Türkçe-İngilizce için İstatistiksel Bilgisayarlı Çeviri Sistemi

İlknur Durgar El-Kahlout ve Kemal Oflazer

Mühendislik ve Doğa Bilimleri Fakültesi
Sabancı Üniversitesi
İstanbul, 34956, Türkiye
ilknurdurgar@su.sabanciuniv.edu, oflazer@sabanciuniv.edu
http://www.hlst.sabanciuniv.edu

Özetçe. Bu bildiride, Türkçe İngilizce dil çifti için istatistiksel bilgisayarlı çeviri sistemi anlatılmaktadır. İki dil arasındaki yapısal farklılıklardan kaynaklanan problemler, biçimbirimsel analiz yapılarak eklerin ayrı gösterimi ile ortadan kaldırılmıştır. Yaklaşım sözcük öbeği tabanlı çözücü ile test edilmiştir. Sistem performansı, eklerin bigram tabanlı gruplandırılması ile iyileştirilmiştir. Önerilen metot ile standart modele kıyasla daha iyi sonuçlar elde edilmiştir. 22000 cümlelik paralel metinler ile oluşturulan sistemin performansı tatmin edici olmasa da bir başlangıçdır.

1 Giriş

Bir dilin (kaynak dil) diğer bir dile (hedef dil) otomatik olarak çevrilmesi diğer adıyla bilgisayarlı çeviri (BÇ) bilgisayar bilimlerinin ve doğal dil işlemenin çok eskiden bu yana ilgilendiği konulardan biridir. Bu tür bir çalışmanın yapılabilmesi için bilgisayarın her iki dili, dillerdeki eş anlamlı sözcükleri, sözcük öbeklerini ve gramerlerini bilmesi gerekir. BÇ için uygulanabilecek yaklaşımlardan biri dilbilimcilerin gerekli bilgileri kurallar kümesi olarak bilgisayara tanımlamasıdır ki bu uzun zaman alacak emek yoğun bir iştir ve de temelde şu ana kadar belli-başlı birkaç dil çifti dışında çok da başarılı olamamıştır. Daha yeni bir yaklaşım ise cümle bazında eşleştirilmiş birbiri ile aynı içeriği taşıyan iki farklı dilde yazılmış paralel metinlerin bilgisayara yüklenmesi ve bilgisayarın istatistiksel metotlar ile bu bilgiden yola çıkarak diğer tüm bilgileri otomatik olarak öğrenmesine dayanmaktadır.

İstatistiksel Bilgisayarlı Çeviri (İBÇ) yaklaşımının popüler olmasının sebebi, paralel metinler dışında ekstra bir dil bilgisine başvurmadan etkili sonuçlar üretmesidir. İBÇ çeviri işlemini gürültülü kanal sinyal geri elde etme problemine benzer olarak çözmektedir [1, 2]. Örneğin, bir İngilizce tümce e, birçok Türkçe tümceye çevrilebilir. İstatistiksel çeviri ilk adımda bütün Türkçe tümcelerin, bütün İngilizce tümcelerin çevirisi olduğunu kabul eder fakat her İngilizce tümcenin, Türkçe tümcenin çevirisi olmasının belirli bir olasılığı vardır. Herhangi bir sözcük öbek çifti (t,e) için, $Pr(t|e)$ verilen İngilizce tümce e'nin, çevrildiği zaman Türkçe tümce t'yi üretme olasılığıdır. Çeviri sisteminin amacı, verilen e için, bir çevirmenin üreteceği en yüksek olasılıklı t öbeğini bulmaktır;

$$t^* = \arg\max_t \Pr(t \mid e) \quad (1)$$

Tümceler doğru yapılanmış ve yanlış yapılanmış olarak iki gruba ayrılabilir. Örneğin, *yağmur yağdığı için maç iptal edildi* ve *he is not here* cümleleri doğru yapılanmış cümleler iken, *için edildi maç yağdığı iptal yağmur* ve *here is he not* cümleleri yanlış yapılanmış cümlelerdir. Doğru bir çeviride sadece kaynak dilde bulunan sözcüklerin doğru hedef dil sözcüklerine birebir çevrilmesi yeterli değildir. Bunun yanısıra doğru sözcük sıralaması beklenir. Hedef dilde üretilecek söz öbeğinin, belirli bir olasılığı olmalıdır. Üretilecek olan söz öbeğinin olasılığını hesaplayabilmek için *Pr(t|e)* Bayes kanunu kullanılarak yeniden yazıldığında, belirli bir İngilizce söz öbeği *e* için payda kaynak söz öbeği *t*'den, bağımsız olduğundan istatistiksel bilgisayarlı çevirinin temel denklemi;

$$t^* = \arg\max_t \frac{\Pr(e|t)\Pr(t)}{\Pr(e)} \cong \arg\max_t \Pr(e \mid t)\Pr(t) \quad (2)$$

şeklinde yazılmaktadır. Denklemdeki *Pr(t)* hedef dildeki söz öbeğinin olasılığını, *Pr(e|t)* ise çeviri olasılığını ifade etmektedir. *Pr(t)* dil modeli, *Pr(e|t)* ise çeviri modeli olarak adlandırılmaktadır.

İlk İBÇ sitemleri, dillerin biçimbirimsel veya sözdizimsel özelliklerine dikkat etmeksizin, salt kelime-tabanlı yaklaşımları kullanan sistemlerdir [2]. Takip eden yaklaşımlar, biçimbirimsel ve sözdizimsel özelliklerin bir şekilde modellere dahil edilmesini göstererek kullanmışlardır [1, 3-5].

İBÇ sistemleri dil ve çeviri modeli parametrelerini cümle bazında eşleştirilmiş paralel metinlerden tahmin eder [1]. Doğru parametreleri elde edebilmek için sistemler olabildiğince çok paralel metine ihtiyaç duymaktadır. Fakat yeterli derecede paralel metin elde etmek bazı diller için mümkün değildir. Bazı dillerde ise yeterli miktarda paralel metin bulunsa bile dilin biçimbirimsel yapısı parametreleri yaklaşık şekilde elde etmeye uygun değildir. Türkçe'nin dil yapısı günümüze kadar geliştirilmiş olan İBÇ sistemlerinde kullanılan dillerden oldukça farklı olduğu için, varolan yaklaşımlar Türkçe için birebir kullanıma uygun değildir. Türkçe'nin eklemeli bir dil olması çeviri modeli parametrelerini doğru şekilde elde etmeye engeldir. Türkçe'ye özel problemlere uygun yaklaşımlar üretilerek özgün bir çalışma yapılması gerekmektedir.

Bu bildiride, Türkçe İngilizce dil çifti için paralel metinlerin eşleştirilmesinden çıkan sonuçlar, sistemin son durumu ve çözümlenmesi planlanan problemler anlatılmaktadır. Çalışmada, Lee'ni çalışmasına [4] benzer bir biçimbirimsel yaklaşım kullanılmış, ve yaklaşım daha da genelleştirilerek daha yaklaşık sonuçlar elde edilmiştir.

2 Türkçe'nin Dil Yapısına Genel Bakış

Türkçe Ural-Altay dil ailesine ait sondan eklemeli bir dildir. Sözcüğün anlamı İngilizce gibi dillere göre oldukça farklıdır. Sözcükler bir çok çekim ve yapım eklerinin kök sözcüğe eklenmesi ile oluşur. Her biçimbirim farklı bir bilgi taşımaktadır. Kök sözcüklere biçimbirimler eklenerek binlerce yeni sözcük türetilebilir. Çeşitli kurallar, biçimbirimlerin değişik sözcüklerde değişik biçimler almasına sebep olur. Türkçe bir sözcük kimi zaman İngilizce bir cümleyi ifade edebilir. Örneğin, sağlamlaştırdığımızdaki[1] sözcüğü sağlam +laş +tır +dığ +ımız +da +ki[2] şeklinde ayrıştırabilir. Bir adım öteye taşıyarak sözcüksel biçimi sağlam +lAş +DHr +DHk +HmHz +DA +ki olarak ifade edilir [6]. Bu gösterimde örneğin +DHr, fiil yapım ekleri *+dır, +dir, +dur, +dür, +tır, +tir, +tur, +tür*'ü temsil etmektedir. Gösterimin amacı, dil kuralları gereği farklı biçimler alan fakat aynı bilgiyi taşıyan biçimbirimleri tek bir şekilde ifade etmektir. Örneğin, yüzey biçimleri farklı olan *defterinde* ve *masasında* sözcükleri, sözcüksel biçimde *defter+sH+ndA* ve *masa+sH+ndA* şeklinde ifade edildiğinde iki sözcüğün aynı biçimbirimler ile türetildiği görülmektedir.

3 Türkçe-İngilizce Paralel Metinlerin Kelime bazında Eşleştirilmesi

"Aradığında okuldan eve gidiyorum" ve *"When he called, i was going home from the school"* cümle çifti için sözcük tabanlı bir eşleştirme yapıldığında, Şekil 1'e benzer bir sonuç elde edilir. Şekilde de görüldüğü üzere, bir Türkçe sözcük İngilizce birçok sözcük ile eşleşmektedir. Bu tarz bir eşleştirmenin problemi, İngilizce paralel metinde yüksek frekanslarda bulunan bir sözcük Türkçe paralel metinde daha düşük frekansta bulunması ya da hiç bulunmamamasıdır. Buna rağmen sözcüğün değişik biçimleri metinde olabilir. Tablo 1 *faaliyet* sözcüğünün farklı biçimlerinin örnek bir İngilizce - Türkçe paralel metinin Türkçe kısmındaki frekanslarını vermektedir.

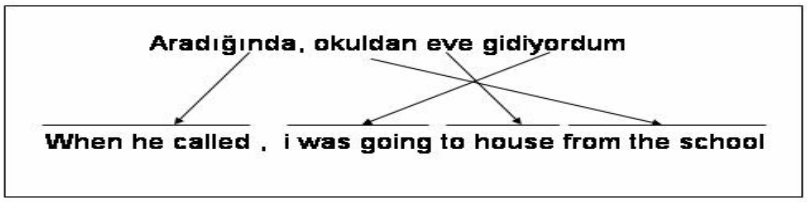
Tablo 1'de, *faaliyet* sözcüğünün değişik biçimleri toplamda 41 kez bulunmasına rağmen, her biri çok düşük frekanslarda bulunur. Sözcük biçimlerinin paralel metinlerde analiz edilmeden kullanılması iki probleme yol açmaktadır. Birincisi, Türkçe sözcüklerin değişik biçimleri yüzünden Türkçe ve İngilizce kök sözcüklerin doğru şekilde eşleştirilememesi, ikincisi ise, İngilizce işlevsel sözcükler ile Türkçe eklerin eşleştirilememesidir.

Her iki problemin çözümü için biçimbirimsel çözümleme yapılması gerekmektedir. Örneğin, *faaliyetleriyle* sözcüğü *faaliyet +ler + i +yle* şeklinde kök sözcük ve yüzey eklerine bölünerek Türkçe paralel metinde ifade edilmelidir. Bu şekilde *faaliyet* sözcüğü ne kadar farklı formlarda bulunursa bulunsun, *activity* sözcüğü ile eşleşme olasılığı çok yüksek olacaktır. Yanısıra, biçimbirimsel analiz yapılmazsa, çevirisi için ele alınan yeni bir cümlede kelimenin sistemde kullanılan paralel metinlerde

[1] İngilizce çevirisi (the thing existing) at the time we caused (something) to become strong şeklinde yapılabilir.

[2] Ekler daha sonraki kullanımlarda kolaylık sağlamak için başlarında '+' işareti ile ifade edilmiştir.

geçmeyen bir formu varsa (örneğin, *faaliyetlerindeki*) kelimenin bu formu çevrilemeyecektir. Biçimbirimsel çözümleme, hem kök kelimenin hem de eklerin birbirinden bağımsız olarak eşleşmesini sağlamakta ve sistemin performansını arttırmaktadır.



Şekil 1. Türkçe ve İngilizce Cümleler için Eşleştirme

Biçimbirimsel çözümlemede yüzey ekler kullanıldığında, aynı bilgiyi taşıyan eklerin farklı biçimlerinin olduğu dikkat çekmektedir. Örneğin, bulunma hal eki dört farklı yüzey biçimi *{+de, +da, +te, +ta}* ile ifade edilmektedir. Eklerin yüzey biçimleri yerine sözcüksel biçimlerinin kullanılması ile aynı bilgiyi taşıyan ekler tek biçimde ifade edilmesini sağlar. Bunu yapmaktaki amaç, eklerin birden çok olan yüzey biçimlerini tek bir sözcüksel biçim ile ifade ederek, hem çeviri olasılıklarını iyileştirmek hem de çeviri sırasında kelime köküne eklenecek ekin yüzey biçimini bulma görevini çözücüye yüklememektir. Kelime kökü ve eklerin sözcüksel biçimleri birbirinden bağımsız olarak çevrildikten sonra, yapılacak ek bir çalışma ile köke uygun ekin yüzey biçimi bulunabilir. Sözcüksel biçimler kullanılarak, faaliyetleriyle sözcüğü, faaliyet +lAr +sH +ylA şeklinde ifade edilir.

Tablo 1. faaliyet kelimesinin değişik formları

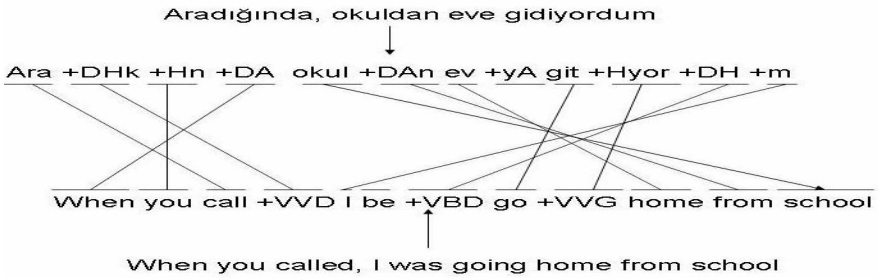
Kelime Formu	Sayı	Anlamı
faaliyet	3	'activity'
faaliyete	1	'to the activity'
faaliyetinde	1	'in its activity'
faaliyetler	3	'activities'
faaliyetlere	6	'to the activities'
faaliyetleri	7	'their activities'
faaliyetlerin	7	'of the activities'
faaliyetlerinde	1	'in their activities'
faaliyetlerine	5	'to their activities'
faaliyetlerini	1	'their activities (accusative)'
faaliyetlerinin	2	'of their activities'
faaliyetleriyle	1	'with their activities'
faaliyette	2	'in (the) activity'
faaliyetteki	1	'that which is in activity'
Toplam	**41**	

Benzer biçimde, İngilizce metinler için de biçimbirimsel analiz yapılmıştır. İngilizce biçimbirimsel analiz TreeTagger [7] kullanılarak yapılmıştır. İngilizce

metinlerde tüm etiketleri kullanmak yerine sadece biçimbirimsel bilgi taşıyan etiketler kullanılmıştır. Örneğin, çoğul eki için *NNS*, geçmiş zaman eki için *VVD* kullanılmıştır. Biçimbirimsel analiz ile *activities* sözcüğü *activity +NNS* şeklinde ifade edilmektedir

Her iki metin için de biçimbirimsel analiz yapılmasının sebebi, kısıtlı olan metinlerden olabilecek en yüksek faydayı sağlamaktır. Analiz tamamlandıktan sonra Şekil 1'de verilen cümle çifti için elde etmeyi planladığımız eşleştirme Şekil 2'de gösterilmektedir.

Biçimbirimsel çözümleme Türkçe ve İngilizce arasındaki yapısal farklılıkları ortadan kaldırmakla beraber, sistem eğitimi sonrasında otomatik olarak yaptığı kelime bazlı eşleştirmelerde, gerçeğe yakın olmayan karmaşık kelime eşleştirmeleri bulunmaktadır. Eklerin ayrı olarak gösterilmesi ile hem Türkçe hem de İngilizce metinlerdeki cümlelerde kelime artışı olmuştur. Detaylı analiz, özellikle Türkçe metinlerde birçok ekin ortaya çıkması ile Türkçe - İnglizce metinlerinde kelime uzunluklarında büyük bir orantısızlığa sebeb olmuş, bu da kelimelerin kaydırma olasılıklarını düşürmüş ve üretkenlik olasılıklarını çok yükseltmiştir. Eklerin gösteriminde bir iyileştirme çalışmasının yapılması gerektiği çok açıktır.



Şekil 2. Ayrıştırılmış ekler ile cümle eşleşmesi

Yapılan çalışmalarda birbirini takip eden Türkçe eklerin, birbirini takip eden İngilizce ek ve işlevsel sözcüklere denk geldiği görülmüştür. Ayrıca ekler, yakınlarında bulunan eklere göre farklı anlamlar ifade etmektedir. Örneğin *+DHr* ekinin *+DHr +mA* ve *+yAcAk +DHr* ek öbeklerinde anlamları farklıdır. Bu gibi ek ve işlevsel sözcük öbeklerini bulmak için söz öbekleri bulma algoritmaları yerine daha basit iki aşamalı bir yöntem kullandık. İlk önce her iki dil metinlerinde bigramlar üretildi. Yüksek frenkslı ek öbekleri elde edilip birleştirildi.

Günümüz İBÇ sistemlerinde söz öbeklerini elde etmek için en fazla dört sözcük kullanıldığı göz önüne alınarak birleştirilmiş paralel metinler için tekrar bigram üretildi ve yüksek frekanslı ek öbekleri tekrar elde edilerek birleştirildi. Sonuç olarak 2, 3 ve 4 ek ve sözcükten oluşan öbekler elde edildi. Basit olmasına karşın, uygulanan metot ile metinlerdeki kelime sayısı azaltılmıştır. Tablo 2 sistemin performansındaki artışı göstermektedir.

4 Performans

Sistem 22000 cümlelik İngilizce - Türkçe paralel metin ile eğitilmiştir. Model parametreleri GIZA++ IBM 4 [8] modeli ile elde edilmiştir. Yeni cümlelerin çevirisi için sözcük öbeği tabanlı [9] çeviriler için kullanılan Pharaoh Decoder [10] kullanılmıştır. Sistemin testi günümüz İBÇ sistemlerinde yaygın olarak kullanılan BLEU [11] değerlendirme aracı ile elde edilmiştir. Sistemin testi için 500 cümle kullanılmıştır. Çözücü çıktısı Türkçe çevirilerde ekler köklere birleştirilirken köke uygun ekler şeçilerek birleştirilmiş, köke eklenmeyen ekler atılmıştır. Tablo 2 sistemin sadece biçimbirimsel analizi ile ve ek öbekleri elde edildikten sonraki performansını gösterilmektedir.

Tablo 2. İngilizce –Türkçe İBÇ sistemi için istatistikler

Metot	**BLEU sonuçları**
Standart	11.33
Biçimbirimsel Analiz	10.58
Biçimbirimsel Analiz + n-gram Analizi	13.41

Tablo 3 sistem eğitildikten sonra test aşamasında denenen bazı İngilizce cümlelerin Türkçe çevirilerini göstermektedir.

Tablo 3. Çeviri Çıktıları

Girdi: international terrorism also remains to be an important issue

Standart çeviri çıktısı: ulus+lararası terörizm de önem+li kal+mış+tır . bir konu ol+acak+tır

Kök birleştirme olmaksızın çeviri çıktısı: ulus+lararası terörizm de ol+ma+ya devam et+mek+te+dir önem+li bir sorun+dur

Kök birleştirmesi ile çeviri çıktısı: ulus+lararası terörizm de önem+li bir sorun ol+ma+ya devam et+mek+te+dir

Referans cümle: ulus+lararası terörizm de önem+li bir sorun ol+ma+ya devam et+mek+te+dir

Girdi: the initiation of negotiations will represent the beginning of a next phase in the process of accession

Standart çeviri çıktısı: müzakere+ler+in gör+üş+me+ler yap+ıl+acak bir der+ken aşama+nın hasar+ı sürec+i başlangıc+ı+nı 15+'i

Kök birleştirme olmaksızın çeviri çıktısı: müzakere+ler temsil ed+il+me+si+nin başlangıc+ı bir aşama+sı+nda katılım sürec+i+nin ertesi

Kök birleştirmesi ile çeviri çıktısı: müzakere+ler+in başla+ma+sı+nın başlangıc+ı+nı temsil ed+ecek+tir katılım sürec+i+nin bir sonra+ki aşama

Referans cümle: müzakere+ler+in başla+ma+sı , katılım sürec+i+nin bir sonra+ki aşama+sı+nın başlangıc+ı+nı temsil ed+ecek+tir

5 İleri Konular

Çalışmanın amacı Türkçe İngilizce dil çifti için başarılı bir istatistiksel bilgisayarlı sistemi geliştirmektir. Sistemin eklerin ayrı gösterilmesini temel almaktadır. Üzerinde çalışılması düşünülen konular iki ana başlıkta toplanabilir. İleride bu yönde çalışmalar yapılacaktır. Birinci olarak, çözücü her ne kadar doğru kelime ve kökleri bulsa da sıralamalarında hatalar vardır. Çözücüden sonra yapılacak ek bir çalışma ile doğru ek ve köklerin birleştirilmesi gerekmektedir. İkinci olarak, değerlendirme kriteri olarak kullanılan BLEU aracı Türkçe için uygun olmadığından değerlendirme kriterleri üzerinde çalışmalar yapılması gerekmektedir.

6 Sonuçlar

Bu bildiride, Türkçe - İngilizce dil çifti için istatistiksel bilgisayarlı çeviri sistemi için yapılan çalışmalar anlatılmıştır. Var olan paralel metinlerden en yüksek verimi almak ve diller arasındaki yapısal farklılıkları en aza indirmek için biçimbirimsel analiz yapılarak ekler kelime köklerinden ayrı olarak ifade edilmiş, kök ve eklerin birbirinden bağımsız olarak çeviri olasılıkları hesaplanmıştır. Türkçe metinlerde birçok ek ortaya çıktığı için sistemin başarı düşmüş, bu düşme birbirini takip eden yüksek frekanslı eklerin birbirine bağlanması ile sistemin başarısı arttırılmıştır.

Ek Bilgi. Yapılan çalışma, 105E025 numaralı "İngilizce-Türkçe İstatistiksel Bilgisayarlı Çeviri Sistemi" projesi dahilinde TÜBITAK tarafından desteklenmektedir.

Referanslar

1. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lafferty, J.D., Mercer, R.L.: Analysis, statistical transfer, and synthesis in machine translation. In: Proceeding of TMI: Fourth International Conference on Theoretical and Methodological Issues in MT. (1992) 83–100
2. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19 (1993) 263–311
3. Yamada, K., Knight, K.: A syntax-based statistical translation model. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse (2001) 00–00
4. Lee, Y.S.: Morphological analysis for statistical machine translation. In: Proceedings of HLT-NAACL 2004 - Companion Volume. (2004) 57–60
5. Niessen, S., Ney, H.: Statistical machine translation with scarce resources using morpho-syntatic information. Computational Linguistics 30 (2004) 181–204
6. Oflazer, K.: Two-level description of Turkish morphology. Literary and Linguistic Computing 9 (1994) 137–148
7. Schmid,H.: Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing (1994).
8. Och, F.J., Ney, H.: Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong (2000) 440–447

9. Och, F.J., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics 30 (2004) 417–449
10. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of HLT/NAACL. (2003)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia (2002) 311-318