Türkçe Tweetlerden Soru İfadelerini Bulmak
Identifying Questions on Turkish Tweets

Celal Cengiz
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
celalcengiz@gmail.com

Banu Diri
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
banu@ce.yildiz.edu.tr

***Özetçe*—Günümüzde Twitter, çok kullanılan sosyal medya ortamlarından birisidir. Twitter sadece bilgi paylaşılan bir ortam değil, aynı zamanda soru sorulan bir ortam haline de gelmiştir. Bu çalışmada, Türkçe atılmış tweetler arasından gerçek soru içeren tweetlerin bulunup çıkarılması üzerine çalışılmıştır. Asıl amaç, bu soruları otomatik olarak bulmak ve ileride sorulara otomatik cevap vermeyi sağlayan bir sistem geliştirmektir. Bu çalışmada öncelikle, gerçek soruların bulunması işlemine yer verilmiştir.**

Anahtar Kelimeler — Türkçe tweet, gerçek soru bulma.

***Abstract*—Nowadays Twitter is one of the most used social media platform. Twitter is not only platform of sharing information it is also become a platform of question asking. In this study, we try to identify questions on tweets which is broadcasting in Turkish. The main goal is identifying questions automatically and in the future answering them automatically. For this reason in this paper, it is given identifying real question process.**

Keywords — Turkish tweet, identifying real questions.

I. GİRİŞ

Twitter dünyada olduğu gibi Türkiye'de de yaygın kullanımı olan, o an ne yapıldığının yazıldığı mikroblog ortamı olmasının yanı sıra, forumlarda olduğu gibi soruların sorulup, cevabının beklendiği bir ortam haline de gelmiştir. Türkiye'nin 2012 yılı twitter raporuna göre yaklaşık 7 milyon twitter kullanıcısı olduğu ve bu kullanıcıların 5 milyon tanesinin günde en az bir adet tweet atarak aktif kullanıcı olduğu bildirilmiştir. Ayda yaklaşık 1.7 milyon adet tweet atıldığı, saniyede 20 tweete karşılık geldiği, tweeter kullanıcılarının %53' ünün erkek olduğu ve en fazla tweetin akşam 21:00 ile 22:00 saatleri arasında atıldığı rapor edilmiştir [1]. Morris [2], Twitter kullanıcılarının %10'dan fazlasının en az bir kez Twitter ortamından soru sorduğunu belirtmiştir. Efron [3] ise, rastgele seçilmiş 2 milyon tweetten oluşan bir derlemin %13'ünün soru cümlesi içeren tweet'lerden oluştuğunu söylemiştir. Biz de bu çalışmada, bu noktadan yola çıkarak Türkçe olarak atılmış tweetler arasından soru içeren tweetleri bulmayı amaçladık.

Tweeter üzerinden soru cümlesi içeren tweetleri bulma üzerine Türkçe dışındaki diğer diller için yapılmış çalışmalar mevcuttur [4, 5]. Fakat, dillerin biçimbirimsel yapısından kaynaklı farklılıklar sebebiyle kullanılan algoritmaların da farklı olması gerekmektedir. Bu sebeple, Türkçe dil bilgisi kuralları incelenmiş, bölüm 3 ve 4'te anlatılmıştır. Bu çalışma, ileride bu konuda yapılacak çalışmalar için bir ön çalışma niteliği taşımaktadır.

Çalışmada yapılan işlemler şu şekilde özetlenebilir. Soru işareti içeren tweetler belirlenerek sözde soru cümlesi olarak etiketlenir. Ancak, Twitter kullanıcıları tweetlerini atarken noktalama işaretlerine dikkat etmediklerinden veya 140 karakter sınırı olduğu için, sadece soru işaretinin kontrol edilmesi bu çalışma için yeterli olmayacağından, içerisinde soru kelimesi geçen tweetler de sözde soru cümlesi olarak etiketlenmiştir. Sözde soru cümlesi olan tweetlerden gerçekten soru cümlesi olan ve bir cevap beklenen tweetler bulunmaya çalışılmıştır. Bunun için Türkçe'ye özgü dilbilgisi kurallarından oluşan kural tabanlı bir yöntem geliştirilmiştir.

Makalenin ikinci bölümünde, Twitter'dan girilen bir sorguyla tweetlerin çekilmesi, üçüncü bölümde bu tweetlerden sözde soruların çıkarılması, dördüncü bölümde sözde sorular arasından gerçek soruların tespit edilmesi ve beşinci bölümde de deneysel sonuçlara yer verilmiştir.

II. TWITTER'DAN SORGUYLA TWEETLERİN ALINMASI

Twitter'dan atılmış tweetleri çekebilmek için twitter sorgu apisi kullanılmıştır. URL olarak verdiğimiz bir sorugunun sonucu bize bir web sayfası olarak döndürülmektedir. Kullanılan URL de, *lang=tr* sorgu kısıtı ile dil olarak sadece Türkçe olan tweetlerin çekilmesi sağlanırken, *geocode=37.781157,30.398720,1000km* yer kısıtı da Türkiye sınırları içinden verilen bir pozisyondan 1000 km uzaklıktaki bir alandan atılan tweetlerin çekilmesi sağlanmıştır. http://search.twitter.com/search.json?&q=*&lang=tr&geocode=37.781157,30.398720,1000km

Sorgu kısıtlarını belirlerken sadece dil kısıtının yeterli olmadığı görülmüştür. Kullanıcı dilinin Türkçe seçilmiş olmasına rağmen, Türkçe olmayan tweetlerin de çekildiği tespit edilmiştir. Bu durumun giderilmesi için yer kısıtı da sorguya eklenmiştir. Konulan bu kısıtlara rağmen yine de Türkçe olmayan tweetlerle karşılaşılmıştır. Bu olumsuz durumun dışında Türkçe yazılmış, ama yazım hataları olan tweetler de bulunmaktadır. Bu hatalı tweetleri düzeltmek için Zemberek kütüphanesinden yararlanılmıştır [6]. Zemberek'in

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

sağladığı *denetle* ve *öner* komutları kullanılarak yazım hatası olan kelimelerin düzeltilmesi yoluna gidilmiştir. Ancak, bu işlemler sonucu yavaşlattığı için *öner* fonksiyonun verdiği ilk öneri doğru kelime olarak kabul edilmiştir.

Örnek tweetlerimizi, Twitter üzerinde arama sorgularını kullanarak 15 adet tweetten oluşan web sayfaları olarak elde ettik. Verilmiş olan sorgu ilk kez 16 Ocak 2013'te, ikinci kez de 13 Mart 2013 tarihinde toplam 70 kere çalıştırılmış ve toplamda 1045 adet tweet çekilmiştir. Sonuç olarak aldığımız web sayfasında sadece tweetler değil, elde edilen her tweet için text başlığı da gelmektedir. Yetmiş web sayfası bu çalışma için düz yazı olarak metin dosyalarına kaydedilmiştir.

III. SÖZDE SORU CÜMLELERİNİN BULUNMASI

Sözde soru cümlelerinin çıkarılması için dosyalara kaydedilen 1045 adet tweet okunarak içerisinde soru işareti geçenler sözde soru cümleleri olarak işaretlenmiştir. Tweetler yazılırken genelde noktalama işaretlerine uyulmadığı için, soru kelimesi içeren tweetler de sözde soru cümlesi olarak etiketlenmiştir.

Tweetler içerisinde geçen soru kelimeleri olarak, Ne, Niçin, Nasıl, Neden, Nere, Kaç, Kim, Hangi kelimeleri ve m[I] eki seçilmiştir. Bir tweetin sözde soru cümlesi olabilmesi için, içindeki kelimelerde bu soru kelimelerinin geçmesi şartı aranmıştır. Arama işlemi kelimenin tamamında veya kök kısmında yapılmıştır. Kelimelerin kök ve eklerine ayrılmasında yine Zemberek kütüphanesinden yararlanılmıştır. Ayrıca, bölüm 2'de bahsedildiği üzere Zemberek kütüphanesi tweetler yazılırken yapılan yazım hatalarının düzeltilmesinde de kullanılmıştır. *Denetle* fonksiyonundan yazım hatası olan kelimelerin yerine yeni bir kelime önermesi istenmiş ve önerilen kelimeler arasından da ilk öneri kabul edilmiştir. Bunun sebebi olarak da denetleme işlem maliyetinin yüksek olması ve denetlemenin gerçek soru cümlesi olup olmadığını belirlemede küçük bir katkısının olması gösterilebilinir.

Kelimenin köklerine ayrılması işlemi ile seçilen soru kelimelerinden türemiş olan tüm soru kelimeleri de sözde soru cümlelerinin bulmasında kullanılmıştır. Örneğin, "Ne" soru kelimesinden türeyen "Neyi", "Neye", "Neden"; "Nere" soru kelimesinden türeyen "Nereye", "Neresi", "Nerede" gibi diğer soru kelimeleri de bulunabilmiştir. Ayrıca, "Ne" soru kelimesini içeren "Ne kadar", "Ne ile", "Ne için" gibi iki kelimeden oluşan ve içerisinde sadece bir soru kelimesinin bulunduğu soru ifadelerinin de sözde soru cümlesi olarak bulunması sağlanmıştır.

Bunların dışında "mi", "mı", "mu" ve "mü" gibi soru takıları da sözde soru cümlesi bulunmasında kullanılmıştır.

Ayrıca, kök olarak içerisinde soru kelimesi içeren fakat, aldığı ek ile soru kelimesi olmayan "Kimse" gibi örnek verebileceğimiz kelimeler de mevcuttur. Bu kelimenin kökü olan "Kim", soru kelimeleri listemizde olsa bile çalışmamızda bu tip kelimeler sözde soru cümlesi olarak kabul edilmemiştir.

IV. GERÇEK SORU CÜMLESİNİN BULUNMASI

Sözde soru cümlesinin gerçek soru cümlesi olup olmadığını anlamak için Türkçe dil bilgisi kurallarından yararlanılmıştır. Kurallar çıkarılmadan önce, sözde soru cümlesi olarak etiketlenen ancak, gerçek soru cümlesi olma ihtimali düşük olan durumlar belirlenerek sözde soru cümleleri içerisinden çıkarılmıştır. Bunlardan biri,URL içeren tweetlerin reklam veya haber sunma ihtimalinin yüksek olması nedeni ile ve gerçek niyetin tam anlaşılmaması yüzünden gerçek soru cümlesi olarak değerlendirilmesi engellenmiştir. İçerisinde sadece soru işareti olan, ama soru kelimesi içermeyen tweetler de gerçek soru cümlesi olarak kabul edilmemiştir. Çünkü soru işareti, atılan tweet içerisinde soru anlamının dışında başka bir anlam vermek için de kullanılmış veya yanlışlıkla da yazılmış olabilir.

Geliştirilen uygulamada kullanılan dilbilgisi kurallarından birincisi zarf olan "Ne" kelimesidir. Sözde soru cümlesi "Ne" kelimesini içerebilir ama bu kelime soru anlamı içermeyip, zarf olarak kullanılmış olabilir. Bu durumda bu cümleyi gerçek soru cümlesi olarak kabul edemeyiz. Bunun için "Ne" kelimesinden sonra gelen ilk kelimenin türüne bakılır. Eğer kelimenin türü sıfat ise "Ne" gerçek soru cümlesi değildir. Kelimenin türünü belirleme işlemleri için yine Zemberek kütüphanesinden yararlanılmıştır. Kurala örnek verecek olursak,*"Ne güzel evim var"* cümlesini gösterebiliriz. Bu cümlede "Ne" kelimesi soru anlamı içermemektedir, güzel kelimesini nitelediği için zarf olarak kullanılmıştır.

İkinci kural olarak, sıfat ikilemesinin arasında geçen *mi* ekinin soru anlamı içermemesi kullanılmıştır. Bu kurala *"Güzel mi güzel bir evim var"* cümlesini verebiliriz. Cümleden görüldüğü gibi burada "mi" soru eki cümlede soru eki olarak değil, pekiştirme anlamında kullanılmıştır. Bu durumda olan sözde soru cümleleri de gerçek soru cümlesi olarak sayılmamıştır.

Üçüncü kural, herhangi bir soru kelimesinden sonra eylem bildiren bir kelime belirtme eki ile beraber kişi sahiplik eki de alıyorsa, bu durumda soru kelimesi yine soru anlamı içermemektedir. Bu kuralı uygulamak için sonraki kelimeye gelen eklere de bakılmıştır. Örnek verecek olursak, *"Nasıl yaptığını biliyorum"*. Bu cümlede görüldüğü gibi gerçek bir soru cümlesi değildir.

Dördüncü kural ise, şart belirtmek için kullanılan soru kelimelerinin tespitinden oluşmaktadır. Örneğin, *"Nereye giderse ben de gideceğim"* cümlesinde kullanılan soru kelimesi soru anlamı içermemekte olup, gerçek soru cümlesi olarak alınmamıştır.

Bir diğer kural, "Neden" ve "Niye" sorularının bulunduğu tweet içerisinde "çünkü" kelimesinin aranmasıdır. Bu durumdaki tweetler hem soru hem de cevabı içerdiği için gerçek soru cümlesi olarak sayılmamıştır. Örnek bir tweet verecek olursak *"erene makarna yapıyorum niye çünkü canı şey etmiş öyle diyo"* verilebilir.

V. DENEYSEL SONUÇLAR

Yapılan çalışmanın başarısını ölçmek için 1045 tweetten oluşan bir veri seti hazırlanmış ve kullanılmıştır. Bu veri seti birinci kişi tarafından '*soru cümlesi*" ve "*soru cümlesi değildir*" diye etiketlenmiş olup, sonrasında ikinci kişi tarafından etiketlerin üzerinden geçirilerek etiketleme işlemi sonlandırılmıştır.

Veri seti içerisinden sözde soru cümlelerini çıkarırken kullandığımız kurallara rağmen, kaçırılan sözde soru cümleleri de olmuştur. Bu cümlelerin geçtiği tweetleri incelediğimizde bazı tweetlerin içeriğinin yazım hatalarından dolayı çözümlenemediği görülmüştür. Yazım hatası olan bir kelime için Zemberek den gelen ilk öneriyi almış olmamızdan kaynaklanan hatalarda vardır. Bu tip problemleri çözmek için iki kelime arasındaki mesafeyi ölçen *–edit distance* gibi yöntemlerin kullanılması çok daha doğru bir seçim yapılmasına neden olacaktır. Tweetlerdeki bir diğer problem ise, *"@gulnur_fb @ForzaFener19o7 estarfllah onedemek öle yaw :))"* örneğinde olduğu gibi kelimeler arasında olması gereken boşluğun bırakılmamış olmasıdır.

Bunların dışında karşılaştığımız bir başka yazım hatası da *"Numara mi degistirdim, hadi sapiklar arayin ulasa bilirseniz."* cümlesinde olduğu gibi bölünmemesi gereken bir kelimenin kullanıcı tarafından bölünerek yazılmış olmasıdır. Bu durum tweetin sözde soru cümlesi olarak alınmasına yol açmıştır.

Sistemin başarısını ölçmek için hata kestirim modeli olarak hata matrisleri (Tablo 1) kullanılmış, Tutturma (Precision) (1), Bulma (Recall) (2) ve F-ölçüm (3) değerleri hesaplanmıştır.

Tablo 1. Hata Matrisi

Tahmin \ Gerçek	**Gerçek Soru**	**Gerçek Soru Değil**
Gerçek Soru	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
Gerçek Soru Değil	Yanlış Negatif (FN)	Doğru Negatif (TN)

$$\text{Tutturma} = TP / (TP + FP) \quad (1)$$

$$\text{Bulma} = TP / (TP + FN) \quad (2)$$

$$\text{F-ölçüm} = (2*\text{Tutturma}*\text{Bulma}) / (\text{Tutturma}+\text{Bulma}) \quad (3)$$

Başarıyı ölçmek için hazırladığımız etiketli veri setinde 1045 adet tweetin 180 tanesi *gerçek soru* cümlesi olarak elle etiketlenmiştir. Sistem ilk olarak 1045 adet tweetin 335 tanesini *sözde soru cümlesi* olarak belirlemiştir. İkinci aşama da ise 335 adet *sözde soru cümlesi*nin 260 tanesini *gerçek soru cümlesi* olarak işaretlemiştir. Tablo 2 bize sayısal veriler ile hata matrisini vermektedir.

Tablo 2. Hata Matrisi

Tahmin \ Gerçek	**Gerçek Soru**	**Soru Değil**
Gerçek Soru	177	83
GerçekSoru Değil	3	782

Tutturma 0.68 olarak elde edilirken, Bulma değeri, 0.98 olarak hesaplanmıştır. Sistemin F-ölçüm değeri de 0.80 olarak belirlenmiştir.

Tutturma değerinin 1'den uzaklaşmasının nedeni kullandığımız kuralların yetersiz oluşu ile ilgilidir. Çalışmada belirttiğimz kuralların dışında yeni kurallar ile sistemi besleyebilirsek yanlış pozitif oranını düşürebiliriz. Gerçekte soru olmayan ancak sistemin soru dediği tweetlere örnek verecek olursak:

"radyo kuzey nedir boyle yahu sanki benim listemi calıyor gibi adamlar. cok iyi."

"yağmurda ne birikmiş aga venedik oldu hep buralar."

"ulan su topkapı habibler hattına birgun aktarma yapcam yolun sonunda ne var merak ediyorum"

Bulma değeri bire oldukça yakındır. Bulmadaki başarıyı düşüren etken olarak yazım hataları karşımıza çıkmaktadır. Çalıştığımız tweetler içinde gerçek soru olup da sistemin gerçek soru değil dediği tweetlerde ayrı yazılması gereken *mi* soru ekinin yanlış olarak kullanıldığı görülmüştür. Bu tweetler için örnek verecek olursak:

"RT @cimanekeni: #TeröristeTörenYapanTekÜlkeyiz ortamı germek için açılmış bir tagdır.Barış adına atılan her adımdan sonra,böyle adımların gelmesi tesadüfmü?",

"@kula9 sen yarın gelcenmi okula ?"

Kural tabanlı yöntemler dışında öğrenme tabanlı yöntemlerin performans artışı sağlayıp sağlamayacağının belirlenmesi açısından denenmesi önerilebilinir.

VI. SONUÇ

Bu çalışmada 140 karakter sınırı olan, kullanıcı tarafından hiçbir kurala dikkat edilmeden yazılan ve sosyal medyada çok önemli bir yeri olan Twitter dan atılan tweetler içerisindeki soru cümlelerinin çıkarılması hedeflenmektedir. İleride soruyu soran tweet sahibine anında cevabı gönderebilen sistemler tasarlanabileceği düşünülürse bu çalışma bir başlangıç niteliğinde olacaktır. Sistemin başarısı çok yüksek değildir. Daha iyi sonuçlar alabilmek için çıkarılan kuralların genişletilmesi ve geliştirilmesi, yazım hatalarının daha yüksek başarı ile düzeltilmesi ve soru kelimesinden sonra gelen ilk kelime ile değil de ilerisindeki kelimelerle de olan ilişkilere bakılması gerektiği düşünülmektedir. Ayrıca, sadece kural tabanlı yaklaşımlar değil öğrenme tabanlı yaklaşımlarla da sistemin geliştirilmesi gerekir.

KAYNAKÇA

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

[2] Morris, M. R., Teevan,J. ve Panovich, K., "What do people ask their social networks, and why?: a survey study of status message Q&A behavior", *In CHI '10*, 2010.

[3] Efron, M. ve Winget, M., "Questions are content: a taxonomy of questions in a microblogging environment", *In Proc. of ASIST '10*, 2010.

[4] Li, B., Si, X., Lyu, M. R., King, I. ve Chang, E. Y., "Question Identification on Twitter", *Proceedings of the 20th ACM international conference on Information and knowledge management,* ACM, New York, NY, USA, 2011.

[5] Wang, K. ve Chua, T. S., "Exploiting salient patterns for question detection and question retrieval in community-based question answering", *In COLING '10*, 2010.

[6] http://code.google.com/p/zemberek/