılında paralel cümlelerin uzunluk oranlarını ölçerek bu işlemi gerçekleştirmeyi önermişlerdir. Fakat uzunluk tabanlı yaklaşımlar sadece uzunluk korelasyonları yüksek dil çiftleri için olumlu sonuçlar vermektedir. Chen ve Nie'nin [24] geliştirdiği İngilizce – Çince paralel metin elde etmek için Web'i tarayan sistemlerinde cümle uzunluk ilişkilerinden ve dil belirleme yöntemlerinden faydalanmışlardır. Resnik ve Smith'in [21] çalışmasında ise çeviri benzerliği ölçüsü toplanılan verileri temizleme işlemi olarak kullanılmıştır. Benzerlik skoru simetrik kelime tabanlı bir modelle hesaplanan, kaynak metinde yer alan kelimelerin çevirilerinin hedef dildeki metinde yer alıp almadığına bakılarak elde edilen bir ölçüttür.

Paralel derlemlerdeki filtreleme işlemi paralel olmayan kaynaklardan paralel cümlelerin elde edimi ve paralel kaynaklardan kaliteli örneklerin seçimi olmak üzere iki başlık altında çalışılmıştır.

Khadivi ve Ney'in [25] 2005 yılındaki çalışmasında uzunluk sabitlerini ve çeviri olasılığı ölçüsünü kullanan kural tabanlı bir sistem geliştirerek Avrupa Birliği sitesinden elde edilen bir paralel derlemdeki cümleleri çeviri olasılığı ölçüsüne göre sıralayarak üstteki %97,5'lik kısımla BLEU puanını 46,8'den 47,2'ye çıkarmayı başarmışlardır. Yasuda ve diğerleri [75] 2008 yılında İngilizce – Çince dilleri arasında bir İBÇ sistemi için her cümle için dil modelleriyle hesaplanan karmaşıklık skoru ile eğitim verisindeki olumsuz örnekleri eleyerek %1,76 BLEU puanı civarında ilerleme kat etmişlerdir. Liu ve Zhou'nun [26] 2010 yılındaki makine öğrenmesi tabanlı çalışmasında paralel cümlelerden çeşitli dilsel özellikleri çıkartarak bir özellik vektörü elde etmiş ve Destek Vektör Makineleri (DVM) yardımıyla problemi bir sınıflandırma problemi olarak ele almıştır. Yanlış yazılan kelime sayısı, dil modeli puanı, sözdizimsel ayrıştırılmış cümledeki bağlantısız sözcük sayısı ve çeviri benzerliği özellikleriyle örnekleri kaliteli ve kalitesiz olarak sınıflandırmışlar ve 40 binlik bir derlemde 0,88 tutturma ve bulma sonuçlarını elde etmişlerdir. Taghipour ve diğerlerinin [76] 2010 yılında yapmış olduğu çalışma da paralel verileri temizlemek için bir sınıflandırma yöntemini içermektedir. Dil

modeline ve IBM çeviri modeline dayalı özellikler, uzunlukla alakalı özellikler kullanılarak geliştirdikleri yöntemde sınıflandırma işlemi için maksimum entropi modeli kullanılmış, 48 binlik Farsça – İngilizce paralel derlem üzerinde yapılan deneyler sonucu sistemlerinin %98.3 doğruluk düzeyinde çalıştığı gösterilmiştir. Cui ve diğerlerinin [77] 2013 yılındaki çalışmasında ise öğreticisiz öğrenme yöntemlerinden Rastgele Yürüyüş (Random Walk) algoritması kullanılarak her cümle çiftinin önem puanı iteratif olarak hesaplanmaktadır. Çince – İngilizce bir paralel derlemde test edilen sistem 0,5~0,1 BLEU puanı aralığında iyileştirme sağlamıştır.

Munteanu ve Marcu tarafından [27] 2006 yılında yaptıkları çalışmada paralel olmayan haber kaynaklarından Çince, Arapça, İngilizce paralel cümleleri bulan bir sistem geliştirilmiştir. Maksimum entropi sınıflandırıcısı ile cümleleri uzunluk, eşleşme ve sözcük hizalama sonucu elde edilen özellikleri kullanarak sınıflandırılan bu sistem ile Arapça – İngilizce bir derlemde 0,94 tutturma ve 0,67 bulma sonuçları elde edilmiştir. Hoang ve diğerlerinin yaptığı çalışmada [28] bir konuya 50 bin cümlelik özel başlangıç derleminin üzerine farklı konulardaki paralel olmayan kaynaklardan uzunluk ve eşleşme bilgilerine göre 95 bin paralel cümle eklenerek BLEU puanı 8,92'den 24,07'ye yükseltilmiştir.

Paralel cümlelerin değerlendirilmesi konusunda yapılan çalışmalar incelendiğinde genellikle bazı dilbilgisel ve istatiksel özellikler, cümle uzunlukları ilişkisi ve sözcüklerin eşleşmesi gibi özellikler kullanılarak makine öğrenmesi yöntemlerinden faydalanıldığı ve paralel cümlelerin 'kaliteli' ve 'kalitesiz' olarak sınıflandırıldığı görülmektedir. Bu çalışma kapsamında geliştirilen, benzer özellikleri kullanarak Türkçe – İngilizce cümle çiftlerinin İBÇ açısından kalitesini değerlendiren paralel cümle filtresi sistem bölüm 3.5'te açıklanmıştır.

3.3 Türkçe - İngilizce Dillerinde Paralel Derlemler

İngilizce – Türkçe dillerinde derlenen paralel derlemlerin kısıtlılığı bu diller arasında gerçekleştirilen tatmin edici sonuçlar veren İBÇ sistemlerinin bulunmamasının önemli sebeplerindendir. Son yıllarda bu alanda artan çalışmaların sonucu olarak yayınlanan İngilizce – Türkçe paralel derlemler bulunmaktadır ve bu derlemler bölüm 3.3.1'de incelenmiştir. Bu çalışmanın amaçlarından biri olan İngilizce – Türkçe paralel derlem

kaynaklarının artmasını sağlamak için yapılan çalışmalar ile yeni derlemler elde edilmiş ve bu derlemler de bölüm 3.3.2'de incelenmiştir. Aynı zamanda Türkçe – İngilizce paralel derlemler BÇ sisteminin başarısı açısından ve diğer yönlerden karşılaştırılmıştır.

3.3.1 Erişime Açık Türkçe – İngilizce Paralel Derlemler

Bu bölümde Türkçe – İngilizce dilleri arasında çeşitli kaynaklardan derlenmiş erişime ve kullanıma açık olan paralel derlemler hakkında bilgiler verilmektedir.

3.3.1.1 SETimes Paralel Derlemi

SETimes (South-East European Times) paralel derlemi [20] Balkan dillerinde (Arnavutça, Bulgarca, Hırvatça, Yunanca, Makedonca, Rumence, Sırpça ve Türkçe) ve İngilizce yayın yapan bir haber sitesi olan "South-East European Times[1]" haber sitesinde toplanılan paralel metinlerden oluşan bir derlemdir. Cümle ayrıştırıcı olarak *SentParBreaker cümle bölüştürücüsü*[2] [78] cümle hizalama için ise sözlük tabanlı bir cümle hizalayıcısı olan *hunalign*[3] [79] kullanılmıştır. Bu derlem tüm dil çiftleri için toplamda 17 milyon 60 bin adet paralel cümle içerirken, Türkçe – İngilizce dil çifti için 207 bin paralel cümle içermektedir. SETimes paralel derlemi erişilebilir ve kullanıma açık bir derlemdir.[4]

3.3.1.2 OpenSubtitles Paralel Derlemi

OpenSubtitles paralel derlemi 59 ayrı dilde film altyazılarından derlenmiş bir paralel derlemdir [18]. Filmlerin altyazılarını sağlayan çevrimiçi veri tabanları mevcuttur. Bu çevrimiçi veri tabanlarına kullanıcılar oluşturdukları altyazıları ekleyebildikleri gibi, altyazı araması da yapılabilmektedir. OpenSubtitles.org sitesi bunlardan zengin içerikli ve güvenilir olanlarından biridir ve bu paralel derlem de bu sitede yer alan 18.900 film altyazısından derlenmiştir.

---

[1] http://www.setimes.com/

[2] http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector

[3] http://mokk.bme.hu/resources/hunalign/

[4] http://opus.lingfil.uu.se/SETIMES2.php

Film altyazıları değişik formatlarda (sub-viewer, microDVD, vb.) hazırlandığı için tüm altyazılar sub-viewer formatına getirilmiş ve cümle hizalama işlemi için geleneksel cümle hizalayıcı yöntemler yerine film altyazılarında bulunan zaman bilgilerini kullanan bir cümle hizalama yöntemi kullanılmıştır [80].

OpenSubtitles paralel derlemi 361 dil çiftinde (Türkçe dili de dahil) her bir dil çifti için 22 milyon paralel cümle içermektedir.

3.3.1.3 Diğer Erişime Açık Türkçe – İngilizce Paralel Derlemler

Örnek çeviri hafızası Tatoeba[1] sitesinden oluşturulan Tatoeba paralel derlemi [18] 129 dilde toplam 2,94 milyon paralel cümle içermektedir. Türkçe – İngilizce dil çifti için ise 107 bin adet paralel cümle bu derlem içerisinde mevcuttur.

Linux tabanlı K Masaüstü Ortamı'nın kararlı son sürümü olan KDE Software Compilation 4 yazılımının yerelleştirme dosyalarından elde edilen KDE4 paralel derlemi[2] 92 dil için toplam 8,89 milyon paralel örnek içerirken, Türkçe – İngilizce dil çifti için 153 bin paralel örnek içermektedir. İnternet için yaratılmış, sunucu taraflı, çok geniş kullanımlı, genel amaçlı, HTML içerisine gömülebilen betik ve programlama dili olan PHP (Üstünyazı Önişlemcisi, Aslen: Personal Home Page - Kişisel Ana Sayfa) dokümantasyonlarından[3] elde edilen PHP paralel derlemi[4] ise 22 dilde toplam 1,38 milyon paralel örnek içerirken Türkçe – İngilizce dilleri arasında 32 bin kadar paralel örnek içermektedir. KDE ve PHP paralel derlemleri teknik dokümanlardan elde edildiği için ve örnekleri cümle olmadığı, daha çok çeşitli teknik tanımlamaların başka bir dildeki ifadeleri olduğu için genel amaçlı BÇ sistemlerinin eğitimine katkı sağladıkları söylenemez.

İnternet üzerinden kitap satışı yapan EU bookshop sitesinden[5] derlenen EUbookshop paralel derlemi[1] 48 dilde toplam 173 milyon paralel cümle içermesine ramen Türkçe –

[1] http://tatoeba.org/

[2] http://opus.lingfil.uu.se/KDE4.php

[3] http://se.php.net/download-docs.php

[4] http://opus.lingfil.uu.se/PHP.php

[5] https://bookshop.europa.eu

İngilizce dilleri arasında içerdiği örnek cümle sayısı 23 bindir. EUbookshop paralel derlemi örneğin İngilizce – Almanca dilleri arasında 8 milyon paralel cümle içermektedir. Yani bazı dil çiftleri için önemli bir kaynak olmasına rağmen Türkçe – İngilizce dil çifti için az sayıda örnek içermesi dolayısıyla değerli bir kaynak değildir. Özellikle Avrupa dillerinde BÇ çalışmaları için çok önemli bir kaynak olan Avrupa Birliği yayınlarından derlenen EuroParl paralel derlemi ise ne yazık ki hiç Türkçe örnek içermemektedir.

3.3.2 Çalışma Kapsamında Oluşturulan Türkçe – İngilizce Paralel Derlemler

Türkçe – İngilizce BÇ çalışmalarına kaynak olabilmesi amacıyla bu çalışma kapsamında da bu diller için yeni paralel derlemler oluşturulmuştur.

3.3.2.1 Yazınsal Derlem

Taşçı ve diğerlerinin [72] çalışmasında toplanılan e-kitaplardan, haber makalelerinden, akademik çalışmalardan ve çeviri şirketlerinin verilerinden toplanılan paralel metinlerden derlenen bu paralel derlem 688 bin paralel cümle içermektedir. Cümle hizalama işlemi için sözlük tabanlı yöntem olan Champollion aracı Türkçe – İngilizce sözlük kullanarak ve bölüm 3.1.3.1'de anlatılan basit kök eşleştirme yöntemiyle birlikte kullanılmıştır.

3.3.2.2 Akademik Derlem

Yüksek Öğrenim Kurulu, Ulusal Tez Merkezi[2] veri tabanından 270 bin adet tez özeti (lisans, lisansüstü ve doktora) İngilizce ve Türkçe Dillerinde çekilerek, cümle bölüştürme ve cümle hizalama işlemleri gerçekleştirilmiş ve toplam 1 milyon 400 bin paralel cümlelik bir derlem elde edilmiştir. Cümle hizalama işlemi için sözlük tabanlı yöntem olan Champollion aracı Türkçe – İngilizce sözlük kullanarak ve bölüm 3.1.3.1'de anlatılan basit kök eşleştirme yöntemiyle birlikte kullanılmıştır.

---

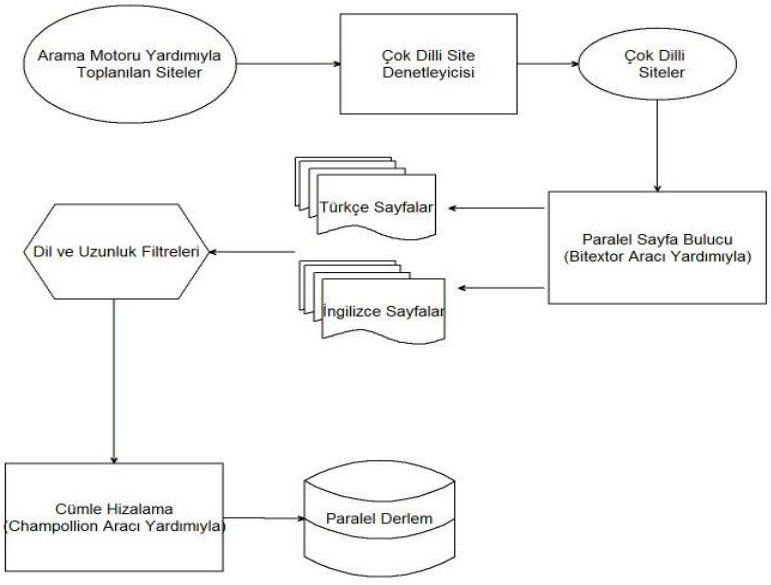
[1] http://opus.lingfil.uu.se/EUbookshop.php

[2] https://tez.yok.gov.tr/UlusalTezMerkezi/

Bu paralel derlem ile eğitilen İBÇ sisteminin konuya özel olarak kullanılabileceği yani akademik ve bilimsel metinlerin çevirisi için başarılı sonuçlar vereceği düşünülmektedir.

3.3.2.3 Web Derlemi

Bu çalışma kapsamında bölüm 3.3.1'de anlatılan internetten paralel metin toplama yöntemlerini kullanarak paralel metin madenciliği yapan bir sistem geliştirilerek çok dilli yayın yapan internet sayfalarından bir paralel derlem elde edilmiş ve Web derlemi ismi verilmiştir. Bu sistem arama motoru yardımıyla elde ettiği sitelerin çoklu dilli yayın yapıp yapmadığını kontrol edip, eğer çoklu dilli yayın yapıyorsa sayfalarını diske kaydetmektedir. Bitextor[1] programı yardımıyla site içerisinde yer alan sayfaları birbirlerinin çevirisiyle hizalayıp daha sonra da hizalı sayfaları 3.1.3.1'de anlatılan Champollion aracının Türkçe için geliştirdiğimiz versiyonu ile cümle bazında hizalama yapılarak parallel derlem elde edilmiştir. Şekil 3.2'de çok dilli yayın yapan internet sitelerinden paralel derlem elde eden sistemin çalışması gösterilmiştir.



Şekil 3.2 Çok dilli yayın yapan internet sitelerinden paralel derlem elde eden sistem

Arama motorları yardımıyla site isimlerinin toplanılması işlemi 'inlinks' sorgu ifadesi kullanılarak linklerde 'İngilizce', 'English' gibi sözcükler sorguya eklenerek

[1] http://bitextor.sourceforge.net/

gerçekleştirilmiştir. Çok dilli site denetleyicisi ismi verilen site linklerini ve sayfaların dillerini kontrol ederek sitenin İngilizce ve Türkçe yayın yaptığını denetleyen bir araç geliştirilmiştir. Çok dilli site denetleyicisi için 100 adet çok dilli ve 100 adet çok dilli olmayan siteler üzerinde testler yapıldı. Test sonuçları şu şekilde elde edildi: 100 adet çok dilli siteden 77 doğru, 23 yanlış; 100 adet çok dilli olmayan siteden 99 doğru 1 yanlış sonuç üretildi. Tutturma değeri 0,987 ve bulma değeri ise 0,77 olarak hesaplandı. Daha sonra geliştirilen bu çok dilli site denetleyicisi kullanılarak Google arama motoruna yaplıan sorgular sonucu elde edilen sitelerden çok dilli olan 80 bin kadar site işaretlenerek bir çok dilli site listesi oluşturuldu. Giriş olarak verilen çoklu dilli sitelerden paralel metin toplayan dilden bağımsız otomatik bir sistem olan Bitextor aracı, bu sistem içerisinde sadece çok dilli olarak işaretlenen siteler içerisindeki paralel sayfaları bulmak amacıyla kullanılmıştır. İngilizce – Türkçe dilleri arasında çalışabilmesi için değişiklikler ve eklemeler yapılan Champollion aracı cümle hizalama işlemleri için kullanılmıştır.

3.3.2.4 Yeminli Sözlük Paralel Derlemi

Örnek çeviri hafızası Yeminli Sözlük[1] sitesinden oluşturulan Yeminli Sözlük paralel derlemi 63 bin Türkçe – İngilizce paralel cümle içermektedir.

3.3.2.5 Wikipedia Delemi

Son olarak kullanıcıların çeşitli konularda bilgi girerek oluşturdukları çevrim içi ansiklopedi olan Wikipedia sitesinde bir çok maddenin farklı dillerdeki açıklamaları mevcuttur. Bu sayfalar birbirinin çevirisi yani paralel olmasa da paralel olan cümleler barındırabilmektedirler. Wikipedia sitesindeki tüm maddeler ele alınarak Bölüm 3.5'te anlatılan filtreleme yöntemi sayesinde paralel olduğu belirlenen 36 bin cümle elde edilmiştir.

[1] http://www.yeminlisozluk.com/

3.3.2.6 İncil

Hıristiyanların kutsal kitabı olan İncilin İngilizce ve Türkçe çevirilerinden 8 bin cümle elde edilmiştir. İncil'de cümleler 'ayet' olarak ayrıştırıldığı ve numaralandırıldığı için cümle hizalama ve bölüştürme işlemlerine gerek duyulmamıştır.

3.4 Türkçe – İngilizce Paralel Derlemlerin Karşılaştırılması

Çizelge 3.2'de Türkçe – İngilizce dilleri arasında erişilebilir olan paralel derlemlerin ve bu çalışma kapsamında oluşturulan paralel derlemlerin çeşitli özelliklerini karşılaştırılmıştır. Derlemin büyüklüğü içerdiği paralel cümle sayısı ile verilmişken, ortalama cümle uzunluğu ise İngilizce taraftaki cümlelerin içerdiği ortalama sözcük sayısı ile verilmiştir. Her bir paralel derlemin %10'u test için ayrılarak İBÇ sistemleri eğitilmiş ve sistem başarıları BLEU puanı cinsinden hesaplanmıştır. Bu çizelgede sistem performansları değerlendirilirken paralel derlemlerin büyüklüklerinin farklı olduğuna ve test kümelerinin derlemlerin kendi içerlerinden seçildiğine dikkat etmek gerekmektedir. Dolayısıyla adil bir karşılaştırma olmadığı, eğitim verilerinin kendi alanı içersindeki başarısının gösterilmiş olduğu söylenebilir.

Çizelge 3.2 Türkçe – İngilizce Paralel Derlemlerin Karşılaştırmalı Özellikleri

Paralel Derlem	Derlem Büyüklüğü (Cümle)	Ortalama Cümle Uzunluğu (Sözcük)	BLEU Puanı (%)
OpenSubtitles Derlemi	21 Milyon	10,41	-
Yazınsal Derlem	688 Bin	17,715	37,06
SETimes Derlemi	207 Bin	17,251	22,08
Web Derlemi	152 Bin	23,10	34,86
Akademik Derlem	1 Milyon 400 Bin	24,33	34,68
Yeminli Sözlük Derlemi	63 Bin	12,60	10,18
Tatoeba Derlemi	107 Bin	6,72	20,70
KDE4	153 Bin	4,5	21,26
EUBookshop	23 Bin	20,8	12,11
İncil	30 Bin	25,4	4,83
Wikipedia	36 Bin	21,7	12,85

OpenSubtitles paralel derleminin büyüklüğü çok fazla olduğundan fiziksel kaynakların yetersizliği dolayısıyla sistem başarısı hesaplanamamıştır. Daha adaletli bir sistem performansı değerlendirmesi için paralel derlemlerin 150 bin cümlelik kısımları rasgele seçilerek, her paralel derlemden örneklerin mümkün olduğunca eşit sayıda yer aldığı ortak bir test kümesi ile yapılan sistem başarıları değerlendirmesi Çizelge 3.3'e bakılabilir.

Çizelge 3.3 Eşit büyüklükte paralel derlemlerin ve ortak test kümesi üzerinde sistem performansları

Eğitim Verisi	Eğitim Verisi Büyüklüğü (Cümle)	BLEU Puanı (%)
SETimes Derlemi$_{150K}$	150 Bin	16,59
OpenSubtitles Derlemi$_{150K}$	150 Bin	4,72
Web Derlemi$_{150K}$	150 Bin	19,56
Yazınsal Derlemi$_{150K}$	150 Bin	9,49
Akademik Derlem$_{150K}$	150 Bin	10,42

3.5 Türkçe – İngilizce Paralel Derlem Filtresi

Gürültü içeren bir paralel derlemi filtrelemek için her cümle çiftinin birbirlerinin uygun çevirileri olup olmadığını belirleyecek güvenilir bir yönteme ihtiyaç duyulmaktadır. Bu değerlendirmeyi gerçekleştirebilmek için paralel cümle çiftlerinden kaliteyle ilgili olduğu düşünülen bazı özellikler çıkartılarak otomatik bir sınıflandırıcının eğitilmesi yolu tercih edilmiştir. Paralel cümlelerin kalitesi sadece hizalamanın doğru olmasına bağlı değildir, dilbilgisel doğruluk, akıcılık ve doğru sözcük kullanımı da kalite açısından önemli özelliklerdir [26].

3.5.1 Sınıflandırma İşlemi için Kullanılan Özellikler

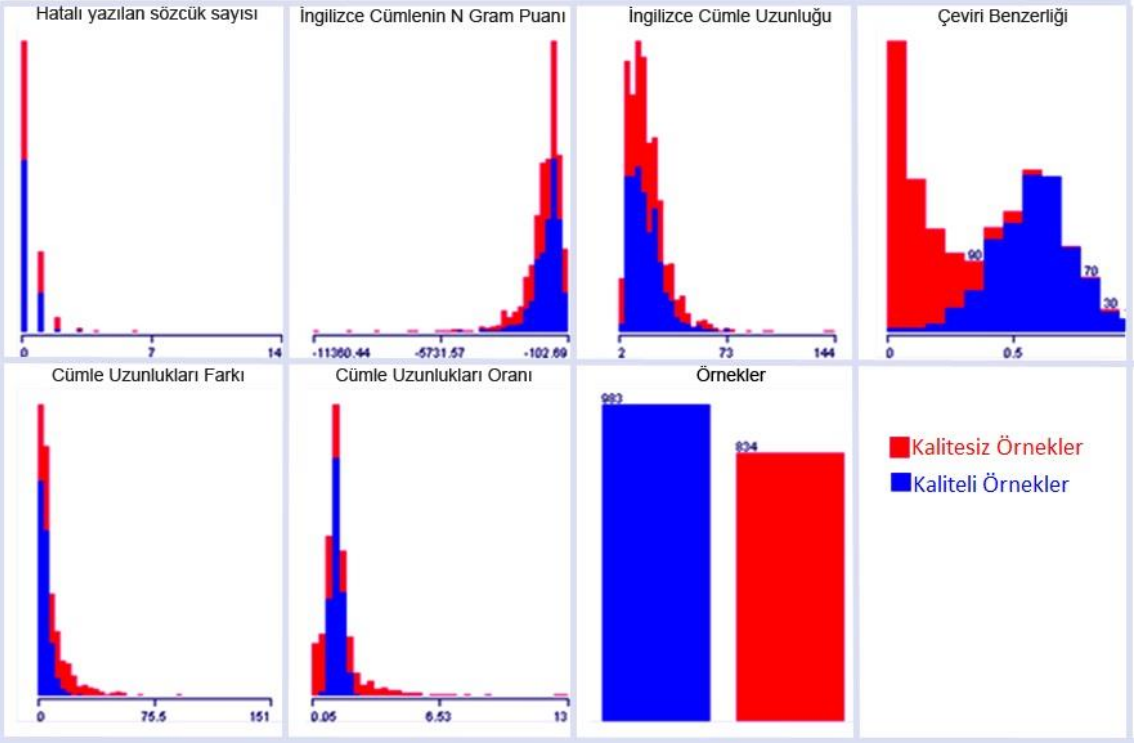
Dilbilgisel olarak hatalı cümleler çeviri modelinde bozulmalara yol açabilmektedir. Bu sebepten yazım kuralları özellikleri cümle çiftlerinin kalitesini değerlendirme konusunda önemli olması sebebiyle İngilizce tarafında yanlış yazılmış sözcük sayısı bir özellik olarak kullanılmıştır.

Dilbilgisel doğrulukla ilgili olduğu düşünülen diğer bir özellik ise dil modeline dayalı bir özelliktir. İngilizce tarafındaki cümlenin dil modelindeki olasılığı akıcılığı ifade eden bir

özellik olarak kullanılmıştır. Uzun cümlelerin dil modeli olasılıklarının düşük olması da dikkate alınarak cümle uzunlukları da özellik olarak kullanılmıştır. İngilizce dil modelini üretmek için BerkeleyLM aracı [81] ve Web 1T[1] derleminden faydalanılmıştır.

Cümle çiftleri uzunlukları arasındaki ilişki cümle hizalama gibi birçok paralel metinlerle ilgili uygulamalarda sıkça kullanılan özelliklerdir ve bu sebepten filtreleme işleminde kullanılan özelliklere cümlelerin uzunluklarının yanı sıra farkları ve oranları da eklenilmiştir.

Filtreleme işlemi için kullanılan son özellik ise sözlüğe göre karşı tarafta çevirileri bulunan sözcüklerin oranıdır. Bu özellik cümlelerin içerikleriyle ilgili bir özelliktir. Bu özelliğin çıkarımında 88.824 İngilizce sözcük ve Türkçe karşılıklarını içeren elektronik bir sözlükten ve bölüm 3.1.3.1'de bahsedilen sözcüğün ilk 5 harfini baş kelime (lemma) olarak ele alan basit kök eşleştirme yönteminden yararlanılmıştır.

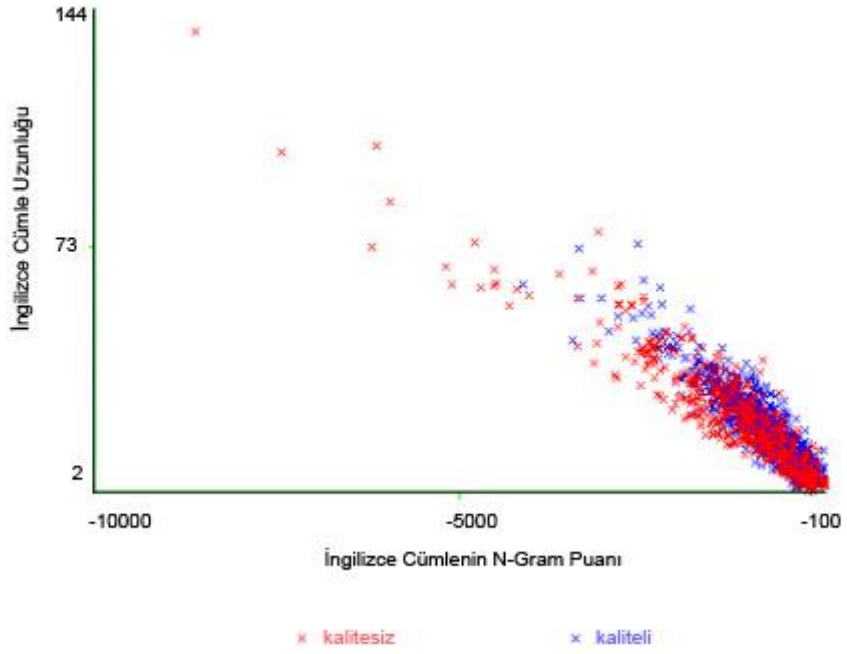


Şekil 3.3 Eğitim için Kullanılan Örneklerin Özelliklere göre Dağılımı

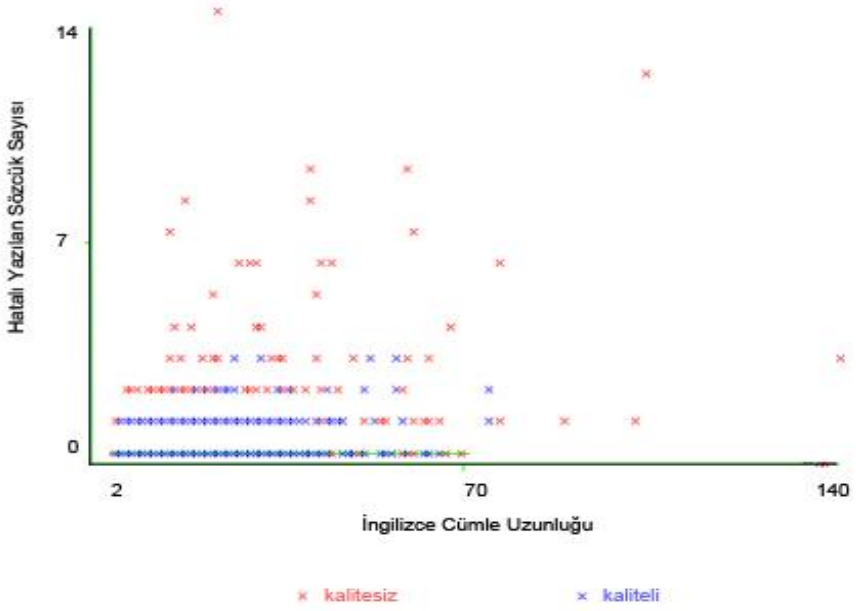
Şekil 3.3'te sınıflandırıcının eğitimi için kullanılan örneklerin özelliklere göre sınıf bilgileri gösterilmektedir. Kırmızı renkteki örnekler 'kalitesiz' olarak etiketlenen

[1] http://get1t.sourceforge.net/

örnekler iken, mavi renkteki örnekler ise ‘kaliteli’ olarak etiketlenen örneklerdir. Türkçe ve İngilizce cümlelerin benzerliği ile hesaplanan çeviri benzerliği özelliğinin sınıfları ayırmak için uygun olduğu gözlemlenirken diğer özelliklerin de kalitesiz örneklerin elenmesinde kullanıla bileceği görülmektedir. İngilizce cümlede hatalı yazılan sözcük sayısı ve dil modelinden faydalanarak hesaplanan N-Gram puanı özellikleri ise İngilizce cümle uzunluğu özelliği ile birlikte değerlendirildiği zaman Şekil 3.4 ve Şekil 3.5’te gözlemlenebileceği gibi sınıf bilgileri için ayırt edici bilgiler sunmaktadır.



Şekil 3.4 İngilizce Cümle Uzunluğu ve N-Gram Puanı Özelliklerine göre Eğitim Verisinin Dağılımı



Şekil 3.5 İngilizce Cümle Uzunluğu ve Hatalı Sözcük Sayısı Özelliklerine göre Eğitim Verisinin Dağılımı

3.5.2 Eğitim Verisi

Otomatik bir sınıflandırıcıyı eğitmek ve değerlendirmek için eğitim ve test verilerine ihtiyaç duyulmaktadır. Eğitim verisi paralel derlemlerden rasgele seçilen örneklerin insan emeğiyle 'kaliteli' ve 'kalitesiz' olarak etiketlenmesiyle oluşturulmuştur. 983 cümle çifti 'kaliteli' olarak işaretlenirken, 160 cümle çifti 'kalitesiz' olarak işaretlenmiştir. 'kalitesiz' olarak işaretlenen bu 160 örnek tamamen faydasız örnekler değildir, kalitelerinin düşük olmasının sebebi dilbilgisel ve yazımsal hatalardan, aslına uygun olmayan çevirilerden kaynaklanmaktadır. Ayrıca, yanlış rasgele hizalamalar sonucu üretilen 674 örnek de 'kalitesiz' olarak işaretlenmiş ve eğitim verisine yapay gürültü olarak eklenmiştir. Çizelge 3.4'te paralel cümle sınıflandırıcısı için hazırlanan eğitim kümesi detaylı bir şekilde verilmiştir.

Çizelge 3.4 Paralel Cümle Sınıflandırıcı Eğitim Kümesi

Eğitim için seçilen örnekler		**Örnek Sayısı**
Kaliteli		983
Kalitesiz	Derlemlerden Seçilen	160
	Yapay Gürültü	674
	Toplam	834
Toplam		1817

3.5.3 Sınıflandırma İşlemi

Sınıflandırma; verideki gizli bilgileri ortaya çıkarmakta kullanılan bir yöntemdir. Sınıflama ile veriler belli özelliklere göre küçük homojen gruplara ayrılır. Sınıflandırma, yeni gelen bir verinin hangi sınıfa ait olduğunu gösteren bir analiz tekniğidir ve bir öğrenme algoritmasına dayanmaktadır. Bu algoritmanın amacı; bir sınıflama modeli oluşturarak, hangi sınıfa ait olduğu bilinmeyen bir veri için sınıf belirlemektir. Burada iyi belirlenmiş değişkenler kilit rolü oynamaktadır. Çeşitli sınıflama yöntemleri bulunmaktadır.

Yapay Sinir Ağları (YSA), insana özgü düşünce ve öğrenme sisteminin taklit edilerek, mevcut verilerden öğrenen ve daha önce karşılaşılmamış durumlarda uygun çıktılar üretecek şekilde yapılandırılan modellerdir. Bu özelliği ile bir yapay zeka türü olan yapay sinir ağları, insan sinir sisteminin bilgisayar ortamındaki benzetimi olarak değerlendirilebilir. Yapay Sinir Ağları, özellikle bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan matematiksel ilişkilerin modellenmesinde kullanılmaktadır. Bu bağlamda Yapay Sinir Ağları, tahminleme, sınıflandırma, kümeleme, sinyal işleme, görüntü ve ses tanıma vb. birçok alanda başarıyla uygulanabilmektedir.

Çok katmanlı YSA'lar sinir hücresi olarak adlandırılan birbirleriyle bağlanmış işlem ünitelerinin katmanlar halinde düzenlenmeleriyle oluşur. Öğrenme sürecinde ağ, ağırlık

değerlerini ayarlar ve böylelikle ağa verilen girdi setine karşılık gelen doğru şekilde tahmin edilmiş veya sınıflandırılmış çıktı değerleri elde edilir [82].

İleri beslemeli yapıda bir tür yapay sinir ağı olan Radyal Tabanlı Fonksiyon Ağları ise özellikle sınıflandırma ve tahminleme problemlerinde kullanılmaktadır. Radyal tabanlı fonksiyon ağları, klasik istatistik yöntemlere göre daha az varsayıma sahip olmaları nedeniyle gerçek hayat problemlerine daha kolay uyum sağlamaktadır.

Karar Ağaçları ise girdi uzayının sorgulama yolu ile art arda iki bölgeye ayrılması esasına dayanmaktadır. Sürekli sorgulamalar yolu ile alt hiper-dikdörtgen uzaylara bölünen girdi uzayı, birbirinden farklı özellikler gösteren girdilerin gruplandığı bölgelere ayrılmış olur. Rassal Karar Ormanları (RKO) ise eğitim kümesindeki örneklerin alt kümleri ile eğitilen birden fazla karar ağacının birleşimden meydana gelmektedir.

Naive Bayes Sınıflandırma yönteminde değişkenlerin birbirinden bağımsız olduğu varsayılır. Yeni örnekler üzerinde sınıflama ise Bayes kuralına göre sınıflandırılacak örneğe en yüksek olasılıkla benzerlik gösteren sınıf seçilerek yapılır. Tüm değişkenlerin birbirinden bağımsız olması gerçek hayatta neredeyse imkânsız olsa da Naive Bayes'in sınıflandırmadaki başarısı çeşitli çalışmalarla kanıtlanmıştır.

Destek vektör makineleri (DVM) ile sınıflandırmada genellikle {-1,+1} şeklinde sınıf etiketleri ile gösterilen iki sınıfa ait örneklerin, eğitim verisi ile elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Söz konusu karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek hiper-düzlem bulunur.

3.5.4 Sınıflandırıcı Seçimi

Sınıflandırıcıyı eğitmek için farklı makine öğrenmesi yöntemleri ile deneyler gerçekleştirilmişti. Radyal Tabanlı Ağlar (RTF – Radial Basis Function), Rassal Karar Ormanı (RKO – Random Forest), Çok katmanlı yapay sinir ağı (YSA – Multilayer Perceptron), Destek Vektör Makineleri (DVM – Support Vector Machine) ve Naive Bayes (NB) tabanlı sınıflandırıcılar WEKA aracı [83] kullanılarak deneyler yapılmış ve Çizelge 3.5'te yer alan sonuçlar elde edilmiştir.

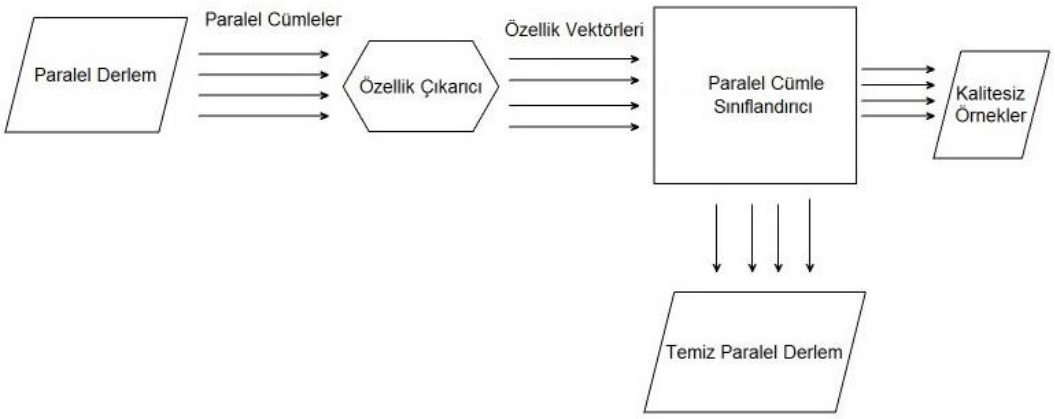
Çizelge 3.5 Sınıflandırma Algoritmalarının Başarıları

Sınıflandırıcı	Tutturma	Bulma	F1 Ölçüsü
RTF Ağları	0,903	**0,964**	0,933
RKO	**0,969**	0,960	**0,965**
Çok Katmanlı YSA	0,938	0,953	0,946
DVM	0,932	0,963	0,947
NB	0,736	0,930	0,822

Çizelge 3.5'te verilen tutturma ölçütüyle sınıflandırıcının yaptığı sınıflandırmaların ne kadarının doğru olduğu ifade edilirken, bulma ölçütüyle gerçekten kaliteli olan örnek cümle çiftlerinin ne kadarının sınıflandırıcı tarafından bulunduğu gösterilmektedir. Paralel derlem filtresi amacı için en iyi tutturma değerini veren sınıflandırıcının uygun olduğu söylenebilir. Sonuçlara bakıldığında karar ağacı tabanlı RKO algoritmasının filtreleme görevi için en uygun sınıflandırıcı olduğu görülmektedir.

3.5.5 Filtre Mimarisi

Bölüm 3.5.1'de detayları verilen eğitim verisi ile Bölüm 3.5.3'te yapılan deneyler sonucunda en iyi sonucu veren Rassal Karar Ormanı (RKO) algoritması kullanılarak sınıflandırıcı eğitilmiş ve paralel derlemler içerisindeki cümle çiftlerini değerlendiren bir sistem geliştirilmiştir. Şekil 3.6'te gösterilen gürültülü paralel derlem filtresinin mimarisinde görüldüğü gibi paralel derlem içerisindeki cümle çiftlerinden hatalı yazılan sözcük saysısı, dil modeli puanı, içerik benzerliği ve cümle uzunlukları ilişkilesi özellikleri çıkarılarak vektörler oluşturulmakta ve bu vektörler sınıflandırıcıya verilmektedir. Sınıflandırıcı cümle çiftlerini kaliteli ve kalitesiz olarak sınıflandırmaktadır. Paralel cümle çiftlerinin kalitesini ölçen bu yöntem ile gürültülü paralel derlemler temizlenebileceği gibi; bu çalışmada oluşturulan Wikipedia derlemindeki gibi paralel olmayan kaynaklardan paralel cümleler elde edilebilinmektedir. Geliştirilen gürültülü paralel derlem filtresinin İBÇ sistemlerinin başarısına olan etkileri Bölüm 4'te gerçekleştirilen deneyler sonucunda verilmektedir.



Şekil 3.6 Gürültülü Paralel Derlem Filtresi Mimarisi

BÖLÜM 4

DENEYSEL SONUÇLARIN DEĞERLENDİRİLMESİ

4.1 Deneylerde Kullanılan Araçlar ve Fiziksel Kaynaklar

İngilizce'den Türkçe'ye İBÇ üzerinde paralel derlemin büyüklüğünün ve kalitesinin sistem performansına etkilerini gözlemlediğimiz deneylerde sözcük öbeği tabanlı İBÇ araçlarından faydalanılmıştır. IBM modelleriyle sözcük hizalama için GIZA++ aracı[1] [50] ve dil modeli için SRILM[2] [84] dil modeli aracı, İBÇ modelinin oluşturulması için MOSES araç takımı[3] [36] kullanılmıştır. Eğitilen İBÇ sistemlerinin başarılarını karşılaştırılmasında ölçüt olarak BLEU ölçüsü [56] kullanılmıştır.

Paralel cümlelerin kalitesini değerlendirmek için makine öğrenmesi tabanlı sınıflandırıcıların kullanılmasında için WEKA aracı [83] ve Java yazılım geliştirme teknolojisi[4] kullanılmıştır.

Deneyler Linux tabanlı bir işletim sistemi olan Ubuntu 11.10[5] işletim sistemi kurulumuna sahip, 64 Gb hafızalı, 16 çekirdek AMD Opteron 6276 işlemciye sahip sunucu bir makine üzerinde yapılmıştır.

[1] https://code.google.com/p/giza-pp/

[2] http://www.speech.sri.com/projects/srilm/

[3] www.statmt.org/moses/

[4] http://www.java.com/

[5] http://www.ubuntu.com/

4.2 Deneylerde Kullanılan Eğitim Verileri

Türkçeden İngilizceye İBÇ sistemlerinde eğitim verisinin büyüklüğünün ve kalitesinin etkilerini gözlemleyebilmek için yapılan deneylerde farklı kalite düzeylerindeki paralel derlemlerden alınan örneklerden oluşan 1 milyon cümlelik bir Türkçe – İngilizce paralel derlemden faydalanılmıştır. Deneylerde kullanılan bu paralel derlem Bölüm 3.3'de detayları verilen Türkçe – İngilizce paralel derlem kaynaklarından film alt yazıları (OpenSubtitles paralel derlemi), haber metinleri (SETimes paralel derlemi), yazınsal metinler (yazınsal derlem) ve Web sayfalarından toplanan metinlerden (Web Derlemi) eşit oranlarda alınan örneklerden oluşmaktadır. Bu paralel derlemler her ne kadar paralel olarak adlandırılsa da cümle hizalama sorunları, yazımsal ve dilbilimsel hatalar içermektedir.

Bölüm 3.3'te detayları verilen bu paralel derlemler hakkında karşılaştırmalı bilgiler Çizelge 3.2'de ve 3.3'de gösterilmiştir.

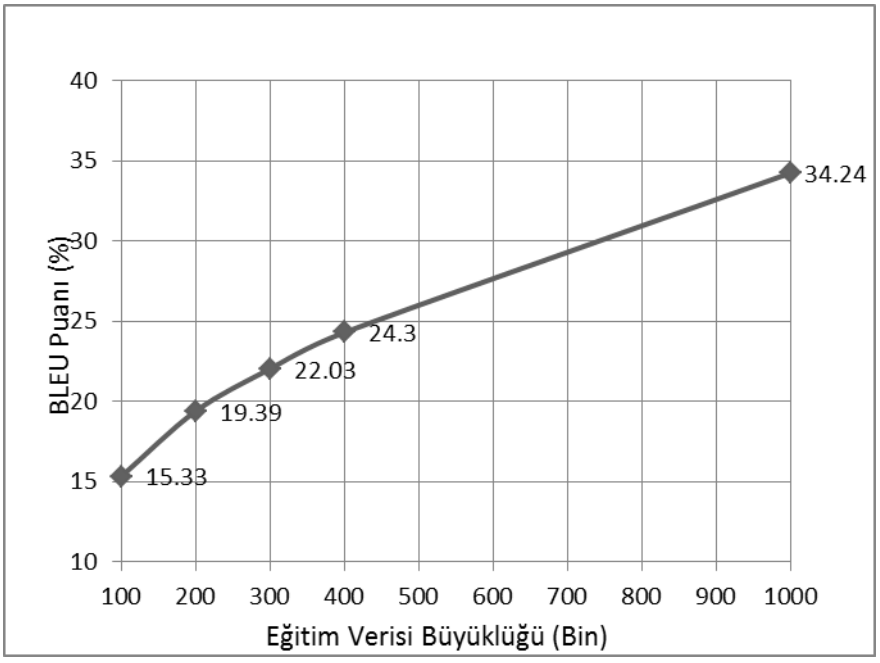
4.3 Türkçe – İngilizce İstatiksel Bilgisayarlı Çeviri Sistemlerinde Paralel Derlem Büyüklüğünün Etkisi

İBÇ sistemlerinde "daha çok veri, daha çok başarı" ilkesi genel olarak kabul görmektedir. Bir önceki bölümde bahsedilen 1 milyon cümle içeren paralel veriden rasgele seçilen farklı sayıdaki örneklerle İBÇ sistemleri eğitilmiş ve Çizelge 4.1'deki sonuçlar elde edilmiştir. Sonuçlar yorumlandığında beklenildiği gibi "daha çok veri, daha çok başarı" ilkesinin geçerli olduğu görülmüştür.

Bu deneylerde eğitim kümesinin %10'u test kümesi olarak kullanılmış ve hedef dil modeli için eğitim verisinin Türkçe tarafı kullanılmıştır. Farklı büyüklüklerdeki eğitim verileriyle yapılan deneylerin sonuçlarını görselleştiren Şekil 4.1'e bakıldığında ise 1 milyon cümleye kadar olan eğitim verilerindeki performanstaki artış ivmesinin hala yüksek olduğu görülmektedir. Buradan yola çıkarak 1 milyon cümlelik eğitim verisi daha da artırılırsa performanstaki artışın da aynı oranla artmaya devam edeceği söylenebilir.

Çizelge 4.1 Farklı büyüklüklerde eğitim verisi ile yapılan deney sonuçları

Derlem Büyüklüğü (Cümle)	BLEU Puanları (%)
100 Bin	15,33
200 Bin	19,39
300 Bin	22,03
400 Bin	24,30
1 Milyon	34,24



Şekil 4.1 Farklı büyüklüklerde eğitim verisi ile yapılan deneylerin sonuçları

4.4 Türkçe – İngilizce İstatiksel Bilgisayarlı Çeviri Sistemlerinde Paralel Derlem Kalitesinin Etkisi

Yapılan deneylerde gözlemlenmek istenen diğer bir olgu ise daha kaliteli çevirilerin daha iyi örneklerle mümkün olup olmadığıdır. Örneklerin kalitesinin İBÇ sistemlerinin performanslarına etkilerini gözlemleyebilmek için ilk yapılan deney Bölüm 3'te detayları verilen paralel derlemlerden (SETimes paralel derlemi, OpenSubtitles Paralel Derlemi, Web Derlemi, Yazınsal Derlem ve Akademik Derlem) rasgele seçilmiş 150 bin cümlelik eğitim verileriyle ve bu verileri Bölüm 3.5'te anlatılan çalışma kapsamında geliştirilen parallel cümle filtreliyici ile kaliteli olarak işaretlenen kısımları ile İBÇ sistemi

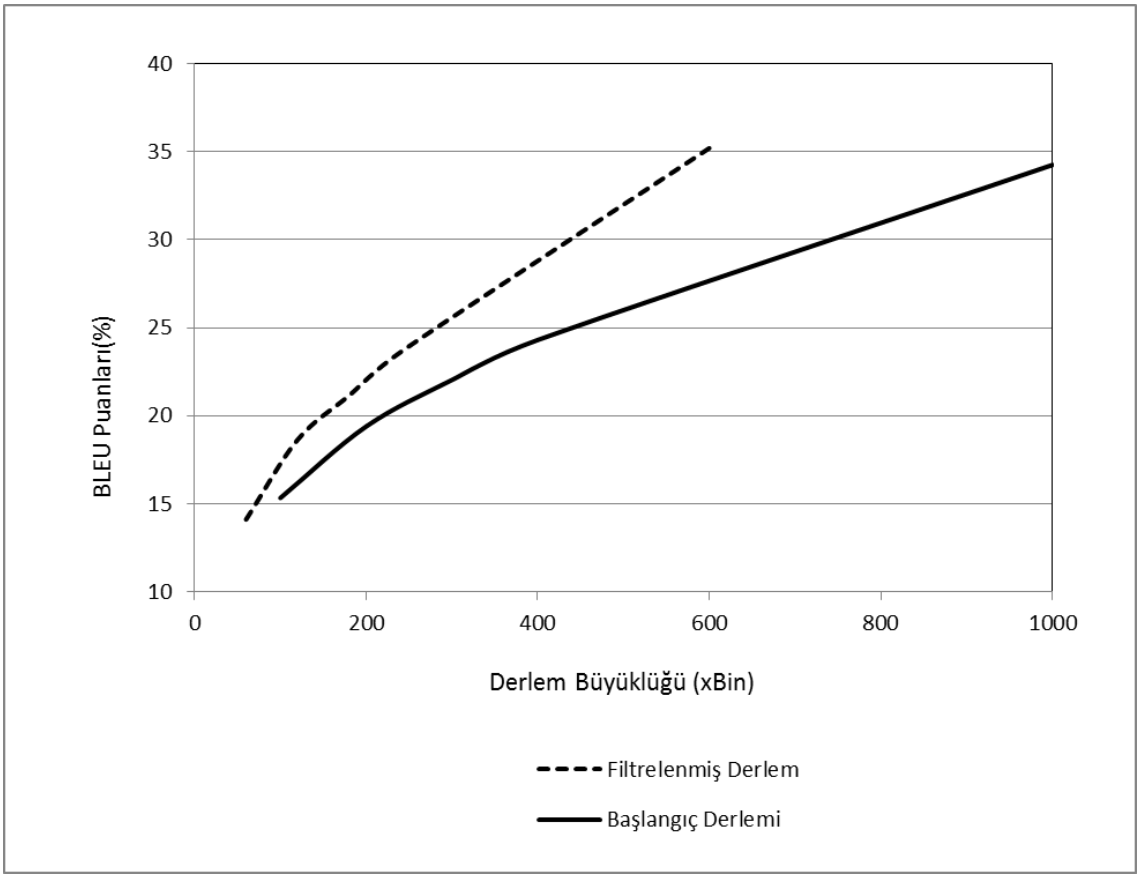
eğitilmiş ve Çizelge 4.2'deki sonuçlar elde edilmiştir. Bu deneylerde test kümesi deneylerde kullanılan parallel derlemlerden biner örnek alınarak oluşturulan 5 binlik ortak bir test kümesidir. Sonuçlar incelendiğinde filtresinin etkisi paralel derlemlere göre değişmektedir. SETimes parallel derleminden alınan 150 bin cümlelik eğitim verisine kıyasla filtrelenmiş 114 bin cümlelik veryi ile %3 civarında bir iyileşme görülmekteyken diğer parallel derlemlerde eğitim verisi büyüklüğü oldukça düşerken sistem performanslarındaki düşüş çok daha küçük oranlarda olmuştur.

Çizelge 4.2 150 Binlik Eğitim verileri ve Filtrelenmiş Kaliteli Kısımları ile Yapılan Deneylerin Sonuçları

Eğitim Verisi (EV)	**EV Büyüklüğü**	**BLEU Puanı (%)**	**Filtrelenmiş EV Büyüklüğü**	**Filtrelenmiş EV BLEU Puanı (%)**
SETimes Parelel Derlemi	150 Bin	16,59	114 Bin	19,43
OpenSubtitles Paralel Derlemi	150 Bin	4,72	72 Bin	3,38
Web Derlemi	150 Bin	19,56	76 Bin	17,17
Yazınsal Derlem	150 Bin	9,49	82 Bin	9,29
Akademik Derlem	150 Bin	10,42	103 Bin	9,55

Çizelge 4.3 Farklı büyüklüklerde Eğitim verisi ve Filtrelenmiş Kısımlarıyla Yapılan Deneylerin Sonuçları

Başlangıç Derlem		**Filtrelenmiş Derlem**	
Derlem Büyüklüğü (Cümle)	BLEU Puanı (%)	Derlem Büyüklüğü (Cümle)	BLEU Puanı (%)
100 Bin	15,33	60 Bin	14,11
200 Bin	19,39	120 Bin	18,63
300 Bin	22,03	181Bin	21,18
400 Bin	24,30	242 Bin	23,68
1 Milyon	34,24	600 Bin	35,19



Şekil 4.2 Farklı büyüklüklerde Eğitim verisi ve Filtrelenmiş Kısımlarıyla Yapılan Deneylerin Sonuçları

Eğitim verisinin kalitesinin ve filtresinin İBÇ sisteminin performansı üzerine etkilerini gözlemleyebilmek için yapılan diğer bir deney ise SETimes paralel derlemi, OpenSubtitles paralel derlemi, Web derlemi ve yazınsal derlemden eşit oranlarda rasgele seçilerek oluşturulan 1 milyon cümlelik karma paralel derlem üzerinde gerçekleştirilmiştir. Bu karma paralel derlemin 100 bin, 200 bin, 300 bin, 400 bin ve 1 milyon cümlelik eğitim verileri oluşturularak İBÇ sistemleri eğitilmiş ve daha sonra kullanılan bu eğitim verileri filtreleme işlemine tabi tutularak kaliteli olarak işaretlenen örneklerle eğitimler yapılarak sonuçlar sistem performansları açısından karşılaştırılmıştır. Yapılan bu deneyin detaylı sonuçları Çizelge 4.3'te gösterilmektedir. Şekil 4.2 ise başlangıç derlemi olarak ifade edilen karma derlemin ve filtrelenmiş derlemin büyüklüğe göre sistem performanslarını göstermektedir. Başlangıç derleminin sadece %60'lık kaliteli kısmı kullanılarak göreceli olarak %2,77'lik bir iyileştirme sağlanmış, 35,19 BLEU puanı elde edilmiştir.

Sonuçlara bakarak, gürültülü paralel derlemin tamamını eğitim verisi olarak kullanmaktansa geliştirilen filtresi yardımıyla paralel derlemin sadece kaliteli örneklerinin kullanılmasının sistem performansı açısından olumlu sonuçlar verdiği ve eğitim verisinin büyüklüğünü azalttığı söylenebilir. Eğitim verisinin büyüklüğünün azalması günler ve haftalar süren eğitim zamanını azaltılmasına da yol açmaktadır. Çizelge 4.4'te ham ve filtrelenmiş verilerin eğitim süreleri verilmiştir.

Çizelge 4.4 Ham ve Filtrelenmiş Verilerin Eğitim Süreleri

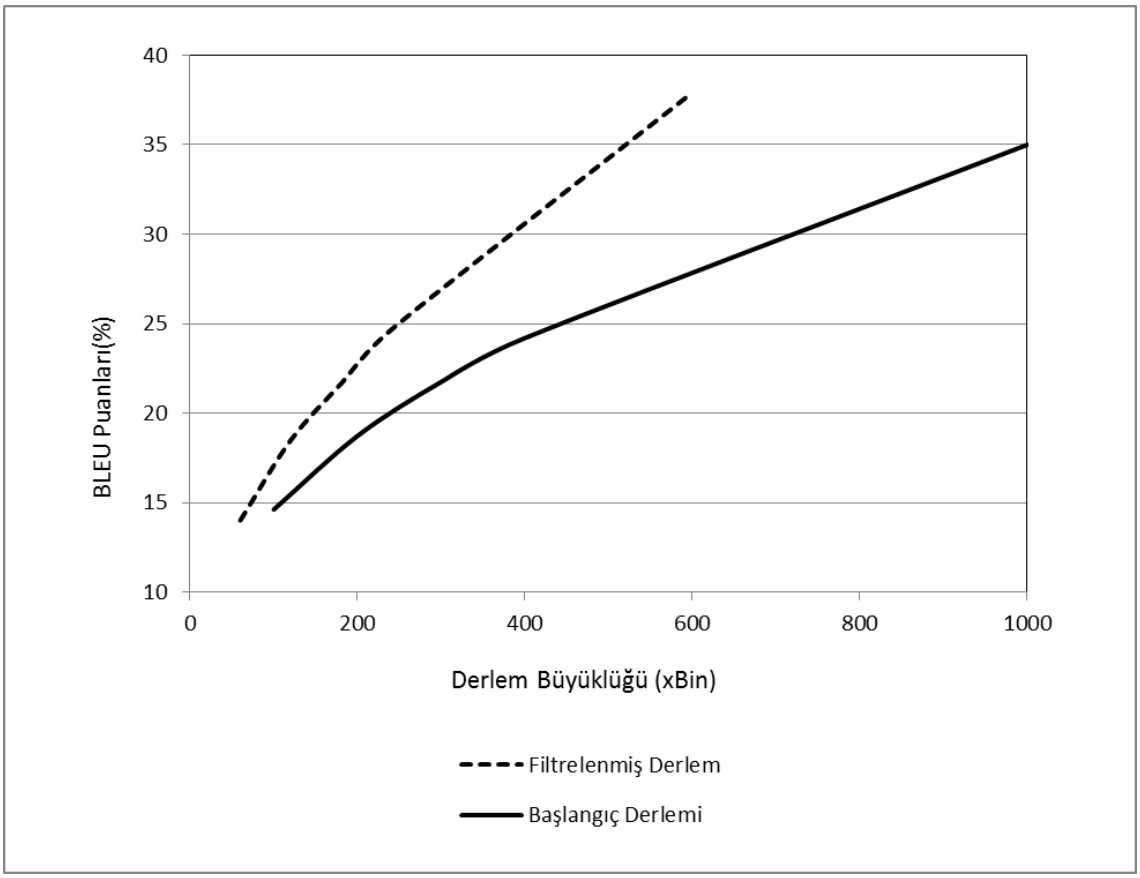
Başlangıç Derlem		**Filtrelenmiş Derlem**	
Derlem Büyüklüğü (Cümle)	Eğitim Süresi	Derlem Büyüklüğü (Cümle)	Eğitim Süresi
100 Bin	25 saat 43 dakika	60 Bin	16 saat 38 dakika
200 Bin	89 saat 11 dakika	120 Bin	38 saat 16 dakika
300 Bin	74saat 49 dakika	181Bin	36 saat 32 dakika

4.5 Türkçe – İngilizce Yönünde Deneyler

İngilizceden Türkçeye İBÇ sistemlerinde eğitim verisinin büyüklüğünün ve kalitesinin etkilerini gözlemlemek için yapılan deneyler; ters yönde yani Türkçe'den İngilizce'ye aynı eğitim verileriyle de yapılarak Çizelge 4.5 ve Şekil 4.3'teki sonuçlar elde edilmiştir. Sonuçlar değerlendirildiğinde filtresinin etkisi ve eğitim verisinin büyüklüğünün etkisi açısından benzer çıktılar görülmektedir.

Çizelge 4.5 Türkçeden İngilizceye Sonuçlar

Başlangıç Derlem		**Filtrelenmiş Derlem**	
Derlem Büyüklüğü (Cümle)	BLEU Puanı (%)	Derlem Büyüklüğü (Cümle)	BLEU Puanı (%)
100 Bin	14,62	60 Bin	14,01
200 Bin	18,73	120 Bin	18,47
300 Bin	21,75	181Bin	21,68
400 Bin	24,20	242 Bin	24,73
1 Milyon	35,74	600 Bin	37,92



Şekil 4.3 Türkçe'den İngilizce'ye Sonuçlar

4.6 Deneysel Sonuçların Değerlendirilmesi

Paralel derlemler İstatiksel Bilgisayarlı Çeviri (İBÇ) modellerinde aktarım modelini oluşturmak için kullanılan eğitim verileridir ve sistem başarısı açısından son derece önemli bir rol oynamaktadır. Church ve Mercer'in "daha çok veri daha iyi başarı" [59]

tavsiyesi bilgisayarlı çeviri alanında istatiksel yöntemler için genel kabul gören bir tavsiyedir.

Bu çalışma içerisinde deneylerde farklı kaynaklardan alınarak oluşturulan 1 milyon cümle içeren bir paralel derlem kullanılmıştır. Yapılan ilk deneyde Church ve Mercer'in tavsiyesi üzerine, yapılan ilk deneylerde İngilizce'den Türkçe'ye İBÇ sistemleri için de geçerliği olduğunu göstermek adına farklı büyüklüklerde eğitim verileri ile denemeler yapılmış ve sonuçları Şekil 4.1 ve Çizelge 4.1'de verilmiştir. Sonuçlara bakıldığında "daha çok veri, daha iyi başarı" ilkesinin geçerliliğinin gösterildiği ve İngilizceden Türkçeye daha yüksek başarılı İBÇ sistemleri geliştirmek için 1 milyon paralel cümle çiftinden daha fazla eğitim verisine ihtiyaç duyulduğu görülmektedir.

Yapılan diğer deneylerin amacı ise, gürültü içeren bütün paralel derlemi kullanmak yerine, sistem başarısına olumlu etkisi olan kaliteli örneklerin seçilerek İBÇ sisteminin eğitilmesi durumunda sistem başarısının gösterdiği değişiklikleri incelemektir. Bölüm 3.5'de detaylı anlatılan ve bu çalışma kapsamında gerçekleştirilen paralel cümle filtresi ile kaliteli örneklerin seçimi yapılmış ve bir önceki deneyde kullanılan eğitim verileri filtrelenerek İBÇ sistemleri eğitilmiştir. Bu deneylerin sonuçları da Çizelge 4.3 ve Şekil 4.2'de yer almaktadır. Bu deneyden çıkan sonuç ise bütün bir derlem yerine gürültüden arındırılmış kaliteli örneklerin kullanılmasının sistem başarısı açısından daha olumlu sonuçlar üreteceğidir. 1 milyon cümlelik eğitim verisi yerine kullanılan 600 bin cümlelik kaliteli örneklerle eğitilen sistemin başarısının göreceli olarak %2,77 BLEU puanının daha yüksek olduğu görülmektedir. Ayrıca, çok fazla zaman alan eğitim sürelerinin de kalitesiz örnekleri ayıklama yoluyla azaltıldığı Çizelge 4.4'te verilen eğitim sürelerine bakılarak anlaşılabilmektedir.

Aynı deneyler, Türkçeden İngilizce yönünde de yapılmış ve Çizelge 4.5 ve Şekil 4.3'te gösterilen benzer sonuçlar elde edilmiştir. Böylelikle paralel derlem filtreleme yaklaşımının sistem başarısına etkisinin ters yönde de olumlu olduğu görülmüştür.

BÖLÜM 5

SONUÇLAR ve ÖNERİLER

Bilgisayarlı Çeviri (BÇ) herhangi bir dilde yazılmış bir ifadenin başka bir dile çevrilmesi işleminin bilgisayar yardımıyla otomatik olarak yapılması işlemidir. Bu çalışmada 1950'li yıllarda başlayan BÇ üzerine çalışmaların günümüze kadar olan evrimi incelenmiş ve İngilizce – Türkçe dilleri arasında başarılı BÇ sistemlerinin geliştirilmesi yolunda yapılacak çalışmalara kaynak olması amaçlanmıştır. BÇ metotları, 1950'li yıllarda bilgisayar bilimcilerinin ve dil bilimcilerin ortak çalışmalarıyla geliştirilen kural tabanlı sistemlerden 1990'lı yıllardan itibaren daha iyi sonuçlar üreten, dilden ve dilbilgisinden bağımsız çalışan istatiksel yöntemler üzerinde yoğunlaşmıştır. İstatiksel yöntemlerin ilk kullanımından bugüne kadar oldukça fazla ilerleme kaydedilmiş olsa da istenen özelliklere ve başarıya sahip, BÇ sistemleri henüz tam anlamıyla gerçekleştirilmediği için bu alanda birçok araştırmacının ilgisi ve çabası sürmektedir. Son yıllarda Türkçe – İngilizce dilleri arasında BÇ çalışmaları yapılmaya başlanmıştır. Bu çalışmada bu diller arasında yapılan çalışmalar, kullanılan yöntemler incelenmiş ve anlatılmıştır.

İstatiksel Bilgisayarlı Çeviri (İBÇ) iki ayrı modelden faydalanmaktadır. Bu modellerden ilki bir cümlenin bir dilde görülme olasılığını içeren dil modelidir ve mümkün olduğunca çok sayıda tek dilli metinlerden oluşan tek dilli derlemlerdir. Diğer bir model ise kaynak dildeki sözcüklerin hedef dile aktarılırken kullanılan sözcüklerin olasılığını içeren aktarım modelidir. Aktarım modeli birbirinin çevirisi olan metinlerden oluşan paralel derlemleri eğitim verisi olarak kullanmaktadır. Paralel derlemler BÇ dışında bilgi çıkarımı, sözcük belirsizliği giderme gibi diğer doğal dil işleme konularında da önemli eğitim verileri olarak kullanılmaktadır.

Paralel derlem oluşturma işlemi için paralel metin elde edimi ve cümle hizalama gibi alt işlemler yapılmalıdır. Paralel metin toplama konusunda yapılan çalışmalarda resmi kurumların yayınları, yazınsal kitaplar, dini belgeler, film altyazıları ve internet sayfalarını kaynak olarak kullanan çalışmalar mevcuttur. İngilizce ve Türkçe dilleri için var olan paralel derlemler genellikle film altyazıları, çok dilli haber siteleri gibi birçok dil çifti için yapılan çalışmalar sonrasında elde edilmiştir ve kısıtlı sayıdadır. Bu diller arasında kaliteli İBÇ sistemleri geliştirmek için aşılması gereken ilk engel eğitim verisindeki bu kısıtlılıktır. Bu sebepten bu çalışmada internetten otomatik paralel metin toplayan bir sistem geliştirilerek internet sayfalarından 150 bin civarında paralel cümle toplanılmış, çeviri kitaplardan ve makalelerden 680 bin paralel cümle, Yükseköğretim Kurulu tez merkezinde yer alan tez özetlerinden 1 milyon 300 bin paralel cümle ve Yeminli Sözlük isimli çeviri hafızasından 60 bin kadar paralel cümle elde edilmiştir. Ayrıca, daha önceki çalışmalarda oluşturulan OpenSubtitles paralel derlemi (film altyazılarından oluşturulmuştur), SETimes paralel derlemi (SETimes haber sitesindeki haber metinlerinden oluşmaktadır) ve diğer erişilebilir paralel derlemler incelenerek karşılaştırılmıştır.

Paralel derlemlerin İBÇ sistemlerinde eğitim verisi olarak kullanılabilmesi için hangi cümlenin hangi cümlenin çevirisi olduğunun işaretlenmiş olması, yani cümle bazında hizalama yapılmış olması gerekmektedir. Aslına uygun olmayan çeviriler, bir cümlenin birden fazla cümleyle çevrilmesi gibi sebeplerden otomatik cümle hizalama yöntemleri %100 başarıyla çalışamamaktadır. Otomatik cümle hizalama yöntemleri cümle uzunluklarından veya sözlük yardımıyla gerçekleşen benzerlik ölçümünden yararlanmaktadır. Türkçe – İngilizce paralel metinlerde cümle hizalamaların başarısı ölçülmüş ve en yüksek başarıyı sözlük tabanlı eşleştirme yöntemi göstermiştir. Sözlük tabanlı yöntem için İngilizce, Çince ve Arapça dilleri için geliştirilen Champollion aracı Türkçe – İngilizce sözlük eklenerek ve sözcüklerin eşleşmesi için ilk 5 harfi sözcük kökü olarak ele alan bir eşleşme yaklaşımı geliştirilerek Türkçe – İngilizce dillerinde cümle hizalama yapılması sağlanmıştır. Türkçe – İngilizce için modifiye edilen bu araç için insan emeğiyle hizalanan paralel cümleler test olarak kullanılmış ve başarım 0,97 tutturma ve 0,94 bulma olarak elde edilmiştir. Çalışma kapsamında oluşturulan paralel derlemler bu yöntem ile cümle seviyesinde hizalanmıştır.

İBÇ sistemlerinde eğitim verisinin büyüklüğü ne kadar artarsa sistem başarısının da o kadar artacağı kabul görmektedir. Yapılan deneylerde 100 bin paralel cümleden 1 milyon paralel cümleye kadar oluşturulmuş eğitim verileri ile yapılan deneylerde bu yaygın kabule uygun sonuçlar elde edilmiştir.

Paralel cümle çiftlerinin kalitesinin otomatik ölçülebilmesi için cümle çiftlerinden çeşitli özellikler çıkaran makine öğrenmesi yöntemlerinden rassal karar ormanları yöntemini kullanarak cümle çiftlerini kaliteli ve kalitesiz olarak sınıflandıran bir sınıflandırıcı geliştirilmiştir. Yapılan deneylerde paralel derlemin kalitesinin sistem başarısına etkilerini gözlemleyebilmek için farklı boyutlardaki her bir derlemin sadece sınıflandırıcının kaliteli olarak işaretlediği örnekleri kullanarak İBÇ sistemleri eğitilmiştir.

Sonuç olarak paralel derlemin boyutu arttıkça daha yüksek başarılara ulaşıldığı gösterilirken; içerisinde hatalı veya kalitesiz örnekleri temizlenmiş daha az sayıda örnek içeren paralel derlemler ile aynı veya daha yüksek başarılara ulaşıldığı gösterilmiştir. Bütün bir eğitim verisinin %60'ı kullanılarak eğitilen sistem ile göreceli olarak %2,77 BLEU puanı ilerleme sağlanmıştır. Eğitim verisindeki bu küçülmenin sistem başarısındaki olumlu etkisinin yanında çok uzun süren eğitim sürelerinde de kısalmaya yol açtığı gözlemlenmiştir.

İngilizce'den Türkçe'ye yapılan deneyler ters yönde yani Türkçeden İngilizceye de yapılmış ve çıkan sonuçlardan aynı yorumları yapmaya uygun olmuştur. Çalışmalar İngilizce Türkçe dilleri arasında yapılmış olmasına rağmen diğer dil çiftlerine kolayca uyarlanabilir durumdadır.

Bu çalışma ve kapsamında yapılan deneyler, incelemeler ve geliştirmeler Türkçe – İngilizce dilleri arasında başarılı İBÇ sistemlerinin geliştirilebilmesi için önemli kaynaklar sunmaktadır.

KAYNAKLAR

[1] Tantuğ, A. C., (2007) "Akraba ve bitişken diller arasında bilgisayarlı çeviri için karma bir model", Doktora Tezi, İstanbul Teknik Üniversitesi, Fenbilimleri Enstitüsü, İstanbul

[2] Hutchinson, J., (1994). "The Georgetown-IBM Demonstration", MT News International, no.8 : 15-18.

[3] Hutchins, W. J., (1995). "Machine translation: A brief history", Concise history of the language sciences: from the Sumerians to the cognitivists : 431-445.

[4] Brown, P. F., Cocke J., Della Pietra, S., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., (1990). "A statistical approach to machine translation", Computational Linguistics, 16 :79–85.

[5] Durgar El-Kahlout, İ., ve Oflazer K., (2006) "Türkçe-İngilizce için istatistiksel bilgisayarlı çeviri sistemi." Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks, Akyaka, Muğla, Türkiye.

[6] Gorgun, O., ve O. T. Yildiz. (2012). "Using morphology in English-Turkish statistical machine translation." Signal Processing and Communications Applications Conference (SIU), 2012 20th. IEEE.

[7] Yeniterzi R., ve Oflazer K., (2010). "Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish." 48. Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

[8] Sagay, Z., (1981). "A Computer Translation from English to Turkish", Yüksek Lisans Tezi, Ortadoğu Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara.

[9] Turhan, C. K., (1997). "An English to Turkish machine translation system using structural mapping." 5. conference on Applied natural language processing.

[10] Hakkani D. Z., Tür G., Oflazer K., Mitamura T., Nyberg E. H., (1998). "An English-to-Turkish interlingual MT system", Machine Translation and the Information Soup, Springer Berlin Heidelberg.

[11] Hamzaoğlu G., (1993). "Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages", Yüksek Lisans Tezi,

Boğazii Üniversitesi, Fenbilimleri Enstitüsü MS Thesis, Bogazici University Bilim ve Mühendislik Enstitüsü, İstanbul.

[12] Tantuğ, A. C., Adalı E., Oflazer K., (2011) "Türkmenceden Türkçeye bilgisayarlı metin çevirisi." İTÜ Dergisi 7.4.

[13] Orhun M., Adali E., Tantuğ A.C., (2011) "Uygurcadan Türkçeye bilgisayarlı çeviri." ITU Journal Series D: Engineering 10.3.

[14] Altıntaş K., (2001) "Turkish to Crimean Tatar machine translation system", Bilkent Üniversitesi, Fenbilimleri Enstitüsü, Doktora Tezi, Ankara.

[15] Koehn P., (2005). "EuroParl: A Parallel Corpus for Statistical Machine Translation", Machine Translation Summit 2005. Phuket, Thailand.

[16] Resnik, P., Olsen, M. B., Diab, M. (1998). "The Bible as a parallel corpus", Annotating the 'Book of 2000 Tongues'.

[17] Resnik, P., Melamed, I. D. (1997). "Semi-automatic acquisition of domain-specific translation lexicons" Fifth Conference on Applied Natural Language Processing, Washington, D.C.

[18] Tiedemann J., (2009). "News from opus - a collection of multilingual parallel corpora with tools and interfaces", Recent Advances in Natural Language Processing, volume V, : 237–248. Amsterdam/Philadelphia

[19] Uszkoreit J., Ponte J. M., Popat A. C., Dubiner M., (2010). "Large scale parallel document mining for machine translation", 23. International Conference on Computational Linguistics : 1101-1109).

[20] Tyers F. M., Alperen M. S., (2010). "SETimes: a parallel corpus of Balkan languages", multiLR workshop at the language resources and evaluation conference, LREC2010, Malta

[21] Resnik, P., ve Noah A. S., (2003) "The web as a parallel corpus." Computational Linguistics 29.3 : 349-380.

[22] Gale, W. A., ve Kenneth W. Church., (1993). "A program for aligning sentences in bilingual corpora." Computational linguistics 19.1 : 75-102.

[23] Yıldız, E., Tantuğ, A.C., (2012). "Evaluation of Sentence Alignment Methods for English-Turkish Par-allel Texts", LREC 2012: The International Conference on Language Resources and Evaluation. Istanbul

[24] Chen, J. ve Nie J. Y., (2000). "Parallel Web text mining for cross-language information retrievalal", Recherché d'Informations Assist ´ee par Ordinateur (RIAO) : 62–77

[25] Khadivi, S., ve Ney H., (2005). "Automatic filtering of bilingual corpora for statistical machine translation." Natural Language Processing and Information Systems. Springer Berlin Heidelberg : 263-274.

[26] Liu X., ve Zhou M., (2010). "Evaluating the quality of web-mined bilingual sentences using multiple linguistic features", Asian Language Processing (IALP), 2010 International Conference on. IEEE.

[27] Munteanu D. S., Marcu D., (2005). "Improving machine translation performance by exploiting non-parallel corpora." Computational Linguistics 31.4 : 477-504.

[28] Hoang, C., Thai N. P., Bao H. T., (2012). "Exploiting non-parallel corpora for statistical machine translation." Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on. IEEE.

[29] Liddy E. D., (2001). "Natural Language Processing" In Encyclopedia of Library and Information Science, New York.

[30] Özdikililer, E., (2005). "Slav Dilleri ve Türkçe arasında çeviri", Trakya Üniversitesi, Fenbilimleri Enstitüsü, Doktora Tezi, Edirne.

[31] Tarcan, A., ve Bekler E., "İngilizce-Türkçe, Türkçe-İngilizce Makİne Çevirisinde yazilimlarin karşılaştırılması", http://ab.org.tr/ab07/bildiri/72.doc , 5 Mart 2014.

[32] Tekin C., (2008). "İngİlİzce'den Türkçe'ye Makİne Çevirisi Modülü",Yüksek Lisans Tezi, Selçuk Üniversitesi, Fenbilimleri Enstitüsü, Konya.

[33] Durgar El-Kahlout, İ., (2009). "A prototype English-Turkish statistical machine translation system".

[34] Büyükaslan Ali., (2005). "Bilgisayar Destekli Çeviri Üzerine Bir İnceleme." V. Dil, Yazın, Deyişbilim Sempozyumu : 24-25.

[35] Wikipedia, The Free Encyclopedia, "History of machine translation", http://en.wikipedia.org/w/index.php?title=History_of_machine_translation&oldid=585871880, 29.05.2014.

[36] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Herbst E. (2007). "Moses: Open source toolkit for statistical machine translation", In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions : 177-180.

[37] Vauquois B., (1968). "A survey of formal grammars and algorithms for recognition and transformation in mechanical translation", In Ifip Congress : 1114–1122.

[38] Kardeş O., Güngör T., Kipman E., (2003) "Bir uygulama alanında Türkçe metnin anlambilimsel gösterimi." http://www.cmpe.boun.edu.tr/~gungort/papers/Bir%20Uygulama%20Alanında%20Turkce%20Metnin%20Anlambilimsel%20Gosterimi.doc, 3 Mart 2014.

[39] Nagao M., (1984). "A framework of a mechanical translation between Japanese and English by analogy principle", In A. Elithorn & R. Banerji (Eds.), Artificial And Human Intelligence. Elsevier Science Publishers.

[40] Somers Harold., (1999) "Review article: Example-based machine translation", Machine Translation 14.2 : 113-157.

[41] Somers H., (2004). "Machine translation and Welsh: The way forward." A Report for the Welsh Language Board, Centre for Computational Linguistics, UMIST, Manchester.

[42] Chunyu, K., Haihua P., Webster J. J., (2002). "Example-based machine translation: A new paradigm", Translation and information technology : 57.

[43] Brown R. D., (2000). "Automated generalization of translation examples", 18. Conference on Computational linguistics'1. Association for Computational Linguistics.

[44] Mayor A., Alegria I., De Ilarraza A. D., Labaka G., Lersundi M., Sarasola, K. (2011). "Matxin, an open-source rule-based machine translation system for Basque", Machine translation, 25(1) : 53-82.

[45] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L. (1993). "The mathematics of statistical machine translation: Parameter estimation", Computational Linguistics 19 : 263–311.

[46] Shannon, C. E., (2001). "A mathematical theory of communication." ACM SIGMOBILE Mobile Computing and Communications Review 5.1 : 3-55.

[47] Jurafsky D., ve Martin J.H., (2000). "An introduction to natural language processing", computational linguistics, and speech recognition : 577-583.

[48] Bahl, L. R., Jelinek, F., Mercer, R. L., (1983). "A Maximum Likelihood Approach to Continous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, 5 : 179-190.

[49] Chen, S. F., ve Goodman J., (1996) "An empirical study of smoothing techniques for language modeling", 34. Association for Computational Linguistics. Association for Computational Linguistics.

[50] Och, F. J., ve Ney H., (2000). "Improved statistical alignment models", Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics.

[51] Resnik, P., ve Park, C. (2006). "Word-Based Alignment, Phrase-Based Translation?", AMTA : 90–99.

[52] Wu, D., (1997) "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora." Computational linguistics 23.3 : 377-403.

[53] Yamada K., ve Knight K., (2001) "A syntax-based statistical translation model." 39. Association for Computational Linguistics.

[54] Kubon V., Hajic J., Hric J., (2001). "Machine Translation of Very Close Languages", Yayımlanmamış makale

[55] Koehn P., ve Hieu Hoang. (2007). "Factored translation models", EMNLP.

[56] Papineni, K., Roukos, S., Ward, T., Zhu, W. J., (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation", ACL.

[57] Tantug, A. C., Oflazer K., Durgar El-Kahlout I., (2008). "BLEU+: a Tool for Fine-Grained BLEU Computation." LREC.

[58] Davis M. W., ve Dunning T.E., (1995). "A TREC evaluation of query translation methods for multi-lingual text retrieval." Fourth Text Retrieval Conference.

[59] Church, K. W., ve Mercer R., (1993). "Introduction to the special issue on computational linguistics using large corpora", Computational Linguistics, 19(1):1–24.

[60] Lee, C. H., ve Yang H. C., (2001). "Text mining of bilingual parallel corpora with a measure of semantic similarity." Systems, Man, and Cybernetics, 2001 IEEE International Conference on. Vol. 1. IEEE.

[61] Tomás J., Sánchez-Villamil E., Lloret L., Casacuberta F., (2005). "WebMining: An unsupervised parallel corpora web retrieval system", Corpus Linguistics Conference.

[62] Li B., ve Liu J., (2008). "Mining Chinese-English Parallel Corpora from the Web", IJCNLP.

[63] Esplà-Gomis, M., ve Forcada M., (2010) "Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor." The Prague Bulletin of Mathematical Linguistics 93.1 : 77-86.

[64] Espla-Gomis M., (2009). "Bitextor, a free/open-source software to harvest translation memories from multilingual websites", Beyond Translation Memories Workshop (MT Summit XII).

[65] Brown, Peter F., Jennifer C. L., Robert L. M., (1991) "Aligning sentences in parallel corpora", 29. Association for Computational Linguistics.

[66] Aksu, B. T. (2006). Bilgisayar destekli Dil Bilimi Çalıştayı bildirileri: 14 Mayıs 2005 (Vol. 868). Türk Dil Kurumu.

[67] Sennrich R., ve Volk M., (2010). "MT-based sentence alignment for OCR-generated parallel texts." The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado.

[68] Li W., Liu T., Wang Z., Li S., (1994) "Aligning bilingual corpora using sentences location information", 3rd ACL SIGHAN Workshop, 141-147

[69] Melamed, I. D., (1996). "A geometric approach to mapping bitext correspondence", Conference on Empirical Methods in Natural Language Processing.

[70] Wu, D. (1994). "Aligning a parallel English-Chinese corpus statistically with lexical criteria", ACL '94.

[71] Ma X., (2006). "Champollion: A Robust Parallel Text Sentence Aligner", LREC 2006: The Fifth Interna-tional Conference on Language Resources and Evaluation

[72] Taşçı, Ş., Güngör, A. M., Güngör, T., (2006). "Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment", International Scientific Conference Computer Science'2006

[73] Moore R. C., (2002). "Fast and accurate sentence alignment of bilingual corpora." Machine Translation: From Research to Real Users. Springer Berlin Heidelberg, : 135-144.

[74] Oflazer, K. (1994). Two-level description of Turkish Morphology. Literary and Linguistic Computing

[75] Yasuda K., Zhang R., Yamamoto H., Sumita E., (2008). "Method of Selecting Training Data to Build a Compact and Efficient Translation Model", International Joint Conference on Natural Language Processing (IJCNLP), Hyderabad, India

[76] Taghipour K., Afhami N., Khadivi S., Shiry S. (2010). "A discriminative approach to filter out noisy sentence pairs from bilingual corpora," Telecommunications (IST), 2010 5th International Symposium on 2010, : 537-541.

[77] Cui, L., Zhang, D., Liu, S., Li, M., & Zhou, M. (2013). "Bilingual data cleaning for smt using graph-based random walk" In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics Vol. 2, : 340-345.

[78] Piao S., ve Tsuruoka Y., (2010). "A Highly Accurate Sentence and Paragraph Breaker.".

[79] Varga D., Halácsy P., Kornai A., Nagy V., Németh, L., Trón, V., (2007). "Parallel corpora for medium density languages", Amsterdam Studies in the Theory and History of Linguistic Science Series 4 : 292-247.

[80] Tiedemann J., (2007). "Improved Sentence Alignment for Movie Subtitles", Int.Conf. on Recent Advances in Natural Language Processing (RANLP 2007), : 582-588. Borovets, Bulgaria

[81] Pauls, A. ve Klein, D., (2011). "Faster and smaller n-gram language models", ACL, Portland, Oregon.

[82] Lisi F., Schiavo R.A., (1999). "A Comparison Between Neural Networks And Chaotic Models For Exchange Rate Prediction", Computational Statistics & Data Analysis, Vol. 30, : 87-102

[83] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Peter, R., Witten, I. H. (2009). "The weka data mining software: An update", SIGKDD Explorations : 10-18.

[84] Stolcke A., (2002). "SRILM-an extensible language modeling toolkit", INTERSPEECH.–1122.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : ERAY YILDIZ

Doğum Tarihi ve Yeri : 28.01.1987 HATAY

Yabancı Dili : İNGİLİZCE

E-posta : yildizeray@hotmail.com.tr

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2014
Lisans	Bilgisayar Mühendisliği	Kocaeli Üniversitesi	2011
Lise	Fen Bilimleri	Kadıköy Anadolu Lisesi	2004

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2014	Demtaş Bilgi Sistemleri	Yazılım Geliştirici
2011	Proline Bilişim	Yazılım Geliştirici

YAYINLARI

1. **Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts**

Eray YILDIZ, Ahmed Cüneyd TANTUĞ

LREC'2012 - First Workshop on Language Resources and Technologies forTurkic Languages

Mayıs 2012

Proje

1. DPT, İTÜ İngilizceden Türkçeye Bilgisayar Çevirisi Projesi (2011 - 2015)