İstatiksel Bilgisayarlı Çeviride Paralel Derlemin Büyüklüğünün ve Kalitesinin Etkileri

Eray YILDIZ[1] A. Cüneyd TANTUĞ[2] Banu DİRİ[3]

[1,3] Bilgisayar Mühendisliği Bölümü
Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, İSTANBUL
[2]Bilgisayar Mühendisliği Bölümü
Bilgisayar ve Bilişim Fakültesi
İstanbul Teknik Üniversitesi, İSTANBUL

Email: yildizeray@hotmail.com.tr tantug@itu.edu.tr banu@ce.yildiz.edu.tr

Özet

Giriş: Birbirinin çevirisi olan metinlerden oluşan ve cümle seviyesinde hizalanmış olan paralel derlemler İstatiksel Bilgisayarlı Çeviri (İBÇ) için temel eğitim verisi görevini görmektedir. Yüksek başarılı çeviriler yüksek sayıda örnek içeren eğitim verileri ile mümkün olmaktadır ve İBÇ sisteminin çıktılarının kalitesi cümle çiftlerinin kalitesiyle doğrudan alakalıdır. Bu çalışma, büyük ve gürültülü bir paralel derlemin veya filtrelenmiş kaliteli bir derlemin eğitim verisi olarak kullanımının etkilerini gözlemlemek amacıyla yapılmıştır. Deneyler İngilizce ve Türkçe dilleri arasında çalışan İBÇ sistemlerinde gerçekleştirilmiştir.

Yöntem: Büyük bir paralel derlemdeki hataları elle gidermek pek mümkün olmadığından, paralel bir cümle çiftinin doğru olup olmadığını belirlemek için otomatik bir değerlendirme yöntemi gerekmektedir. Paralel derlemde yer alan paralel cümle çiftlerini kaliteli veya kalitesiz olarak sınıflandırmak için geliştirilen makine öğrenmesi tabanlı bir sınıflandırıcı, 1 milyon paralel cümle içeren bir paralel derleme uygulanmış ve 600 bin kaliteli paralel cümle elde edilmiştir. Tüm paralel derlem ve filtrelenmiş kaliteli derlem içerisinden seçilmiş farklı büyüklüklerdeki eğitim verileriyle İBÇ sistemleri eğitilip başarıları karşılaştırılmıştır

Sonuç: Paralel derlem büyüklüğünün etkilerini gözlemleyebilmek için ilk olarak farklı büyüklüklerde eğitim verileri ile deneyler yapılmıştır. Beklenildiği gibi deneyler İBÇ sistemlerinin başarısında etkili olan en önemli faktörün eğitim verisinin büyüklüğü olduğunu göstermiştir. Yine deneylerden elde edilen sonuçlara göre filtrelenmemiş derlemin tamamının kullanılmasındansa sadece filtrelenmiş kaliteli kısmı ile daha yüksek başarılar elde edilebildiği gibi uzun çalışma zamanları kısaltılabilmektedir. Geliştirilen filtre kullanılarak derlemin kaliteli olarak işaretlenen %60'ı ile sistem eğitilmiş ve derlemin tamamı ile eğitilen sisteme göre BLEU puanında %2.77 göreceli iyileşme başarılmıştır.

Tartışma: Geliştirilen kalite filtresi sayesinde gürültülü bir paralel derlemden elverişsiz örneklerin ayıklanması ile elde edilen daha küçük ve kaliteli bir derlem ile daha başarılı İBÇ sistemlerinin geliştirilebilmesi mümkün olmaktadır.

THE EFFECT OF PARALLEL CORPUS QUALITY VS SIZE IN ENGLISH-TO-TURKISH SMT

Eray YILDIZ[1] A. Cüneyd TANTUĞ[2] Banu DİRİ[3]

[1,3] Department of Computer Engineering
Yildiz Technical University, ISTANBUL
[2] Department of Computer Engineering
Istanbul Technical University, ISTANBUL

Email: yildizeray@hotmail.com.tr tantug@itu.edu.tr banu@ce.yildiz.edu.tr

Summary

Introduction: A parallel corpus plays an important role in statistical machine translation (SMT) systems. Higher translation accuracy can be achieved when machine translation systems are trained on increasing amounts of training data and the quality of an SMT system output extremely depends on the quality of sentence pairs. This study focuses on the effect of using a large parallel text versus using a filtered high-quality corpus. The experiments are carried out for an English-to-Turkish SMT task.

Method: Automatic parallel sentence quality evaluation method is required because it is not possible to correct the mistakes in a large parallel corpus manually. Therefore, we develop a machine learning based classifier to classify parallel sentence pairs as high-quality or poor-quality. We applied this classifier to a parallel corpus containing 1 million parallel English-Turkish sentence pairs and obtained 600K high-quality parallel sentence pairs. We train multiple SMT systems with various sizes of entire raw parallel corpus and filtered high-quality corpus and evaluate their performance.

Results: Several experiments are conducted with our entire raw parallel corpora and filtered high-quality parallel corpus. To see the impact of the parallel corpus size, we run tests with different corpus sizes up to 1M sentences. As expected, our experiments show that the size of parallel corpus is a major factor in translation performance. However, instead of extending corpus with all available "so-called" parallel data, a better translation performance and reduced time-complexity can be achieved with a smaller high-quality corpus using a quality filter. By using only 60% of the raw training data 2.77% relative improvement in BLEU score can be achieved with a quality classifier based filtering.

Discussion: The experiments show that our filtering method can be used for extracting unsuitable pairs from a noisy parallel corpus and the remaining pairs can still be effective in achieving results as high as the results of the entire raw corpus.

1. Giriş

Birbirinin çevirisi olan metinlerden oluşan ve cümle seviyesinde hizalanmış olan paralel derlemler İstatiksel Bilgisayarlı Çeviri (İBÇ) için temel eğitim verisi görevini görmektedir [1]. Paralel derlemler İBÇ'nin yanı sıra sözcük belirsizliği giderme, bilgi erişimi gibi diğer doğal dil işleme alanlarında da kullanılmaktadır. Yüksek başarılı bir İBÇ sistemi için en çok zaman alan işlem, paralel cümle çiftlerini içeren bir paralel derlem oluşturmaktır. İBÇ'nin bileşenlerinden olan çeviri modeli parametrelerin kestirimi için büyük boyutlarda eğitim verisine ihtiyaç duymaktadır [2]. Bu sebepten yüksek başarılı çeviriler yüksek sayıda örnek içeren eğitim verileri ile mümkün olmaktadır [3]. İBÇ sisteminin çıktılarının kalitesi cümle çiftlerinin kalitesiyle doğrudan alakalıdır. Erişilebilir paralel metin derlemlerinin bir kısmı sınırlı sayıda dil için Birleşmiş Milletler, Avrupa Parlamentosu, Kanada Parlamentosu gibi çok dilli organizasyonlar ve devletler tarafından üretilmektedir. İnsan emeğiyle paralel derlem oluşturmak oldukça maliyetli ve zaman gerektirdiği için paralel derlemler genellikle otomatik yöntemlerle üretilirler. Otomatik üretilen paralel derlemler hatalar ve belirsizlikler içerebilirler. Paralel olmayan cümleler içeren, düşük kalitede bir eğitim derlemi düşük kalitede çevirilere sebep olmaktadır [2]. Otomatik üretilmiş bir derlemdeki gürültü, kaynak ve hedef dokümanlardaki farklılıklardan, aslına uygun olmayan çevirilerden veya cümle hizalama hatalarından meydana gelebilir. Büyük bir paralel derlemdeki hataları elle gidermek pek mümkün olmadığından, paralel bir cümle çiftinin doğru olup olmadığını belirlemek için otomatik bir değerlendirme yöntemi gerekmektedir [4]. Otomatik değerlendirme yöntemi eşleşen sözcükler, cümle uzunluk ilişkisi, dilbilgisel doğruluk, akıcılık gibi etkenlerden faydalanabilir.

Bu çalışma, büyük ve gürültülü bir paralel derlemin veya filtrelenmiş kaliteli bir derlemin eğitim verisi olarak kullanımının etkilerini gözlemlemek amacıyla yapılmıştır. Deneyler İngilizce ve Türkçe dilleri arasında çalışan İBÇ sistemlerinde gerçekleştirilmiştir. Bölüm 2'de mevcut çalışmalardan Bölüm 3'te kullanılan eğitim verisinden, Bölüm 4'te de paralel derlem filtreleme yönteminden ve deneysel sonuçlardan bahsedilmektedir.

2. İlgili Çalışmalar

Paralel olmayan cümle çiftlerinin filtrelenmesi işlemi paralel metin madenciliğinin son işleme adımlarından biri olarak düşünülmektedir. Resnik ve Smith'in [5] geliştirdiği paralel metin elde etmek için Web'i tarayan sistemde çeviri benzerliği ölçüsü adını verdikleri bir özellik, toplanan verileri temizleme işleminde kullanılmıştır. Çeviri benzerliği ölçüsü kaynak metinde yer alan sözcüklerin çevirilerinin hedef dildeki metinde yer alıp almadığına bakılarak elde edilen bir ölçüttür.

Khadivi ve Ney [6] cümle uzunluklarını ve çeviri olasılığı ölçüsünü kullanan kural tabanlı bir model geliştirmiştir ve Avrupa Birliği sitesinden elde edilen bir paralel derlemi bu modelin ürettiği sonuçlara göre sıralayarak üstteki %97,5'lik kısımla BLEU puanını 46,8'den 47,2'ye çıkarmayı başarmışlardır. Yasuda ve diğerleri [7] İngilizce–Çince dilleri arasında her cümle için dil modelleriyle hesaplanan karmaşıklık skoru ile eğitim verisindeki olumsuz örnekleri eleyerek %1,76 BLEU puanı civarında ilerleme kat etmiştir. Liu ve Zhou'nun [4] çalışmasında paralel cümlelerden çeşitli dilsel özellikler çıkartılarak bir özellik vektörü elde edilmiş ve Destek Vektör Makineleri (DVM) yardımıyla problem, bir sınıflandırma problemi olarak ele alınmıştır. Yapılan testler sonucunda 40 binlik bir derlemde 0,88 tutturma ve bulma sonuçlarını elde etmişlerdir. Taghipour [2] ise dil modeline ve IBM çeviri modeline dayalı özellikleri kullanarak geliştirdikleri yöntemde sınıflandırma işlemi için maksimum entropi modelini kullanmışlardır. 48 binlik Farsça–İngilizce paralel derlem üzerinde yapılan deneyler sonucu sistemlerinin %98,3 doğruluk düzeyinde çalıştığı gösterilmiştir. Munteanu ve Marcu tarafından [8] yapılan çalışmada paralel olmayan haber kaynaklarından Çince, Arapça, İngilizce paralel cümleleri bulan bir sistem geliştirilmiştir. Maksimum entropi sınıflandırıcısı ile

cümleleri uzunluk, eşleşme ve sözcük hizalama sonucu elde edilen özellikleri kullanarak sınıflandıran bu sistem ile Arapça–İngilizce bir derlemde 0,94 tutturma ve 0,67 duyarlılık sonuçları elde edilmiştir.

Yine bu çalışmayla ilişkili olarak Schwenk'in [9] tek dilli metinleri çeviren ve sonrasında filtreleyerek eğitim verisine ilave eden yaklaşımı, Matsoukas'ın [10] eğitim verisindeki cümlelere farklı ağırlıklar veren yöntemi ve Axelrod'ın [11] özel bir konuya göre eğitim verisinden örnek seçen yöntemi gösterilebilir.

3. Eğitim Verisi

Bu çalışmada haber metinleri [12], yazınsal metinler [13], film altyazıları [14] ve çok dilli Web sayfaları [15] gibi farklı kaynaklardan alınan 1 milyon paralel cümle çifti içeren bir derlemden faydalanılmıştır. Bu kaynaklar cümle seviyesinde hizalanmış olsa da, önemli hizalama hataları içermektedir. Dolayısıyla, hizalama problemlerini gidermek ve belirli bir kalite seviyesini korumak için her bir derlemdeki paralel cümle çifti ön elemeden geçirilmiştir. Sözlük tabanlı bir cümle hizalama aracı olan Champollion [16] aracı bu ön eleme işlemi için kullanılmış ve bu işlemden sonra deneylerde kullanılan eğitim verisi oluşturulmuştur. Derlemlerin içerdiği cümle sayıları Tablo 1'de gösterilmiştir.

4. Deneyler

Yapılan ilk araştırma, paralel veriyi oluşturan kaynakların kalitesini gözlemlemek için yapılmıştır. Adil bir karşılaştırma yapabilmek için eğitim verisinin büyüklüğünün çeviri kalitesine olan etkisinden kaçınmak gerekmektedir. Bu sebeple paralel verilerden eşit büyüklükte (150K cümlelik) parçalar alınarak İBÇ sistemleri eğitilmiş ve sistemlerin performansları yine her bir kaynaktan eşit sayıda örnek alınarak oluşturulan 4K cümlelik bir test kümesi üzerinde ölçülmüştür. İBÇ sistemlerinin eğitimi için MOSES aracı [17]; başarı ölçüsü olarak sistem çıktıları ile referans çevriler arasındaki benzerliği ölçerek elde edilen BLEU puanı [18] kullanılmıştır. Tablo 2'de her derlemin cümle başına düşen ortalama sözcük sayısı ve çeviri başarıları verilmektedir.

Tablo 2'deki sonuçlar yorumlandığında, çeviri kalitesinde önemli farklılıkların olduğu görülmektedir. Bu duruma yol açan etkenler içerdikleri sözcük sayısının yanı sıra çeviri kaliteleri ve hizalama hatalarıdır. Dolayısıyla, kaliteli cümle çiftlerini kalitesiz olanlardan ayıracak bir sınıflandırıcı faydalı bir ön işleme adımı olarak düşünülebilir.

4.1. Kalite Sınıflandırıcısı

Derlemdeki her bir paralel cümle çifti için birbirlerinin uygun çevirileri olup olmadığını denetleyen bir yönteme ihtiyaç duyulduğu için; çeviri kalitesi ile ilişkili olduğu düşünülen özelliklerden faydalanarak otomatik bir sınıflandırıcı geliştirilmiştir. Çeviri kalitesi sadece hizalamanın doğruluğuna bağlı değildir, aynı zamanda dilbilgisel doğruluk, akıcılık ve doğru sözcük kullanımı da kaliteyi değerlendirmede etkilidir [4].

Dilbilgisi kurallarına aykırı olan bir cümle, çeviri modelinde bozulmalara yol açabilmektedir. Dilbilgisi ile ilişkili olarak kullanılan özellikler: bir yazım denetleyicisi [19] kullanılarak elde edilen İngilizce cümledeki hatalı yazılan sözcük sayısı ve akıcılığı, doğru sıralamayı ölçen dil modeli puanıdır. Uzun cümlelerin dil modeli olasılıklarının düşük olması da dikkate alınarak cümle uzunlukları da özellik olarak kullanılmıştır. İngilizce dil modelini üretmek için BerkeleyLM aracı [20] ve Web 1T[1] derleminden faydalanılmıştır. Filtreleme işleminde kullanılan özelliklere cümlelerin uzunluklarının yanı sıra farkları ve oranları da eklenmiştir. Kullanılan son özellik ise sözlüğe göre karşı tarafta çevirileri bulunan sözcüklerin oranıdır. Bu özellik cümlelerin içerikleriyle ilgili bir özellik olup, 88.824

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

İngilizce sözcük ve Türkçe karşılıklarını içeren elektronik bir sözlükten ve sözcüğün ilk 5 harfini baş kelime (lemma) olarak ele alan basit kök eşleştirme yönteminden yararlanılmıştır [21].

Sınıflandırıcının eğitimi için gereken eğitim verisi paralel derlemlerden rasgele seçilen örneklerin insan emeğiyle 'kaliteli' ve 'kalitesiz' olarak etiketlenmesiyle oluşturulmuştur. 983 cümle çifti 'kaliteli', 160 cümle çifti 'kalitesiz' olarak işaretlenmiştir. 'kalitesiz' olarak işaretlenen bu 160 örneğin kalitelerinin düşük olmasının sebebi dilbilgisel ve yazınsal hatalardan, aslına uygun olmayan çevirilerden kaynaklanmasıdır. Ayrıca, yanlış rasgele hizalamalar sonucu üretilen 674 örnek de 'kalitesiz' olarak işaretlenmiş ve eğitim verisine yapay gürültü olarak eklenmiştir. Tablo 3'te sınıflandırıcının eğitimi için kullanılan eğitim verisi gösterilmektedir.

Sınıflandırıcıyı eğitirken en uygun yöntemi belirlemek amacıyla farklı makine öğrenmesi yöntemleri olarak Radyal Tabanlı Ağlar (RTF), Rassal Karar Ormanı (RKO), Çok katmanlı Yapay Sinir Ağı (YSA), Destek Vektör Makineleri (DVM) ve Naive Bayes (NB) tabanlı sınıflandırıcılar ile WEKA aracı [25] kullanılarak deneyler yapılmış ve Tablo 4'de yer alan sonuçlar elde edilmiştir. Tüm deneyler onlu çapraz doğrulama ile yapılmıştır. Tutturma ölçütüyle sınıflandırıcının yaptığı sınıflandırmaların ne kadarının doğru olduğu ifade edilirken, bulma ölçütüyle gerçekten kaliteli olan örnek cümle çiftlerinin ne kadarının sınıflandırıcı tarafından bulunduğu gösterilmektedir. Sonuçlara bakıldığında; RKO algoritmasının filtreleme görevi için en uygun sınıflandırıcı olduğu görülmektedir.

Derlemlerden seçilerek elle etiketlenen kalitesiz örneklerin sınıflandırılması işleminin yapay üretilen kalitesiz örneklerin sınıflandırılması işlemine göre daha güç olduğu söylenebilir. Derlemlerden seçilen bu 160 kalitesiz örnek üzerinde RKO algoritmasının verimliliğini görmek için derlemlerden seçilen 80 kalitesiz örnek eğitim verisinden çıkartılıp test kümesi olarak kullanılmıştır. RKO algoritması ile 76 örneğin doğru etiketlendiği görülmüştür (0,95 doğruluk).

4.2. Deneysel Sonuçlar

Derlem büyüklüğünün etkilerini gözlemleyebilmek için Tablo 1'de gösterilen veri kümelerinden alınan örneklerden oluşturulan 1 milyon cümle çiftine kadar farklı büyüklüklerde eğitim verileri kullanılarak deneyler yapılmıştır. Eğitim verisinin %10'luk kısmı test kümesi olarak kullanılırken, Türkçe kısmı ile dil modelinin eğitimi gerçekleştirilmiştir. İngilizce-Türkçe ve Türkçe-İngilizce yönlerinde yapılan deneylerde elde edilen çeviri başarıları BLEU puanı olarak Şekil 1 ve Şekil 2'de görselleştirilmiştir.

Beklenildiği gibi, daha çok verinin daha yüksek başarı sağlayacağı tezi doğrulanmıştır. İngilizce'den Türkçe'ye yapılan deneylerde, işlenmemiş başlangıç derlemindeki paralel cümlelerin sayısı 100 binden 1 milyona çıktığında BLEU puanı 15,33'den 34,24'e çıkmıştır (Şekil 1). Grafikteki eğilim dikkate alındığında 1 milyon cümleden daha fazla eğitim kümesiyle daha yüksek başarılar elde edilebileceği söylenebilir. Çıkarılan bir diğer sonuç, yüksek başarılara daha küçük ve kaliteli paralel bir derlem ile ulaşılabileceğidir. Kalite sınıflandırıcısı ile filtrelenerek elde edilen başlangıç derleminin %60'ı kullanılarak BLEU puanı 35,19'a çıkmış, yani %2,77 göreceli ilerleme sağlanmıştır. Eğitim verisinin büyüklüğünün azalması günler ve haftalar süren eğitim zamanında %40 civarında bir azalma sağlamaktadır.

5.Sonuçlar

Paralel derlemin kalitesinin İBÇ sistemlerinin başarısına olan etkilerini gözlemlemeyi amaçlayan bu çalışma da çeviri benzerliği ve dilbilgisel özellikleri kullanarak paralel cümle çiftlerini kaliteli veya kalitesiz olarak sınıflandıran RKO

sınıflandırıcısının kullanılması önerilmektedir. Geliştirilen kalite filtresi sayesinde gürültülü bir paralel derlemden elverişsiz örneklerin ayıklanması ile elde edilen daha küçük ve kaliteli bir derlem ile daha başarılı İBÇ sistemlerinin geliştirilebilmesi mümkün olmaktadır. Geliştirilen paralel derlem filtreleme yönteminin ortak bir test kümesinin olmayışı ve farklı diller üzerinde çalışmaları sebebiyle her ne kadar adil bir kıyaslama yapmak mümkün olmasa da filtreleme üzerine yapılan diğer çalışmalarla karşılaştırıldığında tutturma-bulma değerleri ve Bleu puanlarındaki iyileşme dikkate alındığında benzer başarılar gösterdiği gözlemlenmektedir.

Türkçe ve İngilizce dilleri arasında paralel metin kaynakları ne yazık ki kısıtlıdır. Film altyazıları gibi büyük boyutlarda paralel derlemler var olsa da bu tarz gürültülü derlemleri doğrudan eğitim verisine koymak yanlış olacaktır. Bu çalışmada geliştirilen filtreleme metodu ile böyle gürültülü kaynakların kullanımı sağlanmış olacaktır. Ayrıca geliştirilen filtreleme metodu paralel derlemlerden elverişsiz örneklerin ayıklanması işleminde kullanıldığı gibi Wikipedia sayfaları gibi paralel olmayan kaynaklardan paralel cümlelerin tespiti içinde kullanılabilmekte ve yeni paralel derlemlerin oluşturulmasına katkı sunabilmektedir.

Deneyler Türkçe ve İngilizce dilleri arasında yapılmış olsa da, diğer dillere uyarlanması oldukça kolaydır.

Geliştirilen filtreleme metodunun, Türkçe-İngilizce dilleri arasında geniş kapsamlı ve kaliteli bir paralel derlem oluşturma işleminde kullanılması planlanmaktadır.

6. Kaynaklar

[1] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer ve Paul S. Roossin. (1990) "A statistical approach to machine translation", Computational Linguistics, 16 (2):79–85.

[2] Taghipour, Kaveh, Nasim Afhami, Shahram Khadivi ve Saeed Shiry. (2010) "A discriminative approach to filter out noisy sentence pairs from bilingual corpora", 5th International Symposium Telecommunications (IST),: 537-541.

[3] Koehn, Philipp. (2002) "Europarl: A multilingual corpus for evaluation of machine translation", Information Sciences Institute, University of Southern California.

[4] Liu, Xiaohua ve Ming Zhou. (2010) "Evaluating the quality of web-mined bilingual sentences using multiple linguistic features", Asian Language Processing (IALP)

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

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

[7] Yasuda, Keiji, Ruiqiang Zhang, Hirofumi Yamamoto and Eiichiro Sumita. (2008) "Method of Selecting Training Data to Build a Compact and Efficient Translation Model" IJCNLP, Hyderabad, India

[8] Munteanu, Dragos Stefan, ve Daniel Marcu. (2005) "Improving machine translation performance by exploiting non-parallel corpora", Computational Linguistics 31.4 : 477-504.

[9] Schwenk, Holger. (2008) "Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation". International Workshop on Spoken Language Translation

[10] Matsoukas, Spyros, Antti-Veikko I. Rosti, and Bing Zhang. (2009) "Discriminative Corpus Weight Estimation for Machine Translation". EMNLP, Singapore : 708--717

[11] Axelrod, Amittai, Xiaodong He, ve Jianfeng Gao. (2011) "Domain adaptation via pseudo in-domain data selection." EMNLP

[12] Tyers, Francis M., and Murat Serdar Alperen. (2010) "SETimes: a parallel corpus of Balkan languages", LREC2010, Malta, pp 49–53

[13] Taşçı, Şerafettin, A. Mustafa Güngör, and Tunga Güngör. (2006) "Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment", International Scientific Conference Computer Science

[14] Tiedemann, Jörg. (2009) "News from opus - a collection of multilingual parallel corpora with tools and inter-faces", Natural Language Processing, volume V, Amsterdam/Philadelphia

[15] Yıldız, Eray ve Tantuğ, A.Cüneyd. (2012) "Evaluation of Sentence Alignment Methods for English-Turkish Par-allel Texts", LREC 2012. Istanbul

[16] Ma, Xiaoyi. (2006) "Champollion: A Robust Parallel Text Sentence Aligner", LREC 2006.

[17] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, ... and Evan Herbst . (2007) "Moses: Open source toolkit for statistical machine translation", ACL 2007 : 177-180.

[18] Papineni, Kishore, Salim Roukos, Todd Ward ve Wei-Jing Zhu. (2002) "BLEU: A Method for Automatic Evaluation of Machine Translation", Computational Linguistics :311-318

[19] Idzelis Mindaugas. (2005) "Jazzy: The java open source spell checker", http://jazzy.sourceforge.net/

[20] Pauls, Adam, ve Dan Klein. (2011) "Faster and smaller n-gram language models", Association for Computational Linguistics,, Portland, Oregon.

[21] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann ve Ian H. Witten.(2009) "The weka data mining software: An update", ACM SIGKDD explorations newsletter 11.1 : 10-18.

Tablo 1. Ön-eleme işlemi sonrası eğitim kümesi

EğitimKümesi	**CümleSayısı (Bin)**
Haber	200
Altyazı	1.200
Web	150
Yazınsal	680
Total	**2,230K**

Deneylerde kullanılan eğitim verileri ve içerdikleri cümle sayıları Tablo 1'de gösterilmiştir.

Tablo 2. Her bir derlemden eşit sayıda örnekle eğitilmiş İBÇ sistemlerinin çeviri başarıları

EğitimVerisi	**EğitimVerisi Büyüklüğü (Bin Cümle)**	**OrtalamaCümleUzunluğu (Sözcük)**	**BLEU(%)**
$Haber_{150K}$	150	17,25	16,59
$Altyazı_{150K}$	150	10,41	4,72
Web_{150K}	150	23,1	19,56
$Yazınsal_{150K}$	150	17,71	9,49

Paralel derlemlerden eşit büyüklükte (150K cümlelik) parçalar alınarak İBÇ sistemleri eğitilmiş ve sistemlerin BLEU ölçüsünde çeviri başarıları ve cümle başına düşen ortalama sözcük sayısı Tablo 2'de verilmektedir.

Tablo 3. Sınıflandırıcı için Eğitim Verisi

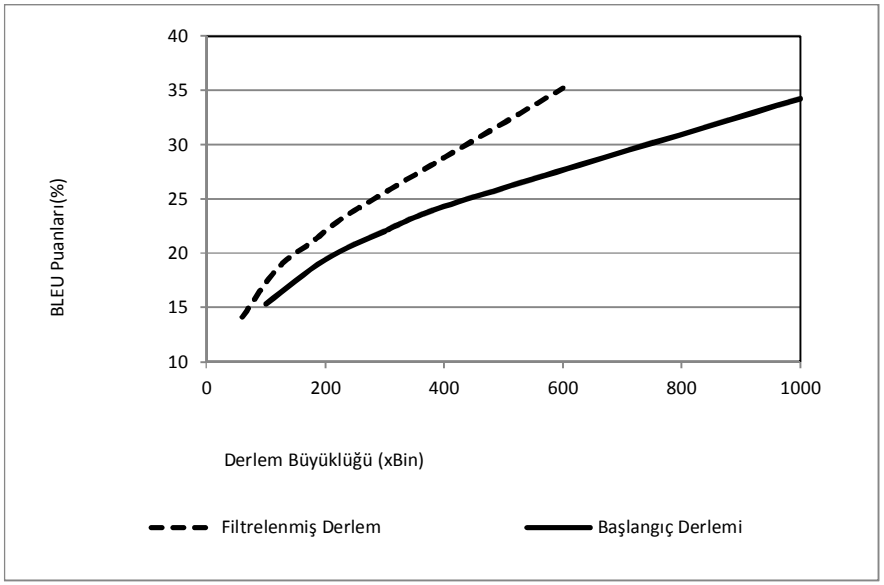
Cümle Çifti Kalitesi	**Adet**
Kaliteli	983
Kalitesiz	
YapayGürültü	674
DerlemlerdenSeçilen	160
Total	**1817**

Sınıflandırıcının eğitimi için gereken eğitim verisi paralel derlemlerden rasgele seçilen örneklerin insan emeğiyle 'kaliteli' ve 'kalitesiz' olarak etiketlenmesiyle oluşturulmuştur. 983 cümle çifti 'kaliteli', 160 cümle çifti 'kalitesiz' olarak işaretlenmiştir. Ayrıca, yanlış rasgele hizalamalar sonucu üretilen 674 örnek de 'kalitesiz' olarak işaretlenmiş ve eğitim verisine yapay gürültü olarak eklenmiştir. Tablo 3'te sınıflandırıcının eğitimi için kullanılan eğitim verisi gösterilmektedir.

Tablo 4. Sınıflandırma algoritmalarının başarıları

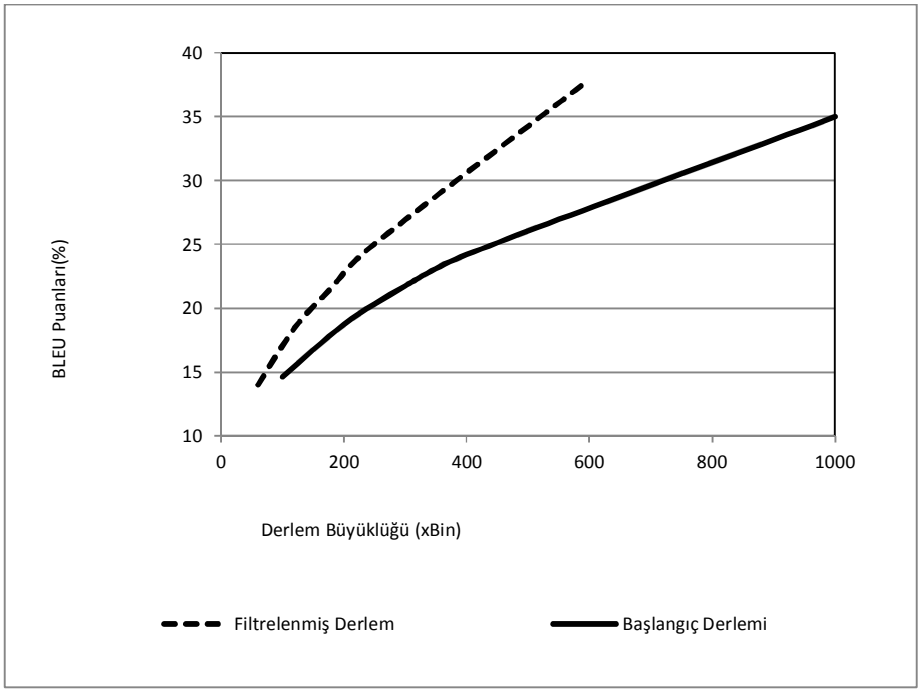
Sınıflandırıcı	**Tutturma**	**Duyarlılık**	**F1**
RTF	0.903	**0.964**	0.933
RKO	**0.969**	0.960	**0.965**
YSA	0.938	0.953	0.946
DVM	0.932	0.963	0.947
NB	0,736	0,930	0,822

Sınıflandırıcıyı eğitirken en uygun yöntemi belirlemek amacıyla farklı makine öğrenmesi yöntemleri kullanılarak deneyler yapılmış ve Tablo 4'de yer alan sonuçlar elde edilmiştir. Tüm deneyler onlu çapraz doğrulama ile yapılmıştır. Sonuçlara bakıldığında; RKO algoritmasının filtreleme görevi için en uygun sınıflandırıcı olduğu görülmektedir.



Şekil 1. İBÇ sistemlerinin eğitim verisi büyüklüğü ve kalitesine göre sonuçları (İngilizce'den Türkçe'ye)

Derlem büyüklüğünün etkilerini gözlemleyebilmek için 1 milyon cümle çiftine kadar farklı büyüklüklerde eğitim verileri kullanılarak deneyler yapılmıştır. Eğitim verisinin %10'luk kısmı test kümesi olarak kullanılırken, Türkçe kısmı ile dil modelinin eğitimi gerçekleştirilmiştir. İngilizce-Türkçe yapılan deneylerde elde edilen çeviri başarıları BLEU puanı olarak Şekil 1'de görselleştirilmiştir.



Şekil 2. İBÇ sistemlerinin eğitim verisi büyüklüğü ve kalitesine göre sonuçları (Türkçe'den İngilizce'ye)

Derlem büyüklüğünün etkilerini gözlemleyebilmek için 1 milyon cümle çiftine kadar farklı büyüklüklerde eğitim verileri kullanılarak deneyler yapılmıştır. Eğitim verisinin %10'luk kısmı test kümesi olarak kullanılırken, Türkçe kısmı ile dil modelinin eğitimi gerçekleştirilmiştir. Türkçe-İngilizce yapılan deneylerde elde edilen çeviri başarıları BLEU puanı olarak Şekil 2'de görselleştirilmiştir.