OTAP Osmanlıca Metinleri Internet Arayüzü
OTAP Ottoman Archives Internet Interface

Emre Şahin[1], Hande Adıgüzel[1], Pınar Duygulu[1], Mehmet Kalpaklı[2]

[1]Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi
06800 Bilkent Ankara
{iesahin, adiguzel, duygulu}@cs.bilkent.edu.tr

[2]Tarih Bölümü
Bilkent Üniversitesi
06800 Bilkent Ankara
kalpakli@bilkent.edu.tr

ÖZETÇE

Osmanlı Metin Arşivi Projesi *kapsamında Osmanlı Türkçesi metinlerinin yüklenmesi, ikilileştirilmesi, satır ve kelime bölütlenmesi, etiketlenmesi, tanınması ve testlerinin yapılması amacıyla bir Genel Ağ arabirimi geliştirilmiştir. Bu arabirim sayesinde Osmanlı arşivleriyle çalışan araştırmacıların uzmanlık yardımının alınması ve geliştirdiğimiz tanıma teknolojilerinin elyazması arşivlere uygulanması mümkün hale gelmiştir.*

ABSTRACT

Within *Ottoman Text Archive Project* a web interface to aid in uploading, binarization, line and word segmentation, labeling, recognition and testing of the Ottoman Turkish texts has been developed. It became possible to retrieve expert knowledge of scholars working with Ottoman archives through this interface, and apply this knowledge in developing further technologies in transliteration of historical manuscripts.

1. Giriş

Osmanlı Devletinin altı yüzyıllık ömründen kalan arşivleri araştırmacıların ilgisini her zaman çekmiştir. Ülkeye yayılmış değişik kütüphanelerde tutulan arşivlerin daha erişilebilir hale getirilmesi, çevrimiçi veritabanlarının kurulması, tarihi metin işleme ve el yazısı tanıma teknolojilerinin kullanımını zorunlu hale getirmektedir. Bilkent Üniversitesi ve University of Washington'ın ortak yürüttüğü OTAP Osmanlı Metin Arşivi Projesinin amacı, hem probleme ilişkin Bilgisayarla Görme bilimsel araştırmalarına zemin sağlamak, hem de Tarih ve Edebiyat gibi alanlarda çalışan araştırmacıların teknolojik imkanları kullanarak arşivlere erişimini kolaylaştırmaktır.

Osmanlıca metin işleme konusunda bu proje kapsamında ve dışında yürütülen çalışmalar algoritma geliştirmeye odaklanmıştır. Çalışmaların olduğu halleriyle Bilgisayar Bilimi dışında çalışan araştırmacıların kullanması zordur. Bu eksikliği gidermek, bugüne kadar bu proje çatısı altında yapılan çalışmalarda edinilmiş ilerlemeyi araştırmacıların kullanımına sunmak amacıyla Genel Ağ üzerinden çalışan bir arabirim geliştirilmiştir.

Bu arabirim yardımıyla yeni Osmanlıca dokümanlar veritabanına kaydedilebilmekte, ikilileştirilmekte, satır bölütleme algoritmaları çalıştırılmakta, uzmanlık bilgisi gerektiren ve gözetimli öğrenme için elzem olan etiketleme işlemleri gerçekleştirilebilmekte, etiketlenen metinler üzerinde kelime araması ve etiketlenmemişler üzerinde resim yoluyla arama yapılabilmektedir.

Bu çalışmada OTAP Kaşifi adını verdiğimiz bu arabirimin tanıtımını yapmaktayız.

2. Yazılım Katmanları

OTAP Kaşifi dört ana bileşenden oluşmaktadır. Kullanıcı önyüzünü oluşturan Genel Ağ katmanı, Genel Ağ isteklerini işleyen ve yönlendiren sunucu katmanı, resim ve üstbilgi barındıran veritabanı katmanı ile Bilgisayarla Görme algoritmalarını çalıştıran algoritmik işleme katmanı.

2.1. Genel Ağ Katmanı

Genel Ağ Katmanı kullanıcının OTAP Kaşifiyle etkileştiği bileşendir. Bu bileşen sayesinde öncelikle kullanıcıların tanıtlama ve yetkilendirme yapılmaktadır.

Kullanıcılar ellerindeki JPEG, PNG veya PDF formatında taranmış sayfaları bu arabirim sayesinde yükleyebilir. Bu sayede araştırmacıların edindikleri dosyaların ortak bir konumda toplanması mümkün olmuştur. Yükledikleri belgelere dair kitap adı, yazar, tarih gibi künye bilgilerini girdikleri zaman bu bilgiler yoluyla dosyaya erişim mümkün olmaktadır.

Bilgisayarla Görme araştırmalarımız açısından hayati bulunan sözcük etiketleme de bu arabirim yardımıyla yapılmaktadır. Elyazmalarında sözcük arası boşlukların belirgin olmayışı nedediyle otomatik olarak kelimelere ayırmak mümkün değildir. Bu nedenle araştırmacıların satırlara ayrılmış metinlerden kelimeleri işaretleyebilecekleri ve etiketleyebilecekleri bir arabi-

Bu çalışma Türkiye Bilimsel ve Teknik Araştırma Kurumu tarafından 109E006 proje numarasıyla desteklenmektedir.

rim sunulmuştur. Kullanıcılar fare yardımıyla, kelimenin altını çizer gibi seçim yapmakta, taşmış olan fazlalıkları silebilmekte ve sözcüğün okunuşunu transliterasyon alfabesiyle yazabilmektedir. OTAP Kaşifi, hem önceki yıllarda geliştirilmiş OTAP Osmanlıca Transliterasyon Alfabesini[1], hem de sanal klavye yoluyla doğrudan Osmanlıca girişini desteklemektedir.

Kullanıcılar açısından diğer önemli işlev aramadır. Kullanıcı arayüzü üç şekilde aramaya olanak sunmaktadır. Bunlardan birincisi etiketler yoluyla sözcüklerin ve onların geçtiği metinlerin bulunmasıdır. Bu sözcükler bulunduktan sonra bağlı parçalar çıkarılabilmekte ve bu sayede Türkçe çekim ekleri sözcük resimlerinden gerektiğinde silinebilmektedir. Üçüncü arama yöntemi serbest metin tabanlı aramadır. Kullanıcı gerek OTAP Transkripsiyon Alfabesiyle, gerek sanal klavye yoluyla metin girebilmekte, bunlar çeşitli Arapça ve Farsça fontlar yardımıyla resme çevrilmekte ve öznitelikleri çıkarılarak arama yapılmaktadır.

2.2. Sunucu Katmanı

Sunucu katmanı Genel Ağ katmanından gelen istekleri yönlendiren ve Genel Ağ katmanına gerekli bilgileri sağlayan katmandır. Kullanıcı arabirimiyle JSON (JavaScript Object Notation) formatında iletişim kurmakta ve gerekli HTML sayfalarını üretmektedir. Python dilini kullanan Django çerçevesi yardımıyla veritabanı iletişimi ve dinamik web dokümanları üretilmektedir.

2.3. Veritabanı Katmanı

Veritabanı katmanı iki farklı biçimde veritabanı bilgilerini saklamaktadır. Yüklenen belgelerin künyeleri, resim parçalarının konumları ve hangi sayfaya ait oldukları, etiketleri gibi üstbilgi bir SQLite veritabanında saklanmaktadır. Sayfa resimlerinin diskteki konumları da veritabanında saklanmaktadır. Sayfaların kendisi ise diskte tutulmakta, bu şekilde resimleri işlediğimiz algoritmaların diğer katmanlardan bağımsız şekilde test edilmesi mümkün olmaktadır. Veritabanı yapısı yine Django çerçevesinde yapılmış ve veritabanı yazılımından bağımsızdır, SQLite'ın yetersiz gelmesi durumunda yüksek talep sayısında daha iyi performans sunan PostgreSQL'e geçiş yapmak sadece veritabanı bağlantı ayarları yapıp, veri aktarımını yaparak mümkündür.

2.4. Algoritmik İşleme Katmanı

Osmanlı metinlerinin işlenmesi konusunda yapılmış bir çok çalışma ve kaydedilmiş ilerleme bu bileşen yardımıyla kullanıma sunulmaktadır. Kendi başına yeterli ve belli bir işi yapan küçük komut satırı programları şeklinde tasarlanmıştır. Grubumuzun [2], [3] ve [4] gibi daha önce yaptığı çalışmalar gerek Matlab ve Octave'la aynen, gerek OpenCV kütüphanesini kullanan C++ ve Python programlarına çevrilerek bu katmanda yer almaktadır.

Değişik ikilileştirme algoritmaları kullanan `otap-binarize`, sayfaları satırlara ayıran `otap-segment`, sözcük resimleriyle öznitelik çalışması yapan `otap-train`, bu özniteliklerle arama yapan `otap-search` ve `otap-spot` programları çeşitli parametreler yardımıyla kullanılmaktadır.

3. Sonuç

Bu tanıtımda OTAP Kaşifi adını verdiğimiz Tarihi Metin İşleme ve El yazısı işleme arabirimini tanıttık. Çeşitli katmanlardan oluşan bu yazılıma `http://retina.cs.bilkent.edu.tr/otap/` adresinden ulaşılabilir. Konuk hesaplarına ilişkin bilgiyi `iesahin@cs.bilkent.edu.tr` adresinden temin edebilirsiniz.

4. KAYNAKÇA

[1] W. G. Andrews, M. Inan, S. Kebeli, and S. Waters, "Rethinking the transcription of ottoman texts: The case for reversible transcription," http://courses.washington.edu/otap/reverse/reverse/o_Reverse_trans_article728.html, 2008.

[2] E. Ataer and P. Duygulu, "Matching ottoman words: An image retrieval approach to historical document indexing," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 341–347.

[3] E. Can and P. Duygulu, "A line-based representation for matching words in historical manuscripts," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1126–1138, June 2011.

[4] D. Arifoglu, P. Duygulu, and M. Kalpakli, "Segmentation of historical documents using cross document word matching," *Pattern Recognition Letters*, In Review.