MANAS Journal of Engineering

MJEN

Volume 3 (Issue 2) (2015) Pages 59-68

Kırgız ve Türkiye Türkçeleri arasında istatistiksel bilgisayarlı çeviri uygulaması ve başarım testi

Nakılay Tayirova
Kırgızistan Türkiye Manas Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, Bişkek, Kırgızistan, nakusyat@gmail.com
Mehmet Tekerek
Kahramanmaraş Sütçüimam Üniversitesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Kahramanmaraş, Türkiye, tekerek@ksu.edu.tr
Ulan Brimkulov
Kırgızistan Türkiye Manas Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bişkek, Kırgızistan, ulan.brimkulov@manas.edu.kg

Received: 29.09.2015; Accepted:10.11.2015

Öz Bilgisayarlı çeviri, doğal diller arası metin çevirmede kullanılan farklı yöntem ve yazılımları araştırmayı amaçlayan bilgisayarlı dilbilim alt alanıdır. Bilgisayarlı çeviri araçlarının yüksek seviyede sözdizimsel ve anlambilimsel analiz sağlayamamasına rağmen; gelişmiş bilgisayarlı yöntemler uygulanarak yaygın kullanılan diller arası çeviride nispeten kabul edilebilir sonuçlara ulaşılmıştır. Son yıllarda, bilgisayarlı çeviride, büyük veri üzerinde istatistiksel analizle kendi kendini eğitebilen yöntemler geliştirilmiştir. Türkçe dil ailesi üzerine yapılan bilgisayarlı çeviri araştırmalarında, doğal dillerin kurallı yapısı çözümlenerek kural tabanlı yöntemlerin uygulandığı, ancak yaygın olarak araştırılan İstatistiksel Bilgisayarlı Çeviri yöntemlerinin ise sınırlı sayıda ve kısmen uygulandığı görülmektedir. Bu çalışmanın amacı, Kırgız Türkçesi ve Türkiye Türkçesi arasında N-GRAM Tabanlı ve İfade Tabanlı İBÇ sistemlerini uygulamak ve sınırlı paralel korpus üzerinde eğitilen İstatistiksel Bilgisayarlı Çeviri sistemlerinin başarımını çeviri örnekleri üzerinde test etmek ve incelemektir. Sonuçta her iki sistemin çeviri kalitesi BLEU değerlendirme yöntemi ile puanlanmıştır. Değerlendirmeye göre, Kırgız Türkçesi ve Türkiye Türkçesi arasında çeviri kalitesi ortalama 0.1 değerinde elde edilmiştir. Çevrisi hiç bulunmayan, ya da insan çevirisine göre uyumsuz durumlar da gözlemlenmiştir. Daha yüksek çeviri kalitesine ulaşma ve sistemler geliştirme amacıyla çeşitli öneriler sunulmuştur.

Anahtar Sözcükler: *N-GRAM , istatistiksel bilgisayarlı çeviri, dil modeli, çeviri modeli, kod çözme, kortej*

Statistical machine translation implementation and performance tests between Kyrgyz and Turkish Languages

Abstract Machine translation is a computational linguistics sub-field that studies translation methods and the use of software to translate text between natural languages. Although machine translation is not as accurate as human translation in terms of syntactic, semantic accuracy criteria, the usage of improved methods has lead to obtaininig comparatively acceptable results. In recent years, self-learning methods with statistical analysis of big data are being developed. The current development in machine translation among Turkish languages has moved towards rule-based methods. However, one of today's leading methods, Statistical Machine Translation, is being poorly applied. For this purpose, in this work N-GRAM based and Phrase based Statistical Machine Translation methods were applied between Kyrgyz and Turkish languages, using limited training data. For both methods, the translation quality was evaluated with BLEU scoring algorithm. According to test results, both applied methods, provided translations with poor translation quality, 0.1 (one hundredths) on average. In most cases, no translations obtained, or human translation incompatible results were observed. In order to reach higher translation qualities, various suggestions were proposed.

Keywords *N-GRAM, statistical machine translation, language model, translation model, decoding, tuple*

GİRİŞ

Günümüzde, özellikle dünya çapında iletişim geliştikçe, her birey için bilgi paylaşımı gerekirken, diller arası çeviriye ihtiyaç da artmaktadır. Bilgi teknolojilerinin gelişimi her alanda olduğu gibi çeviri süreçlerine de katkıda bulunmuştur. Bu, Bilgisayarlı Çeviri (BÇ) olarak adlandırılan ve çekirdeği, çeviri işlemini otomatikleştirme olan sistemlerin geliştirilmesiyle mümkün olmaktadır [1]. Bilgisayarlı Çeviri (BÇ), bir doğal dilden ikinci dile sözlü veya yazılı çeviri yaparken, farklı sözlükler içeren, eş anlamlılar arasından mantıksal seçim yapabilen, eksik sözcükleri sağlayabilen, kelime sırasını yeniden kurabilen yazılımları içeren sistemler olarak tanımlanabilir [2].

İlk BÇ çalışmalarında, çevirideki diller arası yapısal kurallar vurgulanarak kural tabanlı yöntemlere ulaşılmıştır [3, 4, 5]. Kural tabanlı yöntemlerin anlamsal muğlaklık gibi kısıtlarından dolayı, alternatif olarak İstatistiksel Bilgisayarlı Çeviri (İBÇ) yöntemleri kullanılmaktadır ve yararları akademik ve ticari araştırmalarda kanıtlanmıştır [6].

İBÇ iki dildeki metin analizinden türetilen istatistiksel model bazında oluşturulur. İBÇ modelleriyle kaynak (k) dildeki her cümle ihtimallerle birlikte hedef (t) dildeki cümleye çevrilir [7]. Çeviri işlemi, Bayes'in olasılık teoremi [8] kullanılarak Eşitlik 1. deki gibi yazılabilir.

$$Pr(t|k) = \frac{Pr(t)Pr(t|k)}{Pr(k)} \tag{1}$$

Burada, *Pr(k), (k)*'den bağımsız olduğu için, en yüksek olasılıklı çeviri formülü, Eşitlik 2. deki gibi yazılabilir.

$$k = argmax_k Pr(t)Pr(t|k) \tag{2}$$

Pr(t) değeri hedef dil modeli, $Pr(t|k)$ çeviri modeli olarak adlandırılır.

İBÇ'nin temeli sayılan, sözcük tabanlı İBÇ sistemleri Stephan Vogel tarafından tasarlanan 1-5 IBM Hidden Markov modelleri [9] ve Franz-Joseph Och'un IBM 6 modeli [8], sözcük hizalama işleminde kullanılmaktadır. Sözcük tabanlı İBÇ'deki kısıtlamaları azaltmak için ifade tabanlı, N-GRAM tabanlı, vd. [10, 11] modeller geliştirilmiştir [12].

İfade tabanlı bilgisayarlı çeviri yönteminin amacı, uzunlukları farklı olan tüm sözcük dizilerini çevirerek sözcük tabanlı çevirideki kısıtları azaltmaktır [13].

Bayes'in teoremine göre çeviri yönü tersine çevrilir ve *PLM* dil modeli eklenerek ifade tabanlı model için çeviri modeli Eşitlik 3 ile ayrıştırılır.

$$\begin{aligned} t_{best} &= argmax_t P(k|t) = argmax_t P(t|k)P_{LM}(t) \\ &= \prod_{i=1}^{I} \varphi(t_i|k_i)\, d(start_i - end_{i-1} - 1) \end{aligned} \tag{3}$$

İfade tabanlı İBÇ için çeviri modelinde, kaynak cümle ilk önce ifadelere ayrıştırılır ve ifadelerin çeviri olasılıkları hesaplanır. Örneğin, Kırgız Türkçesinde bir *(k)* cümlesi *I* tane (k_i) ifadesine bölünür ve Türkiye Türkçesi (t_i) ifadelerine çevrilir. İfade çevirisi $\varphi(t_i|k_i)$ olasılık dağılımıyla modellenir. Türkiye Türkçesi (t_i) ifadelere yeniden sıralama modeli [13] uygulanır.

Türkiye Türkçesi (t_i) çıkış ifadelerinin yeniden sıralanma olasılığı rölatif bozulma olasılık dağılımı, d ile hesaplanabilir. Yeniden sıralama, bir önceki ifadeye göre gerçekleştirilir. $start_i$'yi, i'nci Türkiye Türkçesi (t_i) çıkış ifadesine çevrilen Kırgız Türkçesi (k_i) giriş ifadesini birinci sözcüğünün yeri olarak tanımlanır ve end_i aynı ifadenin son sözcük yeri olarak tanımlanır. Yeniden sıralama mesafe değeri $start_i - end_i - 1$ olarak hesaplanır[14].

İkinci çeviri yöntemi olarak Türkçe dil ailesi arasında N-GRAM tabanlı İBÇ olarak adlandırılan Marino vd. [12] tarafından geliştirilen N-GRAM İBÇ modeli uygulanmıştır. Bu model, kortej olarak adlandırılan belirli

iki dil biriminden oluşan N-GRAM dil modelini oluşturur. Öğretme cümle çiftleri, $kaynak_1^J, hedef_1^I K$ tane $(t_1, \ldots, t_K)$ kortejlere yani tekli parçalara ayrıştırılır. Bu şekilde çeviri modelinin cümle düzeyindeki olasılığı Kortej N-GRAM'ları (Eşitlik 4) kullanılarak hesaplanır.

$$p(kaynak, hedef) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k | t_{k-2}, t_{k-1}) \qquad (4)$$

İki dil kortej seviyesinde bağlantılı olduğundan bu çeviri modeli tarafından sağlanan içerik de iki dillidir. Genelde, ifade tabanlı İBÇ ve N-GRAM tabanlı İBÇ sistemleri için dil modelleme yöntemi olarak N-GRAM tabanlı dil modelleme kullanılır.

N-GRAM Tabanlı Dil Modelleme

Bilgisayarlı dilbilim ve olasılık alanlarında; N-GRAM, belirli metin ya da konuşma dizisinin *n* tane öğesinin bitişik sekansıdır. İBÇ'de öğeler, sözcükler veya sözcüklerin temel çiftleri olabilir. 1 boyutlu N-GRAM'a 'unigram', 2 boyutlu N-GRAM'a 'bigram', üç boyutlu N-GRAM'a 'trigram' adı verilir.

N-GRAM modeli, bir sonraki öğeyi tahmin etmek için kendisinden önce hangi *(n-1)* sözcükler olduğunu tahmin eder. N-GRAM dil modellemesi, cümledeki her birim için olasılık atama görevini yapar[15]. Bir cümleyi oluşturan N-GRAM grubu birbirlerine göre mantıklı yerleştirilirse, cümle mantıklı veya anlaşılır olabilir sayılır. Verilen Türkiye Türkçesi (t) cümle için $P(t)$ hesaplanmalıdır, yani (t) cümlesi öğeleri insan tarafından anlaşılabilir biçimde yerleştirilme olasılığıdır.

Genel olarak, $s_1 s_2 \ldots s_n$ sözcüklerinden oluşan Türkiye Türkçesi (t) cümlesi için olasılık aşağıdaki gibi (Eşitlik 5) yazılabilir:

$$P(t) = P(s_1 s_2 \ldots s_n) = P(s_1) P(s_2 | s_1) P(s_3 | s_1 s_2) \ldots P(s_n | s_1 s_2 \ldots s_{n-1}) \qquad (5)$$

Kaynak cümle için uygulanan çeviri ve dil modeli olasılıklarından en yüksek olan hedef cümleyi bulmak için kod çözümleme işlemi yapılır.

Kod Çözümleme

İfade ve N-GRAM tabanlı iki İBÇ modeli olasılık puanı atayan matematiksel eşitliklerle tanımlanmıştır. En yüksek olasılık dağılımlı çeviri değerini bulmak için çözücü kullanılır. Çözümleme, belirli giriş cümlesi için en iyi tercüme adayı, üstel sayıda olduğundan zor bir işlemdir. Çözümleme problemi *NP-Tamlık* tır (polynomial time) [13], yani sonuç arama alanı büyük olan problemlerde ek işlemler yapılarak arama alanı daraltılabilir.

Çözümleme için kullanılan başlıca algoritmalar; kaba kuvvet (brute force), arama (Beam search), yığın (stack) arama, artırımlı (Greedy increment) algoritmalarıdır.

Her çözümleme algoritmasının kendine özgü avantajları ve dezavantajları vardır. Hangi algoritmanın kullanılacağı İBÇ yöntemine ve donanıma bağlıdır.

Morfoloji ve Bilgisayarlı Çeviri

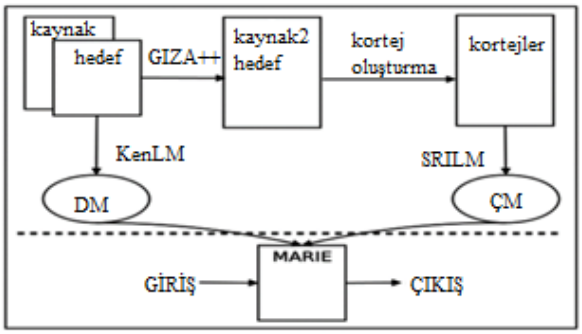
BÇ' de kaynak ve hedef dil yapısı çeviri kalitesini olumlu yada olumsuz yönde etkiler. Kelimeler arası ilişkiler, sözcük eki, bağlaçlar vb. morfolojik değişimler ile ifade edilir. Bu çalışmanın dilsel bileşenleri olan Kırgız Türkçesi ve Türkiye Türkçesi, Altay dillerine ait Türkçe Dil Ailesi üyesidir [16]. Her iki dil morfolojik ve sözdizimsel yönden birbirine benzerler, kelime türleri ve cümle öğeleri birbiriyle hemen hemen aynı seviyede farklı adlandırılarak gruplandırılmıştır [17, 18]. Sözdizimsel yandan da Kırgız Türkçesi ve Türkiye Türkçesi cümle öğelerinin sırası birbirine benzerdir [19, 20].

Çeviri dilleri olarak seçilmiş Kırgız Türkçesi ve Türkiye Türkçesi dillerinin sözdizimsel özellikleri göz önüne alındığında, çevrilen sözcüklerin yer değiştirilmesine, yada yeniden düzenlenmesine ihtiyaç olmadığı

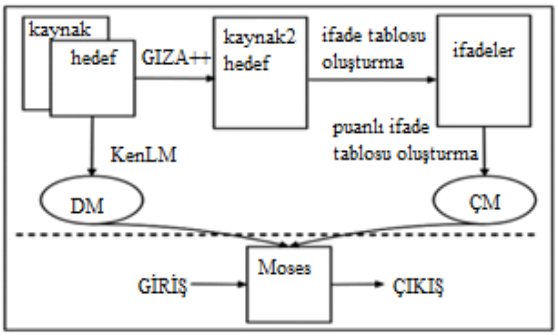
varsayılabilir. Bu iki dilin morfolojik özelliği dikkate alınırsa, kaliteli çeviri için büyük miktarda veriye ihtiyacın karşılanması gerektiği varsayılır.
Bu çalışma kapsamında Türkçe akraba dillerine N-GRAM Tabanlı ve İfade Tabanlı İBÇ sistemlerinin uygulanması ve çeviri performansı için BLEU başarım testi gerçekleştirilmesi hedeflenmiştir. Bu doğrultuda Kırgız Türkçesi ve Türkiye Türkçesi arasında hazırlanan sınırlı paralel korpus üzerinde eğitilmiş İBÇ sistemlerinin uygulanması ve araştırmaya konu edilen İBÇ sistemlerinin çeviri kalitesi ve farklarının araştırılması amaçlanmıştır.

GEREÇ ve YÖNTEM

Bu çalışmada, çeviri sistemi; Giza++ sözcük hizalama, N-GRAM tabanlı İBÇ için Stanford Research Instutite Language Model (SRILM) dil modelleme ve MARİE kod çözücü araçlarıyla kurulmuştur (Şekil 1.a). MOSES İBÇ sistemi için ise Ken Language Model (KenLM) dil modelleme ve MOSES kod çözücü araçlarıyla kurulmuştur (Şekil 1.b).
Her iki İBÇ yönteminin sonuçları Kırgız Türkçesi ve Türkiye Türkçesi dilleri arasında karşılaştırılmış ve BLEU başarım testi uygulanmıştır.



Şekil 1 a) N-GRAM İBÇ Sistem mimarisi.



Şekil 1 b) Moses İBÇ Sistem mimarisi.

N-GRAM tabanlı İBÇ sistemi talimatların sırayla takip edilmesiyle, sonraki bileşen bir öncekinin sonucundan faydalanarak kurulur.
İfade tabanlı İBÇ sistemi için Moses İBÇ sisteminin İfade tabanlı İBÇ bileşeni kullanılmıştır.

Deney Ortamının Kurulması

Deney Intel Core, 2.40GHz dört çekirdekli CPU, 8 Gb bellekli, Ubuntu 14.04 Uzun Vadeli Destek (Long Term Support, LTS) işletim sistemi kullanan bilgisayar üzerinde gerçekleştirilmiştir.
N-GRAM tabanlı ve İfade tabanlı İBÇ sistemleri birbirinden bağımsız olarak iki ayrı çeviri sistemi üzerinde denenmiştir.
Tam N-GRAM tabanlı İBÇ sistemini oluşturmak için;
a) Çeviri birimleri ayıklamak için EGYPT araç takımı parçası olan *GİZA++* sözcük-sözcüğe hizalama aracı kullanılmıştır. Ayıklama işleminin tam olarak uygulanması için program bileşenleri: *GIZA++*, *snt2cooc.out, mkcls* ikili olarak dosyalaştırılmıştır.
b) Ardından, kortej olarak adlandırılan N-GRAM tabanlı İBÇ'nin çeviri modelini oluşturan birimleri çıkarmak (üretmek) için MARİE araç takımı parçası olan *extract-tuple* ikili dosyası, kullanılan çevre için yapılmıştır.
c) Üretilen kortejleri kullanarak hedef ve iki dilli modelleri öğrenmek için SRILM dil modelleme aracı kurulmuştur.
d) N-GRAM tabanlı İBÇ sistemi için özel geliştirilmiş MARIE kod çözücü yazılımı *gcc 3.1.1* sürümü kullanılarak derlenmiştir.
Tam İfade tabanlı İBÇ sistemini oluşturmak için ise;

a) Herhangi İBÇ sisteminin temelini oluşturan sözcük-sözcüğe hizalanmış birimleri üretmek için aynı *GİZA++* aracı kullanılmıştır.
b) BÇ'ye İfade tabanlı istatistiksel yaklaşım aracı olarak iki ana bileşenden oluşan, bilgi kaynağı ve kod çözücü olarak Moses adlı İBÇ sistemi kurulmuştur. Moses sistemi, *g++-4.8* GNU C++ derleyicisi ile C++ programlama dili için, Boost kütüphaneler toplamı ek olarak kurulmuştur. Bilgi kaynağı bileşeni olarak sayılan *GİZA++* aracı Moses sistemine entegre edilmiştir.
c) Dil modelleme aracı olarak Moses sisteminin bileşik KenLM [21] aracı kullanılmıştır.

Deneylerin Gerçekleştirilmesi
Korpus Hazırlanması ve Önişleme:
Herhangi İBÇ sistemini eğitmek için cümle düzeyinde hizalanmış paralel veri (iki farklı dile çevirilen metin) gerekmektedir. Bu uygulamada, toplam olarak, Tablo 1'de gösterilen veri kaynağı ve miktarı kullanılmıştır.

Tablo 1 Sistemi öğretmek için kullanılan veri çeşidi ve sayısı

Kaynak	Cümle sayısı	Cümledeki sözcük sayısı (ortalama)
OPUS [33]	11.220	5
Gazete Manas	3.000	15
Sözlük [34]	84.342	1
Sözlükten alınmış İfade	27.600	3
TOPLAM	126.142	-

Genellikle, toplanmış verilerin ham, başka deyişle, bitişik noktalama işaretli, büyük harfli sözcükler ile oluşturulmuş cümlelerin ön işlenmesi tercüme kalitesini yükseltmek için önerilir.
Her iki, N-GRAM tabanlı ve İfade tabanlı İBÇ sistemi için korpus ön işlenmesi Moses İBÇ sisteminin bileşenleri; bölünme (tokenization) için *tokenizer.perl,* büyük-küçük harf düzenleme (truecasing) işlemi için *train-truecaser.perl,* ve *truecase.perl,* temizeleme (clean) için *clean-corpus-n.perl* araçları kullanılmıştır.
Dil modeli hedef dilde sözdizimsel ve anlamsal olarak daha düzgün çıktı sağlamak için kullanılır.
KenLM dil modelleme aracıyla, ön işlenmiş dosyalardan *.arpa* uzantılı 3-GRAM dil modeli üretilmiştir. Dosyanın çevirme zamanında daha hızlı yüklenmesi için Moses sisteminin yerleşik *build_binary* aracıyla ikili dosyası yapılmıştır.
İfade tabanlı İBÇ sisteminin eğitilmesi için, *GİZA++* aracıyla sözcük hizalama, ifade çıkartma ve olasılık dağılım puanlama, yeniden sıralama tablosunu oluşturma ve Moses yapılandırma dosyasını çıkartma işlemleri *train-model.perl* aracıyla yapılmıştır.
N-GRAM Tabanlı İBÇ Sisteminin Eğitilmesi için, kortej çıkartma ve kortejlerden iki dilli sözlük oluşturma aşamaları yapılmıştır. N-GRAM modelini eğitmek için çıkartılmış kortej, iki dilli birimlerin cümlesi olarak kabul edilebilir. SRİLM dil modelleme aracıyla iki dilli N-GRAM üniteleri 3-GRAM için çıkartılmıştır. Sonuçta, Tablo 2'de gösterilen N-GRAM kortej sayısı elde edilmiştir.

Tablo 2 N-GRAM kortej sayısı

Diller	Unigram	Bigram	Trigram
Türkçe → Kırgızca	115497	225936	14681
Kırgızca → Türkçe	124304	239626	11071

Kod Çözümleme
Çeviri için aday cümlelerin en iyisi, çeviri ve dil model olasılıklarını maksimize eden cümleyi bulmak kod çözümleme işlemi olarak adlandırılır.
a. İfade tabanlı kod çözümleme: İfade tabanlı İBÇ sistemi için Beam Search algoritmasına dayalı Moses kod çözücüsü kullanılmıştır. Kod çözümleme sırasında çeviri hipotezleri lineer model ile değerlendirilmiştir. Ön tanımlı model bileşenleri dil, ifade tablosu, yeniden sıralama tablosu olasılıklarıdır. Bu uygulama için kod çözücünün dil, İfade tablosu, yeniden sıralama tablosu ön tanımlı model bileşenleri olasılıkları değerlendirilerek gerçekleştirilmiştir [14].

b. N-GRAM tabanlı kod çözümleme: Beam Search algoritmasına dayalı N-GRAM tabanlı kod çözücü aracı olan MARİE kullanılmıştır [15]. Çeviri model bileşenleri olarak kortejlerden oluşan iki dilli sözlük ve dil modeli kullanılmıştır.
Bu uygulamadaki ifade ve N-GRAM tabanlı İBÇ sistemleri nispeten daha az veri üzerinde eğitilmiştir. Büyük miktardaki veri korpusları için uygulanmış durumlarda kod çözücü performansının yükseltilmesi için hipotez kısaltma, yada uzun giriş cümlelerinin çıktı cümlesi sözcük sayısını ayarlama gibi ek model bileşenleri için özel ayarlama yapılmamıştır. Dolayısıyla her iki sistem için uygulama, ek model bileşenleri ayarlanmamıştır.

Tercüme Kalitesini Test Etme
Bu uygulamadaki İBÇ sistemlerinin çeviri kalitesi BLEU(BiLingual Evaluation Understudy) algoritmasına dayalı olarak değerlendirilmiştir.
BLEU değerlendirme puanı her zaman 0 ile 1 arasındadır. Bu değerlendirme puanı 1'e yakın olması, bilgisayarlı çeviri, insan çevirisine yakın olduğunu gösterir [22]. BÇ sistemlerinde, BLEU değerleri yüzde olarak ifade edilir [23]. Bu nedenle, örneğin, 0.0370 BLEU değeri 3.70 olarak gösterilmiştir ve BÇ çıktı kalitesi değerlendirme için iBLEU aracı [23] kullanılmıştır.
Belirli bir alan üzerinde eğitilmiş sistemlerde, başka alandan oluşturulan cümle çevrisi düşük olur. Bu uygulamadaki çeviri sisteminde, örneğin, hukuk alanındaki cümleler çevrilirse, cümleler için hiç çeviri bulunmaması büyük ihtimaldir. İBÇ sistemi, bütün alandaki mevcut korpus ile eğitildiği zaman, her çeşit cümle için yüksek kalitedeki çeviri üretebilir. Dolayısıyla, bu çalışmada, hedeflenmiş çeviri alanı güncel ve edebi konu varsayılarak, deneme verisi olarak, güncel ifadeler seçilmiştir. Yansıra, iki farklı İBÇ sistemi kod çözücü araçlarını test etme amacıyla deneme verisi uzun ve kısa cümle olarak gruplandırılmıştır.

BULGULAR

Moses sisteminde İfade tabanlı ve Marie kod çözme sisteminde N-GRAM tabanlı İBÇ sistemlerin eğitiminde kullanılmamış olan kısa ve uzun cümleler için çeviri BLEU kalite testi yapılmıştır. Kısa cümleler için Kırgız Türkçesi ve Türkiye Türkçesi diller arasında iBLEU aracı değerlendirme sonuçları Tablo 3 ve Tablo 4'te gösterilmiştir.

Tablo 3. N-GRAM tabanlı İBÇ sistem eğitiminde kullanılmayan kısa cümle çevrilerinin iBLEU test sonuçları (%)

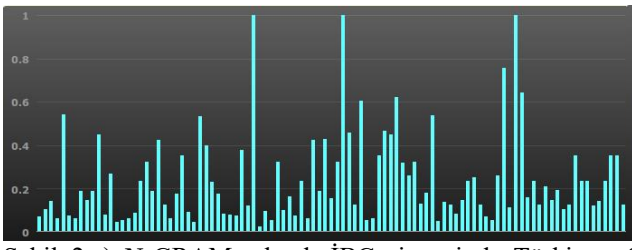
Cümle sayısı	Kırgızca→Türkçe	Türkçe→Kırgızca	Kırgızca→Türkçe (Ortak)	Türkçe→Kırgızca (Ortak)
100	4.26	6.22	16.41	18.07

Tablo 3'te Kırgız Türkçesi ve Türkiye Türkçesi için önce Kırgız Türkçesi kaynak ve Türkiye Türkçesi hedef dil seçilerek, sonra tam tersi sözcük hizalama işleminin ve iki taraflı hizalanmaların ortak noktalarının seçilmesinin etkinlik farkı görülebilir. Ortak noktaların seçilmiş durumdaki çeviri kalitesi daha yüksek olmuştur.

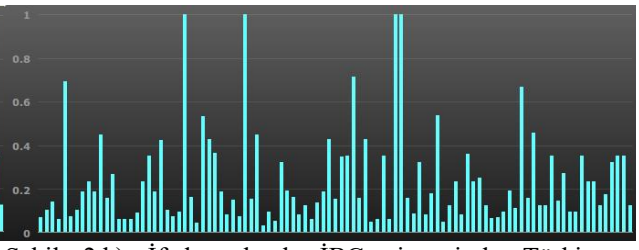
Tablo 4. İfade Tabanlı İBÇ sistem eğitiminde kullanılmayan kısa cümle çevrilerinin iBLEU test sonuçları (%)

Cümle sayısı	Kırgızca→Türkçe	Türkçe→Kırgızca
100	17.19	20.07

Tablo 4'te İfade tabanlı İBÇ sisteminde gerçekleştirilmiş Kırgız Türkçesinden Türkiye Türkçesine yapılan çeviri kalitesi, Türkiye Türkçesinden Kırgız Türkçesine yapılan çeviri kalitesinden daha düşük olmuştur.
Tablo 3 ve Tablo 4'te gösterilen N-GRAM tabanlı ve İfade tabanlı İBÇ sistemleriyle çevrilen kısa 100 cümlenin, iBLEU aracıyla elde edilen test sonuç grafikleri de Şekil 2.a ve Şekil 2.b'de gösterildiği gibi oluşmuştur.



Şekil 2.a) N-GRAM tabanlı İBÇ sisteminde Türkiye Türkçesinden Kırgız Türkçesine doğru çeviri (ortak nokta) için iBLEU değerlendirme grafiği



Şekil 2.b) İfade tabanlı İBÇ sisteminde Türkiye Türkçesinden Kırgız Türkçesine doğru çeviri için iBLEU değerlendirme grafiği

Yukarıda gösterilen grafiklerde, dikey değerler İBÇ sistemleri tarafından çevrilmiş hipotez cümlenin, insan tarafından çevrilmiş cümle ile denk gelme oranıdır. Örneğin, cümlelerin tam denk gelme değeri 1'dir. Yatay çizgideki dikey çizgiler ise çevrilmiş cümle dizisidir.

Kısa cümleler için çeviri, değişik alandan seçilmiş güncel konuşmada kullanılan cümleler üzerinde yapılmıştır. Cümlelerin kısa, özlü ve güncel olmasından çeviri kalitesi ortalama 0,137 (yüzde on) başarımla çevrilmiştir. Tablo 5'te Türkiye Türkçesinden Kırgız Türkçesine kısa cümleler için çeviri örneği gösterilmiştir.

Tablo 5. Türkiye Türkçesinden Kırgız Türkçesine kısa cümleler için çeviri örneği

Kaynak cümle	Bu çalışmadaki sistem çevrisi	İnsan çevrisi
Ama en yakın eczane 300 metre ileride.	бирок эң жакын дарыкана 300 метр ары жакта.	бирок эң жакын дарыкана 300 метр ары жакта.
İki kilo üzüm, üç kilogram şeftali, iki kilogram patates ve bir kilo fasulye verin lütfen?	эки килограмм жүзүм, үч килограмм шабдаалы, эки килограмм картошка жана бир килограмм фасоль verin сураныч?	эки килограмм жүзүм, үч килограмм шабдаалы, эки килограмм картошка жана бир килограмм төө буурчак бериңиз.
Çok kötü hissediyorum.	абдан жаман сезип жатам.	өтө жаман сезип жатам.
Hangi renkten hoşlanıyorsunuz?	кайсы өңдү hoşlanıyorsunuz?	кайсы өңдү жактырасыз?
Mavi renkten hoşlanırım.	көк өңдү hoşlanıyorum.	көк өңдү жактырам.

Tablo 6'da Kırgız Türkçesinden Türkiye Türkçesine kısa cümleler için çeviri sonucu örneği gösterilmiştir.

Tablo 6. Kırgız Türkçesinden Türkiye Türkçesine kısa cümleler için çeviri örneği

Kaynak cümle	Bu çalışmadaki sistem çevirisi	İnsan çevirisi
Анткени ар бир органдын өзүнчө кызматы бар .	çünkü her bir органдын kendince hizmeti var.	çünkü her organın ayrı ayrı görevleri var.
Гүл алгым келет кайсынысын сунуш кыласыз?	çiçek алгым gelir кайсынысын teklif yaparsınız ?	çiçek almak istiyorum ne tavsiye edersiniz?
Алгач үй айбандары бөлүмүнө саякат кылалы .	ilk hayvan бөлүмүнө seyahat yapalım.	önce evcil hayvanlar bölümüne gezelim.
Кайсы өң көрктүү ?	hangi yüz güzel?	hangi renk daha güzel?
Бул жазууну окуй аласынбы ?	bu yazıyı her türlü bilgiyi аласынбы ?	o yazıyı okuyabilir misin?

Tablo 5 ve Tablo 6'daki çeviri örnekleri, İBÇ sistemleri ve Türkçe dil ailesi özelliklerine, iyi bir örnek olabilir. Örneğin, Tablo 5'teki *"mavi renkten hoşlanırım"* cümlesindeki sözcükleri herhangi kural tabanlı BÇ sisteminde büyük ihtimalle *"көк түстөн/өңдөн жактырам"* olarak çevrilirdi ve anlamsal muğlâklık oluşturması ihtimaldir. Bu uygulamadaki İBÇ sistemlerini eğitme korpusunda "*renkten*" sözcüğü birkaç kez karşılaştırılmıştır ve büyük olasılıkla "mavi" ve "hangi" sözcükleriyle birlikte geldiği zaman "өңдү" olarak, örnek cümlenin "*mavi renkten*" sözcükleri semantik açıdan doğru çevrilmiştir. Bu yönden, İBÇ sistemleri içeriğin semantik aktarması özelliğiyle daha tercihlidirler.

Tablo 5'teki örneklerdeki "*hoşlanıyorsunuz*" ve "*hoşlanırım*" sözcükleri Kırgız Türkçesi ve Türkiye Türkçesi dillerinin morfolojik karmaşıklığına bir örnektir. "Hoşlanmak" sözcüğünün türetim ekleriyle iki ayrı sözcüğe dönüşmesi İBÇ sistemlerinde veri eksiklik sorunu yaratır ve bu sözcüklerin çevrilebilmesi daha büyük eğitme korpusuna ihtiyaç duyar.
Tablo 5'teki örnekte, "*bir kilo fasulye*" ifadesinin Kırgız Türkçesinde "төө буурчак" yerine, "*фасоль*" olarak bir Rusça sözcüğe çevrilmesi, İBÇ sisteminin eğitme korpusuyla doğrudan ilişkide bulunma örneğidir. Eğitme korpusu berrak olmazsa, yani metinler dilbilgisi ve semantik açıdan güvenilir olmazsa, İBÇ sistemlerinin çıktısı güvenilmez ve kalitesi düşük olabilir.
Tablo 5 ve Tablo 6'da değerleri gösterilen sınırlı veriler üzerinde kurulan İBÇ sistemlerinin kısa günlük cümleler için neredeyse her cümlede en az bir sözcüğün çevrildiği görülmüştür. Ek olarak, bu uygulamadaki İBÇ sistemleri kod çözme performansını daha ayrıntılı test etmek için sistem eğitiminde kullanılmayan uzun cümlelerin çevirisi test edilmiştir. Kırgız Türkçesi ve Türkiye Türkçesi dilleri için uzun cümle çevirilerinin iBLEU test sonuçları Tablo 7 ve Tablo 8'de gösterilmiştir.

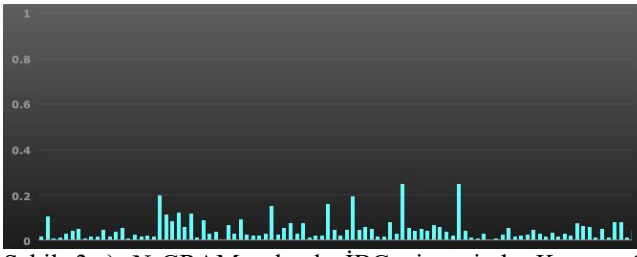
Tablo 7. N-GRAM tabanlı İBÇ sistem eğitiminde kullanılmayan uzun cümle çevrilerinin iBLEU test sonuçları (%)

Cümle sayısı	Kırgızca→Türkçe	Türkçe→Kırgızca	Kırgızca→Türkçe (Ortak)	Türkçe→Kırgızca (Ortak)
100	1.26	1.34	1.86	2.81

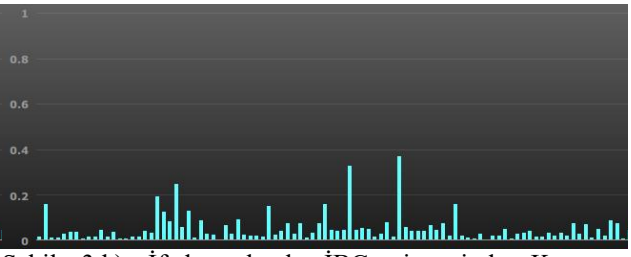
Tablo 8. İfade tabanlı İBÇ sistem eğitiminde kullanılmayan uzun cümle çevrilerinin iBLEU test sonuçları (%)

Cümle sayısı	Kırgızca→Türkçe	Türkçe→Kırgızca
100	3.08	2.41

Tablo 7 ve Tablo 8'de gösterilen N-GRAM tabanlı ve İfade tabanlı İBÇ sistemleriyle çevrilen uzun 100 cümlenin, iBLEU aracıyla elde edilen test sonuç grafikleri de Şekil 3.a ve Şekil 3.b'de gösterildiği gibi oluşmuştur.



Şekil 3.a) N-GRAM tabanlı İBÇ sisteminde Kırgız Türkçesinden Türkiye Türkçesine uzun cümle çevrisi (ortak nokta) için iBLEU değerlendirme grafiği



Şekil 3.b) İfade tabanlı İBÇ sisteminde Kırgız Türkçesinden Türkiye Türkçesine uzun cümle çevrisi için iBLEU değerlendirme grafiği

İBÇ sistemlerinin uzun ve edebi cümleler için, çeviri kalitesi düşük veya çeviri sonucu olmayan sonuçlar verdiği görülmüştür. Tablo 9'da Kırgız Türkçesinden Türkiye Türkçesine uzun cümle çevirisi örneği verilmiştir.

Tablo 9. Kırgız Türkçesinden Türkiye Türkçesine uzun cümleler için çeviri örneği

Kaynak cümle	көрсөтүүнүн жүрүшүндө студенттер депутатка саясий темадагы суроолор менен бирге жеке турмушуна байланыштуу, кесиптик жана интеллектуалдык деңгээлди текшерген суроолорду да беришти.
Bu çalışmadaki sistem çevirisi	көрсөтүүнүн жүрүшүндө öğrenci депутатка siyasî konu суроолор ile birlikte özel турмушуна bağlı, meslek ve интеллектуалдык деңгээлди текшерген суроолорду de беришти.
İnsan çevirisi	programda öğrenciler vekile siyaset dışında entelektüel, mesleki ve özel hayatı ile ilgili sorular sordular.

Tablo 9'daki çeviri, uzun cümleler için genel örnektir. Çoğu cümlelerde bir sözcük için çeviri bulunmuştur, ya da hiç bulunmamıştır.
Tablo 10'da Türkiye Türkçesinden Kırgız Türkçesine uzun cümle çevrisi örneği verilmiştir.

Tablo 10. Türkiye Türkçesinden Kırgız Türkçesine uzun cümleler için çeviri örneği

Kaynak cümle	17 mart 2015 tarihinde kırgızistan-türkiye manas üniversitesi akademisyen ve öğrencileri için edebiyat fakültesi tarafından düzenlenen ve baken kıdıkeeva kırgız devlet gençlik tiyatrosu ve yönetmen elvira ismailova özbek tiyatro yazarı ahmad saidin gelinlerin isyanı konulu komedisini sahteleştirdi.
Bu çalışmadaki sistem çevirisi	17 март 2015 tarihinde кыргыз-түрк манас университетинин академик жана студенттери үчүн адабият факультетинин тарабынан уюштурулган жана baken kıdıkeeva кыргыз мамлекеттик жаштар tiyatrosu жана кино elvira ismailova өзбек театр жазуучусу ahmad saidin gelinlerin isyanı темалуу komedisini sahteleştirdi.
İnsan çevirisi	2015 - жылдын 17 - мартында кыргыз-түрк « манас » университетинин окутуучу жана студенттери үчүн гуманитардык факультети тарабынан уюштурулган бакен кыдыкеева атындагы кыргыз мамлекеттик жаштар театры автору саид ахмат жана режиссёру эльвира ибраимова тарабынан сахналаштырылган " келиндер козголоңу " аттуу комедия спектаклин коюшту .

Daha önce bahsedildiği gibi, eğitme korpusunu oluşturan alanla ilişkili cümlelerin bulunma ihtimali, Tablo 10'daki çeviride gözlemlenmiştir. Bu uygulamadaki eğitme korpusunun kısmen Kırgızistan Türkiye Manas Üniversitesi gazetesinden oluşturulmasıyla, "kırgızistan", "türkiye", "manas" gibi sözcüklere çeviri bulunmuştur.
Bu uygulamadaki eğitme korpusu nispeten küçük olduğu için, her kısa ve uzun cümle için çeviri kalitesi nispeten düşük olmuştur.

4. TARTIŞMA ve SONUÇLAR

Bu çalışmada kullanılan İBÇ yöntemleri, güncelliği ve ayrıca Türk dilleri için uygulanmasında yenilik olduğu için seçilmiştir. Bu güne kadar Türkçe dil ailesi için, özellikle Kırgızca ve Türkçe için bir örnek çalışma olmadığından, bu çalışmada uygulanan sistemler, İBÇ'nin temel özelliklerini içermekte ve sınırlı veriler üzerinde uygulanmıştır.
Çalışmada tasarlanan N-Gram Tabanlı ve İfade Tabanlı İBÇ sistemlerini eğitme sonucunda; Kırgız Türkçesi ve Türkiye Türkçesi dilleri arasında çeviri işlemi gerçekleştirilmiştir. Eğitme verilerinin sınırlılığı ve belirli alana ait olması sebebiyle, deney verileri uzun ve kısa cümle olarak çeşitlendirilmiştir. Her iki İBÇ sistemi için deney çıktıları, iBLEU aracıyla değerlendirilmiş ve farklı sonuçlar elde edilmiştir.
N-GRAM tabanlı İBÇ sistem çevrisinde, kortejlerin tek taraflı (kaynak dilden hedef dile) hizalanması yerine, iki taraflı yani kaynak dilden hedef dile ve tam tersine, hedef dilden kaynak dile hizalanma gövdesinin ortak noktasından oluşturulması çeviri kalitesinin daha yüksek olmasını sağladığı [12], doğrulanmıştır.
İfade Tabanlı İBÇ sistem ile her bir kısa ve uzun cümle çevirisinde, N-GRAM Tabanlı İBÇ sisteminde karşılaştırmalı daha yüksek kaliteli çevriler elde edilmiştir. Ancak, İfade tabanlı İBÇ sisteminin çeviri ve dil modelleme araçlarının Moses İBÇ sistem araç takımıyla bütünleşik ve ayarlanmış olması, bunun aksine, N-GRAM tabanlı İBÇ sistemi öneri ve talimatları uygulayarak kurulması ve ayarlanması gereğinin dikkate alınması gerekir.
Son olarak, çeviri yönlerinin değişmesine ait (kaynak dil olarak Kırgız Türkçesini ya da Türkiye Türkçesini belirtmek) kararlı çeviri kalitesi yükselmesi, yada azalması gözlemlenmemiştir. Bu kararlılığı daha güvenilir gözlemleyebilmek için İBÇ sistemlerinin büyük korpus üzerinde eğitilmesi gerekir.
Elde edilen bulgulara dayanarak, bu çalışmanın daha sonra gelişmesi için; eğitme verilerini çoğaltmak, N-GRAM Tabanlı ve İfade Tabanlı İBÇ yöntemlerinin iyileştirme özelliklerinin ayarlanması önerilebilir. Büyük eğitme korpusunun eksik olmasıyla sözcük için hiç çeviri adayı bulunmayan durumlarda, İBÇ

sistemlerini kural tabanlı BÇ yöntemleri ile bütünleştirerek karma BÇ sistemlerinin uygulanması da bir diğer öneri olabilir.
Sonraki çalışmalarda, uygulanan İBÇ sistemlerinin gelişmesi için, İBÇ temelini oluşturan ve çeviri kalitesine doğrudan etkileyen paralel metin korpusu oluşturma ve toplama sistemi geliştirmek amaçlanabilir.

KAYNAKLAR

[1] Hutchins, W. J. (1986). *Machine translation: past, present, future* (p. 66). Chichester: Ellis Horwood.
[2] Gökgöz, E., Kurt, A., Kulamshaev, K., & Kara, M. (2011). Two-Level Qazan Tatar Morphology.
[3] Chéragui, M. A. (2012). Theoretical Overview of Machine Translation. *Proceedings ICWIT*, 160.
[4] Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation* (Vol. 362). London: Academic Press.
[5] Delavenay, E., & Delavenay, K. M. (1960). *An introduction to machine translation*. London: Thames and Hudson.
[6] Sadler, L. (1992, July). Rule-Based Translation as Constraint Resolution. In *Proc. FGNLP Workshop, S. Ananiadou (ed.)* (pp. 1-21).
[7] Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, *40*(3), 8.
[8] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19-51.
[9] Vogel, S., Ney, H., & Tillmann, C. (1996, August). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2* (pp. 836-841). Association for Computational Linguistics.
[10] Chiang, D. (2007). Hierarchical phrase-based translation. *computational linguistics*, *33*(2), 201-228.
[11] Marcu, D., & Wong, W. (2002, July). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 133-139). Association for Computational Linguistics.
[12] Marino, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., & Costa-Jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics*, *32*(4), 527-549.
[13] Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
[14] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48-54). Association for Computational Linguistics.
[15] Dunning, T. (1994). *Statistical identification of language* (pp. 10-03). Computing Research Laboratory, New Mexico State University.
[16] Slobin, D. I., & Zimmer, K. (Eds.). (1986). *Studies in Turkish linguistics* (Vol. 8). John Benjamins Publishing.
[17] Abduvaliev, I. (2008). Kyrgyz tilinin morfologiyasy. "Kyrgyz tili jana adabiyaty" adistigi boyuncha jogorku okuu jailardyn studentteri uchun okuu kitepteri. Bishek
[18] Korkmaz, Z. (2003). Türkiye Türkçesi grameri şekil bilgisi. Atatürk Kültür, Dil ve Tarih Yüksek Kurumu, Türk Dil Kurumu, Ankara.
[19] Akunova, A., Raimbekova, M., Karamendeeva, Ch. (2010). Azyrky Kyrgyz tili. Sintaksis. Jogorku okuu jaidyn studentteri uchun. Bishkek.
[20] Lewis, G. L. (1985). *Turkish grammar*. Oxford University Press, USA.
[21] Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 187-197). Association for Computational Linguistics.
[22] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.
[23] Madnani, N. (2011, September). iBLEU: Interactively debugging and scoring statistical machine translation systems. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on* (pp. 213-214). IEEE.