Türkçe Soru Cevaplama Sistemlerinde Kural Tabanlı Odak Çıkarımı
Rule-Based Focus Extraction in Turkish Question Answering Systems

Caner Derici, Kerem Çelik,
Arzucan Özgür, Tunga Güngör
Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi
{caner.derici, kerem.celik2, arzucan.ozgur, gungort}@boun.edu.tr

Ekrem Kutbay, Yiğit Aydın,
Günizi Kartal
Eğitim Teknolojileri
Boğaziçi Üniversitesi
{ekrem.kutbay, yigit.aydin, gunizi.kartal}@boun.edu.tr

***Özetçe* —Bu çalışmada, lise öğrencilerinin eğitimlerinde yardımcı olması için tasarlanan Türkçe soru cevaplama sisteminin soru analizi kısmında kullanılmak üzere geliştirilen kural tabanlı yöntem sunulmuştur. Soru içerisinden, cevabın türü ve niteliğini belirten kelimelerin (odak) elde edilmesinde kullanılan bu yöntem, uzmanlar tarafından elle toplanıp işaretlenmiş coğrafya soruları kümesi üzerinde deneye tabi tutulmuş, sonuçları belirtilmiştir. Kurallar coğrafya alanına bağlı kalınmayıp, genel Türkçe soru kalıpları üzerine kurulduğu için bu yöntem, alandan bağımsız olarak herhangi bir Türkçe soru cevaplama sisteminde kullanılabilir. Hem kaynak kodları, hem de hazırlanan soru kümesi, çalışmanın tekrarlanabilmesi ve daha da geliştirilebilmesi adına elektronik olarak sağlanmıştır.**

Anahtar Kelimeler—soru cevaplama sistemleri, soru analizi, kural tabanlı bilgi özütleme, odak

***Abstract*—This study describes a rule-based approach that is employed in the question analysis module of the Turkish question answering system we design for high-school students to assist their education. The technique is used to extract parts of the question (the focus) that indicate the type and the character of the answer. Evaluation is performed on a set of Geography question data that are collected and annontated by human experts, and the results are provided. Manually tailored rules are based on general Turkish question patterns, thereby not dependent on the domain. Therefore, this technique can be used by any Turkish question answering system, regardless of the domain. Both the source codes and the annontated data set are electronically provided for both reproducability and future work.**

Keywords—question answering systems, question analysis, rule-based information extraction, focus

I. GİRİŞ

Soru Cevaplama sistemleri, doğal dil ile üretilen bir sorunun cevabını otomatik olarak bulma, sentezleme ve sonunda üretmeyi amaçlar. Son yıllarda dünyada doğal dil işleme ve bilgi çıkarımı gibi konularda olan gelişmeler, ülkemizde soru cevaplama sistemlerine olan ilgiyi de arttırmış, öyle ki, soru sıklığı yüksek kurumlar[1] ya da mevzuatlar[2] ile ilgili bilgilendirme amaçlı, kamu kullanımına açık yarı-otomatik çevrim içi soru-cevap sistemleri oluşturulmaya başlanmıştır.

Akıllı sistemlerin de yaygınlaşmasıyla, genel olarak anahtar kelime tabanlı aramaya dayanan soru cevaplama sistemleri, yerlerini daha akıllı çözümlemeler yapan gelişmiş sistemlere bırakmıştır. Bu tür sistemlerin kurgulanması ve meydana getirilmesi, içerisinde soru analizi, bilgi çıkarımı ve sentezlenmesi, mantıksal analiz ve çıkarım gibi karmaşık problemler barındırdığından, kapsamlı ve detaylı bir inceleme gerektirmektedir. Akıllı soru cevaplama sistemlerinde soru analizi bu bağlamda son senelerde önem arz etmeye başlamıştır.

Genel olarak bir soru cevaplama sisteminde, ilk olarak verilen bir soru analiz edilip, içerisinden sistemin ileriki safhalarında kullanılmak üzere gerekli temel bilgiler çıkartılır. Daha sonra bu bilgiler kullanılarak, daha önce oluşturulmuş bilgi bankasından ilgili dökümanlar tespit edilip, bu dökümanlar üzerinde bilgi çıkarım yöntemleri ile aday olabilecek cevaplar belirlenir. Bu aday cevaplar yine önceden belirlenmiş ölçütler ya da önceden eğitilmiş istatistik tabanlı modeller kullanılarak notlandırılır. Son olarak, en yüksek olasılığa (ya da puana) sahip aday, başta verilen sorunun cevabı olarak seçilir ve bu cevabın bulunduğu döküman kullanılarak, istenilen kriterlere göre (örneğin bağlamlı ya da bağlamsız olarak) en son cevap üretilir ve kullanıcıya iletilir [1], [2]. Bu bağlamda, bir soru cevaplama sisteminin en önemli kısımlarından biri, verilen sorunun analizinin yapıldığı soruda tam olarak neyin sorulduğunun belirlendiği kısımdır. Örneğin, aşağıdaki altı çizili kelimeler ("doğal bitki örtüsü"), sorunun tam olarak neyi sorduğunu (odak) belirten kısmı oluşturmaktadır.

"Akdeniz bölgesinin doğal bitki örtüsü nedir?"

Sorunun odağı, blok halinde kelimeler bütünü olmak zorunda değildir. Aşağıdaki örnekte görüldüğü gibi, odağı oluşturan kelimeler yan yana değil, sorunun farklı kısımlarında olabilir. Bu örnek için odak, "araca" ve "ad" kelimelerinden oluşmaktadır, zira soruda bir aracın adı aranmaktadır.

Hava sıcaklığını ölçen araca ne ad verilir?

Bu bildiri, Türkçe soruların analizinde kullanılabilecek, kural tabanlı bir yöntem önerip, uzmanlar tarafından el ile hazırlanmış soru kümesi üzerinde değerlendirmeye tabi tutmaktadır. Bilgimiz dahilinde, ülkemiz literatüründe daha önce soru analizi ile ilgili yapılmış bir çalışma yoktur. Dolayısıyla, bu çalışma hem konu, hem de sağlanan hazır veri itibarıyla Türkçe soru cevaplama sistemleri konusunda gelecek araştırmalara ortam hazırlamaya katkıda bulunmaktadır.

Dökümanın II. bölümü konu ile ilgili literatürdeki çalışmaları, III. bölümü bu çalışmada önerilen yöntemi, IV. bölümü

[1] www.tesk.org.tr
[2] soru-cevap.memurlar.net

978-1-4673-5563-6/13/$31.00

önerilen yöntemin sınandığı deneyleri açıklamaktadır. Son olarak V. bölümü çalışmayı neticelendirmektedir.

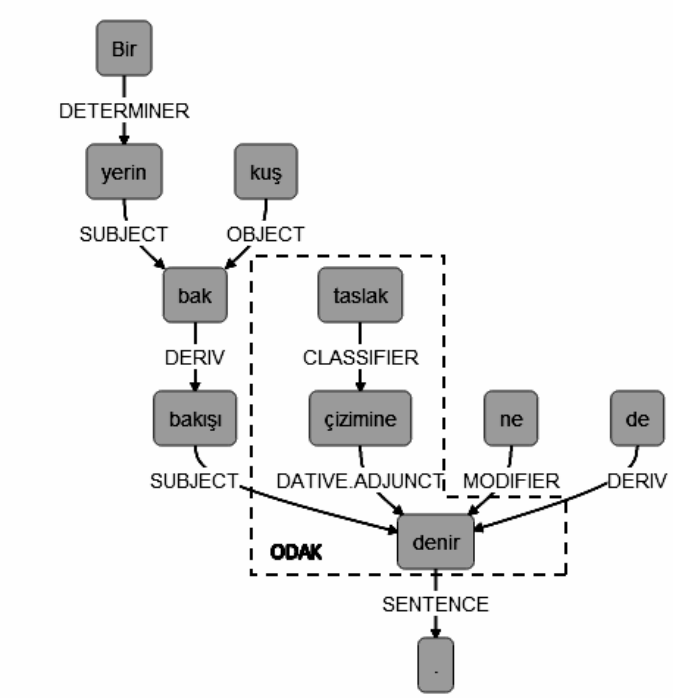
II. Literatür Özeti

Daha önce önerilmiş ve uygulamaya konulmuş soru cevap sistemleri, genellikle tamamen istatistik tabanlı olduklarından ve sadece biçimsel analiz yaptıklarından, soru analizi gibi sistemi anlamsal öğelerle de destekleyecek adımlar uygulamamışlardır. Biçimsel analizin en tipik örneklerinden bir tanesi, kelime sıklığı hesabının kullanılmasıdır. Örneğin, İlhan ve dig. [3], soru ve cevap metinleri içerisinden çıkartılan anahtar kelimelerin tf-idf değerlerinden oluşturulan vektörler arası kosinüs benzerliğini kullanarak, soru ile en çok örtüşen cevabı bulmaya çalışmıştır. Bu çalışmanın en temel eksikliği anlamsal bilginin hiç kullanılmamış olması, bunun sonucunda da örneğin bahsedilen vektör benzerliğinin bir soru için bütün cevap verileri ile ayrı ayrı yapılması gerekmesidir. Bu da hesapsal olarak işlemi tüm cevap verilerinin sayısına göre doğrusal boyutta tuttuğu için, soru ve cevap çeşitliliği, dolayısıyla da kullanılan kaynaklar ve cevap sayısı arttığında sistemi kullanışsız bir duruma sokacaktır. Benzer bir yöntemi kullanan bir diğer çalışmada ise Zheng [4], soru içerisinde geçen kelimelerin, önceden belirlenmiş genel amaçlı arama motorlarının (Google, Yahoo, AltaVista, vb.) döndürdüğü sayfa sayısı (hit) değerlerine göre, soruya en uygun arama motorunu seçip, bu arama motorundan sorunun tamamı için dönen dökümanları kullanıcıya döndürmüştür. Bir önceki çalışmada belirtilen hesapsal yükün kaldırılmasında bu durumda arama motorlarına bel bağlanmıştır. Ayrıca bu çalışmada cevabın kendisini bulma işlemi kullanıcıya bırakılmış, soru cevaplama sisteminin birincil amacını tam olarak karşılamayan bir sistem kurgulanmıştır.

Anlamsal verileri de kullanan soru cevaplama sistemleri, genellikle cevabı aramadan önce sorunun içerisinden anlamsal çıkarımlar yapabilmek için öncelikle soru analizi denilen adımı gerçekleştirirler. Soru analizinde, soru içerisinden doğrudan anlamsal ilişkiler çıkarılabildiği gibi, sorunun önceden belirlenmiş (ne, nerede, ne zaman gibi) soru tiplerine sınıflandırılması ile arama uzayının daraltılması sağlanabilmektedir [5], [6]. Arama uzayının daraltılması için sorunun sınıflandırılmasından daha çok, cevabın sınıflandırılması, aynı zamanda daha akıllı bir cevaplamanın kurgulanmasında önemli rol oynamaktadır. Bu bağlamda yapılan en önemli çalışmalardan birinde Bunescu ve dig. [7], bir önceki bölümde belirtilen anlamda sorunun odağını belirlemek için kural tabanlı analiz ile çıkartılan bilgileri destek vektör makineleri (support vector machines) için birer özellik (feature) olarak kullanmıştır.

III. Önerilen Yöntem

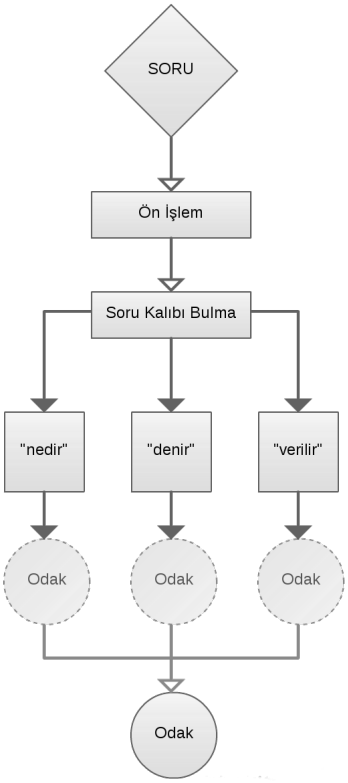
Verilen bir sorunun odağının bulunmasında ilk akla gelen kural tabanlı yöntemlerden biri, orijinal soru metni üzerinde düzenli ifadelerle (regular expressions) eşleştirme yapmaktır. Oysa Türkçe'nin serbest nizam olması ve çok zengin bir bitişken biçimbirimsel yapıya sahip olması (sondan ekleme ile yeni kelimelerin oluşturulabiliyor olması), düzenli ifadelerle yapılacak statik bir analizi zorlaştırmaktadır. Bu nedenle, doğrudan soru metni üzerinde çalışmak yerine, sorunun ayrıştırılmış halini analiz etmek, genelliği korumak ve uygulanabilirlik adına daha uygun olmaktadır. Bu bağlamda, soru cümlesinin odağının bulunması için öncelikle ayrıştırma



Şekil 1: "Bir yerin kuş bakışı taslak çizimine ne denir?" sorusunun bağlılık ağacı.

(syntactic parsing) yapılması gereklidir. Bu çalışma, Türkçe soru cümlelerinin ayrıştırılmasında [8], [9] ve [10] çalışmalarında belirtilen Türkçe bağlılık ayrıştırıcısını (dependency parser) kullanmaktadır. Bağlılık ayrıştırıcısı, verilen cümledeki sözcüklerin arasındaki ikili bağlılık ilişkilerini, sözcük diziliş sıralamasından bağımsız olarak çıkartır. Bu özellik, bağlılık ayrıştırıcılarını Türkçe gibi serbest nizamlı diller için oldukça kullanışlı kılmaktadır. Bağlılık ayrıştırması işlemi ile tümce metninden, her sözcüğün birer boğum (node) ve sözcükler arasındaki ikili bağlılık etiketlerinin de birer kenar (edge) olduğu bir bağlılık ağacı (dependency tree) meydana getirilir. Şekil 1'de örnek bir bağlılık ağacı, sözcük bağlılıkları ve etiketleri ile gösterilmiştir. Bağlılık etiketleri [8] çalışmasında belirtilen Türkçe bağlılık AğaçBanka'sında tanımlanmıştır, ve içerisinde SUBJECT (özne), CLASSIFIER (belirtisiz isim tamlaması), POSSESSOR (belirtili isim tamlaması), MODIFIER (zarf/sıfat tamlaması), SENTENCE (yüklem) gibi etiketleri barındırır. Bağlılık ilişkilerinin yanında, ayrıştırıcı aynı zamanda bütün sözcüklerin kökleri ve öğe bilgilerini (part-of-speech) de kelimelerin kendileri ile birlikte vermektedir. Bu da örneğin, "araca" ve "ad" kelimeleri ile bir araç adı arandığı bilgisine ulaşmada yarar sağlamaktadır.

Ön işlem olarak, sorudan tire (-) dışındaki noktalama işaretlerinin hepsi çıkartılmış, kesme işaretli (') kelimeler (örneğin İstanbul'da) birleştirilmiştir. Buna yalnız sayılar tabi tutulmamıştır, zira örneğin 1.6kg (ya da 1,6kg) ile 16kg anlamsal olarak ciddi farklar yaratabilmektedir. Benzer şekilde oran ifade eden sayıların başındaki yüzde (%) işareti de kaldırılmamış, olduğu gibi alınmıştır. İleriki çalışmalarda bu tür detayların metinden ziyade meta bilgilerde toplanması



Şekil 2: Soru analizi akışı

amaçlanmaktadır.

Alandan (domain) bağımsız bir şekilde Türkçe sorular incelendiğinde, kullanılan belli başlı soru kalıpları göze çarpmaktadır. Bu kalıpların her biri, Türkçe'nin türevsel zenginliğinden dolayı (anlam değişse bile) farklı şekillerde kullanılabilir. Örneğin, aşağıdaki soruların cevabı farklı bile olsa, iki soruda da aranan şey bir rüzgarın adıdır.

"İstanbul'da güneydoğudan esen rüzgara ne ad verilir?"

"Güneydoğudan esen rüzgara İstanbul'da ne ad verilir?"

Bu farklılıktan dolayı, her bir soru kalıbı için, bağlılık ağacı üzerinde inceleme yapabilecek kurallara ihtiyaç vardır. Model teşkil etmesi bakımından bu çalışmada üç tane kalıp incelenmiş ("nedir", "verilir" ve "denir") ve her biri için odağı bulacak ayrı kurallar belirlenmiştir.

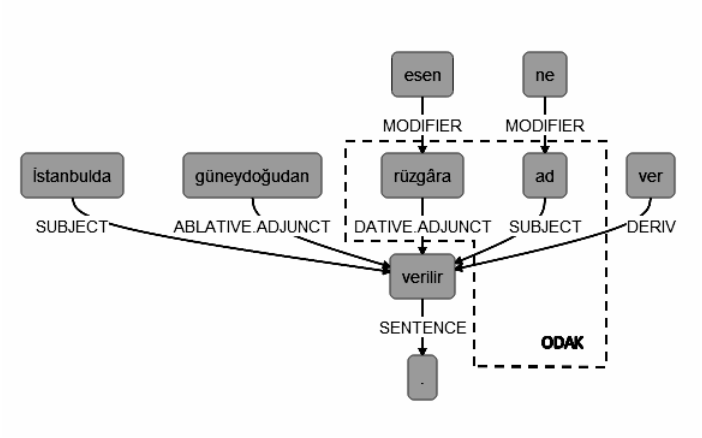
"nedir"

"Bitki ve hayvan topluluklarını inceleyen bilim dalı nedir?"

Bu gibi sorularda, ilk başta cümlenin yüklemine bağlı özne bulunur. Daha sonra bu özneye bağlı belirtisiz isim tamlaması varsa, bu tamlayan üzerinden devam edilerek, müteakip bütün tamlayanlar toplanır. Varsa toplanan tamlayanlar ile beraber özne, sorunun odağı olarak belirlenir.

"denir"

"Sıvı haldeki astenosfer malzemesine ne denir?"



Şekil 3: "İstanbul'da güneydoğudan esen rüzgara ne ad verilir?" sorusunun bağlılık ağacı.

Bu tip sorularda ise, özne olmaksızın, yükleme bağlı yönlem bağlacı bulunur. Yönlem bağlacının olduğu dal üzerinde, varsa yönleme bağlı art arda belirtisiz isim tamlamaları toplanır. Sonunda yönlem bağlacı ve varsa toplanmış tamlayanlar soru tümcesinin odağı olarak çıkartılır.

"verilir"

"Yeryüzündeki büyük kara parçalarına ne ad verilir?"

Burada ilk olarak yükleme bağlı özne bulunur. Öznenin yanında yine yükleme bağlı yönelme bağlacı (DATIVE.ADJUNCT) aranır. Daha sonra ikisinin dallarından, varsa art arda takip eden belirtisiz isim tamlamaları toplanır. Sonunda özne ve yönlem ile birlikte varsa toplanan tamlayanlar, sorunun odağını teşkil eder. Şekil 3'de "verilir" kalıbıyla ilgili örnek bir bağlılık ağacı ve yukarıda anlatılan yöntem ile çıkartılan odak kısmı görülmektedir.

Şekil 2'de görüldüğü gibi, sisteme verilen herhangi bir soru için, ön işleme sürecinden sonraki ilk adım, sorunun kalıbının tespit edilmesidir. Daha sonra soru, bu tespit edilen kalıba uygun kurala göre işlenecek ve bu kural vasıtasıyla odağı çıkartılacaktır. Kalıbın belirlenebilmesi, bu üç kuralın da tüm soru kümesi üzerinde aynı anda yürütülebilmesine olanak sağlamaktadır. Bu yöntem ayrıca birden fazla soru kalıbı hakkında bilgi sahibi olabilecek kuralların yapılabilmesine ve bir soru hakkında bu tür kuralların birbirleri arasında gerçekleşebilecek güven (confidence) tabanlı bir fikir alışverişine alt-yapı hazırlamıştır.

IV. Deneyler ve Değerlendirme

Önerilen yöntemin deneyleri, uzmanlar tarafından elle toplanıp, ikişerli gruplar halinde ayrı ayrı işaretlenerek oluşturulan 500 soruluk soru kümesi kullanılarak yapılmıştır. Sorular, veritabanında

<Soru Metni> | <Odak Sözcükler>

şeklinde tutulmaktadır. Örneğin:

Akdeniz bölgesinin doğal bitki örtüsü nedir | doğal bitki örtüsü

III. bölümde anlatılan yöntem dahilindeki kurallar, çalışmanın ilk aşamasında toplanıp işaretlenen 107 soru üzerinde yapılmış, sonradan toplanan sorular üzerindeki deneme işlemleri sırasında (küçük ayarlamalar dışında) kurallarda bir değişiklik yapılmamıştır. Buna böre, 107 soru geliştirme kümesi, geriye kalan 393 soru ise doğrulama kümesi olarak kullanılmıştır.

Ayrıca belirtmek gerekir ki, deneylerin yapıldığı soruların veritabanındaki metinlerine soru işareti (?) dahil değildir, oysa soru soracak kullanıcının bahsi geçen işareti kullanması doğaldır. Kullanılan soru işaretleri de diğer işaretler ile birlikte, ön işlem sırasında noktalama işaretlerinin elenmesi aşamasında atıldığı için, yeni verilen soru bu çalışmanın yapıldığı soruların şekli ile aynı hale gelmektedir.

	Soru Sayısı	**Kesinlik**[3]	**Duyarlılık**[4]	**F-Ölçütü**[5]
nedir	150	0.86	0.82	0.84
denir	96	0.81	0.72	0.76
verilir	72	0.89	0.88	0.88
toplam	500	0.73	0.74	0.74

Tablo I: Soru tipleri ve başarım değerleri

Tablo I'de ayrı ayrı soru kalıpları için oluşturulan kuralların hem kendi başlarına, hem de bütün halinde soru kümesi üzerindeki sonuçları gösterilmektedir. Bu tablodan anlaşıldığı gibi, her bir soru kalıp kuralı, kendi soruları içerisinde ayrı ayrı değerlendirilmeye sokulmuştur. Bunun yanında toplam soru kümesi üzerinde de sonuçlar belirtilmiştir. Belirtmek gerekir ki, bu üç soru kalıbının yanında, soru kümesi içerisinde (ne zaman, kim, nerede gibi) farklı kalıplar kullanılarak ifade edilmiş sorular da (182 tane) mevcuttur. Bu kalıpların her biri için kural oluşturacak kadar çeşitli sorular soru kümesi içerisinde olmadığından, o kalıpların ele alınması yeterli soru çeşitliliğine ulaşıldığında yapılacaktır. Ayrıca, bir kalıba oturtulamayan sorularda ve kalıp kuralının kural dışı bir yapı ile karşılaştığı durumlarda harekete geçecek genel bir kuralın oluşturulması da ileriki çalışmalarda amaçlanmaktadır.

V. Sonuç

Bu çalışma, lise öğrencilerinin eğitimlerine teknik destek sağlamak amacıyla yeni tasarlanan bir Türkçe soru cevaplama sisteminin ilk adımı olan soru analizi problemine ışık tutmaya çalışmıştır. Soru analizinde, soruda tam olarak neyin sorulduğunun (odak) anlaşılabilmesi sağlanmaya çalışılmıştır. Bunun için, model olarak belirlenen soru kalıplarının her biri (nedir, denir, verilir) için bağlılık ayrıştırıcısından (dependency parser) çıkan bağlılık ağacı üzerindeki ilişkilerden çıkarım yapabilen kurallar belirlenmiştir. Bu kurallar, uzmanlar tarafından elle derlenip (odağı) işaretlenmiş 500 coğrafya sorusu üzerinde denenmiş ve sonuçlar rapor edilmiştir.

İleriki çalışmalarda, soru çeşitliliğinin arttırılmasıyla birlikte, diğer (ne zaman, kim, nerede gibi) kalıpların da ele alınıp onlar için de kural belirlenmesi, ayrıca birden fazla kalıp hakkında fikir sahibi olabilecek kapsayıcı kurallar ve bu kurallar arasında güven (confidence) tabanlı iletişim üzerine çalışılması amaçlanmaktadır. Bu sayede, soru analizi tamamlanacak ve soru cevaplama sisteminin ileriki kısımları, soruda anlamsal olarak neyin sorulduğuna dair bir fikir sahibi olacağından dolayı, daha akıllı ve kapsamlı analizler sonucunda hesapsal anlamda verimli bir şekilde daha doğru cevaplar bulabilecek, dahası bilgi dağarcığında var olmayan cevapları üreterek, kendi bilgi dağarcığını da genişletebilecektir.

VI. Teşekkür

Bu çalışma, 113E036 proje numarası kapsamında TÜBİTAK tarafından desteklenmiştir.

Kaynakça

[1] J. Kaur, V. Gupta, "Effective Question Answering Techniques and their Evaluation Metrics", International Journal of Computer Applications, 2013.

[2] P. Gupta, V. Gupta, "A Survey of Text Question Answering Techniques", International Journal of Computer Applications, 2012.

[3] S. İlhan, N. Duru, Ş. Karagöz, M. Sağır, "Metin Madenciliği ile Soru Cevaplama Sistemi", ELECO, 2008.

[4] Z. Zheng, "AnswerBus Question Answering System", HLT, 2002.

[5] B. Katz, "Annotating the World Wide Wed Using Natural Language Processing", RIAO, 1997.

[6] Y. Pan, Y. Tang, Y.M. Luo, L.X. Lin, G.B. Wu, "Question Classification Using Profile Hidden Markov Models", International Journal on Artificial Intelligence Tools, 2009.

[7] R. Bunescu, Y. Huang, "Towards a General Model of Answer Typing: Question Focus Identification", CICLING, 2010.

[8] G. Eryiğit, "The impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish", LREC, 2012.

[9] J. Nivre, J. Hall, J. Nilsson, A. Marinov, E. Marsi, "MaltParser: A Language-Independent System for Data-Driven Dependency Parsing", Natural Language Engineering Journal, 2007.

[10] G. Eryiğit, J. Nivre, K. Oflazer, "Dependency Parsing of Turkish", Computational Linguistics, 2008.