İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

İNGİLİZCE'DEN TÜRKÇE'YE İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE ALAN UYARLAMASI İLE BAŞARININ ARTIRILMASI

YÜKSEK LİSANS TEZİ

Ezgi YILDIRIM

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

HAZİRAN 2014

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

İNGİLİZCE'DEN TÜRKÇE'YE İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE ALAN UYARLAMASI İLE BAŞARININ ARTIRILMASI

YÜKSEK LİSANS TEZİ

Ezgi YILDIRIM
(504111515)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yard. Doç. Dr. Ahmet Cüneyd TANTUĞ

HAZİRAN 2014

İTÜ, Fen Bilimleri Enstitüsü'nün 504111515 numaralı Yüksek Lisans Öğrencisi **Ezgi YILDIRIM**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"İNGİLİZCE'DEN TÜRKÇE'YE İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE ALAN UYARLAMASI İLE BAŞARININ ARTIRILMASI"** başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Yard. Doç. Dr. Ahmet Cüneyd TANTUĞ** .....
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ** .....
İstanbul Teknik Üniversitesi

Yrd. Doç. Dr. Gülşen ERYİĞİT .....
İstanbul Teknik Üniversitesi

Doç.Dr. Deniz YÜRET .....
Koç Üniversitesi

Teslim Tarihi : **05 Mayıs 2014**
Savunma Tarihi : **04 Haziran 2014**

Anneme ve abime,

ÖNSÖZ

Tez çalışmam boyunca benden bilgisini ve yardımını esirgemeyen danışmanım Yard. Doç. Dr. Ahmet Cüneyd Tantuğ'a ve bu süreçte bana olan güvenleri ve gösterdikleri anlayış dolayısıyla sevgili aileme, anneme ve abime, sonsuz teşekkürlerimi sunarım.

Haziran 2014

Ezgi YILDIRIM
Bilgisayar Mühendisi

İÇİNDEKİLER

Sayfa

KISALTMALAR

DDİ : Doğal Dil İşleme
İBÇ : İstatistiksel Bilgisayarlı Çeviri
ÇM : Çeviri Modeli
DM : Dil Modeli
BLEU : Bilingual Evaluation Understudy
METEOR : Metric for Evaluation of Translation with Explicit ORdering
IALA : International Auxiliary Language Association
AÖ-İBÇ : Alana Özgü İstatistiksel Bilgisayarlı Çeviri
BM : Birleşmiş Milletler
MT : Machine Translation

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

İNGİLİZCE'DEN TÜRKÇE'YE İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ SİSTEMLERİNDE ALAN UYARLAMASI İLE BAŞARININ ARTIRILMASI

ÖZET

Doğal dildeki bir metni veya bir konuşmayı diğer bir doğal dile insan gözetimli veya gözetimsiz olarak bilgisayarların yardımıyla çevirme işlemi *bilgisayarlı çeviri* olarak bilinmektedir. Bilgisayarlı çeviri, doğal dil işlemenin en temel, en etkin ve tarihi en eskiye dayanan uygulama alanlarından biridir. 1950'lerde temelleri atılan bilgisayarlı çeviri alanında, önceleri çoğunlukla çeşitli dilbilgisel düzeylerde (biçimbilimsel, sözdizimsel, anlamsal) bilgi aktarımını sağlayan kural tabanlı yöntemler kullanılmıştır. 1990'lardan sonra geliştirilen sistemlerde ve çalışmalarda ise ses tanıma ve işlemede başarısı gözlenen istatistik biliminin desteğine başvurulmuştur. Kültürler arası etkileşimin ve erişilebilir dil kaynaklarının artması ile bilgisayarlı çeviri probleminin çözümünde son yirmi yıldır istatistiksel yaklaşımların kullanımı oldukça artmıştır. Kural tabanlı yöntemlerde gelişmesi sınırlanan bilgisayarlı çeviri çalışmalarına bu gelişmeyle yeni bir başlangıç yapılmıştır.

İstatistiksel yaklaşımlar, emek yoğun bir iş olan kural tanımlama yerine, hizalanmış metinler üzerinden otomatik olarak çeviri parametrelerini öğrenirler. Bazı diller için, bu alanda çok sayıda başarılı çalışma yapılmasına rağmen Türkçe için yapılan çalışmalar oldukça kısıtlıdır. Bu tez çalışmasında, İngilizce'den Türkçe'ye gelişmiş ve kabul gören bir yöntem olan istatistiksel bilgisayarlı çeviri sistemlerinde farklı alan uyarlaması yöntemlerinin etkileri incelenmiş ve sonuçları sunulmuştur. Buradan elde edilen bilgiler ışığında, farklı alanlara uyum sağlayabilen genel amaçlı bir istatistiksel bilgisayarlı çeviri sisteminin modeli oluşturulmuştur.

İngilizce'den Türkçe'ye farklı alanlara uyum sağlayabilen genel amaçlı bir istatistiksel bilgisayarlı çeviri modeli oluşturmanın pek çok açıdan zorlukları bulunmaktadır. En önemli problem, farklı dil ailelerine mensup bu iki dilin birbirinden oldukça farklı yapısal özelliklere sahip olmasından kaynaklanmaktadır. İngilizce oldukça sınırlı bir biçimbilimsel yapıya sahipken, Türkçe oldukça zengin, üretken, türetimsel ve bükümlü bir biçimbilimsel yapıya sahiptir. Öyle ki, İngilizce'de bir çok sözcükten oluşan bir ifade Türkçe'de tek bir sözcükle rahatlıkla oluşturulabilmektedir. Bunun yanında, İngilizce cümleler özne-yüklem-nesne şeklinde sabit bir bileşen sıralamasına sahipken, Türkçe cümleler genellikle özne-nesne-yüklem sırasının tercih edilmesiyle birlikte oldukça esnek bir bileşen sıralamasına sahiptir. Bir diğer problem de istatistiksel yöntemler için gerekli olan dil kaynaklarının Türkçe için yetersiz olmasıdır. Bu yetersizlik Türkçe'nin zengin biçimbilimsel yapısı nedeniyle daha belirgin olmaktadır.

Bu çalışma, dil kaynağı bakımından dezavantajlı ve zengin biçimbilimsel yapısı nedeniyle de veri seyrekliği probleminden şiddetle etkilenen Türkçe için, istatistiksel bilgisayarlı çeviri sistemlerinde başarıyı artırmaya ve genel amaçlı, farklı alanlara uyum sağlayabilen sistemler için bir ön model oluşturmaya odaklanmaktadır. Bununla birlikte, literatürdeki diğer yöntemlerin Türkçe'ye uygulanabilirliğini (muhtemelen

benzer diğer dillere de) ve çeviri başarımına etkilerini açıklayarak bu alanda öncü olmakta, daha önce Türkçe için incelenmemiş olan ve değerlendirilmesi gereken bu etkili uygulama alanındaki ileri çalışmaların da önünü açmaktadır. Bu tez çalışmasında, öncelikle incelenen yöntemlerin kıyaslanabilmesi için bir *yalın sistem* oluşturulmuştur. Bu yalın sistem, elde edilen tüm alanlardaki verilerin kullanılması ile tek bir istatistiksel bilgisayarlı çeviri sistemi eğitilerek elde edilmiştir. Alan uyarlaması çalışmalarında ilk olarak, her biri kendi alanına ait verilerle eğitilmiş, dolayısıyla her biri kendi verisinin temsil ettiği alana uyum sağlamış, farklı istatistiksel bilgisayarlı çeviri sistemleri oluşturulmuş ve bir metin sınıflandırıcı ile bu sistemler birleştirilmiştir. Böylece çevrilmesi istenen giriş cümleleri uygun sistemlere yönlendirilmekte ve sahip olduğu alana sadık kalarak hedef dile çevirisi yapılabilmektedir. Bu yöntemin daha kapsamlı bir çeviri modeli ile iyileştirildiği ek bir uygulaması yapılmıştır. Referans amacıyla kullandığımız yalın sistem, bu sistemin yetersiz kaldığı noktalarda danışılmak üzere bir geri çekilme modeli olarak kullanılmıştır. Bir diğer alan uyarlaması değerlendirmesi, faktörlü çeviri modeli çatısından faydalanarak alan bilgisinin çeviri aşamasına doğrudan katılması ile gerçekleştirilmiştir. Çeviri modelindeki eşleşmiş her bir sözcük öbeği çifti elde edildikleri alanın etiketini kendileriyle birlikte taşımaktadırlar. Bu ek bilgi ile çeviri opsiyonlarının değerlendirildiği çözümleme aşamasında alanların bilincinde bir seçim yapılacağı öngörülmüştür. Son olarak, istatistiksel bilgisayarlı çeviri sistemi bileşenlerinden olan dil modeli aracılığıyla bir alan uyarlaması modeli gerçeklenmiştir. Her biri kendi alanına ait verilerle oluşturulmuş dil modelleri yalın sistemdeki genel dil modelinin yerine kullanılmış, bu yöntemle dil modeli ile farklı alanlara uyarlaması yapılmış sistemler bir metin sınıflandırıcı yardımıyla bütün bir sistem oluşturmak amacıyla bir araya getirilmiştir. Böylece alana özgü dil modeli kullanmanın çeviri kalitesine etkisi gözlemlenmiştir. Çalışmaların sonuçları bir bilgisayarlı çeviri otomatik değerlendirme ölçütü olan BLEU ile değerlendirilmiştir.

Yapılan çalışmalar göstermektedir ki, İngilizce'den Türkçe'ye bilgisayarlı çeviri sistemlerinde en iyi alan uyarlaması performansı dil modeli uyarlaması ile elde edilmektedir. Bu yöntemle birlikte çeviri başarısı 27,36 BLEU puanından 29,89 BLEU puanına yükselmiştir. Yalın istatistiksel bilgisayarlı çeviri sistemine kıyasla %9,25 oranında göreceli iyileşme gözlemlenmiştir.

EVALUATION OF DOMAIN ADAPTATION APPROACHES ON ENGLISH-TO-TURKISH STATISTICAL MACHINE TRANSLATION SYSTEMS

SUMMARY

Machine Translation (MT) is the automatic translation of texts or speeches from one natural language into another with or without human assistance. It is useful for different purposes and application environments. MT is practical for the interaction, dissemination and assimilation of information. It is used for not only producing "printable" quality texts, but also editing of "raw" outputs. Raw defines immature production which requires post-editing. Although the ideal goal of a machine translation system is to be able to produce high-quality translations, in practice translation outputs are generally revised. One should note that post editing outputs do not differ from the outputs of human translators with the advantage of less labor by a human translator. The correction of misspellings, the detection of domains or languages, and the classification of documents are in the scope of MT. MT can support individual users in the case of sufficient clarity of translation, such as reading/writing e-mails, surfing the web, basic writing in a foreign language. MT may also be used by embedding in a different system for information retrieval, information extraction, transliteration, summarization, question answering (cross-language) or authoring software.

MT is one of the major, oldest and the most active areas of natural language processing. The initial research in this area started in the 1950s primarily on the rule-based methods, which transfer the information within different levels of linguistic knowledge (morphological, syntactical, semantical). Since the 1990s, after the success of the statistics is recognized in the speech recognition and speech processing, MT research shifted to the statistics-based approaches. In the last two decades, with the increase of interaction between different cultures and increasing number of available language resources, the usage of statistical approaches gathered pace.

Statistical approaches are based on machine learning of the translation probabilities from the aligned parallel texts instead of the labor intensive rule definitions. Although there has been quite extensive work in this area for some fortunate languages, there has not been enough research for Turkish. In this thesis, the effects of different domain adaptation methods on a state-of-the-art English-to-Turkish statistical machine translation system are researched, then results are reported. In the light of these results, we constructed a prototype of a general-purpose statistical machine translation system adaptable to different domains. The majority of studies in the literature show the effect of domain adaptation on a specific domain, whereas this study shows the positive effect of domain adaptation on general translation quality.

There are several challenges of building that kind of an English-to-Turkish model in many aspects. The major challenge is that these two languages belong to different language families and have distant typologies. While English has a limited

morphological structure, Turkish has a rich, productive, derivational and inflectional morphological structure. A single word in Turkish can be stated in English with a phrase composed of many words. For example, the word "*güldürebilmiştim*" can be translated into English in a complete sentence "*I had been able to make somebody laugh.*". While English has a fixed constituent order like subject-verb-object (SVO), Turkish has a free constituent order (subject-object-verb (SOV) is generally preferred). The sentences "*Bozulan bilgisayarımı abim tamir ettirdi.*" (Object-S) and "*Abim bozulan bilgisayarımı tamir ettirdi.*" have completely the same meaning ("*My brother had my broken computer repaired.*").

This paper focuses on the usage of different domain adaptation methods to build a general purposes statistical machine translation (SMT) system for languages with limited parallel training data. Turkish prominently suffers from data sparsity problem because of its morphologically rich nature. In a morphologically rich language, one stem can have multiple surface representations, that is many words can be derived from one root. Hence, it is quite difficult to build a corpus that includes all possible surface representations in the respective language. In this research, the usability and the effects of domain adaptation methods on the English-Turkish SMT are investigated on behalf of other similar disadvantaged languages. This study is carried out using four different sources of domain data namely literature, news, web and subtitles. The data in this study consists of sentence-aligned English-Turkish translations, which is called parallel data in the literature. This research shows the first results of domain adaptation for Turkish, so it will be the pioneer of this valuable research subject for future studies.

The acknowledged domain adaptation methods in the literature are the ones based on the domain-specific data, the translation model, the language model, and the factor translation models framework. In this thesis, a baseline system is built to compare other methods to a reference point. This baseline is trained on all available parallel data from all domains, in this way a single statistical machine translation system is constructed. The translation model of the baseline translation system is obtained from all available parallel data and the language model is obtained from the monolingual data set in the target language of the same parallel corpus. In the first domain adaptation method, four domain specific SMT systems are built. The language and translation models of these systems are obtained from data of their own domains. Then, they are combined together with a text classifier. The classifier sends the input sentences to appropriate domain-specific SMT system, so the complete system can translate sentences in compliance with the domains. As an extension of this method, the baseline system is used as a back-off solution in case it fails to produce any translation options. If a translation option cannot be found in the domain-specific translation model, the domain-adapted system looks for a possible translation in the general translation model. The translation option obtained from the general translation model is better than not having any translation. Thus, this back-off method is expected to increase the general translation quality. The other domain adaptation method used in this thesis is to use the domain information as a factor in the framework of factored translation models. Every phrase pair in the translation model is extracted with its domain information from the parallel data. With the insertion of these domain tags directly into the translation process, the system is capable to select the best options in the consciousness of domains. Finally, a domain adaptation model is formed by the language model as one of the statistical machine translation system components. For this purpose, four different domain-specific language models are built from the monolingual data of

their own domains. These domain-specific language models constructed four different domain-adapted-systems by combining with a general translation model, which is the same model used in the baseline system. So that, the effect of using domain-specific language models on translation quality can be observed. The results of this research are evaluated by BLEU metric which is the well-known machine translation evaluation metric.

One of the results of this study is that domain adapted systems are not quite successful at translating out-of-domain sentences. Second, in case of insufficient data, domain adapted systems based on domain specific data fail to produce systems representing that domain. Hence, if sufficient domain specific data is not available, to build a compact translation system out of all data is more appropriate than to combine domain specific systems. The use of factored translation models to convey domain information directly into the translation process did not increase the overall translation quality in this study. It is shown that adapting translation model is a promising domain adaptation method; especially, through the multiple decoding paths and back-off models. In the conclusion of all experiments, our comparative experiments show that the language model adaptation gives the best domain adaptation performance on the English-to-Turkish statistical machine translation system. With the use of language model adaptation, translation success increased with a relative 9.25% improvement yielding 29.89 BLEU points on multi-domain test data.

1. GİRİŞ

Dünya üzerinde farklı coğrafyalarda yaşayan insanlar, kendi aralarında iletişim sağlayabilmek için ihtiyaçları doğrultusunda dil adını verdiğimiz iletişim araçlarını geliştirmişlerdir. Fakat her dilin yapısı geliştiği coğrafyaya bağlı olarak farklılık göstermektedir. Yeryüzünde 136 dil ailesi ve 7 binden fazla yaşayan dil bulunmaktadır [1]. Bu diller arasında aktarımı sağlamak için bilgisayar biliminin yeteneklerinden faydalanılmaktadır. Her dil kendi problem uzayına sahip olduğu için, dilin bilgisayarlarla işlenmesinde de kendine özgü yöntemler geliştirilmektedir.

Dönemin gerekliliklerine, ticari çıkarlara ve ihtiyaca uygunluğa yönelik olarak bazı diller (İngilizce, Almanca, Fransızca, Çince gibi) bilgisayarlı çeviri alanında yoğun olarak çalışılırken, Türkçe için yapılan çalışmalar oldukça kısıtlı kalmıştır. Üstelik dilbilimsel özellikleri zengin bir dil olması ve kullanılabilir veri miktarının oldukça az olması nedeniyle, Türkçe çalışması zor ve yoğun emek isteyen bir dildir. Fakat, Türkçe üzerine yapılan çalışmalar benzer özelliklere sahip Altay dil ailesine bağlı olan diğer Türk dillerinde (Azerice, Türkmence, Özbekçe, Kırgızca, Kazakça gibi) veya zengin biçimbilimsel yapıya sahip diğer dezavantajlı dillerde (Fince, Macarca, Çekçe, Tamilce, İbranice gibi) yapılan çalışmalara da katkı sağlamaktadır.

Bu çalışmada, veri yetersizliği nedeniyle dezavantajlı olan diller için iyileştirme sağlaması öngörülen alan uyarlaması yöntemlerinin İngilizce'den Türkçe'ye istatistiksel bilgisayarlı çeviri sistemlerindeki etkilerinin değerlendirilmesi yapılmakta ve birden fazla alana uyum sağlayabilen genel amaçlı bir sistemin prototipi oluşturulmaktadır.

1.1 Doğal Dil İşleme ve Bilgisayarlı Çeviri

Doğal dil işleme, ana görevi bir doğal dili otomatik olarak çözümlemek, anlamak, yorumlamak ve üretmek olan bilgisayar sistemlerinin tasarım ve gerçeklenmesini araştıran bilim ve mühendislik dalıdır. Yapay zeka (artificial intelligence) ve dilbilimin (linguistic) bir alt alanıdır.

Hızlı problem çözme ve kalıcı öğrenme yeteneklerinden dolayı, günlük yaşam da dahil olmak üzere pek çok alanda bilgisayarlardan faydalanılmaktadır. Bilgisayarlarla iletişimde, insanların bilgisayarın anlayacağı dilden konuşması gerekmektedir. Bu gereklilik ise bilgisayarların tercih edilirliğini azaltmaktadır. Bu nedenle, insanların kullandığı doğal yollarla, yani konuşma ya da yazma ile bilgisayarlarla iletişim kurmak için doğal dil işleme tekniklerinden faydalanılır. Doğal dil işleme; konuşma tanıma, otomatik yanıtlama, yazılı metin anlamlandırma, özetleme, gruplandırma, seslendirme, bir metni başka dile çevirme, konuşma üretme, yazım hatası düzeltme, veritabanı sorgusu oluşturma gibi pek çok alanda uygulanabilmektedir. Genel olarak insan-bilgisayar ve hatta insan-insan etkileşimini artırmaya yönelik çalışmalar doğal dil işlemenin uygulama alanlarıdır.

Doğal dil işleme çalışmalarındaki en büyük engel bir dilin modellenmesindeki karmaşıklıktır. Dilin doğal yapısında pek çok belirsizlik bulunmaktadır. Bazen insanlar tarafından bile anlaşılamayan, deneyim ve diğer çevresel etmenlerle yorumlanabilen bu belirsizlikleri bilgisayarların öğrenmesi oldukça zordur. *"Annem telefonunu düşürdü."* cümlesinde konuşmacının annesinin kendi telefonunu mu, yoksa konuşmacının konuştuğu kişinin telefonunu mu düşürdüğü anlaşılamamaktadır. Yazılı olarak bile anlaşılmayan bu cümleyi bilgisayarların kolayca anlamasını beklemek haksızlık olacaktır.

Doğal dil işlemenin bir uygulama alanı da bilgisayarlı çeviri sistemleridir. Yazılı metinler üzerinde dil çevirisi için geliştirilen ilk sistemler dili ifade eden pek çok kuralın sisteme tanımlanması ile gerçeklenmiştir. Fakat geniş ölçekli paralel derlemlerin erişilebilirliğinin artması ile istatistiksel bilgisayarlı çeviri (İBÇ), en umut veren bilgisayarlı çeviri (BÇ) yöntemi olmuştur. İstatistiksel bilgisayarlı çeviri sisteminin performansı paralel derlemdeki eğitim verisinin miktarıyla doğrudan ilişkilidir. Son yıllarda sahip oldukları paralel veri miktarının artması sayesinde, bilgisayarlı çeviri alanındaki çalışmalar çoğunlukla İngilizce, Almanca, Arapça, Çince gibi sınırlı sayıda dil üzerine odaklanmaktadır. Fazla miktarda paralel derleme sahip olmayan diller için, çeviri kalitesi dilbilimsel bilginin çeviri işlemine eklenmesi, daha iyi sözcük ve cümle hizamalama yöntemlerinin uygulanması, alan uyumlu sistemlerden faydalanılması gibi farklı çalışma alanlarıyla artırılabilmektedir.

1.2 Türkçe'nin Yapısı

Doğal dil işleme üzerine yapılan çalışmaların sayısı son yıllarda hızla artmaktadır. Fakat bu çalışmalar başta İngilizce olmak üzere Hint-Avrupa dilleri yoğunluklu olarak yapılmaktadır. Ural-Altay dil grubuna dahil olan Türkçe için ise yeterli çalışma bulunmamaktadır. Bunun önemli bir nedeni eklemeli (agglutinative) diller olarak adlandırılan dillerde kök durumundaki sözcüğün, sahip olduğu eklerle anlam ve yüzeysel biçim (surface representation) değişimine uğramasıdır. Bu durum, başarısı kanıtlanmış çalışmaların Türkçe üzerine uygulanmasını zorlaştırmaktadır.

Türkçe dili, sahip olduğu biçimbilimsel (morphological) zenginlik ile başka dillerde bütün bir cümleyle ifade edilen bir anlamı tek bir sözcükle ifade edebildiği için kavraması zor bir yapıya sahiptir. Örneğin *güldürebilmiştim*[1] sözcüğü İngilizce'ye *"I had been able to make somebody laugh."* cümlesi olarak aktarılmaktadır.

Türkçe'nin türetimsel (derivational) yapısı, yapım eklerinin (derivational morpheme) kullanımıyla bir kökten pek çok farklı sözcük elde edilmesine imkan vermektedir. İsimden isim, isimden fiil, fiilden isim ve fiilden fiil olmak üzere dört ana kategoride türetme yapılabilmekte ve elde edilen yeni anlamlarıyla cümledeki görevleri değişen farklı sözcükler elde edilebilmektedir. Aşağıdaki örnekte olduğu gibi *gözetmenlik* sözcüğü *göz* sözcüğünün çeşitli yapım ekleri alarak farklı sözcük türlerine dönüşmesi ile oluşmuştur.

gözetmenlik
göz (isim)
göz + et (isim-fiil)
göz + et + men (fiil-isim)
göz + et + men + lik (isim-isim)

Ünlü uyumu (vowel harmony) ve sesbirim değişiklikleri (phoneme alternation) nedeniyle ekler bağlandığı sözcüğe göre ya da sözcükler aldıkları eklere göre değişebilmektedir. Örneğin, *kitap* sözcüğü belirtme durumunda (accusative case) kullanıldığında *kitap+ı→kitabı* olmaktadır. *Burun* sözcüğü iyelik eki (possessive suffix) aldığında, en sondaki ünlü harf düşmekte ve *burunum* yerine *burnum* haline dönüşmektedir. Çoğul eki (plural suffix) olan *-ler/-lar* eki ise *fındık* sözcüğüne

[1] gül+dür+ebil+miş+ti+m şeklinde eklerine ayrılmaktadır.

eklendiğinde ünlü uyumu nedeniyle *fındıklar* olarak kullanılırken, *peçete* sözcüğüyle birlikte *peçeteler* şeklinde kullanılmaktadır.

Türkçe serbest sözcük sıralamasına (free word order) sahiptir, yani özne, yüklem ve nesnelerin cümle içindeki yerleri belirli ve sabit değildir. Bu nedenle aynı anlama gelen bir ifadeyi söylemenin birden fazla yolu vardır. Örneğin *"Bozulan bilgisayarımı abim tamire verdi."* ile *"Abim bozulan bilgisayarımı tamire verdi."* cümleleri arasında anlam yönünden bir farklılık bulunmamaktadır. Buna karşılık aynı sözcük cümlede kullanıldığı yere göre farklı anlamlar içerebilmektedir. *"Kafasını sert zemine vurdu."* ile *"Kafasını zemine sert vurdu."* cümleleri arasında anlam farklılığı bulunmaktadır.

Türkçe, dünya üzerinde 70 milyondan fazla kişi tarafından ana dil olarak konuşulmaktadır [1] (Dünya üzerinde en çok konuşulan diller hakkında bilgi için bknz. Ek A.2). Oldukça üretken biçimbilimsel yapısı sayesinde yüksek miktarda yüzeysel forma sahiptir. Türkçe yaklaşık 30,000 kök sözcük ve yaklaşık 150 farklı ek barındırmaktadır. Bu durum ciddi veri seyrekliği (data sparsity) problemlerine yol açmaktadır. Türkçe sözcüklerin yüzeysel biçimleriyle eğitilecek yetkin bir İBÇ sistemi için gereken paralel derlem miktarı basit biçimbilimsel yapıya sahip diğer diller için gerekenden çok daha fazladır. Beklentinin aksine erişilebilir Türkçe paralel derlem miktarı ise diğer dillere kıyasla oldukça kısıtlıdır. Bu şartlar altında Türkçe'ye veya Türkçe'den başka dillere İBÇ zorlayıcı bir araştırma alanıdır.

1.3 Çalışmanın Amacı

Kültürler arası etkileşimler arttıkça diller arası çevirilerin gerekliliği artmaktadır. Küreselleşen ve hızla gelişen dünyada bu çevirileri çevirmenlere yaptırmak maliyeti ve zaman kısıtı dolayısıyla neredeyse imkansızlaşmıştır. Fakat internetten bilgi almak, rezervasyon yapmak, ürün satın almak gibi işler oldukça yüzeysel bir yabancı dil bilgisi ile yapılabileceğinden mükemmel çeviriye ihtiyaç duyulmamaktadır. Bu gibi durumlarda kusursuz olmasa bile otomatik bir sistemin üreteceği çeviriler oldukça faydalı olacaktır. Günümüzde bilgisayarlı çeviri sistemleri özellikle bazı diller için oldukça iyi sonuçlar üretebilmektedir.

Sınırlı bir kapsamda, belirli bir amaç için geliştirilen sistemler genel sistemlere göre epey başarılı olabilmektedir. Fakat her amaca ve her bağlama uygun sistemler

geliştirmek oldukça zor ve kullanışsızdır. Bu nedenle geliştirilen genel sistemlerin belirli alanlara uyarlanması önerilmektedir. İstatistiksel bilgisayarlı çeviride alan uyarlamasının, yeterli eğitim verisinden yoksun olan diller için çeviri kalitesini artırabileceği düşünülmektedir. Bu çalışma, alan uyarlaması yöntemlerinin paralel veri miktarı oldukça kısıtlı olan Türkçe'deki başarımlarını göstermek ve olası çalışmaların önünü açmak amacıyla gerçekleştirilmiştir. Alan uyarlaması yöntemlerinin belirli alanlardaki başarımlarının değerlendirilmesinin yanında, şimdilik dört alandan oluşan ancak genişletilebilir çok alanlı bir çeviri sisteminin örneklemesi yapılmıştır. Daha sonra elde edilecek veriler ve sistemlerle bu yapının genişletilebilmesi ve daha genel, güvenilir ve başarılı bir sistem haline dönüştürülmesi mümkündür.

1.4 Önceki Çalışmalar

Farklı diller arasında yapılacak çevirilerde bilgisayarların kullanılması yarım yüzyılı aşkın bir süredir araştırmacıların çalışmalarında yer verdikleri bir konudur. Bu amaçla uzun zamandır gerçeklenen çalışmalar, dilin doğası gereği sahip olduğu karmaşıklığa ve diller arasındaki farklılıklara rağmen bilginin çeşitli düzeylerde aktarımının yapılmasına hizmet etmektedir. Günümüzde çalışmaların geldiği noktada, birbirine yapısal olarak benzer olan diller arasında ve belirli bir konuyla sınırlandırılmış alanlarda otomatik sistemler kabul edilebilir sonuçlar üretebilmektedirler. Fakat ekonominin öncelikli olarak teşvik ettiği ve dünya genelinde yaygın olarak kullanılanlar dışındaki diller için bilgisayarlı çeviri çalışmaları yeterli seviyeye ulaşamamıştır.

GİRİŞ bölümünde kısaca değinilen Türkçe'nin özel yapısı ve zorluklarından dolayı bilgisayarlı çeviri alanında Türkçe için yapılan çalışmalarda özel çaba gösterilmesi gerekmektedir. Bilgisayarın daha iyi öğrenmesi ve modelleyebilmesi için kısıtlı miktarda veriyle dilin biçimbilimsel yapısının analiz edilmesi ile gerçeklenen çalışmalarda başarının artırılabildiği gösterilmiştir [2, 3]. Biçimbilimsel bilgiyi çeviri sürecine dahil eden bu çalışmalarda, İngilizce-Türkçe dilleri arasında her iki çeviri yönünde bazı eklerin bağlı oldukları sözcüğe bitişik bırakılması ve uygun olan bazılarının ise bağımsız birer sözcük gibi ayrı yazılmasının sözcükler arasında hizalamayı kolaylaştırdığı ve başarıyı yükselttiği görülmüştür.

Bu çalışmanın odağını oluşturan istatistiksel bilgisayarlı çeviri sistemleri veriye bağımlı sistemlerdir. Verinin kalitesi ve çokluğu sistemin iyileşmeni sağlar. Elde edilebilen tüm veriyi kullanarak oluşturulan genel amaçlı istatistiksel çeviri sistemleri ancak ortalama bir başarıya ulaşabildiğinden, alan uyarlaması başarının artırılması için önemli bir etmendir. Sözcük öbeği temelli istatistiksel bilgisayarlı çeviri sistemleri, çeviri modeli (ÇM) ve dil modeli (DM) olmak üzere iki temel bileşenden oluşmaktadır. Bu bileşenler eğitim verisinden farklı olan geliştirme verisi üzerinde optimize edilir. Çoğu alan uyarlaması çalışması bu bileşenlerin farklı alanlara uyarlanması üzerine gerçeklenmektedir.

Bu bileşenler üzerinde farklı test koşulları ile yapılan bir çalışmada en iyi performans alternatif çözümleme yolları ile iki çeviri modeli kullanılarak elde edilmiştir [4]. Ayrıca bu çalışmada dil modelinin de alan uyarlaması için etkili bir bileşen olduğu görülmüştür.

Çeviri modellerinin adaptasyonunun alan uyarlamasına etkilerinin incelendiği bir başka çalışmada farklı yöntemler kıyaslanmış ve çeviri modelinin karmaşıklığının azaltılmasına dayanan bir yöntem önerilmiştir. Karma modellemede model ağırlıklandırma katsayılarını belirlemek için çeviri modellerinin karmaşıklıkları test edilmiş ve en iyi başarıyı sağlayan katsayılar ise ilgili modellerle ilişkilendirilmiştir [5].

Literatürdeki çalışmalar çoğunlukla bir sistemin belirli bir alana uyarlanması ile o alandaki başarıyı artırmaya yönelik olsa da farklı alanlardaki sistemlerin birleştirilmesi ile de genel başarının artırılabileceği gösterilmiştir. Farklı alanlara özgü çeviri sistemlerini bir araya getirmek için sınıflandırıcılardan faydalanılmıştır [6]. Alan sınıflandırıcısı çevirisi yapılacak olan metnin hangi alana ait olduğunu belirlemektedir. Böylece metin ait olduğu alana uygun olarak çevrilebilmektedir. Bu yöntemle iki farklı alana ait sistemin birleşimi ile İngilizce-Çince dilleri arasında yapılan çevirinin kalitesinin artırıldığı gözlemlenmiştir. Önerilen yöntem bu çalışmanın gerçeklendiği koşullarda, daha önceki bir çalışmada [4] başarılı bulunmuş olan iki farklı çeviri modelinin alternatif çözümleme yolu ile birlikte kullanılmasından daha iyi sonuç vermiştir.

Alan uyarlamasında kullanılmış bir diğer yöntem ise çeviri sisteminin faktörlü çeviri modelleri oluşturulmasıdır. Moses uygulama yazılımının bir parçası olan faktörlü

modellerin kullanım alanı çoğunlukla dillerin morfolojik özelliklerinden faydalanmaya yöneliktir. Ancak, alan bilgisinin bir faktör olarak çeviri sistemine ilave bilgi olarak verilmesi de çeviri kalitesinin artırılmasını sağlamaktadır [7]. Alan bilgisi, kaynak metnin ait olduğu alana özgü bir biçimde hedef dile aktarılmasını sağladığından sistemi iyileştirici bir etmen olabilmektedir.

Çift dilli verinin elde edilmesinin maliyetli olması ve çift dilli verilerin yetersiz olmasından dolayı tek dilli verilerin de alan uyarlamasında kullanılması önem kazanmıştır. Alana özgü tek dilli veriler o alanı temsil eden büyük dil modellerinin oluşturulması için kullanılmaktadır [8]. Ayrıca, alana özgü çift dilli veri tek dilli veriler yardımıyla zenginleştirilebilmekte ve böylece genişletilmiş sözcük öbeği tablosu ve sözlüksel yeniden sıralama modelleri elde edilebilmektedir [9]. Kaynak dildeki tek dilli verinin cümle seviyesinde sözdizimsel olarak farklı ifade edilişleri ile elde edilen çift dilli verilerin var olanlarla birlikte kullanılması ilgili alandaki çeviri başarısını artırmaktadır. Ayrıca var olan otomatik sistemler ile daha az emek harcayarak alana özgü çeviri kalitesi artırılabilmektedir. Sözcük öbeği temelli istatistiksel bilgisayarlı çeviri sistemlerinde alana özgü tek dilli verinin çevrilmesinden elde edilen sentetik çift dilli derlemin ilave bilgi olarak varolan derlemle kullanılması farklı alan uyarlaması yöntemlerinde başarıyı yükseltmiştir [10].

1.5 Tezin Bölümleri

Bu tez çalışmasında, BİLGİSAYARLI ÇEVİRİ bölümünde literatürdeki bilgisayarlı çeviri yaklaşımları, bilgisayarlı çevirinin aşamaları, bilgisayarlı çeviride karşılaşılan zorluklar ve bu sistemlerinin başarısını ölçmek için kullanılan ölçütler anlatılmaktadır. ALAN UYARLAMASI bölümünde alan uyarlaması için kullanılan yöntemler tanıtılmakta ve alan uyarlamasının gerekliliğinden, hangi koşullarda ihtiyaç duyulduğundan bahsedilmektedir. Kuramsal olarak anlatılan yöntemlerin kullanılması ile İngilizce'den Türkçeye bilgisayarlı çeviri sistemlerinde alan uyarlaması için önerilen sistemler ise İNGİLİZCE'DEN TÜRKÇE'YE ÇOKLU ALAN UYUMLU İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ bölümünde detayları ile birlikte verilmektedir. Uygulanan yöntemlerin sonuçları ve başarım değerlendirmeleri karşılaştırmalı olarak UYGULAMA VE SONUÇLAR bölümünde sunulmaktadır. DEĞERLENDİRME VE ÖNERİLER bölümü ise çalışma ile ilgili değerlendirmeleri ve önerileri içermektedir.

2. BİLGİSAYARLI ÇEVİRİ

Bilgisayarların kullanılmaya başlandığı ilk zamanlardan itibaren, insanlar arasındaki iletişimi artırmak amacıyla bilgisayarlardan faydalanılması bir araştırma konusu olmuştur. Diller arasında çevirilerin yapılması için çeşitli sistemler tasarlanmıştır. Fakat büyük bir heves ve inançla başlanan bilgisayarlı çeviri çalışmalarından elde edilen bilgilere göre, bilgisayarlı çeviri probleminin yapay zeka alanında çözülmesi zor olan problemleri ifade etmek için kullanılan *AI-complete* bir problem olduğu görülmüştür. Bununla birlikte problemin çözümü için yapılan çalışmalar çeşitli uygulama alanlarında beklentileri karşılayabilmektedir. Yetkin bir bilgisayarlı çeviri sisteminin sahip olması gereken üç temel özellik bulunmaktadır:

- *Otomatiklik* İnsan müdahalesine gerek kalmadan sonuç verebilme
- *Kalitelilik* Anlaşılabilir ve aslına uygun sonuçlar üretebilme
- *Geniş Kapsamlılık* Konudan bağımsız olarak pek çok alanda sonuç üretebilme

Günümüzde gerçeklenen sistemler incelendiğinde, bu gereksinimlerden en fazla ikisinin aynı anda sağlanabildiği görülmektedir. Bu bağlamda, üç farklı sistem oluşturulabilmektedir.

- Otomatik ve Kaliteli

  Bu tip sistemler konunun, metin türünün ve hatta dilbilgisi yapılarının sınırlandırılması ile gerçeklenmektedir. Örneğin, borsa bilgilerini çeşitli dillere çeviren bir sistemde kullanılan cümlelerin yapısı sabittir ve kullanılan sözcük sayısı oldukça sınırlıdır. Bu çeşit sistemlerde otomatik yöntemlerle yüksek kalite yakalamak mümkündür. Bu özellikteki sistemlerin en eski örneği olan Météo, hava tahmin raporlarını İngilizce ve Fransızca dilleri arasında çevirebilmektedir [11]. Bu sistem 1981'den 2001'e kadar uzun yıllar kullanılmıştır.

- Otomatik ve Genel Kapsamlı

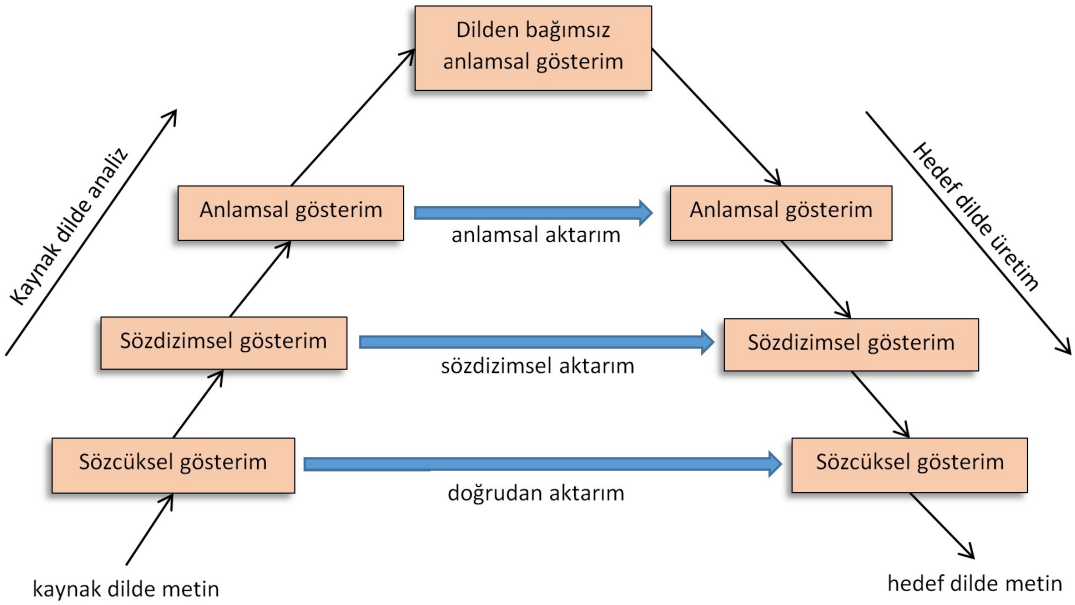
  Özellikle *bilgi elde etme* (information retrieval) amacıyla kullanılan sistemlerde yetkin bir sistemin özelliklerinden olan yüksek kalite koşulu aranmamaktadır. Bu gibi sistemler daha yüzeysel çeviriler yaparak farklı dildeki veriye erişim sağlayabilmektedir. Doğrudan bilgi edinmenin dışında bilgisayarlı çevirilerin aracılık ettiği farklı uygulamalar için de bu tipteki bir sistem kabul görebilmektedir.

- Kaliteli ve Genel Kapsamlı

  Bilgisayarlı çeviri sistemlerinin farklı bir kullanım alanı ise insan tarafından gerçekleştirilen ve emek yoğun bir iş olan klasik çeviri işleminin yükünü azaltmaktır. Bu bağlamda, bilgisayarlı çeviri sisteminin ürettiği sonuçlar daha sonra insan tarafından işlenecek olan ham bilgiyi oluşturmaktadır. Örneğin, bilgisayar tarafından üretilen sonuç bir çevirmen tarafından düzeltilmekte, böylece insan tarafından yapılan çeviri maliyeti azaltılmaktadır. Bu yöntemle elde edilen bir çeviri sonucu kaliteli ve geniş kapsamlı olabilmektedir.

2.1 Bilgi Tabanlı Çeviri Sistemleri

Doğal dil ile ifade edilen bir cümle pek çok bilgi düzeyinde ifade edilebilmektedir. Bilgisayarlı çeviride Vaugouis üçgeni [12] olarak bilinen bilgi düzeylerinin gösterimi Şekil 2.1'de gösterilmektedir.



Şekil 2.1: Bilgi düzeylerinin gösterimi - Vaugouis Üçgeni

2.1.1 Doğrudan aktarım

Doğrudan aktarım en temel ve basit aktarım türüdür. Kaynak dildeki sözcüklerin karşılıklarının bulunması problemi olarak çözülür. Bu yöntemde karşılaşılan en temel sorun kaynak dildeki sözcüğün karşılığının bulunamaması durumudur. Ayrıca kaynak dildeki bir kavramın, hedef dilde aralarında anlam ayrımı olan birden fazla ifade edilişi varsa; bunlardan hangisinin seçileceği de farklı bir sorun teşkil etmektedir. Örneğin, İngilizce'deki *uncle* sözcüğünün Türkçe'de daha özelleşmiş anlamları vardır; *dayı*, *amca* veya *enişte* sözcüklerinden hangisinin seçileceği belirsizdir. Özellikle eklemeli dillerde, kaynak dildeki bir sözcüğün hedef dilde birden fazla anlamı olabilmektedir[1]. Bunlardan hangisinin doğru olduğuna karar vermek için sözcüksel belirsizliğin giderilmesi gerekir. Sözcük bazında aktarım yapan sistemlerde en önemli bileşen aktarım sözlüğüdür. Daha gelişmiş olan aktarımlarda, sözcük yerine sözcük öbeklerinin aktarılması ile bu sözlük genişletilebilmektedir. Bu aktarım düzeyinde biçimbilimsel aktarım da kullanılabilmektedir.

2.1.2 Sözdizimsel aktarım

Bilgi tabanlı aktarım yöntemlerinden birisi de sözdizimsel aktarımdır. Bu aktarım yönteminde kaynak dildeki metnin sözdizimsel analizi ve hedef dildeki sözdizimsel yapıdan sözcüklerle birlikte metin üretimi yapılmalıdır. Çeviri sistemi kaynak dilde sözdizimsel analizi yapılmış ağaç yapısına uygun hedef dildeki ağaç yapısını bulmaya çalışır. Ağaç yapısının oluşturulmasından sonra, doğrudan aktarım yönteminde olduğu gibi bir aktarım sözlüğü ile sözcüklerin hedef dildeki karşılıkları bulunur. Ortaya çıkan sözlüksel belirsizlikler için kaynak metnin çözümlenmesi sırasında anlamsal belirsizlik giderici yöntemler kullanılabilmektedir.

2.1.3 Anlamsal aktarım

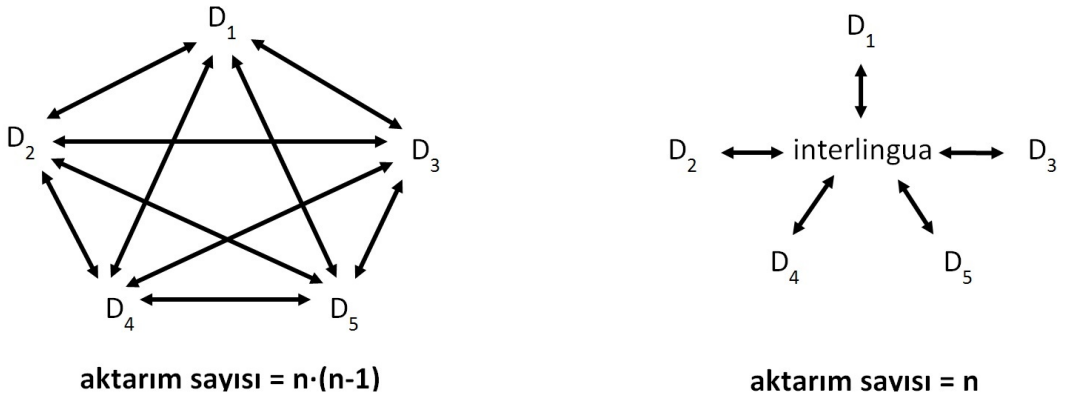
Daha gelişmiş ve daha detaylı analizlere ihtiyaç duyan bir aktarım yöntemi de anlamsal aktarımdır. Sözdizimsel çözümlemesi yapılmış olan cümledeki ayrıştırılan yapılara anlamsal görevlerin yüklenmesi ve bu görevler üzerinden çevirinin yapılmasıdır. Anlamsal çevirisi elde edilen hedef dildeki cümle, sırasıyla sözdizimsel gösterime ve

[1] *Kalemi* sözcüğü bu duruma örnek gösterilebilir: *kalemi* → kale (rook) + iyelik eki *-m* (possessive) + belirtme durumu *-i* (accusative); *kalemi* → kalem (pencil) + belirtme durumu *-i* (accusative)

sözcüksel gösterime dönüştürülür. Bu yöntem, sözdizimsel aktarımında karşılaşılan yapı uyuşmazlıklarını da çözebilmektedir.

2.1.4 Dilden bağımsız anlamsal aktarım

Uluslararası Yardımcı Dil Derneği (IALA-International Auxiliary Language Association) tarafından 1951 yılında tasarlanmış yapay bir dil olan *interlingua* düzeyinde yapılan aktarımdır. İnterlingua, genellikle latin alfabesiyle temsil edilmektedir ve Roman, Cermen dillerinden ortak ve yaygın sözcüklerden, basitleştirilmiş dilbilgisi kurallarından oluşmaktadır. Bu dillerden herhangi birini bilen eğitimli bir kişi tarafından kolaylıkla anlaşılabilecek kavramsal bir dildir. Bilgisayarlı çevirisi yapılacak olan metnin sırasıyla gerekli analiz aşamalarından (biçimbilimsel, sözdizimsel, anlamsal) geçtikten sonra bu gösterime dönüştürülmesi ve hedef dilde yeniden üretilmesi gerekmektedir.



Şekil 2.2: Dilden bağımsız anlamsal düzeyde ve diğer bilgi düzeylerinde gerekli aktarım sayısı

Dilden bağımsız anlamsal aktarım bilgisayarlı metin çevirileri için varılması beklenen hedef noktasıdır. Bu düzeyde aktarım gerekli aktarım işlemlerinin sayısını da oldukça azaltmaktadır. Örneğin, n adet dil arasında kurulacak birebir çeviri sistemleri için, aktarım düzeyi ne olursa olsun, $n \cdot (n-1)$ adet aktarım yapılması gerekir. Her bir aktarım, kaynak dilde analiz ve hedef dilde üretim safhalarını içermektedir. Bu safhalarda kullanılan araçların oluşturulması ve varsa geliştirilmesi gerekliliği, bilgi tabanlı bilgisayarlı çeviri alanında çalışılması gereken pek çok konu olduğunu göstermektedir. Fakat, dilden bağımsız interlingua kullanımı ile çeviri yapıldığında bu sayı n adet aktarıma azaltılabilmektedir. Dilden bağımsız anlamsal aktarım yapan ve

diğer aktarım yöntemlerini kullanan sistemlerin 5 adet dil için örneklemesi Şekil 2.2'de gösterilmektedir. Bu gösterimde kullanılan $\longleftrightarrow$ sembolleri iki dil arasındaki analiz ve üretim aşamalarından oluşan aktarım safhasını bir bütün olarak temsil etmektedir.

2.2 Örnek Tabanlı Çeviri Sistemleri

İlk olarak 1984 yılında Nagao tarafından geliştirilen örnek tabanlı yöntemde, çeviri sistemi birbirinin çevirisi olan iki dildeki paralel cümlelerden *örneksemeyle çeviri* (translation by analogy) yapmayı öğrenir [13]. Varolan çevirilerin, pek çok çeviri probleminin çözümünü içinde barındırdığı düşünülmektedir [14]. Bu nedenle, dilbilimsel kurallar yerine örneklerden öğrenme yoluyla çeviri yapmak bir araştırma konusu olmuştur. Bu yöntem,

- örnekleri parçalara ayırma
- parçaların hedef dile çevrilmesi
- parçalardan sonuç cümlesi üretme

adımlarından oluşmaktadır.

Örnek tabanlı çeviride kaynak dildeki bir sözcük farklı koşullar altında, yani farklı sözcüklerle olan birlikteliklerinde, hedef dile farklı sözcükler olarak çevrilmektedir. Örnek tabanlı çevirinin bu özelliği, *durum temelli akıl yürütme* (case-based reasoning) olarak adlandırılmaktadır. Örneğin, İngilizce'deki *eats* sözcüğü *acid* ve *metal* sözcükleri ile birlikte geçiyorsa, "aşındırmak"; *squirrel* ve *nut* sözcükleri ile birlikte geçiyorsa "yemek" anlamı taşımaktadır.

Squirrel eats nut. $\leftrightarrow$ Sincap fındık yer.
Acid eats metal. $\leftrightarrow$ Asit metali aşındırır.

Örnek tabanlı sistemler, istatistiksel sistemlerden farklı olarak test cümlesini derleminde barındırıyorsa aynı çıktıyı üretmeyi garanti eder. Herhangi bir ön-işleme gerektirmez. Ayrıca derleminde uygun örnekleri bulabildiği sürece düzgün çıktılar üretebilmektedir.

Örnek tabanlı çeviri sistemleri için oluşturulan derlemler özel derlemlerdir ve genellikle birbirinden birer sözcük farklılık gösteren örnek kümelerinden oluşurlar. Böylece sistem alt parçaları daha kolay öğrenebilmektedir.

En yakın cami nerededir? $\leftrightarrow$ Where is the closest mosque?
En yakın müze nerededir? $\leftrightarrow$ Where is the closest museum?

Örneğin yukarıdaki çeviri örneklerinden sistem aşağıdaki kalıpları ve bilgileri öğrenebilir.

En yakın X nerededir? $\leftrightarrow$ Where is the closest X?
cami $\leftrightarrow$ mosque
müze $\leftrightarrow$ museum

2.3 İstatistiksel Çeviri Sistemleri

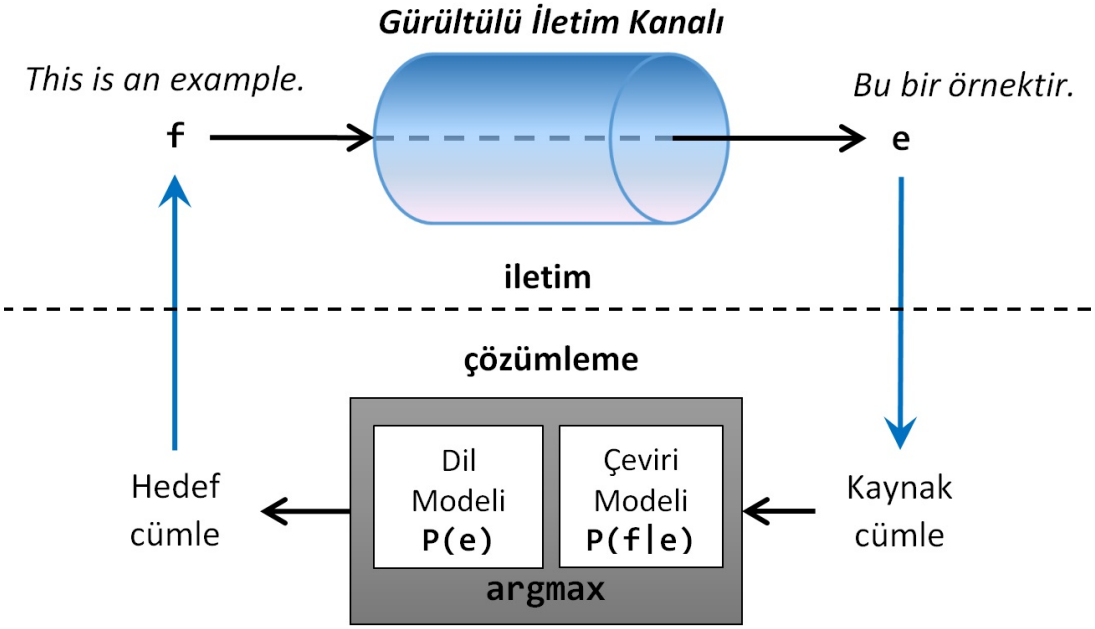
İstatistiksel bilgisayarlı çeviri, doğal dillerin çevirisinin bir makina öğrenmesi problemi olarak modellenmesidir [15]. İlk olarak 1949 yılında Warren Weaver tarafından önerilmiş [16], 1990'ların başında IBM tarafından yürütülen çalışmalarla da istatistiksel bilgisayarlı çevirinin temelleri atılmıştır [17, 18]. İstatistiksel bilgisayarlı çeviri algoritmaları, nasıl çeviri yapacağını insan tarafından oluşturulmuş örneklerden öğrenmektedir. Elektronik ortamda yer alan birbirinin çevirisi olan metinlerin ve bilgisayarların yeteneklerinin artması ile bilgi aktarımı için kurallar geliştiren sistemlerden istatistik bilimine başvuran sistemlere bir yönelme olmuştur. Bunu daha sonraki zamanlarda internetin yaygınlaşması da desteklemiştir. Sözcük öbeği temelli istatistiksel çeviri sistemlerinin geleneksel sözcük temelli istatistiksel çeviri sistemlerinden daha başarılı olması nedeniyle [19], günümüzde araştırmacılar sözcük öbeği temelli yaklaşımlara yönelmişlerdir. İstatistiksel yöntemler, son zamanlarda bilgisayarlı çeviri alanında en çok araştırılan ve en çok tercih edilen yöntemlerdir.

Bilgi tabanlı sistemlerin amacı, hangi bilgi seviyesinde (sözcüksel, sözdizimsel, anlamsal gösterim gibi) bilginin aktarılacağını belirlemek ve bu aktarımın en doğru biçimde gerçekleşmesini sağlamaktır. İstatistiksel sistemler ise sonuçta oluşacak çıktının kalitesine odaklanır. Bilginin hangi yolla ve nasıl aktarılacağı ile ilgilenmez.

Bu noktada istatistik biliminin yardımına başvurur ve 2.1 nolu denklemde görülen koşullu olasılığı maksimize etmeye çalışır.

$$\begin{aligned} \Theta_{en\ iyi} &= arg\ max_e\ p(e|f) \\ &= arg\ max_e \frac{p(f|e) \cdot p(e)}{p(f)} \\ &= arg\ max_e\ p(f|e) \cdot p(e) \end{aligned} \tag{2.1}$$

Bu denklemde görülen ve maksimize edilmeye çalışılan $p(e|f)$ terimi kaynak dildeki bir f cümlesinin hedef dile e cümlesi olarak çevirilme olasılığıdır. Bu denklem Bayes kuralına göre yeniden yazıldığında denklemdeki $p(f)$ olasılığı f cümlesinin görülme olasılığıdır. Fakat bu olasılık denklemde değerlendirilecek bütün durumlar için sabit olduğundan en iyileme denkleminde bulunmasına gerek yoktur. Denklemin düzenlenmiş ve sadeleştirilmiş son halinde yer alan $p(f|e)$ ve $p(e)$ olasılıkları istatistikel sistemlerde sırasıyla ***çeviri modeli*** (ÇM) ve ***dil modeli*** (DM) olarak adlandırılan temel bileşenleri temsil etmektedir.



Şekil 2.3: Gürültülü Kanal Modeli

Bu durum *Shannon Teoremi* olarak da bilinen *Gürültü Kanalı Modeli* (Noisy Channel Model) ile temsil edilir [20]. Bu yaklaşıma göre kanala giren f cümlesi kanaldaki gürültü nedeniyle bozularak e cümlesi olarak kanaldan çıkmaktadır. Problem kanaldan çıkan e cümlesinin aslında ne olabileceğini bulma problemidir. İletim ve çözümleme

olarak adlandırdığımız bu iki durum Şekil 2.3'te gösterilmiştir. Günümüzde pek çok konuşma tanıma sistemi de bu yaklaşımla çalışmaktadır.

Kanaldan çıkan e cümlesi için pek çok olası f cümlesi oluşturulur. Oluşturulan çok sayıda çözüme rağmen, bugün kullanılan pek çok sistem Denklem 2.1'i en iyileştiren tek bir çeviri sonucunu kullanıcılara sunmaktadır. İstatistiksel bilgisayarlı çeviri sisteminin kullanım amacına bağlı olarak diğer olası çeviriler de kullanıcılara sunulabilmektedir. *En iyi n listeleri* (n-best lists) adı verilen bu olası sonuçlar listesi farklı başarı kriterleri ile değerlendirdirildiğinde, sistemin kullandığı iyileştirme algoritmasının her zaman en iyi sonuçları seçmediği de bilinmektedir [21]. Denklemi en iyileyen çevirinin, insan tarafından yapılan değerlendirmelere göre de en iyi çözüm olması sistemin doğru çözümlemeyi bulmadaki yüksek başarısını göstermektedir. Doğru çözümlemeyi bulmak çeviri başarısını artırılmasını sağlamaktadır.

2.3.1 Dil modeli

Dil modellemede asıl amaç $p(e_1^L)$ fonksiyonunun[2] en uygun gösterimini elde etmektir. *Üretici modeller* (generative models) bunun için olasılıksal araçları kullanmaktadırlar. Bu araçlardan birisi olan *zincir kuralı* (chain rule) aşağıdaki gibi formülleştirilmektedir.

$$P(e_1^L) = \prod_{i=1}^{L} P(e_i|e_1^{i-1}) \tag{2.2}$$

Denkleme göre, e_1^L cümlesinin koşullu olasılığı, her biri birer sözcükle ilişkili olan pek çok koşullu olasılığın ürünüdür. Modeli basitleştirmek için yapılan basit bir varsayım ile, e_i sözcüğünün üretilme olasılığının sadece kendisinden önce gelen (preceding) $n-1$ sözcüğe (e_{i-n}^{i-1}) bağımlı olduğu, bunların dışındaki sözcüklerden bağımsız olduğunu gösterebiliriz. Örneğin, x ve y değişkenleri birbirinden bağımsız değişkenler ise, $P(x|y) = P(x)$ olması gerekir. Bu eşitlik y'nin bilinmesinin x'in olasılık dağılımını etkilemediğini söylemektedir. Buna olasılık teorisinde, *koşullu bağımsızlık* (conditional independence) denilmektedir. Dil modellemesi için de cümledeki her bir sözcüğün diğerleriyle birebir bağımlı olmadığı varsayımıyla, sadece kendisinden önce gelen $n-1$ adet sözcüğe bağımlı olduğunu varsayalım. Bu varsayımdaki n

[2] e hedef dildeki L sözcüklü cümleyi temsil ederken, 1 ve L değerleri bu hedef dilin sınırlarını ifade etmektedir. e_1^L ise bu cümlenin 1'den L'e kadar olmak üzere tüm sözcükleridir.

değerinin seçimi oldukça kritiktir, çok büyük seçilmesi performanı düşürürken, çok küçük seçilmesi de bağlamın yakalanmasını ve dili modellemeyi zorlaştırmaktadır. Bu bağımsızlık varsayımına dayalı oluşturulan dil modelinde e_{i-n}^{i} sözcük öbeği *n-gram* olarak adlandırılmaktadır. Denklem 2.2'yi bir n-gram dil modeli olarak yeniden yazarsak;

$$P(e_1^L) = \prod_{i=1}^{L} P(e_i|e_1^{i-1}) = \prod_{i=1}^{L} P(e_i|e_{i-n}^{i-1}) \tag{2.3}$$

eşitliğini elde ederiz.

2.3.2 Çeviri modeli

Çeviri modeli iki dilde birbirinin çevirisi olan paralel eğitim derlemlerinden oluşturulur. 2.1 numaralı denklemde görüldüğü gibi $p(f|e)$ parametresinin modellenmesidir.

Olası her cümlesinin eğitim derleminde yer aldığı mükemmel dünyada, daha önceden sistem tarafından öğrenildiği için her f cümlesinin doğru ve dile uygun bir çevirisi yapılabilir. Fakat gerçek dünyada bir dilde oluşturulabilecek her cümleyi içerecek büyüklükte bir derlem yoktur. Bu nedenle, paralel derlemdeki cümleler küçük *çeviri birimlerine* (translation units) bölünür. Bu sayede, çeviri olasılık dağılımı daha kolay modellenir. Sözcük öbeği temelli istatistiksel sistemlerde bu birimler sözcük öbeklerine karşılık gelir.

Birbirinin çevirisi olan cümleler çeviri birimlerinden oluşur, fakat kaynak dildeki hangi birimin hedef dildeki hangi birime karşılık geldiğini bilinmemektedir. Birimlerin hangilerinin birbiri ile ilişkili olduğunun bilinmesi, yani x kaynak dildeki çeviri birimini, y ise hedef dildeki çeviri birimini temsil etmek üzere $p(x|y)$ parametrelerinin bulunabilmesi için *beklenti maksimizasyonu* (expectation maximization) algoritması kullanılmaktadır. Buna göre, model parametrelerine başlangıç değerleri[3] atanır. Her bir yinelemede karşılaşılan örneklere göre bazı ilişkiler güçlenirken bazıları zayıflar. Buna göre bir çeviri biriminin hangi birimlere, hangi olasılıklarla çevrilebileceği öğrenilir.

[3]En temel yöntemde başlangıç için tüm eşleme ilişkileri için eşdeğer (uniform) olasık değerleri kullanılır.

Modellemenin son aşaması olarak çevirisi yapılan birimler yeniden sıralanarak hedef dilin yapısına uygun hale getirilir. Dillerin cümle içerisindeki sözcük sıralamaları birbirinden farklıdır. Kaynak dildeki cümle yapısının hedef dildeki cümle yapısına nasıl çevrileceği derlemden istatistiksel olarak öğrenilir. Dil modeli *"hedef dilde neyin anlamı olduğunu"* söylerken, çeviri modelinin son aşaması olan yeniden sıralama adımı *"kaynak dildeki cümle yapısının hedef dildeki cümle yapısı ile nasıl eşleştirileceğini"* bildirir.

2.3.3 Aşamaları

İstatistiksel bilgisayarlı çeviri sistemlerinde gerçeklenen temel adımlar ise şunlardır:

- Verinin Hazırlanması

  Derlemdeki farklı yüzeysel biçime sahip her bir sözcük istatistiksel hesaplamalarda farklı semboller olarak temsil edilir. Yani *kediler* ile *Kediler* sözcükleri çeviri problemi uzayındaki farklı noktalardır ve aralarındaki yakın ilişki kestirilemez. Bu nedenle, eğitimden önce tüm sözcüklerdeki harfler küçük harflere çevrilmektedir (lowercasing). Aynı durum nokta (.), virgül (,) gibi sözcüklere bitişik yazılan noktalama işaretlerinin yer aldığı sözcükler için de geçerlidir. Sözcüklere bitişik yazılan bu gibi noktalama işaretlerinin ilgili sözcükten ayrılması ile problem uzayının istenmeyen şekilde büyümesinin ve olasılık dağılımlarının bu durumdan olumsuz etkilenmesinin önüne geçilmiş olur (tokenization). Bu durumda *"yaptı."* sözcüğü *"yaptı"* ve *"."* olmak üzere iki farklı sözcükle temsil edilir.

- Eğitim

  Çift dilli hizalanmış cümleler üzerinde yürütülen eğitim ile istatistiksel olarak birbirinin çevirisi olan sözcük ve sözcük öbeği çiftleri olasılıksal olarak öğrenilir. Ayrıca, bu çevirilerin hedef dile aktarımında uygulanması gereken yeniden sıralama olasılıkları de eğitimle elde edilir.

- İyileştirme

  Eğitimde elde edilen model, sistemin eğitim sırasında görmediği bir iyileştirme test verisi üzerinde çalıştırılır. Model parametreleri, bu veri üzerinde en iyi sonuçları elde edecek şekilde yeniden hesaplanır.

- Çözümleme

  Eğitimde öğrenilen eşleşmelerden faydalanarak kaynak dildeki cümlenin hedef dildeki en iyi temsilinin bulunmasına çalışılır. En iyi çözümlemenin bulunması için çeviri modeli tarafından önerilen en olası çevirilerin seçilmesi yeterli değildir, aynı zamanda çıktının hedef dile uygunluğu da değerlendirilmelidir. Çıktının hedef dilde üretilme olasılığını ise dil modeli ile elde edilir. Çözümlemenin temel amacı çeviri modeli ve dil modeli bileşenlerinden elde edilen bu olasılık değerlerini maksimize etmektir.

- Çevirinin Sonucunun Hazırlanması

  Elde edilen çeviriye, verinin hazırlanması aşamasında yapılan işlemlerin etkilerini geri çevirecek şekilde yeniden büyük/küçük harf bilgisi eklenir ve bitişik yazılması gereken noktalama işaretleri ilgili sözcüğe bitiştirilir.

2.3.4 Faktörlü çeviri

Modern istatistiksel bilgisayarlı çeviri sistemleri veride yer alan sözcükler ve sözcük öbekleri üzerinden öğrenmeyi gerçekleştirir. Bu nedenle, yetersiz veri koşulunda eğer herhangi bir dilbilimsel bilgi (morfolojik, sözdizimsel, anlamsal gibi) kullanılmazsa yetenekleri küçük metin parçacıklarını eşleştirmeye yeter [22]. Fakat dilbilimsel bilginin *ön işleme* (pre-processing) veya *sonradan işleme* (post-processing) aşamalarıyla sisteme dahil edilmesinin başarıyı artırdığı bilinmektedir.

Dilbilimsel bilginin doğrudan çeviri modeline düzgün yapılı ve sağlam bir biçimde dahil edilmesi iki temel neden için istenmektedir:

1. Çeviri modeli sözcüklerin *yüzeysel biçimi* (surface form) yerine örneğin gövdesi gibi daha genel gösterimler üzerinden elde edilebilir. Bu durumda aynı sözcüğün farklı yüzeysel gösterimleri aynı noktaya eşleşeceği için daha güçlü istatistikler elde edilebilir ve eğitim verisinin az olduğu durumlarda ortaya çıkan veri seyrekliği probleminin üstesinden gelinebilir.

2. Çeviri hakkındaki çoğu bakış açısı biçimbilimsel, sözdizimsel veya anlamsal düzeylerde en iyi açıklanabilir. Çeviri modeline bu bilginin sağlanması bu bakış açılarının doğrudan modellenmesini sağlar. Örneğin, sözcük seviyesinde *yeniden*

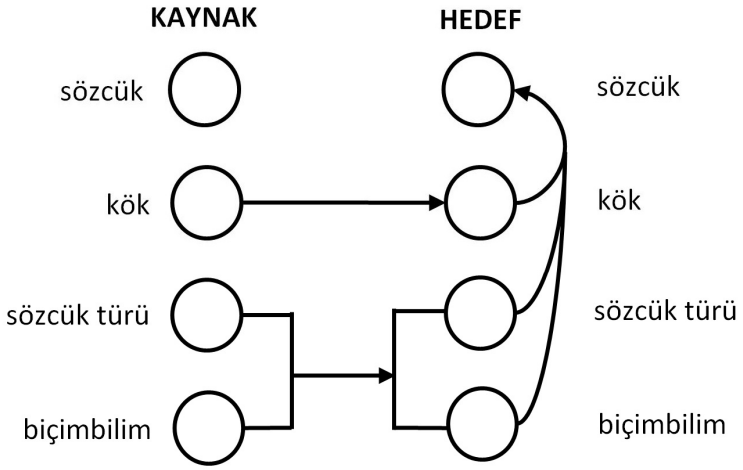
sıralama (reordering) genel sözdizimsel kurallardan, *yerel anlaşma kısıtları* (local agreement constraints) ise biçimbilimden anlaşılmaktadır.

Bu nedenlerle, literatürde dilbilimsel bilginin desteğini de alan *faktörlü çeviri modelleri* (factored translation models) kullanılmaya başlamıştır [23–26]. Bu yeni yaklaşım, sözcük seviyesinde ek işaretlemeler yapmamıza olanak verir. Böylece bir sözcük sistemde tek bir simge olarak değil, farklı seviyeden işaretlemeleri içeren bir faktör vektörü olarak temsil edilir. Örneğin, *öğrenciler* ve *öğrenci* sözcükleri standart istatistiksel çeviri modellerinde birbirinden farklı simgeler olarak temsil edilmektedir ve birbirinden tamamen bağımsız sözcüklerdir. *Öğrenci* sözcüğünün eğitim verisinde bulunması ve öğrenilmiş olması daha önce hiç görülmemiş *öğrenciler* sözcüğünün çevirisinin bilinmesine herhangi bir katkı sağlayamamaktadır. Özellikle Arapça, Almanca, Çekce, Türkçe gibi biçimbilimsel açıdan zengin olan dillerde sözcükler aldıkları eklerle birlikte anlamları değişmese de pek çok yüzeysel gösterime sahip olabilirler. Bu nedenle, örneğin sözcüklerin kökleri ve biçimbilimsel özellikleri ayrı birer bileşen olarak sisteme verilirse, bileşenlerin eşleşmesine dair istatistikler kuvvetlenecek ve daha kolay öğrenilecektir. Kök ve biçimbilimsel özellik olarak çevirisi tamamlanmış sözcükler hedef dil tarafında yeniden üretme yoluyla olması gereken yüzeysel biçime dönüştürülecektir.

Şekil 2.4'de olası bir faktörlü çeviri modelinin örneklemesi bulunmaktadır. Bu örneğe göre bir sözcük, kökü, sözcük türü ve biçimbilimsel bilgilerinden oluşan bir faktör vektörü ile temsil edilmektedir. Bu faktörlerden kaynak dildeki kök hedef dildeki kökle birebir eşleşmekte, sözcük türü ve biçimbilimsel bilgi bir bütün olarak hedef dildeki sözcük türü ve biçimbilimsel bilgi ile eşleşmektedir. Yani kaynak dildeki bir sözcük, hedef dilde üç farklı faktör ile ifade edilmektedir (kök + sözcük türü + biçimbilimsel bilgi). Faktörlü çeviri modellerinde hedef dildeki bu faktörlerden yüzeysel biçimi üretme (generation) aşaması bulunmaktadır. Örnekte kök, sözcük türü ve biçimbilimsel bilgiden yüzeysel biçim oluşturulmaktadır.

2.4 Çeviri Kalitesinin Değerlendirilmesi

Bilgisayarlı çeviriler üzerine pek çok çalışma yapılmaktadır. Yapılan çalışmaların başarısını değerlendirmek ve bu bilgiyle daha başarılı sistemler geliştirebilmek için



Şekil 2.4: Faktörlü çeviri modeli örneği

çeviri sistemlerinin ürettiği çıktıların kalitesini tarafsız bir şekilde değerlendirmek gerekir.

Çeviri kalitesini ve doğruluğunu ölçmek için en bilinen ve basit yol çıktıların insanlar tarafından değerlendirilmesi ve puanlandırılmasıdır. Dile hakim uzmanlar, bir çevirinin kalitesini sisteme verilen cümlenin içerdiği anlamın eksiksiz ve doğru bir şekilde aktarılması ve sonucun hedef dildeki akıcılığı bakımından değerlendirirler. Fakat bu işlem oldukça emek yoğun ve maliyetli bir işlemdir. Ayrıca oldukça uzun zaman gerektirir ve daha önceki değerlendirmelerin bir sonrakine katkısı olmadığından, sistemde yapılan her değişiklikte çıktıların yeniden değerlendirilmesini, aynı zaman ve maliyetin yeniden harcanmasını gerektirir. Bu nedenle, insanların yaptığı değerlendirme pratikte pek kullanılmaz.

Bilgisayarlı çevirinin kullanım alanlarından biri, çıktıların çevirmenler tarafından düzeltilerek uygun hale getirildiği, böylece sıfırdan yapılacak bir çeviri işlemine göre maliyetin azaltıldığı yarı otomatik sistemlerdir. Değerlendirme maliyetlerini de azaltmak için, bu çalışma ilkesine dayanarak, sistemin ürettiği çıktının olması gerekene ne kadar yakın olduğu bir kalite ölçütü olarak kullanılmaktadır. Çevirmenin sistem çıktısı üzerinde yaptığı değişiklikler, çevirmenin harcadığı çabanın bir göstergesi olarak çevirinin kalitesini söylemektedir. Bu değişiklikler, çevirmenin kaç tuşa bastığı, ne kadar zaman harcadığı ile ölçülebileceği gibi, çevirmenin en uygun hale getirdiği çeviri ile sistemin ürettiği çeviri arasındaki farklar gözetilerek de yapılabilir. Aradaki değişimin bulunması için yine harf veya sözcük bazında *en kısa değişim uzaklığı* (minimum edit distance) algoritması ile ölçüm yapılabilir.

İnsanlar tarafından yapılan yoğun iş gücü gerektiren bu değerlendirme yöntemleri yerine, referans çeviriler yardımıyla uygulanan otomatik kalite değerlendirme yöntemleri de bulunmaktadır. Bunlardan günümüzde en sık kullanılanları BLEU/NIST, sözcük hata oranı (word error rate), F ölçütü ve METEOR'dur.

2.4.1 Sözcük hata oranı

Çok basit bir ölçüt olan sözcük hata oranı, konuşma tanıma ve bilgisayarlı çeviride sıklıkla kullanılmaktadır. Levenshtein uzaklığından türetilen bu ölçütte *sesbirim* (phoneme) yerine sözcükler üzerine hesaplama yapılmaktadır. Aynı sistemin çeşitli iyileştirmelerinin yanında birbirinden bağımsız farklı sistemleri değerlendirmek için de iyi bir metriktir. Fakat sistemin hatalarını anlamada yardımcı olamamaktadır, bu nedenle hatanın kaynağının bulma ve hataya odaklanma gerektiren durumlar için iyileştirilmesi gerekmektedir. Farklı uygulama alanları için farklı versiyonları bulunan sözcük hata oranı temel olarak aşağıdaki gibi hesaplanmaktadır.

$$\textit{Sözcük Hata Oranı} = \frac{y+s+e}{n} \tag{2.4}$$

y: yerine koyma yoluyla değiştirilen sözcük sayısı

s: silinen sözcük sayısı

e: eklenen sözcük sayısı

n: referanstaki toplam sözcük sayısı

2.4.2 BLEU/NIST

IBM tarafından önerilen bu yöntem, sistem çıktısının daha önceden çevirmenler tarafından oluşturulmuş n adet referans çeviriyle olan benzerliğini ölçmektedir [27]. Benzerlik sözcük ve sözcük öbeklerinin eşleşmesi ile ölçülür. Temelde *kesinlik* (precision) hesabına dayananır. Kesinlik hesabı, aday cümlede yer alan ve aynı zamanda referans cümlede/cümlelerde de yer alan toplam sözcük (unigram) sayısının aday cümledeki toplam sözcük sayısına bölünmesiyle elde edilir. Fakat bu hesaplama, aşağıdaki 1 numaralı örnekte olduğu gibi, çeviride bulunması gereken sözcüklerin tekrarından oluşan ve aslında kötü bir çeviri olan adayların yüksek puan almalarına

neden olmaktadır. Bu nedenle, *değiştirilmiş n-gram kesinliği* olarak adlandırılan 2.5 numaralı denklemdeki p_n değeri ölçümlerde esas alınır.

$$p_n = \frac{\sum_{C \in Adaylar} \sum_{Ngram \in C} Adet_{bulunan}(Ngram)}{\sum_{C' \in Adaylar} \sum_{Ngram' \in C'} Adet(Ngram')} \tag{2.5}$$

Örnek 1:

Aday: `bir bir bir bir bir bir bir.`

Referans 1: `Bir kedi bir yumağı çeviriyor.`

Referans 2: `Yumağı bir kedi döndürüyor.`

Bu örnekte standart sözcük kesinliği (unigram precision) 7/7 iken, daha doğru değerlendirme sağlayan değiştirilmiş n-gram kesinliği 2/7'dir[4]. Eşleşen n-gramların sayısı hesaplanırken, her bir sözcüğün eşleşme sayısı en fazla herhangi bir referansta eşleştiklerine eşit olarak alınmıştır.

Örnek 2:

Aday: `herhangi bir.`

Referans 1: `O gördüğüm herhangi bir çocuktan farklıydı.`

Referans 2: `Gördüğüm herhangi bir çocuk gibi değildi.`

2 numaralı örnekte görülen durumda ise oldukça kısa olan aday cümle değiştirilmiş n-gram kesinliği ile 2/2 unigram kesinliğe, 1/1 bigram kesinliğe sahiptir. Tam tersi olarak, oldukça uzun olan aday cümle farklı referanslardan kaynaklanan pek çok eşleşmeye sahip olabilir, fakat bu durum ilgili adayın kalitesini değil, aksine zayıflığını gösterir. Bu gibi durumları cezalandırmak için literatürde kullanılan *gerigetirim* (recall) yöntemi hesaplamada birden fazla referans kullanılabildiği için uygun değildir. Bu nedenle, aday cümlenin BLEU puanı aşağıda görüldüğü gibi *uzunluk cezası* (brevity penalty) adı verilen bir katsayı ile ağırlıklandırılmaktadır. Bu ağırlıklandırma derlem bazında uygulanan bir cezalandırma yöntemidir. İlgili referans cümlelerin ortalama uzunluğuna göre cümle seviyesinde cezalandırmak kısa cümlelerin çok sert bir şekilde cezalanmasına neden olabilmektedir. Cümle seviyesinde biraz esneklik sağlamak için derleme genel bir ceza uygulamanır. Bunun için önce aday derlemin toplam uzunluğu c ve *etkin referans uzunluğu* (effective reference length) r hesaplanır.

[4]Hesaplamaya etkisi olan sözcüklerin altları çizilmiştir.

Etkin referans uzunluğu, test kümesindeki aday cümlelerin en çok uyum gösterdiği referans çeviri uzunluklarının toplamıdır.

$$uzunluk\ cezası = \begin{cases} 1, & \text{eğer } c < r \\ e^{(1-r/c)}, & \text{eğer } c \leq r \end{cases} \tag{2.6}$$

Uzunluk cezası katsayısı ile ağırlıklandırılan BLEU puanı hesabı 2.7 nolu denklemde görüldüğü gibidir. Temel olarak, test derlemindeki değiştirilmiş n-gram kesinliklerinin geometrik ortalamasının bir uzunluk cezası ile çarpılmasından elde edilir. [0,1] arasında olabilen BLEU değerinin 1'e yakın olması aday çevirinin referanslardan en az biriyle oldukça uyuşması, 0'a yakın olması aday çevirinin referanslardan hiç biriyle uyuşmaması anlamını taşımaktadır.

$$BLEU = uzunlukcezası \cdot exp\left(\sum_{n=1}^{N} w_n \cdot logp_n\right) \tag{2.7}$$

Benzer bir yaklaşım olan NIST hesabında da BLEU'dan farklı olarak geometrik ortalama yerine aritmetik ortalama kullanılır [28]. Eşleşen n-gramları sıklıklarına göre değerlendiren bu yöntemde, çok sık geçen bir bigram ile çok nadir geçen bir bigramın değeri aynı değildir. Derlemde nadir geçen n-gramlar daha kıymetli olmaktadır.

Otomatik değerlendirme araçları üzerine yapılan bazı çalışmalar ise BLEU hesabının olumsuz olabilecek yanlarını ortaya koymaktadır. Buna göre, farklı yapıdaki sistemleri (istatistiksel ile kural tabanlı gibi) değerlendirirken BLEU puanlamasının güvenilir olmayabileceği gösterilmiştir [29]. Ayrıca sadece BLEU puanı artışına göre iyileştirilen sistemlerde, insan gözüyle yapılan değerlendirmeler sonucunda, aslında çeviri kalitesi açısından bir iyileşmenin garanti edilemeyeceği görülmüştür [29]. BLEU hesaplaması aday n-gramın referans n-gramlarla birebir örtüşmesine göre ölçülmektedir, oysa dilbilimsel bilgi kullanılarak daha detayı analizler yapılabilir ve buna göre sistem iyileştirilebilir. Özellikle biçimbilimsel açıdan zengin dillerde büyük fayda sağlayacak dilbilimsel analizler için BLEU hesabının bu eksikliğini dikkate alarak geliştirilen uygulamalar bulunmaktadır [30]. Olumsuz tarafları da bulunmasına rağmen, insan gözüyle yapılan değerlendirmelerin maliyeti düşünüldüğünde, BLEU hesabı benzer sistemlerin kıyaslamasında kabul gören ve yaygın olarak kullanılan bir ölçüttür.

2.4.3 F ölçütü

F ölçütü (F-measure), kesinlik ve gerigetirim değerlerinin harmonik ortalaması olan bir doğruluk ölçüsüdür. *Bilgi çıkarımı* (information retrieval) alanında yaygın olarak kullanılmaktadır. F ölçütü bilgisayarlı çeviri dışında da doğal dil işlemenin çeşitli alanlarında kullanılan basit ve temel bir performans değerlendirme kriteridir.

$$F = \frac{2 \cdot kesinlik \cdot gerigetirim}{kesinlik + gerigetirim} \tag{2.8}$$

$$kesinlik(A|R) = \frac{|A \cap R|}{|A|}, \qquad gerigetirim(A|R) = \frac{|A \cap R|}{|R|} \tag{2.9}$$

A: Aday

R: Referans

F ölçütünün bilgisayarlı çeviriler için tanımlanmasındaki temel sorun kesinlik ve gerigetirim değerlerinin hesaplanmasında kullanılan aday ve referans arasındaki keşisim kümesine karar vermektir. En uygun çözüm, ikisi arasındaki en uzun eşleşmelerin yer aldığı kümeyi bulmaktır. Hatta yapılan çalışmalar, bu basit ve bilgisayar bilimlerinde sık kullanılan değerlendirme ölçütünün bilgisayarlı çevirilerin performansının değerlendirilmesinde BLEU puanından daha güvenilir olabileceğini göstermektedir [31].

2.4.4 METEOR

BLEU ölçütünün eksiklerini kapatmak için tasarlanan METEOR ölçütü kesinlik ve gerigetirim değerlerinin ağırlıklı harmonik ortalamasıdır [32]. Gerigetirim değerinin bilgisayarlı çeviride kullanılan otomatik ölçütler için daha önemli olduğu bilinmektedir [33]. Bu nedenle, kesinlik ve gerigetirim değerlerinin harmonik ortalaması alınırken aşağıdaki formülde görüldüğü gibi gerigetirim değeri kesinliğe oranla 9 kat daha fazla ağırlıklandırılmıştır.

$$F_{ort} = \frac{10 \cdot kesinlik \cdot gerigetirim}{kesinlik + 9 \cdot gerigetirim} \tag{2.10}$$

Ayrıca bu ölçüt, ancak derlem seviyesinde sonuç verebilen BLEU ölçütünden farklı olarak cümle veya metin parçacıkları seviyesinde de değerlendirme yapabilmekte ve insan değerlendirmelerine daha yakın sonuçlar vermektedir. METEOR ölçütü diğer değerlendirme ölçütlerinde bulunmayan dilbilimsel süreçlerden de faydanarak sadece sözcük eşleşmelerine göre değil, aynı zamanda gövdelerin veya eşanlamlıların da eşleşmelerini değerlendirerek sonuç verebilmektedir.

3. ALAN UYARLAMASI

Doğal diller pek çok açıdan değişiklik gösterir. İlk farklılık dilin zaman içerisindeki değişiminden kaynaklanır [34]. Dil ihtiyaçlara göre şekillenen, kendini yenileyen canlı bir yapıdır. Bazı kavramlar zaman içinde önemini yitirip kullanılmazken, ihtiyaç ile birlikte yeni terimlerin ve kavramların tanımlanmasına da gerek duyulur. Örneğin, bilgisayarların hayatımıza girmesi ile daha önce var olmayan *bilgisayar* sözcüğü dilimizde yer almaya başlarken, bazı sözcükler de dilimizde silinmiş ya da önemini yitirmiştir.

İkinci olarak, farklı alanlardaki metinlerin sözcük birlikteliği istatistikleri birbirinden farklıdır. Örneğin, finans haberlerini içeren metinlerde geçen *faiz oranı* sözcükleri çocuk masallarını içeren metinlerde birlikte yer almazlar. Bazı anlamı belirsiz olan sözcükler ise bulunduğu alana göre anlam kazanabilir. Örneğin, *kale* sözcüğü spor haberlerinde *"takımla oynanan bazı top oyunlarında topun sokulmasına çalışılan yer"* anlamını taşırken, tarih belgelerinde *"düşmanın gelmesi beklenilen yollar üzerinde, askerî önem taşıyan şehirlerde, geçit ve dar boğazlarda güvenliği sağlamak için yapılan kalın duvarlı, burçlu, mazgallı yapı"* anlamında kullanılmaktadır.

Üçüncü neden ise kişilerin kulandıkları dilin yapısını yazdıkları yazının amacına göre belirlemeleridir. Resmi bir makama yazılan yazının dili ne kadar kurallı ve resmi ise, bir arkadaşa yazılan e-postanın dili o kadar konuşma diline yakın ve kuralsız olabilir.

Bir diğer neden, kişilerin kullandıkları dilin, içinde bulundukları sosyo-ekonomik durum ve ruh haline bağlı olarak farklılık göstermesidir. Bu etki daha çok konuşma diline yansıyor olsa bile, istatistiksel çeviride kullanılan eğitim verisinin kaynağına bağlı olarak yazılı sistemleri de etkileyebilir.

İBÇ için kullanılan verinin sistematik ve kapsayıcı biçimde artırılmasının çeviri kalitesini artıracağı bilinmektedir. Fakat veriyi limitsiz bir şekilde artırdığımızı varsayarsak sözcüklerin bazı ender kullanılan anlamları aynı sözcüğün farklı anlamları tarafından ezilecek ve çeviri sistemi tarafından seçilemeyecek kadar düşük olasılıklara

sahip olacaktır. Bu durum da alana özgü ifadelerin, terimlerin seçilmesini engelleyebileceği için alan uyarlamasının gerekliliğini göstermektedir.

Tüm bu nedenler dolayısıyla, daha kaliteli çözümleme yapabilmek için belirsizliklerin sözcüksel, sözdizimsel ve anlamsal özelliklerinin belirlenmesi ve böylece belirsizliklerin giderilmesi için alan uyarlaması istatistiksel çeviri sistemlerini iyileştiren bir etmendir.

Bilgisayarlı çeviride alan uyarlaması motivasyonunu sağlayan bulgu, genel çeviri sistemleri farklı alanlardan cümleleri ortalama bir kalite ile çevirebilirken, belirli bir alanda eğitilmiş bir sistemin giriş cümleleri bu alandan olduğu sürece daha yüksek kaliteli çeviriler yapabilmesidir. Fakat alan uyarlaması için standart bir uygulama yoktur. İstatistiksel bilgisayarlı çeviride alan uyarlaması yaklaşımları; ayrık ve farklı alanlara özgü sistemlerin birleştirilmesi, ya da sadece çeviri modeli bileşeninin veya dil modeli bileşeninin uyarlanması ile gerçeklenmektedir.

3.1 Alana Özgü Veri ile Uyarlama

İstatistiksel bilgisayarlı çeviri yöntemleri veriye bağımlı ve veriden öğrenen yöntemlerdir. Bu nedenle, alan uyarlaması için de en etkili ve en basit yöntem ilgili alana özgü verileri kullanarak bir sistem oluşturmaktır. Alana özgü çeviri modeli ve alana özgü dil modeline sahip sistemlerin başarılı sonuçlar verdiği gösterilmiştir [4].

Alana özgü sözcükler, kalıp ifadeler, sözcük sıralamaları ve ifade ediş biçimleri gibi belirleyici özellikleri barındıran veri, sistemin eğitimi ve iyileştirilmesi için kullanıldığında, sistem bu alana daha uygun çeviriler üretebilmektedir. Örneğin, İngilizce *bank* sözcüğü Türkçe'ye finans alanında *banka* olarak çevrilirken, doğa güzelliklerinden bahseden bir seyahat rehberinde *göl kıyısı* olarak çevrilmelidir. Türkçe *kale* sözcüğü de kullanıldığı alana göre İngilizce'de farklı sözcüklere çevrilmelidir (castle, rook, goal).

Alana özgü veri ile uyarlama, güvenilir ve başarılı olmasına rağmen, paylaşımda olan alana özgü verilerin yetersiz olması zayıf yönüdür. Çift dilli paralel derlemleri oluşturmak maliyetli ve zaman gerektiren işlemlerdir. Bu nedenle, paralel derlemler genellikle alan gözetmeksizin toplanan ve genel diye nitelendirdiğimiz verilerden oluşturulmaktadır. Yetersiz veri ile istatistiksel bilgisayarlı çeviri sistemi oluşturmak

ise test cümlelerini temsil edememe riski taşımaktadır. Cümle içerisinde daha önce hiç görülmemiş bir sözcük ile karşılaşıldığında, sistem bunun için herhangi bir çeviri opsiyonu üretemeyecek ve test cümlesinde bulunduğu haliyle bırakacaktır. Bu durum, genel alanda eğitilmiş sistemin üretebileceği herhangi bir opsiyondan da mahrum kalmak anlamına gelmektedir. Bu nedenle alana özgü veri ile uyarlama sağlandığında bu risk göz önüne alınmalı ve test edilmelidir. Alana özgü paralel derlemlerin amaca uygun olarak yeterli olması durumunda bu yöntem uygulanmalıdır.

3.2 Dil Modeli ile Uyarlama

Dil modeli hedef dili ne kadar iyi temsil ederse, çeviri modelinin önerdiği çeviri opsiyonlarından oluşan cümlenin dile uygunluğu o kadar doğru tespit edilebilir (bknz. Denklem 2.1). Bu nedenle, iyi bir dil modeli kaliteli çeviri anlamına gelmektedir.

Dil kullanıldığı alana bağlı olarak farklılıklar gösterebilmektedir. Örneğin, teknik dokümanlarda edilgen fiiller kullanılırken, e-posta, internet günlüğü gibi resmi olmayan dokümanlarda ise etken fiil yapısı kullanılır. Ayrıca seçilen sözcükler kullanılan alana özgü olabilmekte ve sözdizimsel yapı değişiklik gösterebilmektedir. Tıp alanında yazılmış bir metinde bolca Latince terim bulunurken, seyahat kitaplarından alınmış bir metinde bolca yer ve mekan isimleri bulunmaktadır. Bu nedenle, çeviri sisteminin kullanılacağı alana özgü dil modelleri alan uyarlamasında etkili yöntemlerden birisidir.

En basit yöntem belirli bir alandan seçilmiş verileri kullanarak o alana özgü dil modeli oluşturmak ve ilgili alana ait çeviri sistemlerinde bu modeli kullanmaktır. Dil modelini oluşturacak hedef dildeki verilerin elde edilebileceği çift dilli veriler genellikle yeterli olmadığı için dil modelleri genel kapsamlı ve büyük tek dilli veriler üzerinden oluşturulur. Alana özgü bu tek dilli verilerin miktarını artırmak için çeşitli yöntemler denenmektedir. Bilgi elde etme yöntemleri ile yeni alanlara ait dokümanların bulunması ve bunlarla dil modelleri oluşturulmasına yönelik çalışmalar vardır [35, 36]. Bunun yanında yeni alanlarda veri toplamak için, varolan çeviri sistemlerini kullanan çalışmalar da yapılmıştır [37]. Varolan sistemlerin ürettiği sonuçlar, çeşitli sorgularla benzer cümleleri bulmak için kullanılmaktadır.

Belirli alanlara ait verilerin toplanmasındaki zorluğun yanında bunların nasıl kullanılacağı da kesinlik kazanmış değildir. Uygulanan yöntemler arasında ilgili alanı temsil eden veriden elde edilen alana özgü dil modeli ve genel bir dil modelinin lineer aradeğerleme ile birleştirilmesi yer almaktadır [38]. Bu yöntemde, dil modelleri Denklem 3.1'de olduğu ağırlıklandırılmakta ve alana özgü dil modeli önceliklendirilerek alana daha uygun sonuçların elde edilmesi sağlanmaktadır.

$$\begin{aligned} p_{aradeğerleme}(e) &= \lambda_1 \cdot p_1(e) + \lambda_2 \cdot p_2(e) \\ \lambda_1 &+ \lambda_2 = 1 \\ p_1 &: \textit{genel dil modeli} \\ p_2 &: \textit{alana özgü dil modeli} \end{aligned} \tag{3.1}$$

Farklı dil modellerini bir arada kullanarak alana uyum sağlamak için yapılan bir diğer yöntem, geri çekilme yöntemi ile dil modellerini birleştirmektedir. Denklem 3.2'te görüldüğü gibi e n-gramının alana özgü dil modelinde geçmediği durumlarda, genel dil modeline bir λ cezalandırma katsayısı ile başvurulmaktadır.

$$p_{\textit{geri çekilme}}(e) = \begin{cases} p_{\textit{alana özgü}}(e), & \textit{e görülmüşse} \\ \lambda \cdot p_{genel}(e), & \textit{aksi takdirde, } \lambda\textit{: katsayı} \end{cases} \tag{3.2}$$

Katsayı temelli bu yöntemlerle birlikte maksimum düzensizlik kriterine göre eğitilmiş üstel modeller ve minimum ayrımsama bilgisi hesabı ile oluşturulan modeller de kullanılmaktadır [39].

3.3 Çeviri Modeli ile Uyarlama

İstatistiksel bilgisayarlı çeviride temel sorunlardan biri kaynak dildeki sözcük veya sözcük öbeklerinin eğitim verisinde yer almamasıdır. Eğitim kümesinde bulunanlarınsa çok seyrek görülmüş olması veya ilgili sözcükler eğitim verisinde bulunsa bile, bunların birlikteliğinden oluşan sözcük öbeklerinin bulunmaması da çeviri kalitesini olumsuz etkilemektedir. Daha uzun olan sözcük öbeklerinin çeviri modelinde bulunamaması durumunda daha kısa olanlara ve hatta sözcüklere başvurulmaktadır. En kötü şartlarda sözcüklerin de bulunamaması ile bu bilinmeyen sözcükler sahip oldukları yüzeysel biçim ile hedef dile aktarılmaktadırlar. Bu durum,

üretilen çevirilerin çözümleyicinin en iyileştirme denklemindeki parametrelerinden olan dil modeli tarafından da cezalandırılması anlamına gelmektedir.

Seyrek veri problemine bağlı olarak ortaya çıkan bu durum, sistemin eğitildiği örneklere benzemeyen, farklı alanlardan gelen cümleler bu sistemde çevrilmek istendiğinde de oluşmaktadır. Kullanılabilen çift dilli verilerin belirli alanlarla ve belirli dillerle sınırlı olması (örneğin, Avrupa Parlamentosu ve Birleşmiş Milletler çok dilli dokümanları), ayrıca bunların dönemsel gelişmelere, yeni eğilimlere ve terimlere hizmet edememesi çeviri modeli üzerinde alan uyarlamasının gerekliliğini göstermektedir. Bu nedenle, var olan çift dilli paralel veri miktarını özellikle ilgili alanlardaki verilerle artırmak ve çeviri modelini bu alanlarda zenginleştirmek için çalışmalar yapılmaktadır. Çift dilli verileri elde etmek maliyetli olduğu için, literatürdeki çalışmalar tek dilli verilerden faydalanarak eğitim verisini zenginleştirmeye yönelmiştir [40–43].

Yapılan çalışmalar alana özgü paralel verinin doğrudan çeviri modeli oluşturmak için kullanılmasının ve var olan genel çeviri modelini zenginleştirmek için kullanılmasının çeviri kalitesini artırdığını göstermektedir [4, 44].

3.4 Faktörlü Gösterim ile Uyarlama

Her alan için farklı çeviri sistemleri oluşturmak ve bunları bir araya getirmek yerine araştırmacılar daha genelleştirilebilir ve uygulaması kolay yöntemler arayışındadırlar. Farklı sistemleri bir araya getirmek, genel sistem çıktılarının, çeviri kayıplarının yanında sınıflandırma hatalarından da olumsuz etkilenmesine neden olmaktadır. Bu nedenle, her alan için farklı modeller oluşturmak yerine, bir modeli farklı alanlarda kullanılabilir hale getirmek için yapılan bir çalışmada [44] sözcük öbeği çiftlerinin çıkarılması aşamasında alan bilgisi eklenmiştir. Yani her bir çeviri çifti, hangi alandan çıkarıldığı bilgisi ile birlikte sözcük öbeği tablosuna kaydedilmiştir.

Bu yaklaşımı daha gelişmiş olan ve dilbilimsel bilginin de yardımını alan bir çeviri modeli çatısı olan faktörlü çeviri modelleri ile gerçeklemek mümkündür. Faktörlü gösterimde dilbilimsel bilgiyi kullanmak yerine alan bilgisinin kullanılması çeviri modelinin hangi sözcük öbeği çiftinin hangi alandan çıkarıldığını bilmesini ve sistemin uyumlu çeviri çiftlerini seçmesini sağlayabilir. Bunun için paralel derlemdeki her bir

sözcüğün alan bilgisi ile zenginleştirilmesi yeterlidir. Aşağıdaki örnekte görüldüğü gibi *kalesini* sözcüğünün hangi alana ait veriden elde edildiğinin bilinmesi uygun anlamın bulunmasını (rook, goal) ve buna göre çevrilmesini kolaylaştırmaktadır.

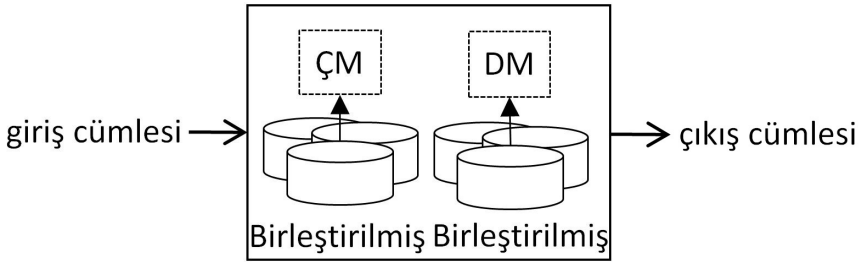
- **Kalesini**|Oyun **ikinci**|Oyun **hamlede**|Oyun **kaptırdı**|Oyun .|Oyun
- **Rakip**|Spor **takım**|Spor **kalesini**|Spor **iyi**|Spor **koruyamadı**|Spor .|Spor

4. İNGİLİZCE'DEN TÜRKÇE'YE ÇOKLU ALAN UYUMLU İSTATİSTİKSEL BİLGİSAYARLI ÇEVİRİ

Bu bölümde, tez çalışması kapsamında İngilizce'den Türkçe'ye çeviri yapan ve birden fazla alana uyum sağlayabilen bir istatistiksel bilgisayarlı çeviri oluşturmak için uygulanan yöntemler tanıtılmaktadır. Oluşturulan bu geniş kapsamlı sistemler ALAN UYARLAMASI bölümünde anlatılan alan uyarlaması yöntemlerini kullanmaktadırlar. Bu yöntemler ile elde edilen sonuçlar UYGULAMA VE SONUÇLAR bölümünde anlatılmaktadır.

4.1 Yalın Sistem

Adil bir değerlendirme yapabilmek için alan uyarlaması yöntemlerini birbiri ile kıyaslamak yerine bir referans sisteme göre değerlendirmek gerekmektedir. Elde edilecek alan uyumlu sistemlerin geleneksel sistemlere katkısı böyle bir temel sistem ile ölçülmelidir. En sezgisel yöntemle, erişilebilen tüm paralel veri bu amaçla tek bir sistemi eğitmek için kullanılmıştır. Oluşturulan istatistiksel bilgisayarlı çeviri sistemi alan uyarlaması olmadan, geleneksel istatistiksel yöntemlerle elde edilebilecek kapsamı en geniş çeviri sistemidir. Bu referans sistem Şekil 4.1'de gösterilmektedir.

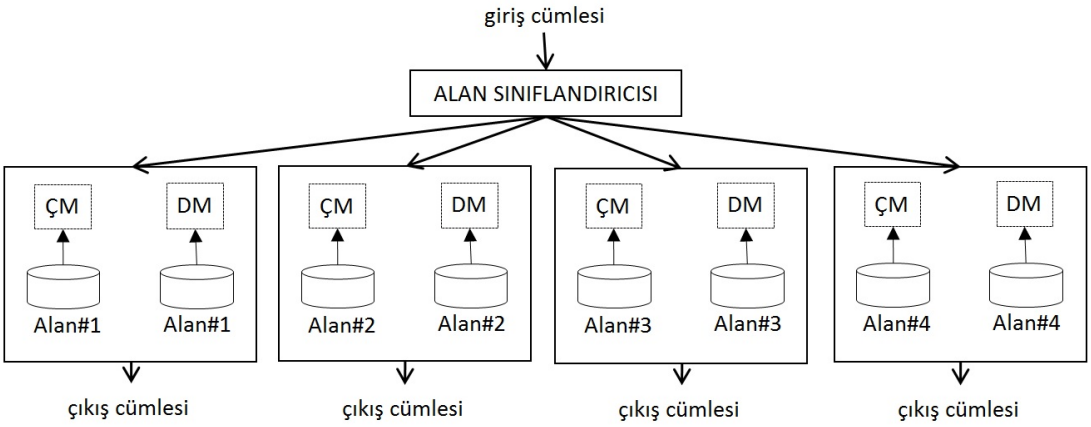


Şekil 4.1: Yalın sistem

Bu sistem farklı alanlara ait verilerin bir araya getirilmesi ile çalışmanın genelinde *birleştirilmiş veri* diye adlandırılan veri üzerinde eğitilmiş ve iyileştirilmiştir. Eğitim verisindeki çeşitlilik, bu referans sistemi kabul edilebilir bir çeviri kalitesi ile genel amaçlı çevirilere uygun hale getirmektedir.

4.2 Alana Özgü Sistemlerin Birleştirilmesi

Birden fazla alana uyum sağlayabilen bir istatistiksel bilgisayarlı çeviri sistemi oluşturmanın en basit yolu, her biri kendi alanında ayrı ayrı eğitilmiş belirli alanlara özgü farklı İBÇ sistemlerinin bir metin sınıflandırıcı ile bir araya getirilmesidir. Alana özgü istatistiksel bilgisayarlı çeviri (AÖ-İBÇ) sistemi olarak adlandırılan bu yöntem Şekil. 4.2'de görülmektedir.



Şekil 4.2: Alana özgü sistemlerin birleştirilmesi

Alana özgü sistemlerin çeviri modelleri ve dil modelleri alana özgü veri üzerinde eğitilmekte ve iyileştirilmektedir. Uygulama sırasında, metin sınıflandırıcısı kaynak dildeki giriş cümlesinin[1] hangi alana ait olduğuna karar verir ve ilgili alanda eğitilmiş olan alana özgü sistemi seçer.

4.2.1 Genelleme ile iyileştirme

Bir önceki bölümde tanıtılan alana özgü sistemlerin bir sınıflandırıcı yardımıyla birleştirilmesi alana özgü sistemlerin yetenekleri ile sınırlıdır. İlgili alan için kullanılan veri o alanı temsil etmeye yetecek düzeyde değil ise oluşturulan İBÇ sistemlerinin çeviri yetenekleri azalır. Bu çalışmada, alana özgü sistemlerin bileştirilmesi ile uyarlama sağlanırken aynı zamanda bu eksikliği gidermek için alternatif çözümleme yolları [24] kullanarak iyileştirilmiş bir model önerilmektedir. Bu yöntem ile alana özgü paralel veride hiç geçmediği için çeviri modelinde de bulunmayan ve dolayısıyla çevirisi yapılamayan girdilerin genel kapsamda çevirisinin yapılması

[1] Bu sınıflandırma paragraf veya doküman seviyesinde de yapılabilir.

hedeflenmektedir. Yönlendirilen alana özgü İBÇ sisteminde aranan girdi için hiç bir çeviri opsiyonu bulunmadığında, alternatif çözümleme yolları ile genel bir çeviri modeline başvurulur ve eğer mümkünse çözümlemesi bu genel model aracılığıyla yapılır. Bunun için alana özgü çeviri modeli ve birleştirilmiş veri çeviri modeli olmak üzere iki çeviri modeli *alternatif çözümleme yolu ve geri çekilme modeli*[2] (alternative decoding paths and back-off model) ile bir araya getirilmiştir. Önerilen birleştirilmiş sistem öncelikle alana özgü çeviri modelinde arama yapar. Eğer olası bir çeviri bulamamışsa daha sonra genel çeviri modelinde uygun çeviriyi arar. Genel çeviri modeli olan ikinci model sadece ilk modelde bulunamayan bilinmeyen sözcük ve sözcük öbekleri için bir geri çekilme modelidir.

Literatürde, dil modelleri alan uyarlaması için çeviri modellerinden daha etkili bulunmuştur [45]. Bu nedenle, bu yöntemde sadece genel kapsamlı çeviri modeli ile alana özgü dil modeli geri çekilme yöntemiyle birleştirilmemiş, aynı zamanda genel kapsamlı dil modelinden de faydalanılmıştır. Alana özgü dil modeli ve genel dil modeli eşit ağırlıklı olarak kullanılmıştır.

Alternatif çözümleme yolları, diğer tekli çeviri modellerinden teorik olarak daha başarılıdırlar. Çünkü birincil model bir sözcük öbeği için hiç bir çeviri opsiyonu sağlamazsa, genel sistem ek opsiyonlar öğrenmek için bir şansa daha sahip olur. Bu yöntemle, alana özgü ve genel kapsamlı çeviri modellerini birleştirerek alan uyarlaması için de aynı avantajın sağlanması hedeflenmektedir.

4.3 Alan Bilgisinin Faktör Olarak Kullanılması

Faktörlü istatistiksel bilgisayarlı çeviri, çeviri işlemine ek bilgi dahil edebilmemize olanak tanır. İstatistiksel bilgisayarlı çeviride faktör kullanımına, biçimbilimsel, sözdizimsel veya anlamsal bilgi gibi ek dilbilimsel özellikleri kullanarak çeviri kalitesini artırma fikri yön vermiştir. Türkçe için yapılan çalışmalar bu yöntemin başarılı olduğunu göstermektedir [46]. Bunun gibi, alan bilgisinin de ek bilgi olarak sisteme verilmesinin başarıyı artırması beklenmektedir. Alanın bilinmesi sözcük seçimini, sözcük sıralamasını vb. çeviriyi etkileyen etmeni değiştirebilir. Bu nedenle, Moses [47] sisteminin faktörlü çeviri çatısında olduğu gibi alan bilgisi de bir faktör

[2] http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc21
Son erişim: Haziran 2014

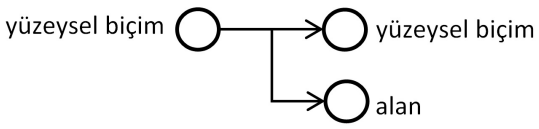
Çizelge 4.1: Alan bilgisinin faktör olarak kullanıldığı çeviri örnekleri

P(interest|faiz, Haber) > P(interest|ilgi, Haber) İngilizce: The credit interest rate is announced. Türkçe: Kredi faiz oranları açıklandı. P(interest|faiz, Altyazı) < P(interest|ilgi, Altyazı) İngilizce: The child didn't show an interest in his new toy. Türkçe: Çocuk yeni oyuncağına ilgi göstermedi.

olarak çeviri modelinin eğitiminde kullanılmıştır. Elde edilen çeviri modeli içerdiği her bir çeviri opsiyonunun hangi alandan elde edildiğini de bilmektedir.

Örneğin, *interest* sözcüğü farklı alanlarda kullanıldığında Türkçe'ye farklı şekilde çevirilmektedir. Finans haberleri alanında *faiz* olarak çevrilmesi, *ilgi* olarak çevrilmesinden daha olasıdır. Diğer taraftan, *ilgi* çevirisi altyazı alanında muhtemelen daha iyi bir çeviridir. Bu yüzden, *"The credit interest rate is announced."* cümlesi *interest* sözcüğünün *rate* ve *credit* sözcükleri ile birlikte kullanımından dolayı haberler alanına ait olmaya daha yatkındır. Bu örnekler Tablo 4.1'te görülmektedir.

Örneklerde olduğu gibi, alan belirteci çeviri sisteminde yüzeysel biçim ile bir bütün olarak kullanılmıştır. Bu çalışmanın amacı alan uyarlaması olduğu için, sadece sözcük yüzeysel biçimi ve alan bilgisi faktör olarak seçilmiştir. Kullanılan çeviri faktörleri Şekil 4.3'de gösterilmektedir.



Şekil 4.3: Faktörlü çeviri modelinde kullanılan çeviri faktörleri

Çeviri faktörleri kaynak dilde yüzeysel biçim, hedef dilde yüzeysel biçim ve alan bilgisidir. Üretim adımı (generation step) olmaksızın sadece tek bir çeviri adımı (translation step) bulunmaktadır.

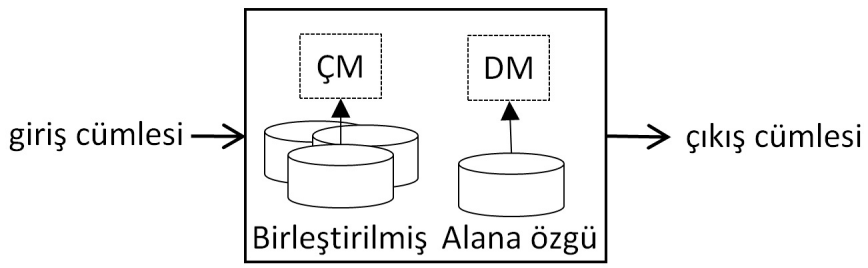
Kök, sözcük türü etiketleri (part-of-speech tags) gibi daha fazla dilbilimsel bilginin eklenmesi ile çeviri kalitesinin artırılabileceği açık olsa da bu çalışmanın kapsamının dışındadır.

Kaynak dilin sadece yüzeysel biçimi eğitimde yer aldığı için, bu model yalın sistemde olduğu gibi yüzeysel biçimli ve çok alanlı test verisi ile test edilmiştir.

4.4 Dil Modeli Uyumlu Sistemlerin Birleştirilmesi

Alana özgü sistemlerin en temel dezavantajı, alana özgü eğitim verilerinin göreceli olarak az olmasından dolayı veri seyrekliği problemi yaşamalarıdır. Özellikle Türkçe gibi eklemeli ve biçimbilimsel açıdan zengin diller üretken biçimbilimsel yapılarıyla sayısız denecek ölçüde yüzeysel forma ve oldukça geniş bir sözlüğe sahiptirler. Bu durumda olan diller genellikle veri seyrekliği probleminden müzdariptirler. Bu nedenle, toplam paralel derlemi alana özgü eğitim kümelerine bölmek, verinin yetersiz kalmasına ve giriş cümlesi için uygun çeviri opsiyonlarının bulunamamasına neden olabilir. Daha önceki bir çalışmada [48] belirtildiği gibi, çeviri sistemini belirli bir alana uyarlamada en önemli mücadele bilinmeyen sözcüklere karşı verilmektedir. Diğer bi taraftan, eğitim verisini ve dolayısıyla çeviri modellerini artan bir şekilde genişletmek eşseslilerin alan dışı çeviri opsiyonlarıyla aşırı uyum göstermesine neden olabilir (overfitting problem). Fakat, alana özgü dil modeli kendisi için uygunsuz olan bu opsiyonları çeviri modelindeki en olası adaylar olsalar bile cezalandırmaktadır. Çeviri sistemi ilgili alan için en uygun adayları seçebildiği için, çeviri sistemi alana özel sözcükleri ve sözcük öbeklerini üretebilecektir.

Interest sözcüğünün çeviri opsiyonları *faiz* ve *ilgi* sözcüklerinde olduğu gibi, alana özgü dil modelleri kullanıldığında çeviri sistemi sözcük öbeklerini bu alana özgü şekilde çevirme eğilimindedir. Bu durumda, haber alanında *faiz* çevirisinin daha olası olması beklenmektedir.



Şekil 4.4: Dil modeli uyumlu alana özgü sistem

Şekil 4.4'de görüldüğü gibi genel bir çeviri modeli ve alana özgü bir dil modeli kullanılarak çeviri sisteminde alan uyarlaması yapılmıştır. Bir önceki modele benzer

şekilde, girdi cümlesinin dahil olduğu alanı tespit etmek için bir metin sınıflandırıcısı görevlendirilmiştir. Bu sistem, dil modeli uyumlu istatistiksel bilgisayarlı çeviri (DM uyumlu İBÇ) sistemi olarak adlandırılmıştır.

5. UYGULAMA VE SONUÇLAR

Yapılan deneylerde, açık kaynak kodlu bir istatistiksel bilgisayarlı çeviri sistemi olan Moses uygulama yazılımı [47] kullanılmıştır. Ayrıca, faktörlü model çatısı da bazı deneylerde kullanılmıştır (bknz. Bölüm 4.3 ve Bölüm 4.4). SRILM dil modelleme uygulama yazılımı [49] da Good-Turing yumuşatması (Good-Turing smoothing) ve aradeğerleme (interpolation) ile birlikte kullanılmıştır. GIZA++ [50] ise simetrikleştirilmiş sözcüğe sözcük hizalamaları (symmetrized word-to-word alignment) oluşturmak için kullanılmıştır (*grow-diag-final-and* opsiyonu ile birlikte). Sonuçlar BLEU [27] otomatik çeviri değerlendirme ölçütü ile verilmiştir. Tüm deneyler küçük harfe çevrilmiş ve noktalama işaretleri ayrı yazılmış veri üzerinde gerçekleştirilmiş, fakat sonuç değerlendirmesi yeniden büyük küçük harfleri korunmuş ve noktalama işaretleri birleştirilmiş veri üzerinde yapılmıştır.

Deneyler öncelikle kahin (oracle) sınıflandırıcı ile elde edilen test sonuçları ile değerlendirilmiştir. Bunun nedeni, kullanılan sınıflandırıcının başarı oranının yapılan çalışmanın sonuçlarına yansıtılmasının istenmemesidir. Kahin sınıflandırıcı her bir girdi cümlesinin doğru sınıfı, yani alanı bilinseydi elde edilebilecek en yüksek puanı sağlamaktadır. Bu nedenle, yöntemler arasında en açık ve güvenilir değerlendirme kahin sınılandırıcı ile yapılabilmektedir. Deneylerin sonunda en başarılı bulunan yöntem Bölüm 5.2'de tanıtılan gerçek sınıflandırıcı ile yeniden değerlendirilmiştir.

5.1 Veri

Bu çalışma, haber, edebiyat, altyazı ve internet olmak üzere dört farklı alan derlemi üzerinde gerçekleştirilmiştir. Haber derlemi [51] ve altyazı derlemi tüm araştırmacıların erişimine açıktır[1]. Edebiyat derlemi roman, hikaye, siyaset ve benzeri alanlardan bir araya getirilmiş metinlerden oluşurken [52], internet derlemi çeviri internet sitelerin içeriklerinden oluşmaktadır [53]. Tüm alanlardaki cümle çiftlerinden üç ve üçten az sözcükten oluşan cümleler eşleriyle birlikte derlemden çıkarılmıştır.

[1] `http://opus.lingfil.uu.se/`, Son Erişim: Haziran 2014

Bu eleme işleminden sonra, her derlemin içerdiği cümle sayısı, sözcük sayısı ve tekil sözcük sayısı Çizelge 5.1'de gösterilmektedir. Her derlem için 2,5K iyileştirme ve 2,5K test cümlesi ayrılmıştır[2].

Çizelge 5.1: Veri detayları

Alan	Cümle Sayısı			Sözcük Sayısı		Tekil Sözcük Sayısı	
	Eğitim	İyileştirme	Test	EN	TR	EN	TR
Edebiyat	624.446	2,5K	2,5K	11.854.879	8.464.621	73.933	136.770
Haber	201.090	2,5K	2,5K	4.327.374	3.764.320	57.350	97.456
Altyazı	742.495	2,5K	2,5K	6.514.838	4.704.216	106.835	207.935
İnternet	141.467	2,5K	2,5K	3.083.162	2.591.270	72.367	106.369

5.2 Sınıflandırıcının Performansı

Deneyler için bir girdi cümleyi ilgili çeviri sistemine yönlendirebilmek için lineer bir metin sınıflandırıcısı kullanılmıştır. Crammer ve Singer tarafından önerilen çok sınıflı bir Destek Vektör Makinesi (DVM) eğitilmiştir [54]. Sınıflandırıcının yüksek performansa sahip olması için, doğru özellik ve ön işleme adımlarını belirleyebilmek amacıyla pek çok test yapılmıştır. Sonuç olarak, gereksiz sözcükleri çıkarmadan (no stopword removal) ve gövdeleme yapmadan (no stemming) ikili (bigram) özelliklerin kullanılması en iyi performansı vermiştir. Bu nedenle, sınıflandırıcı kullanılarak gerçeklenen deneylerde bu opsiyonlarla eğitilmiş sınıflandırıcıya yer verilmiştir.

Tüm alanlardaki toplam eğitim ve iyileştirme verisi (yaklaşık olarak 1,7M cümle) sınıflandırıcıyı eğitmek için kullanılmıştır. Sınıflandırıcının performansını değerlendirmek için alana özgü test kümelerindeki cümleler üzerinde sınıflandırma yapılmıştır. Sınıflandırma sonuçları Çizelge 5.2'de verilmektedir.

Çizelge 5.2: Alana özgü ve çok alanlı test kümeleri ile DVM sınıflandırıcısının doğruluğu

Test Kümesi	Örnek Sayısı	Doğru Tahmin	Yanlış Tahmin	Doğruluk (%)
Edebiyat	2,5K	2405	95	96,2
Haber	2,5K	2168	332	86,72
Altyazı	2,5K	2345	155	93,8
İnternet	2,5K	2376	124	95,04
Çok alanlı	10K	9294	706	92,94

[2] `http://ddi.itu.edu.tr/resources/domainData_en-tr.zip/` adresinden erişilebilir. Son Erişim: Haziran 2014

Sınıflandırıcı sınıflandırma hatalarına neden olabildiği için, bölümün başında belirtildiği gibi, ilk testler her zaman doğru sınıfı seçen kahin sınıflandırıcı ile gerçekleştirilmiştir. En umut verici yönteme karar verildikten sonra, sadece bu yöntem DVM sınıflandırıcısı ile tekrar test edilmiştir.

5.3 Alan Uyarlaması Sonuçları

İlk deneyde, daha önce yalın sistem olarak adlandırılan ve tüm alana özgü verilerin birleşiminden eğitilmiş basit bir çeviri ve bir dil modelinden oluşan sistemin başarısı ölçülmüştür. Bu sistemin otomatik çeviri değerlendirme aracı BLEU ile sağladığı başarı yine BLEU puanı cinsinden 27,36 olarak elde edilmiştir. Bu yalın sistem alan uyarlaması yöntemlerini değerlendirmek için bir referans noktası oluşturmaktadır.

Alana özgü istatistiksel bilgisayarlı çeviri sistemlerinin birleştirilmesi değerlendirilmeden önce, her bir alana özgü sistem hem alana özgü test verisi hem de alandan bağımsız çok alanlı test verisi ile ayrı ayrı değerlendirilmiştir. Bu deneylerin sonuçları Çizelge 5.3'te verilmiştir. Örneğin, edebiyat alanındaki AÖ-İBÇ sistemi, 2,5K alana özgü (edebiyat) test cümlesi ile 36,87 BLEU puanına sahip olurken, 10K genel test cümlesinde önemli ölçüde başarı kaybederek 7,78 BLEU puanına gerilemiştir.

Çizelge 5.3: Alana özgü sistemlerin başarısı

Alan	**ÇM**	**DM**	**Test Kümesi**	**BLEU**	**N-gram kesinliği**			
					1-gr	**2-gr**	**3-gr**	**4-gr**
Edebiyat	Alan	Alan	Alana özgü	36,87	53,9	41,4	32,6	25,4
			Çok alanlı	7,78	18,2	8,3	5,7	4,2
Haber	Alan	Alan	Alana özgü	17,17	46,6	23,3	12,5	7,3
			Çok alanlı	3,98	22,8	5,9	2,1	1,0
Altyazı	Alan	Alan	Alana özgü	7,63	27,1	10,3	4,8	2,6
			Çok alanlı	7,96	26,7	10,4	5,1	2,8
Internet	Alan	Alan	Alana özgü	33,17	47,0	35,8	29,3	24,5
			Çok alanlı	10,66	21,3	11,1	8,2	6,6

Beklenildiği üzere, Çizelge 5.3'teki sonuçlar alana özgü sistemlerin alana özgü girdi verisi ile yüksek kaliteli çıktılar üretebildiğini göstermektedir. Fakat alandan bağımsız genel kapsamlı veri ile yapılan testlerde çıktıların kalitesinde ciddi düşüşler yaşanmaktadır. Bu sonuçlar alan uyarlamasının önemini ve alan dışı cümlelerde uygunsuz çeviri ve dil modelleri kullanmanın başarı üzerindeki şiddetli etkisini göstermektedir.

AÖ-İBÇ sistemlerinin birleşiminden oluşan alan uyarlamalı sistemin kahin sınıflandırıcı ile başarısı 27,92 BLEU puanıdır. Kahin sınıflandırıcı daima doğru alanı tahmin edebildiği için, bu deneyde hiç bir sınıflandırıcı hatası olmayacağı göz önünde bulundurulmalıdır.

Birleştirilmiş AÖ-İBÇ sistemlerinin geliştirilmiş bir versiyonu olarak, Moses faktörlü çeviri modeli çatısındaki alternatif çözümleme yolları ve geri çekilme modelleri kullanılmıştır. Alternatif çözümleme yolları kullanılarak, her AÖ-İBÇ sistemi geri çekilme modeli olan genel amaçlı sistemle (yalın sistem) birleştirilmiştir. Birincil tabloda (alana özgü çeviri modelinde) bulunamayan 4-gram uzunluğa kadar olan bilinmeyen sözcük öbekleri için ikincil tabloda (genel çeviri modelinde) çözümleme araması yapılmıştır.

Geri çekilme genellemesi[3] ile dört AÖ-İBÇ sisteminin çeviri performansı Çizelge 5.4'te gösterilmektedir. Çizelgeden görüleceği üzere, bu genelleme edebiyat ve internet alanlarında etkili olurken, haber ve altyazı alanlarında daha kötü sonuçlar üretmektedir. Bu bozulmanın nedeni geri çekilme aramalarında 4-gramın kullanılması olabilir. AÖ-İBÇ sistemlerinin alana özgü test kümesi üzerindeki performansı (bknz. Çizelge 5.3) bu alanlara ait verinin diğer iki alanla (edebiyat ve internet) kıyaslandığında, test kümelerindeki örnekleri kapsayacak kadar yeterli olmadığını göstermektedir. Test kümesindeki bir 4-gramı bulabilmesi düşük bir olasılık olduğu için, bu alana özgü sistemler genelleştirme için kullanılan yalın sisteme daha bağımlı olurlar. Genel modelde bulunan daha uzun sözcük öbekleri alanın kapsamı dışında olsa bile, bu uzun sözcük öbekleri alana özgü sistemlerdeki kısa sözcük öbeklerinden baskın çıkmaktadırlar.

Geri çekilme ile genelleştirilmiş dört AÖ-İBÇ sistemi bir arada kullanılmış ve kahin sınıflandırıcı ile test edilmiştir. Test için 10K adet çok-alanlı cümle kullanılmış ve birleştirilmiş sistemin genel performansı 29,36 olarak bulunmuştur (bknz. Çizelge 5.6).

Alan bilgisi için faktörlü modellerin kullanılması da bu çalışmadaki bir diğer alan uyarlaması yöntemidir. Bu yöntem temelde yalın sistemin alan etiketlerinin faktör olarak kullanılması ile genişletilmiş versiyonudur. Eğitim, iyileştirme ve test verisi

[3]Genel veriden elde edilen çeviri modelinin geri çekilme yöntemi ile kullanılması çizelgelerde "+ $\text{Genel}_{\text{gç}}$" gösterimi ile belirtilmiştir.

Çizelge 5.4: Alana özgü sistemlerin geri çekilme ile başarısı

Alan	ÇM	DM	Test Kümesi	BLEU	N-gram kesinliği			
					1-gr	2-gr	3-gr	4-gr
Edebiyat	Alan + $\text{Genel}_{\text{gç}}$	Alan + Genel	Alana özgü	44,10	67,2	53,8	44,0	35,8
Haber	Alan + $\text{Genel}_{\text{gç}}$	Alan + Genel	Alana özgü	15,00	48,8	24,2	13,3	7,9
Altyazı	Alan + $\text{Genel}_{\text{gç}}$	Alan + Genel	Alana özgü	5,90	26,5	9,9	4,9	2,8
Internet	Alan + $\text{Genel}_{\text{gç}}$	Alan + Genel	Alana özgü	38,16	56,7	44,4	37,4	32,1

hedef dildeki yüzeysel biçimlerin alan etiketleri ile zenginleştirilmiş olmasının dışında yalın sistemle tamamen aynıdır. Alan belirteçleri, her alan için uygun çeviri opsiyonlarının eğitim aşamaında öğrenilebileceği varsayımı ile kaynak dil tarafında kullanılmamıştır [7]. Bu yöntemin sağladığı temel avantaj herhangi bir sınıflandırıcıya gerek duymamasıdır. Faktörlü çeviri ile yapılan bu deneyde 26,17 BLEU puanı elde edilmiştir. Alan faktörleri, alana özgü terimleri içeren sözcük dizilerinin[4] çevirilerinde etkili olmaktadır. Alana özgü bir terim yakalandığında, bu terimi çevreleyen sözcük ve sözcük öbekleri aynı alana sadık kalınarak çevrilmektedir. Diğer bir taraftan, alanı faktör olarak kullanma amacına rağmen, yüzeysel biçim ve alan faktörlerinin hedef dildeki zorunlu birlikteliği toplam olasılıkları koşullu olasılıklara dağıtmaktadır. Örneğin, *interest* sözcüğünün çevirisinin haber alanındaki *faiz* olarak yapılması veya başka bir alandaki *faiz* olarak yapılması gibi çeviri olasılıkları koşullara bağlanmaktadır. Bu yüzden, sistem tüm koşullu seçeneklerden, test girdisinin alanı için en iyi sonuç olmasa bile, en olası çeviriyi seçer. Bu durum alan bilgisini fazladan bir faktör olarak kullanan çeviri sisteminin başarısını kötüleştirmektedir.

Son deney grubu da alana özgü dil modellerinin kullanımı üzerine yapılmıştır. Bu deneylerde, tüm alanlara ait verinin tamamından büyük bir çeviri modeli ve her alan için ayrı ayrı alana özgü dil modelleri eğitilmiştir. Dört adet DM uyumlu İBÇ sistemi oluşturulmuş ve hangi sistemin kullanılacağına karar vermesi için kahin sınıflandırıcı kullanılmıştır. Bu sistemlerin alana özgü çevirilerdeki performansını araştırmak için, sistemler kendi alanlarına ait test kümelerinde değerlendirilmiş ve sonuçları Çizelge 5.5'te verilmiştir. DM uyumlu İBÇ AÖ-İBÇ ile kıyaslandığında, DM uyumlu

[4]Diğer alana özgü verilerde yer almayan belirli sözcük veya sözcük öbeklerini ifade edilmektedir.

sistemlerin AÖ-İBÇ sistemlerinden daha başarılı olduğu görülmektedir. Örneğin, edebiyat alanındaki AÖ-İBÇ 36,87 BLEU puanına sahip olurken, aynı alandaki DM uyumlu İBÇ sistemi 40,74 puan elde etmektedir. Bu çıkarımdaki tek istisnai durum haber alanında gözlenmektedir. Çizelge 5.1'den de görüleceği üzere, haber alanı en az tekil sözcük sayısına sahip olan derlemdir ve bu durum bu derlemden oluşturulan dil modelinin genel bağlamda yetersiz olmasına neden olmaktadır.

Çizelge 5.5: Dil modeli uyumlu alana özgü sistemlerin başarısı

Alan	**ÇM**	**DM**	**Test Kümesi**	**BLEU**	**N-gram kesinliği**			
					1-gr	**2-gr**	**3-gr**	**4-gr**
Edebiyat	Genel	Alan	Alana özgü	40,74	58,8	45,7	36,3	28,3
Haber	Genel	Alan	Alana özgü	16,68	47,6	23,8	12,8	7,5
Altyazı	Genel	Alan	Alana özgü	8,11	26,6	10,5	5,5	3,4
Internet	Genel	Alan	Alana özgü	36,88	52,1	39,9	32,7	27,3

Çizelge 5.6 tüm deneylerin sonuçlarını özetlemektedir. Çizelgeden anlaşılacağı üzere, dil modeli uyarlaması diğer alan uyarlaması yöntemlerinden daha üstün performans göstermekte ve başarıyı daha fazla artırmaktadır. Bu nedenle, DVM temelli gerçek sınıflandırıcı, sadece en iyi performans gösteren yöntem olan DM uyumlu sistemlerin birleştirilmesi ile kullanılmıştır.

Çizelge 5.6: Çeşitli alan uyarlaması modellerinin genel değerlendirmesi

Adaptasyon Yöntemi	**Test Seti**	**Sınıflandırıcı**	**BLEU**	**Göreceli İyileşme**
(1) Yalın Sistem	Çok alanlı	N/A	27,36	N/A
(2) Alana Özgü Sistemlerin Birleştirilmesi	Çok alanlı	Kahin	27,92	2,05%
(3) Alana Özgü Sistemlerin Birleştirilmesi + Genelleme ile İyileştirme	Çok alanlı	Kahin	29,36	7,31%
(4) Alan Bilgisinin Faktör Olarak Kullanılması	Çok alanlı	N/A	26,17	4,35%
(5) DM-Uyumlu Alana Özgü Sistemlerin Birleştirilmesi	Çok alanlı	Kahin	**30,16**	**10,23%**
(6) DM-Uyumlu Alana Özgü Sistemlerin Birleştirilmesi	Çok alanlı	DVM	29,89	9,25%

Çeşitli alan uyarlaması deneylerinin genel değerlendirmesi gösteriyor ki, İngilizce'den Türkçe'ye istatistiksel bilgisayarlı çeviri sistemlerinin alan uyarlaması için dil modelleri en etkili bileşenlerdir. Yalın sistem 27,36 BLEU puanı elde ederken,

DM uyumlu İBÇ sistemlerinin gerçek bir metin sınıflandırıcısı ile birleştirilmesi 29,89 BLEU puanı elde etmektedir. DM uyumlu sistemlerin birleşimi çok alanlı test verisinde %9,25 göreceli iyileşmeye neden olan 2,53 BLEU puanı kazancı sağlamaktadır.

6. DEĞERLENDİRME VE ÖNERİLER

İstatistiksel bilgisayarlı çeviri için uygulanan çoğu alan uyarlaması çalışması sadece tek bir alana odaklanmaktadır. Bu nedenle, tek bir alan üzerinde yapılan iyileştirmelerle yetinmekte ve daha genel bir sistem için çalışmaların katkılarını gösterememektedir. Bu çalışma kendini uygun alana uyarlayabilen, İngilizce'den Türkçe'ye çeviri yapan bir sistemin çeşitli alan uyarlaması yöntemleri ile ulaştığı başarımları göstermektedir. Ayrıca Türkçe için alan uyarlaması ile ilgili daha önce böyle bir çalışma yapılmadığından hangi yöntemlerin daha başarılı olacağına dair yol gösterici bir çalışma niteliğindedir.

İngilizce'den Türkçe'ye yapılan bilgisayarlı çevirilerde, çeviri kalitesini en çok artıran alan uyarlaması uygulaması dil modeli ile sağlanmaktadır. Dil modeli ile alan uyarlaması yapılan sistemde başarı 27.36 BLEU puanından 29.89 BLEU puanına yükselmiştir. Yalın istatistiksel bilgisayarlı çeviri sistemine kıyasla %9.25 oranında göreceli iyileşme gözlemlenmiştir.

İstatistiksel bilgisayarlı çeviride alan uyarlaması üzerine bundan sonra yapılacak çalışmalarda, kullanılan sınıflandırıcının başarısını artırmak için çalışılabilir. Oluşturulacak sistemin kullanım alanına göre, her zaman cümle çevirisi yapılması beklenmeyebilir. Hatta çoğu uygulama alanında çeviri sistemleri bir cümleden daha uzun olan metinleri çevirmek için kullanılmaktadır. Bu durumda cümle seviyesinde sınıflandırma yapmaya gerek olmadığından doküman seviyesinde sınıflandırma yapılabilir. Bu durum sınıflandırıcının, karar vermesini kolaylaştıracağından daha yüksek başarı elde edilebilir. Cümle seviyesinde sınıflandırmak gerekiyorsa bile sınıflandırıcının kesin karar veremediği ve bu nedenle yanıldığı durumları müsamaha edebilmek için güven aralıkları belirlenebilir. Güven aralığında yer almayan sınıflandırmalar için genel kapsamlı sistemde üretilen çeviriler kullanılabilir. Ayrıca sınıflandırıcı yerine dil modeli denetim birimi ile cümlenin hangi alana ait olduğu bilgisi de elde edilebilir. İlgili alanlardan oluşturulmuş dil modelleri sisteme giren cümlenin kendi alanlarına ait olma ihtimallerini değerlendirebilirler.

En başarılı bulunan dil modeli ile uyarlama yönteminde kullanılan alana özgü dil modelleri için paralel çift dilli derlemin hedef dil tarafındaki tek dilli veri kullanılmıştır. Bunun yerine ilgili alanlara özgü tek dilli verilerin çoğaltılması ile daha kapsamlı ve güvenilir dil modelleri elde edilebilir. Bu durum başarının artırılmasını sağlayacaktır. Veri yetersizliği nedeniyle, bu aşamada bırakılan çalışmanın uygulamaya konulması halinde tek dilli veri miktarında artışa gidilmelidir.

6.1 Çalışmanın Uygulama Alanı

Bu çalışmadan elde edilen çıktılar ile İngilizce'den Türkçe'ye istatistiksel bilgisayarlı çevirilerde başarının artırılmasına yönelik alan uyarlaması yöntemleri değerlendirilmiş ve dil modeli ile uyarlamanın daha iyi sonuç verdiği görülmüştür. Bu yöntem kullanılarak, alan uyarlamalı ve genel amaçlı bir istatistiksel bilgisayarlı çeviri sistemi oluşturulabilecektir. Verinin az olduğu ve yapısı dolayısıyla daha fazla veriye ihtiyaç duyan diller için daha kaliteli çeviriler üretilebilecektir.

Oluşturulan çok alanlı ve alan uyarlamalı sistem, bilgi çıkarımı, sınıflandırma gibi amaçlarla bilgisayarlı çevirinin yeteneklerinden yararlanan diğer disiplinlerde bir araç olarak kulanılabileceği gibi, doğrudan bilgisayarlı çeviri amacına uygun olarak metin çevirisi amacıyla kullanılabilecektir.

KAYNAKLAR

[1] **Lewis, M.P.**, **Simons, G.F. ve Fennig, C.D.**, Ethnologue: Languages of the World, Seventeenth edition., `https://www.ethnologue.com/`, Son Erişim: 25.04.2014.

[2] **Oflazer, K. ve El-Kahlout, I.D.** (2007). Exploring different representational units in English-to-Turkish statistical machine translation, *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, s.25–32.

[3] **Bisazza, A. ve Federico, M.** (2009). Morphological pre-processing for Turkish to English statistical machine translation, *Proc. of the International Workshop on Spoken Language Translation*, s.129–135.

[4] **Koehn, P. ve Schroeder, J.** (2007). Experiments in domain adaptation for statistical machine translation, *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, s.224–227.

[5] **Sennrich, R.** (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, s.539–549.

[6] **Banerjee, P.**, **Du, J.**, **Li, B.**, **Kumar Naskar, S.**, **Way, A. ve Van Genabith, J.** (2010). Combining multi-domain statistical machine translation models using automatic classifiers, Association for Machine Translation in the Americas.

[7] **Niehues, J. ve Waibel, A.** (2010). Domain adaptation in statistical machine translation using factored translation models, *Proceedings of EAMT*.

[8] **Wu, H.**, **Wang, H. ve Zong, C.** (2008). Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora, *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA, s.993–1000.

[9] **Nakov, P.** (2008). Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing, *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, Association for Computational Linguistics, Stroudsburg, PA, USA, s.147–150.

[10] **Bertoldi, N. ve Federico, M.** (2009). Domain Adaptation for Statistical Machine Translation with Monolingual Resources, *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, Association for Computational Linguistics, Stroudsburg, PA, USA, s.182–189.

[11] **Chandioux, J.** (1976). MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public, *Meta: Journal des traducteursMeta:/Translators' Journal*, ***21***(2), 127–133.

[12] **Vauquois, B.** (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation., *Ifip congress (2)*, s.1114–1122.

[13] **Nagao, M.** (1984). A framework of a mechanical translation between Japanese and English by analogy principle.

[14] **Isabelle, P.**, **Dymetman, M.**, **Foster, G.**, **Jutras, J.M.**, **Macklovitch, E.**, **Perrault, F.**, **Ren, X. ve Simard, M.** (1993). Translation analysis and translation automation, *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, IBM Press, s.1133–1147.

[15] **Lopez, A.** (2008). Statistical machine translation, *ACM Computing Surveys (CSUR)*, ***40***(3), 8.

[16] **Weaver, W.** (1955). Translation, *Machine translation of languages*, ***14***, 15–23.

[17] **Brown, P.F.**, **Cocke, J.**, **Pietra, S.A.D.**, **Pietra, V.J.D.**, **Jelinek, F.**, **Lafferty, J.D.**, **Mercer, R.L. ve Roossin, P.S.** (1990). A statistical approach to machine translation, *Computational linguistics*, ***16***(2), 79–85.

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

[19] **Koehn, P.**, **Och, F.J. ve Marcu, D.** (2003). Statistical Phrase-based Translation, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, s.48–54.

[20] **Shannon, C.E.** (2001). A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, ***5***(1), 3–55.

[21] **Yildirim, E. ve Tantug, A.** (2013). The feasibility analysis of re-ranking for N-best lists on English-Turkish machine translation, *2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, s.1–5.

[22] **Koehn, P. ve Hoang, H.** (2007). Factored Translation Models., *EMNLP-CoNLL*, s.868–876.

[23] **Koehn, P.**, **Federico, M.**, **Shen, W.**, **Bertoldi, N.**, **Bojar, O.**, **Callison-Burch, C.**, **Cowan, B.**, **Dyer, C.**, **Hoang, H.**, **Zens, R.** *ve diğerleri* (2006). Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding, *Final Report of the 2006 JHU Summer Workshop*.

[24] **Birch, A.**, **Osborne, M. ve Koehn, P.** (2007). CCG Supertags in Factored Statistical Machine Translation, *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, Association for Computational Linguistics, Stroudsburg, PA, USA, s.9–16.

[25] **Sridhar, V.K.R.**, **Bangalore, S. ve Narayanan, S.S.** (2008). Factored translation models for enriching spoken language translation with prosody., *INTERSPEECH*, s.2723–2726.

[26] **Avramidis, E. ve Koehn, P.** (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation., *ACL*, s.763–770.

[27] **Papineni, K.**, **Roukos, S.**, **Ward, T. ve Zhu, W.J.** (2002). BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, s.311–318.

[28] **Doddington, G.** (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics, *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, s.138–145.

[29] **Callison-Burch, C. ve Osborne, M.** (2006). Re-evaluating the role of BLEU in machine translation research, *In EACL*, Citeseer.

[30] **Tantug, A.C.**, **Oflazer, K. ve El-Kahlout, I.D.** (2008). BLEU+: a Tool for Fine-Grained BLEU Computation., *LREC*.

[31] **Melamed, I.D.**, **Green, R. ve Turian, J.P.** (2003). Precision and Recall of Machine Translation, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, Association for Computational Linguistics, Stroudsburg, PA, USA, s.61–63.

[32] **Banerjee, S. ve Lavie, A.** (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, s.65–72.

[33] **Lavie, A.**, **Sagae, K. ve Jayaraman, S.**, (2004). The significance of recall in automatic metrics for MT evaluation, Machine Translation: From Real Users to Research, Springer, s.134–143.

[34] **Bellegarda, J.R.** (2004). Statistical language model adaptation: review and perspectives, *Speech Communication*, ***42***(1), 93 – 108, adaptation Methods for Speech Recognition.

[35] **Iyer, R.M. ve Ostendorf, M.** (1999). Modeling long distance dependence in language: Topic mixtures versus dynamic cache models, *Speech and Audio Processing, IEEE Transactions on*, ***7***(1), 30–39.

[36] **Mahajan, M.**, **Beeferman, D. ve Huang, X.D.** (1999). Improved topic-dependent language modeling using information retrieval techniques, *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, cilt 1, IEEE, s.541–544.

[37] **Zhao, B.**, **Eck, M. ve Vogel, S.** (2004). Language Model Adaptation for Statistical Machine Translation with Structured Query Models, *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA.

[38] **Seymore, K. ve Rosenfeld, R.** (1997). Using story topics for language model adaptation., ***G. Kokkinakis, N. Fakotakis ve E. Dermatas,*** *(düzenleyenler), EUROSPEECH*, ISCA.

[39] **Chen, S.**, **Seymore, K. ve Rosenfeld, R.** (1998). Topic adaptation for language modeling using unnormalized exponential models, *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, cilt 2, s.681–684 vol.2.

[40] **Eidelman, V.**, **Boyd-Graber, J. ve Resnik, P.** (2012). Topic models for dynamic translation model adaptation, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, s.115–119.

[41] **Su, J.**, **Wu, H.**, **Wang, H.**, **Chen, Y.**, **Shi, X.**, **Dong, H. ve Liu, Q.** (2012). Translation model adaptation for statistical machine translation with monolingual topic information, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Association for Computational Linguistics, s.459–468.

[42] **Lambert, P.**, **Schwenk, H.**, **Servan, C. ve Abdul-Rauf, S.** (2011). Investigations on translation model adaptation using monolingual data, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, s.284–293.

[43] **Snover, M.**, **Dorr, B. ve Schwartz, R.** (2008). Language and translation model adaptation using comparable corpora, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, s.857–866.

[44] **Wang, W.**, **Macherey, K.**, **Macherey, W.**, **Och, F. ve Xu, P.** (2012). Improved Domain Adaptation for Statistical Machine Translation, *AMTA-2012*.

[45] **Foster, G. ve Kuhn, R.** (2007). Mixture-model Adaptation for SMT, *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, Association for Computational Linguistics, Stroudsburg, PA, USA, s.128–135.

[46] **Yeniterzi, R. ve Oflazer, K.** (2010). Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, s.454–464.

[47] **Koehn, P.**, **Hoang, H.**, **Birch, A.**, **Callison-Burch, C.**, **Federico, M.**, **Bertoldi, N.**, **Cowan, B.**, **Shen, W.**, **Moran, C.**, **Zens, R.** *ve diğerleri* (2007). Moses: Open source toolkit for statistical machine translation, *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Association for Computational Linguistics, s.177–180.

[48] **Daumé, III, H. ve Jagarlamudi, J.** (2011). Domain Adaptation for Machine Translation by Mining Unseen Words, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, s.407–412.

[49] **Stolcke, A.**, **Zheng, J.**, **Wang, W. ve Abrash, V.** (2011). SRILM at sixteen: Update and outlook, *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, s. 5.

[50] **Och, F.J. ve Ney, H.** (2003). A systematic comparison of various statistical alignment models, *Computational linguistics*, ***29***(1), 19–51.

[51] **Tyers, F.M. ve Alperen, M.S.** (2010). South-east european times: A parallel corpus of Balkan languages, *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, s.49–53.

[52] **Taşçı, Ş.**, **Güngör, A.M. ve Güngör, T.** (2006). Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment, *International Scientific Conference Computer Science*.

[53] **Yıldız, E. ve Tantuğ, A.C.** (2012). Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts, *First Workshop on Language Resources and Technologies for Turkic Languages*, s. 64.

[54] **Crammer, K. ve Singer, Y.** (2002). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines, *J. Mach. Learn. Res.*, ***2***, 265–292.

EKLER

EK A.1

Bu bölümde, bu alanda yapılan diğer çalışmalarla uyumun sağlanabilmesi için, bahsedilen Türkçe terimlerin İngilizce karşılıkları verilmektedir.

Çizelge A.1: Türkçe terimlerin İngilizce karşılıkları

Türkçe	**İngilizce**
Aday Çeviri	Candidate Translation
Alan Adaptasyonu	Domain Adaptation
Alana Özgü	Domain Specific
Alternatif Çözümleme Yolu	Alternative Decoding Path
Anlamsal	Semantic
Aradeğerleme	Interpolation
Aşırı Uyum Gösterme Problemi	Overfitting Problem
Belirtme Durumu	Accusative Case
Biçimbilimsel	Morphological
Bilgi Çıkarımı	Information Retrieval
Bilgi Elde Etme	Information Retrieval
Bilgisayarlı Çeviri	Machine Translation
Bilinmezlik	Perplexity
Bükümlü	Inflectional
Çeviri Modeli	Translation Model
Çift Dilli	Bilingual
Çok-alanlı	Multi-domain
Çözümleme	Decoding
Destek Vektör Makinesi (DVM)	Support Vector Machine (SVM)
Dil Modeli	Language Model
Doğal Dil İşleme	Natural Language Processing
Durum Temelli Akıl Yürütme	Case-based Reasoning
Eklemeli	Agglutinative
En İyi N Listeleri	N-best Lists
En Kısa Değişim Uzaklığı	Minimum Edit Distance
Eşsesli	Homonym
Etkin Referans Uzunluğu	Effective Reference Length
Faktör	Factor
Faktörlü Çeviri Modeli	Factored Translation Model
Gereksiz Sözcük	Stopword
Geri Çekilme Modeli	Back-off Model
Gerigetirim	Recall
Good-Turing Yumuşatması	Good-Turing Smoothing
Göreceli İyileşme	Relative Improvement
Gürültülü Kanal Modeli	Noisy Channel Model
Hedef Dil	Target Language

Çizelge A.1 (devamı): Türkçe terimlerin İngilizce karşılıkları

Türkçe	**İngilizce**
Hizalama	Alignment
İkili	Bigram
İstatistiksel Bilgisayarlı Çeviri	Statistical Machine Translation
İyelik Eki	Possessive Suffix
Kahin Sınıflandırıcı	Oracle Classifier
Karma Modelleme	Mixture Modelling
Kaynak Dil	Source Language
Kesinlik	Precision
Kısalık Cezası	Brevity Penalty
Koşullu Bağımsızlık	Conditional Independence
Koşullu Olasılık	Conditional Prpbability
Levenshtein Uzaklığı	Levenshtein Distance
Makina Öğrenmesi	Machine Learning
Maksimum Düzensizlik	Maximum Entropy
Olasılık Dağılımı	Probability Distribution
Olasılıksal	Probabilistic
Ön İşleme	Pre-processing
Önce Gelen	Preceding
Örneksemeyle Çeviri	Translation by Analogy
Referans Çeviri	Reference Translation
Serbest Sözcük Sıralaması	Free Word Order
Sesbirim	Phoneme
Sesbirim Değişimi	Phoneme Alternation
Sonradan İşleme	Post-processing
Sözcük Hata Oranı	Word Error Rate
Sözcük Öbeği Tablosu	Phrase Table
Sözcük Öbeği Temelli	Phrase-based
Sözcük Türü Etiketleri	Part of Speech (POS) Tags
Sözcüksel	Lexical
Sözdizimsel	Syntactic
Sözlüksel Belirsizlik	Lexical Ambiguity
Tek Dilli	Monolingual
Tekil Sözcük Sayısı	Unique Word Count
Tekli	Unigram
Türetimsel	Derivational
Üçlü	Trigram
Ünlü Uyumu	Vowel Harmony
Üretici Modeller	Generative Models
Üstel	Exponential
Veri Seyrekliği	Data Sparsity
Yapım Eki	Derivational Morpheme
Yapısal Biçim	Lexical Form
Yeniden Sıralama	Reordering
Yüzeysel Biçim	Surface Form
Zincir Kuralı	Chain Rule

EK A.2

Bu bölümde, dünya üzerindeki diller hakkında bilgiler yer almaktadır. Çizelge A.2 dünyada en çok konuşulan 23 dili ana dil olarak konuşulduğu ülke, toplam konuşulduğu ülke sayısı ve konuşan sayısı bilgileriyle birlikte sıralı olarak sunmaktadır [1].

Çizelge A.2: Dünya üzerinde en çok konuşulan diller

Sıra	**Dil**	**Ana Dil Olduğu Ülke**	**Toplam Ülke**	**Konuşanlar (milyon)**
1	Çince	Çin	33	1,197
2	İspanyolca	İspanya	31	414
3	İngilizce	Birleşik Krallık	99	335
4	Hintçe	Hindistan	4	260
5	Arapça	Suudi Arabistan	60	237
6	Portekizce	Portekiz	12	203
7	Bengalce	Bangladeş	4	193
8	Rusça	Rusya	16	167
9	Japonca	Japonya	3	122
10	Cavaca	Endonezya	3	84.3
11	Lahnda	Pakistan	6	82.6
12	Almanca	Almanya	18	78.2
13	Korece	Güney Kore	5	77.2
14	Fransızca	Fransa	51	75.0
15	Telugu	Hindistan	2	74.0
16	Marathi	Hindistan	1	71.8
17	Türkçe	Türkiye	8	70.8
18	Tamil	Hindistan	6	68.8
19	Vietnamca	Vietnam	3	67.8
20	Urdu	Pakistan	6	63.9
21	Italyanca	İtalya	10	63.7
22	Malay	Malezya	13	59.5
23	Farsça	İran	29	56.6

ÖZGEÇMİŞ

Ad Soyad: Ezgi Yıldırım

Doğum Yeri ve Tarihi: Kadıköy, 19.05.1988

E-Posta: ezgiyildrm@gmail.com

Lisans: İstanbul Teknik Üniversitesi Bilgisayar Mühendisliği Lisans Programı (2006)

Mesleki Deneyim ve Ödüller:

2011-2013:	Proline Bilişim Sistemleri A.Ş. / Yazılım Mühendisi İngilizceden Türkçeye Bilgisayarlı Çeviri Projesi
2013-...:	Turkcell Global Bilgi A.Ş. / Yazılım Geliştirme Uzmanı Sosyal Medya Takip Uygulaması

Yayın ve Patent Listesi:

Yıldırım, E. ve Tantuğ, A. C.(2013). The Feasibility Analysis of Re-ranking for N-best lists on English-Turkish Machine Translation, *2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, s.1–5.

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- **Yıldırım E.**, Tantuğ A. C., (2014). Evaluation of Domain Adaptation Approaches to Improve the Translation Quality, *2014 International Conference on Computer Communication and Informatics (ICCCI)*.