Çok Dilli Metin Analizinde Alan Bağımlı Değerlendirme Verisinin Oluşturulması
Domain-specific Evaluation Dataset Generator for Multilingual Text Analysis

Emrah İnan[1], Vahab Mostafapour[1], Fatih Tekbacak[2]
[1]Bilgisayar Mühendisliği Bölümü, Ege Üniversitesi, İzmir, Türkiye
{emrah.inan, vahab.mostafapour}@ege.edu.tr
[2]Bilgisayar Mühendisliği Bölümü, Adnan Menderes Üniversitesi, Aydın, Türkiye
ftekbacak@adu.edu.tr

***Özetçe*—Web, insanlar, organizasyonlar, sinema filmleri ve onların özellikleri ile ilgili belirli varlıklar için gerekli bilgilerin edinilmesini sağlamaktadır. Bununla beraber birçok Web kaynağı genel olarak yapısal olmayan biçimde durmaktadır ve bu durum belirli varlıklar ile ilgili kritik bilginin bulunmasını zorlaştırmaktadır. Tanımlı Varlık Çıkarımı ve Varlık Bağlama gibi metin analizine dayalı yaklaşımlar varlıkların etiketlenmesi ve verilen bilgi tabanı kaynağındaki ilgili varlıklarla bağlanmasını amaçlamaktadır. Böyle yaklaşımları test etmek için çok fazla genel amaçlı test kümeleri bulunmaktadır. Ancak alan bağımlı yaklaşımları test etmek alana özgü veri kümelerinin eksikliğinden dolayı zorlaşmaktadır. Bu çalışma, çok dil destekli test verisini Vikipedi kategori sayfaları ve DBpedia hiyerarşisini kullanarak belirli alanlar için üreten WeDGeM aracını sunmaktadır. Aynı zamanda, Vikipedi anlam ayrımı sayfaları, üretilen test metinlerinin anlam karmaşıklığı seviyesini ayarlamak için kullanılmaktadır. Üretilen bu test verisinde, Türkçe metinlerini destekleyen tanınmış Varlık Bağlama araçları sinema alanında test edilmiştir.**

Anahtar Kelimeler—varlık bağlama; tanımlı varlık çıkarımı; test kümesi; Dbpedia; Wikipedia.

***Abstract*—Web enables to retrieve concise information about specific entities including people, organizations, movies and their features. Additionally, large amount of Web resources generally lies on a unstructured form and it tackles to find critical information for specific entities. Text analysis approaches such as Named Entity Recognizer and Entity Linking aim to identify entities and link them to relevant entities in the given knowledge base. To evaluate these approaches, there are a vast amount of general purpose benchmark datasets. However, it is difficult to evaluate domain-specific approaches due to lack of evaluation datasets for specific domains. This study presents WeDGeM that is a multilingual evaluation set generator for specific domains exploiting Wikipedia category pages and DBpedia hierarchy. Also, Wikipedia disambiguation pages are used to adjust the ambiguity level of the generated texts. Based on this generated test data, a use case for well-known Entity Linking systems supporting Turkish texts are evaluated in the movie domain.**

Keywords—entity linking; named entity recognition; evaluation dataset; Dbpedia; Wikipedia.

I. GİRİŞ

Web üzerindeki kaynakların sayısının artmasıyla birlikte insanların organizasyonlar, sinema filmleri ve onların özellikleri ile ilgili belirli varlıklar için gerekli bilgileri edinmesi sağlanmıştır. Buna rağmen, birçok Web kaynağı genel olarak yapısal olmayan biçimde durmaktadır. Anlamsal Web ve Bağlı Veri'nin ortaya çıkması ile Web üzerindeki yapısal veri kaynakları hızla büyümektedir. Büyüyen Bağlı Açık Veri bulutunda LinkedMDB [1] ve KnowLife [2] gibi alana özgü birçok bilgi tabanı bulun- maktadır. Bu bilgi kaynaklarından anlamsal arama, soru cevaplama ve bilgi çıkarımı gibi çeşitli anlamsal veri madenciliği [3] yöntemlerinin geliştirilmesinde yararlanıl- maktadır. Bu yöntemler metinlerin yapısallaştırılması için en temel adım olan Varlık Bağlama sistemlerine ihtiyaç duymaktadır.

Varlık Bağlama, verilen bilgi tabanındaki varlıkların metindeki atıflarla eşleşmesini amaçlamaktadır. Tanımlı Varlık Çözümleme, Varlık Bağlama yöntemleri için anlam karmaşıklığı olan aday varlıklar arasında en uygun atıf-varlık çiftinin çözümlenmesini amaçlayan önemli bir alt görevdir. Bu yöntemleri test etmek için başta MSNBC [4], IITB [5] ve Wikilinks [6] olmak üzere birçok genel amaçlı veri kümeleri bulunmaktadır. MSNBC ve IITB kümeleri elle etiketlenmiş alan bağımsız popüler web dokümanlarını içermektedir. Elle etiketlenmiş kümelerdeki problem, belli başlı varlık türlerini etiketleme ve çok az sayıda ortak karara göre işaretleme yapılması olarak gösterilebilinir. Bu

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

eksikliklerin giderilmesi için Wikilinks çalışması Vikipedi kaynağından geniş ölçekli otomatik etiketlenmiş veri kümesi sunmaktadır. Ayrıca Wikilinks etiketli varlıklara ait insan, organizasyon ve yer özelliklerini de içermektedir.

Navigli [7] çalışmasında alana özgü bilgi kaynaklarının, Varlık Bağlama yöntemlerinin daha etkin performansa ulaşabilmesi için gerekliliğini ve öne çıkan bir süreç olduğunu vurgulamıştır. Ancak bu aşamada farklı dillere ve farklı alanlara özgü bilgi tabanlarının eksikliğinden dolayı yöntemlerin karşılaştırılabileceği bir ortam yada veri kümesi bulunmamaktadır. Bu bölümde bu eksikliğin üstesinden gelebilmek için geliştirilen WeDGeM aracının alana ve dile özgü değerlendirme veri kümesi üretilmesi için katkıları aşağıda listelenmiştir:

- WeDGeM aracı Vikipedi kategori sayfaları ve DBpedia taksonomisindeki konu sınıflandırmasından faydalanarak farklı alanlara özelleşmiş etiketli veri kümesi oluşturma desteği vermektedir.
- WeDGeM ayrıca çok dil destekli ve alana özgü varlık etiketleme araçlarını değerlendirmek için Vikipedi ve DBpedia destekli doğal dillerden oluşturulabilecek veri kümesi de sunmaktadır.

WeDGeM aracı Vikipedi anlam ayrımı sayfalarını kullanarak yeterli anlam karmaşıklığını [8] sağlamakta ve varlık etiketleme araçlarının adil ve nesnel bir şekilde karşılaştırılmasına olanak tanımaktadır. WeDGeM aracı ile oluşturulan sinema alanına özgü etiketlenmiş veri kümesi Türkçe ve İngilizce dilleri için hazırlanmıştır. Basit bir şekilde hazırlanan bu veri kümesi daha sonra çok bilinen varlık etiketleme yaklaşımları için değerlendirme çerçevesi olan GERBIL aracına monte edilerek Türkçe dil desteği veren Babelfy [9] ve DBpedia Spotlight [10] olmak üzere iki yaklaşımda denenmiştir. Bu denemenin amacı seçilen yaklaşımları karşılaştırmak yerine oluşturulan veri kümesinin uygulanabilirliğini göstermektir.

Anlamsal Bilgi Çıkarımı, bilgi tabanı kullanan yada açık veri üzerine çalışmalar yapanlar için önemli bir araştırma alanıdır. Anlamsal Bilgi Çıkarımı, Varlık Bağlama ve İlişki Çıkarımı alt görevlerini içermektedir. Bu çalışmada İlişki Çıkarımı alt bölümü üzerinde durulmaktadır. Kişi, organizasyon veya lokasyon gibi varlıklar arasında ilişkiler bulunmaktadır. Bu ilişkilerin oluşturulabilmesi için varlıkların anlamsal özelliklerine uygun biçimde etiketlenmesi gerekir. Dolayısıyla bir doğal dil metninde bulunan etiketlenmiş varlıklar arasında bulunan anlamsal ilişkilerin çıkarılması gerekli bir adımdır. İlişki Çıkarımı, yapısal bilgi tabanlarının oluşturulmasında önemli bir rol oynar. Oluşan yapısal veri, WordNet [11] gibi sözlüklere veya DBPedia gibi bilgi tabanlarına eklenerek uygulamalar tarafından kullanılır hale getirilir.

Anlamsal Bilgi Çıkarımı yöntemlerinin ana amacı yapısal olmayan web sayfalarındaki içeriği yapısallaştırarak makina okunur hale getirmektir. Bu amacı sağlayabilmek için yapısal olmayan metinde varlıkların ve onlarla birlikte geçen ilişkilerin belirlenmesi gerekmektedir. Örnek olarak "Bill Gates, Microsoft şirketinde çalışmaktadır." cümlesinde "Bill Gates" ve "Microsoft" varlıkları "çalışır(BillGates,Microsoft)" anlamsal ilişkisine sahiptir.

Bilgi Çıkarımı doğal dilde yazılmış metnin tanımlanması ve ilgili bilgilerin çıkarılması gibi analiz süreçleri ile uğraşmaktadır. Bilgi Çıkarımı sistemleri, Açık Bilgi Çıkarımı ve Ontoloji Tabanlı Bilgi Çıkarımı olmak üzere iki kategoride toplanmaktadır. Açık Bilgi Çıkarımı teknikleri açık verideki korpuslardaki fiil tabanlı ilişkilerle ilgilenirken Ontoloji Tabanlı Bilgi Çıkarımı alana özgü bilgi tabanı kaynaklarını kullanan modele dayanmaktadır. Bu kategorideki ontolojiler bilgi çıkarım sürecine rehberlik ederek gerekli bilginin tanımlanmasını sağlamaktadır. Ayrıca ontolojiler çizge tabanlı yapıdaki anlamsal kapsamı göstermektedir.

İlişki Çıkarımı varlık çiftleri arasındaki ilişkilerin bulunmasıyla ilgilenmektedir. Varlıklar arasında taksonomik ve taksonomik olmayan şeklinde iki tür anlamsal ilişki kurulmaktadır. Örnek olarak "is-a" ilişkisi taknosomik bir ilişkiyi, "located-in" ilişkisi de taksonomik olmayan bir ilişkiyi göstermektedir.

II. WEDGEM ARACININ YAKLAŞIMI

WeDGeM aracının kullanılmasında ilk adım Vikipedi kategorisinden hangi alan özgü etiketli veri kümesi oluşturulacağının ve hangi doğal dil için üretileceğinin seçilmesidir. Seçilen dile ve Vikipedi kategorisine göre bilgi kutularındaki atıflara ait Vikipedi dokümanlarında yer alan çıpa bağlantıları çekilerek bu bağlantılar, DBpedia bilgi tabanındaki o alana özelleştirilmiş varlıklarla eşleştirilmek- tedir. Örneğin doğal dil olarak Türkçe seçildiğini ve sinema alanında test verisi üretilmek istendiğini düşünelim. Örnek olarak "Wicker Park" atıfı göz önünde bulundurulduğunda bu atıf DBpedia ile zenginleştirilmiş sinema ontolojisindeki ilgili varlıkla eşleşmektedir.

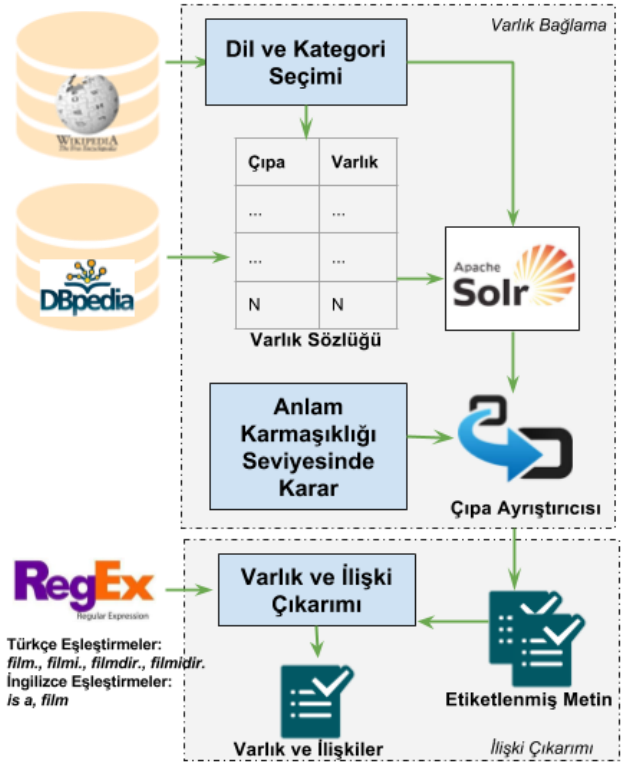
DBpedia bilgi tabanındaki güçlü hiyerarşiden yararlanılarak ve eşleşen varlıkların tiplerine bakarak mevcut metnin bağlamıyla örtüşüp örtüşmediğinin sağlaması da yapılmaktadır. Ayrıca insan, organizasyon ve yer bilgisi gibi varlıkların genel tiplerini de tutarak Varlık Bağlama yöntemlerinin yanında etiketleme yaklaşımlarının da değerlendirilmesine olanak sağlanmıştır. Bütün bu özellikleri çekilmiş varlıklar daha sonra tanımlı varlık ve kavram sözlüğünde saklanmaktadır. "Wicker Park" atıfına ait Vikipedi sayfasında bilgi kutusunda "Josh Hartnett" adlı oyuncunun bu filmin yıldızlarından biri olduğu belirtilerek varlık olarak gerekli tip bilgileri ve URI bağlantısı DBpedia kaynağından çekilmiştir.

Şekil 1 deney veri kümesi üretim aracının genel işleyiş yapısını göstermektedir. Vikipedi bilgi kutularından ve DBpedia sınıflandırmasından üretilen varlık sözlüğü Apache Solr[1] ile indekslenerek arama ve geri çağırma

[1] http://lucene.apache.org/solr/

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

sürelerini hızlandırmaktadır. Seçilen alana ve dile göre



Şekil 1. WeDGeM aracının ana yapısı

bulunan Vikipedi dokümanları paragraflara ayrılmıştır. Her bir paragraf için indekslenen varlık sözlüğünde aranarak mevcut ontolojideki varlıkların paragrafta yer alıp almadığına bakılmaktadır. Eğer paragrafta bir yada daha fazla varlık etiketli halde bulunuyorsa ve bu paragraf daha önce etiketlemiş metin listesinde yer almıyorsa bu listeye eklenir. Bu işleme paralel olarak her bir Vikipedi makalesinde ilgili varlık için anlam ayrımı sayfası aranır. Eğer böyle bir sayfa içeriyorsa bu anlam karmaşıklığına yol açacak etiketli metinler de anlam ayrımı listesinde tutulur.

Vikipedi makaleleri W_a Vikipedi yığınlarından[2] seçilen her bir doğal dil için Vikipedi kategori sayfası elde edilir. Bu metinlerdeki bilgi kutularından DBpedia ile doğrulanarak varlık sözlüğü verilen alana özgü üretilir. Üretilen bu varlık sözlüğü indekslenip her bir Vikipedi makalesindeki etiketli paragraflarda sorgulanır. Alana özgü varlıklara ait etiketlenmiş ve daha önce listede yer almayan metinler etiketli metin listesine eklenir. Mevcut metinlerdeki varlıklar varlık listesine eklenerek Vikipedi anlam ayrımı sayfalarında aranır. Vikipedi anlam ayrımı sayfaları da normal sayfalara uygulanıldığı gibi paragraflara ayrılarak yine etiketli metinlerde varlık sözlüğündeki varlıkların varlığı sorgulanır. Eğer anlam karmaşıklığı içeren sayfa varsa bu metin bu sefer etiketli anlam ayrımı metin listesinde tutulur. Son adımda belirlenen anlam karmaşıklığı oranına göre iki listedeki metinler harmanlanarak istenilen karmaşıklık elde edilir. Algoritmadan elde edilen veri kümeleri bağlantıda[3] yer almaktadır ve GERBIL[4] değerlendirme çerçevesinde istenilen varlık etiketleme yöntemi için kullanıma hazır biçimde bulunmaktadır.

A. *Etiketli Test Metninin Yapısı*

WeDGeM aracını kullanarak Vikipedi metinlerinde eşleşmesi kolay olduğu için DBpedia varlıklarını içeren seçili dil ve alana özgü etiketli test metninin üretilmesine olanak sağlanmıştır. Aynı zamanda anlam karmaşıklığı seviyesi belirlenerek adil bir ortamda varlık etiketleme yöntemlerinin karşılaştırılmasına olanak verilmiştir.

Anlam karmaşıklığını yükseltmek için çekilmiş varlıklara ait Vikipedi anlam ayrımı ve yeniden yönlendirme bağlantıları kullanılmaktadır. Değerlendirme metin kümesinde birçok etiketli metin bulunmaktadır. Bu metinlerin birinin resmi biçimi Liste 1 olarak gösterilmiştir. Örnek metinde NIF[5] doğal diller arası değişim biçimi olarak gösterilmektedir. Bunun amacı RDF/OWL tabanlı bir biçim olarak her türlü varlık etiketleme aracı için standart oluşturması ve metin analiz araçları içinde halihazırda kullanılmasıdır.

Her bir doküman ID parametresine sahiptir ve bu da çekilmiş metindeki Vikipedi sayfa id bilgisine ek olarak kaçıncı paragraf olduğunu da göstermektedir. Liste 1'de görülen *Wicker Park* örnek metinde ID olarak 293 verilmiş ve bu metnin 66 karakterden oluştuğu gösterilmiştir. Referans URI olarak *http://siege.ege.edu.tr/* isim uzayı örnek alınmıştır ve her tür belge başlangıç ve bitiş indekslerine sahip olarak belge boyutunu belirlemektedir. Her bir etiketli metin birden fazla varlık içerebilir ve bunlara ait atıflar başlangıç ve bitiş indeksleri ile gösterilir. *Wicker Park* örneğinde bu atıfın başlangıç indeksi 0 yani metnin başında yer aldığını gösterirken bitiş indeksi de 11 karakteri yani atfın uzunluğunu gösterir. Ayrıca bu atıf verilen bilgi tabanındaki ilgili varlık ile eşleşerek onun URI bilgisi de belgede tutulur. Varlık etiketleme araçlarında da kullanılması için atıfa ait insan, organizasyon ya da yer ile ilgili sınıf türü de belgede saklanır. Bu çalışmada DBpedia bilgi tabanındaki sınıf taksonomisi alana uygun varlıklara ait etiketli metinlerin barındırılması için doğrulama amacıyla kullanılmıştır. Bu nedenle alan tipi DBpedia içinde sorgulanan her bir varlığa ait alan parametresi olarak belgeye eklenmiştir.

B. *İlişki Çıkarımı*

Bir metindeki olası ilişkiler iki aşamalı örüntü tabanlı ilişki çıkarımı yöntemi ile bulunmaktadır. İlk aşamada el ile tanımlanmış desenler ve kök ilişkiler dikkate alınarak ilişkiler çıkarılmaktadır. E (entity) varlığı ifade ettiği anda,

[2] https://dumps.wikimedia.org/
[3] https://github.com/einan/eval4J
[4] http://aksw.org/Projects/GERBIL.html
[5] http://persistence.uni-leipzig.org/nlp2rdf/

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

```
@prefix rdf: <http://www.w3.org/1999/02/
                    22-rdf-syntax-ns#> .
@prefix  xsd:<http://www.w3.org/
                         2001/XMLSchema#> .
@prefix itsrdf:<http://www.w3.org/2005/
                              11/its/rdf#> .
@prefix nif: <http://persistence.uni-
leipzig.org/nlp2rdf/ontologies/nif-core#>.
@prefix rdfs:<http://www.w3.org/
                2000/01/rdf-schema#>.

<http://siege.ege.edu.tr/
                    document293#char=0,11>
a  nif:RFC5147String, nif:String
        , nif:Phrase;
nif:anchorOf  "Wicker Park"^^xsd:string;
nif:beginIndex "0"^^xsd:nonNegativeInteger;
nif:endIndex "11"^^xsd:nonNegativeInteger;
nif:referenceContext <http://siege.ege.
        edu.tr/document293#char=0,66>;
itsrdf:taClassRef
<http://dbpedia.org/ontology/Film>;
itsrdf:taIdentRef <http://dbpedia.org/
         resource/Wicker_Park_(film)>.

<http://siege.ege.edu.tr/
        document293/#offset_12_16>
a  nif:Phrase , nif:OffsetBasedString ;
nif:anchorOf "is a" ;
nif:beginIndex"12"^^xsd:nonNegativeInteger;
nif:endIndex "16"^^xsd:nonNegativeInteger;
nif:referenceContext
<http://siege.ege.edu.tr/
        document293/#offset_0_66> ;
itsrdf:taAnnotatorsRef
<http://siege.ege.edu.tr/trex> ;
itsrdf:taIdentRef
    <http://www.w3.org/1999/02/
                22-rdf-syntax-ns#type> .

<http://siege.ege.edu.tr/
        document293/#offset_59_63>
a nif:Phrase , nif:OffsetBasedString;
nif:anchorOf "film" ;
nif:beginIndex"59"^^xsd:nonNegativeInteger;
nif:endIndex "63"^^xsd:nonNegativeInteger;
nif:referenceContext
    <http://siege.ege.edu.tr/
            document293/#offset_0_66> ;
itsrdf:taAnnotatorsRef
<http://siege.ege.edu.tr/trex> ;
itsrdf:taIdentRef
<http://dbpedia.org/ontology/Film> .

<http://example.org/document293#char=0,66>
 a   nif:RFC5147String, nif:String
        ,nif:Context;
nif:beginIndex "0"^^xsd:nonNegativeInteger;
nif:endIndex "66"^^xsd:nonNegativeInteger;
nif:isString "Wicker Park is a 2004
American psychological drama
    mystery film."^^xsd:string.
```

Liste 1. Wicker Park metninin NIF biçiminde gösterimi

"E* oynar *E" deseni ile aktör olan insan sınıfındaki sinema filminde oynayan varlıklar çıkarılmaktadır. Tanımlı ilişki ile oynar(İnsan, Film) örneklerinin çıkarımı yapılmaktadır. Bu yöntem KnowITALL [12] ve DIPRE [13] çalışmalarından esinlenilerek yapılmaktadır. Ek olarak, bootstrapping ve mesafe öğreticili (distant supervisor) yöntemler ile kurallı ifadeler sayesinde varlık-varlık çiftlerine ait ilişkiler mevcut ontolojide aranarak yapılabilmektedir. Örneğin, "(George Clooney, Suriye)" varlık çiftinin "E* ödülKazandi *E" örüntüsü üzerinden "George Clooney Suriye filminden ödül kazandı." cümlesindeki "ödülKazandı" ilişkisi bulunmaktadır. Bu çalışmada el ile tanımlanmış desenler, eldeki metin üzerinde sorgulanarak uygun varlıklar veya ilişkiler bulunmaya çalışılmıştır. Çalışma alanı olan sinemaya, uygun İngilizce ilişki desenlerinin listesi:

E "is a" *E*

E "film" *E*

E "is directed by" *E*

E "starring at" *E*

Türkçe desenlerin listesi:

E "filmidir."*

E "filmi."*

E "filmdir."*

E "film."*

E "yönetmenliğini" E* "yaptığı"*

E "başrollerini" E* "paylaşmaktadır."*

şeklindedir.

İngilizce verisetinde Liste 1'deki örnekte olduğu üzere "Wicker Park" filmine dair "is a" ve "film" atıfları metin içerisinde bulunup dokümanın ID parametresi, başlangıç ve bitiş indeksleri elde edilerek "Wicker Park" varlığının bir filme karşılık geldiğine dair çıkarım yapılmaktadır. Aynı çıkarım ilişki açısından İngilizcede "is a" ifadesi ile sağlanmaktadır. Bu çıkarımlar yapılırken öncelikle metin içerisinde bulunan "is a" ve "film" gibi ontolojik sınıfı temsil eden ifadelerin başlangıç ve bitiş indeksleri bulunur. Verisetinde bulunan atıfların bu indekslerdeki ilgili varlıklarla arasındaki ilişkilere bakılarak etiketli test metni içerisindeki atıfların hangi tipte ve metnin hangi noktasında olduğu çözümlenir. Örneğin "is a" yapısını kontrol ederken verisetindeki metinler gözönüne alındığında "is a"den önce gelen ifadelerin "Film" tipinde olduğu görülmektedir. Bu doğrultuda, ilgili atıfların, başlangıç ve bitiş indeksleri ile dokümanın ID parametresi kullanılarak dokümana ait atıf yapıları üretilir.

Türkçe verisetinde dilin yapısına uygun olarak özellikle fiillere yapılan ekler gözönüne alınmaktadır. "G.O.R.A., senaryosunu Cem Yılmaz'ın yazdığı, yönetmenliğini Ömer Faruk Sorak'ın yaptığı, bilim kurgu ve komedi türlerindeki

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

2004 çıkışlı Türk filmidir." cümlesini ele alırsak yukarıda verilen Türkçeye dair örüntü listesindeki film ismini bulma veya yönetmenini elde etmeye dair çıkarımlar yapılabilmektedir. Örneğin, etiketli test metinlerinden yola çıkarak metnin ilk cümlesinin sonunda filme dair yukarıda gösterilen ifadelerden biri (filmidir., filmi., filmdir., film. gibi) mevcutsa o cümlenin başındaki varlığa dair bir "Film" atfı indeks değerleri kullanılarak oluşturulur. Ayrıca etiketli metnin ilk cümlesinden sonraki cümlelerde bulunan Türkçe desen ile o desenden önce gelen ilk nokta (önceki cümlenin sonunu ifade etmek için) arasındaki ilk varlık bir "Film" atfı olarak oluşturulur.

III. DEĞERLENDİRME

Öncelikle Türkçe[6] ve İngilizce[7] dillerine ait Vikipedi yığınları etiketli metin oluşturmak için indirilmiştir. Daha sonra etiketli metinlerin elde edilme süreci sinema alanında başlatılarak İngilizce ve Türkçe filmler kategorilerine ait metinler Vikipedi yığınlarından ayrıştırılmıştır. Anlam karmaşıklığı ortamını sunmak için müzik ve yer gibi alanlara ait anlam ayrımı oluşturabilecek alanlar seçilmiştir. Bu alanlar sinema filmleri alanına benzer varlıklar içerdiği için ele alınmıştır. Örneğin *Wicker Park* atıfı Vikipedi anlam ayrımı sayfasında[8] kendisine ait *WickerPark_(film), WickerPark_(soundtrack)* ve *WickerPark_(ChicagoPark)* olmak üzere 3 farklı anlama gelebilmektedir. Bu anlam ayrımı sayfaları sinema, müzik ve lokasyon alanlarının iç içe geçtiğini göstermektedir. Bu anlam ayrımı sayfalarının ana amacı sinema alanındaki aday varlıkların sayısını artırarak varlık çözümleme sürecini zorlaştırmaktır.

Tablo I'de görüldüğü gibi dil, etiketli belge sayısı ve varlık sayıları, sinema alanı için gösterilmektedir. Bilgi tabanı bağımsız Varlık Bağlama sistemlerinde Türkçe desteği olmadığı için İngilizce veri kümesi ele alınıp bu veri kümesinde 945 etiketli doküman bulunmuştur. Vikipedi bilgi kutularından çekilerek kolay olması açısından DBpedia destekli sinema alanına özgü bilgi tabanındaki eşleştirilmiş ilgili varlıklar film isimleri, yönetmenler ve oyuncular ile ilgilidir. Bu varlıklara ait varsa anlam ayrımı sayfalarından üretilmiş etiketli metinle müzik ve lokasyon gibi farklı alanlar da bulunmaktadır. Vikipedi yığınından etiketli metnin oluşturulmasına geçen bütün adımları kapsayan ortalama süreler de saniye cinsinden verilmiştir. Örneğin, metinler için indirilmiş Vikipedi yığınından etiketli metin çıkarımı her iki dilde ortalama her belge için 0.431 saniye sürmektedir.

Anlam karmaşıklığı oranı anlam karmaşıklığı olan her bir varlık için bütün Vikipedi anlam ayrımı sayfalarının bütün belgelere bölünmesi ile elde edilmiştir. Örnekte

[6] https://dumps.wikimedia.org/trwiki/20170420/
[7] https://dumps.wikimedia.org/enwiki/20170420/
[8] https://en.wikipedia.org/wiki/Wicker_Park

Alan	Dil	#Belge	#Varlık	Süre(sn)	Anlam Ka.(%)
Sinema	EN	945	3648	418	28.51
Sinema	TR	824	3182	345	25.86

Tablo I. Değerlendirme kümelerinin özellikleri

görüldüğü gibi *Wicker Park* atıfı sinema alanında *WickerPark_(film)* varlığı ile eşleşirken 2 alternatif aday varlığa sahiptir. Mevcut durumda İngilizce metinlerin anlam karmaşıklık seviyesi yüzde 28.51 olarak belirlenmiştir. Ellis ve arkadaşlarının çalışması [14] yüzde 13 karmaşıklık gösterirken bu çalışma kapsamında elde edilen değer daha yüksektir.

Babelfy [9], GERBIL'de kullanılan seçili varlık bağlama sistemlerinden biridir ve yapısal olmayan Türkçe metinler dahil olmak üzere birçok doğal dili destekler. Babelfy, Wikipedia ve WordNet'i [11] bağlayan çok dilli ansiklopedik bir sözlük olan BabelNet [15] üzerinde oluşturulmuştur. BabelNet'ten yararlanan çizge tabanlı bir anlam karmaşıklığı algoritması kullanır ve verilen atıf için aday varlıklar ile oluşturulmuş en yoğun altçizgeyi bulur. Daha sonra en iyi atıf ve varlık çiftini eşleştirmek için en yoğun altçizgeyi üretir.

DBpedia Spotlight [10], bir diğer İngilizce ve Türkçe metinleri destekleyip iyi bilinen Varlık Bağlama sistemidir ve GERBIL'de mevcuttur. DBpedia Spotlight, bir çokboyutlu kelime uzayının her bir varlık için gösterime sahip olduğu DBpedia varlık oluşlarını içeren bir Vektör Uzay Modeli (Vector Space Model - VSM) kullanır. DBpedia Spotlight'ın anlam karmaşıklığı görevi, Ters Terim Frekansını (Inverse Term Frequency - ITF) Ters Aday Frekansına (Inverse Candidate Frequency - ICT) dönüştürür. Ters Aday Frekansı, terimlerden çok aday varlıklara bağlıdır ve Vektör Uzay Modeli'ndeki kelimelerle ilişkili aday varlıklar ile ters orantılıdır. DBpedia Spotlight ile bundan önce Hollandaca ve İngilizce etiketli metinler üzerinde değerlendirmeler yapılmıştır [16]. Ayrıca, Almanca, Fransızca ve İtalyanca gibi Wikipedia tarafından desteklenen 7 adet dil için de sonuçlar elde edilmiştir.

Üzerinde durulması gereken önemli bir nokta, iyi bilinen Varlık Bağlama sistemlerinin Türkçe metinlerde alan bağımlı herhangi bir değerlendirme yapılmamış olmasıdır. Bildiğimiz kadarıyla WeDGeM, Türkçe değerlendirme verisetini GERBIL'e ekleyen ilk çalışmadır. Ayrıca bu çalışmada, Türkçe ve İngilizce dilleri için spesifik bir çalışma alanında GERBIL'de çalıştırılan DBpedia Spotlight ve Babelfy gibi araçlar değerlendirilmiştir. Ellis ve arkadaşlarının çalışması [14] %13 oranında anlam belirsizliği olduğunu göstermektedir ve aldıkları referans açık alan değerlendirme verisetinde [17] anlam belirsizliğinin %18 olduğu ifade edilmiştir. Çalışmamızda önerilen araç Tablo I'de görüldüğü üzere Türkçe metinler için %25.86 oranında anlam karmaşıklığı üretmektedir.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

EL system	Dataset	Category	Lang.	Mi-F1	Mi-P	Mi-R	Ma-F1	Ma-P	Ma-R
Babelfy	Movie	D2KB	EN	0.923	0.978	0.873	0.886	0.932	0.853
Babelfy	Movie	D2KB	TR	0.945	0.991	0.903	0.879	0.904	0.866
DBpedia Spotlight	Movie	D2KB	EN	0.906	0.972	0.848	0.844	0.931	0.795
DBpedia Spotlight	Movie	D2KB	TR	0.612	0.974	0.446	0.491	0.616	0.436

Tablo II. GERBIL'de bulunan Varlık Bağlama sistemlerinin değerlendirme sonuçları

Cornolti ve arkadaşları [18] genel F1 ölçümlerini makro-(Ma-) ve mikro-(Mi-) ölçümlere genişletmiştir. Ma-ölçümleri, tüm etiketli dokümanlardaki her doküman üzerinde uyumlu ölçümün ortalamasını hesaplarken Mi-ölçümü tüm etiketleri beraberce ele alır. Bu yüzden Mi-ölçümü dokümanların daha fazla etikete sahip olmasına önem verir. Tablo II, Varlık Bağlama görevinin tüm sonuçlarını göstermektedir. Bu sonuçlarda kesinlik (precision), hatırlama (recall) ve F1-skoruna göre üretilmiş değerlendirme kümesi ölçülmektedir. F1-skorlarının gösterdiğine göre, Babelfy hem Türkçe hem de İngilizce verisetleri için sinema çalışma alanında DBpedia'dan daha iyi sonuçlar vermektedir. İngilizce verisetinin Türkçe verisetinden daha yüksek anlam karmaşıklığı olmasına rağmen her iki Varlık Bağlama sisteminin de genel performansı İngilizce değerlendirme verisetlerinde daha iyi sonuçlar üretmektedir.

IV. LİTERATÜR ÖZETİ

Varlık Bağlama sistemlerinin karşılaştırılması için veri kümeleri genel olarak elle yada otomatik olmak üzere iki şekilde elde edilmektedir. Elle elde edilen veri kümeleri elle etiketlenmenin yapılmasından dolayı çok fazla vakit harcanan bir işlem olduğu için genelde küçük boyutlarda metinler içermektedir. Bu çalışmanın ana hedefi alan bağımlı Varlık Bağlama sistemlerinin analizi olarak belirlendiği için değerlendirme veri kümeleri, alana özgü ya da alan bağımsız olarak iki ana kategoride incelenmiştir. Ayrıca oluşturulan veri kümesi üzerinde ilişki çıkarımı ile ilgili çalışmalar da yapıldığı için bu konuyu ele alan veri kümelerinin literatür özeti de eklenmiştir.

A. Alan Bağımsız Veri Kümeleri

Alan bağımsız veri kümelerinden ilki ACE 2004 [19] elle oluşturulmuş bir veri kümesi olarak farklı alanlardan 253 atıfın etiketlendiği sadece 57 haber metnini içermektedir. CONLL [20] yine haber metinlerinden elde edilmiştir. AGDISTIS ve DoSeR çalışmaları bu veri setini deney ortamında kullanmıştır.

Genel bilgi kaynaklarının internet üzerinde gelişmesiyle otomatik olarak veri kümesi üretme işlemine olanak sağlanmıştır. Wikilinks [6] ve Spitkovsky ve Chang [21] çalışmaları Vikipedi üzerinden otomatik veri kümesi oluşturan öncü çalışmalardır. Wikilinks İngilizce dili için otomatik veri kümesi oluşturma metodolojisi gösterirken çok dil desteği sunmamaktadır. Spitkovsky ve Chang [21] çalışmasında çok dil destekli bir yaklaşım ile otomatik etiketli veri kaynaklarını üretebilmektedir. Ancak bu iki çalışma da Vikipedi anlam ayrımı sayfalarını kullanarak anlam karmaşıklığının ayarlanmasını yapmamışlardır.

Li ve arkadaşları [8] diller arası Varlık Bağlama yöntemleri için çok dil destekli dil kaynağı oluşturan bir yöntem sunmuşlardır. Dil kaynağı oluştururken tanımlı varlıkların etiketlenmesinde kalite standardı olan anlam karmaşıklığı ve çeşitliliği kavramlarını göz önünde bulundurmuşlardır. Anlam karmaşıklığı bir atıfın birden fazla aday varlığa etiketlenmesi durumunda ortaya çıkarken çeşitlilik kavramı aynı atıfa ait metinde geçebilecek birden fazla isim çeşitliliğini göstermektedir. Ayrıca Li ve arkadaşları yeniden yönlendirme ve anlam ayrımı sayfalarını kullanması ve anlam karmaşıklığı seviyesini belirlemişlerdir.

Metin Analizi Konferansları (TAC)[9] Bilgi Tabanı Üretimi (KBP) olmak üzere belirli kategorilerde düzenlenmektedir. KBP kategorisinde yapısal olmayan metinlerden bilgi tabanlarının üretilmesi amaçlanmaktadır. Bu kategoride Varlık Keşfi ve Bağlanması (EDL) görevi ile tanımlı varlıkların atıflara bağlanması ve bu varlıkların insan, organizasyon ya da yer gibi tiplerin belirlenmesi ile ilgilenilmektedir. Bu görevin ana amacı, Varlık Bağlama ve etiketleme çalışmalarının yüksek anlam karmaşıklığı içeren alan bağımsız veri kümelerinde ve farklı dillerde karşılaştırılması ile ekipler arasında fikir alışverişi sağlamaktır. Bu sebeple, farklı diller ve farklı alanlar için yeterli seviyede anlam karmaşıklığı içeren veri kümelerinin oluşturulması önem kazanmaktadır.

Son olarak GERBIL aracı [22] bünyesinde, çevrimiçi veya çevrimdışı varlık etiketleme çalışmalarının karşılıklı değerlendirmesini yapabilmek için alan bağımsız Varlık Bağlama ve tanımlı varlık çıkarımı yapan araçlar çalışır halde bulunmaktadır. GERBIL aracının ilham aldığı çalışmada [18] yer alan farklı deney setlerine örnek olarak bilgi tabanına çözümleme (D2KB) ve bilgi tabanına etiketleme (A2KB) gösterilebilir. Bu çalışmada, ilgili iki set ile hem atıfların etiketlenmesi hem de çözümlenmesi aşamalarında ilgilenilmektedir. Bünyesinde entegre halde bulunan Varlık Bağlama sistemlerine ek olarak GERBIL aracında halihazırda alan bağımsız ACE 2004 ve CONLL gibi birçok veri kümesi de bütünleştirilmiştir. Aynı zamanda kullanıcı tarafından üretilmiş veri kümeleri ve varlık etiketleme sistemleri de mevcut araca entegre edilebilmektedir. Ancak, bu araçta alana özgü ve popüler olmayan bir dile özgü açık veri kümeleri bulunmamaktadır. Bu çalışmanın bir diğer amacı da alana

[9] https://tac.nist.gov/2017/

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

ve dile özgü veri kümeleri üreterek alan bağımlı Varlık Bağlama yöntemlerinde kullanılmasını sağlamaktır.

B. Alan Bağımlı Veri Kümeleri

Alan bağımlı veri kümeleri özellikle biyoinformatik alanında çokça rastlanmakla birlikte tamamen varlık etiketlenmesi yerine anlamsal ilintililik yada benzerlik ile ilgili daha özelleştirilmiş durumlar için bulunmaktadır. Pedersen çalışmasında [23] 29 biyoinformatik kavramının elle yapılmış anlamsal ilintililik için özelleştirilmiş bir veri kümesi bulunmaktadır. Buna ek olarak UMLS [24] veri kümesinde 566 tıp terimine ait yine elle hesaplanmış anlamsal benzerlik değerleri bulunmaktadır.

Örnekleri az olsa da bilgi teknolojileri alanına özgü Bitter Corpus [25] veri kümesi Linux işletim sistemleri için hazırlanmış kullanım kılavuzlarında etiketlenmiş 628 İtalyanca ve İngilizce terimleri içerirken 637 tane de iki dille ilgili alana özgü kavramları barındırmaktadır. Biyoinformatik çalışmaları dışında diğer alanlara özelleştirilmiş açık kaynak halinde yayınlanan ve doğrudan varlık etiketlenmesi için sunulmuş veri kümeleri bizim bildiğimiz kadarıyla bulunmamaktadır. Bununla birlikte diller arasında paralel [26] yada iki dil içinde aynı şekilde barındırdığı etiketleme veri kümeleri [27] ve tamamen doğal dil işleme alanına özel biçimsel yapıları içeren [28] veri kümeleri haricinde literatürde Türkçe diline özgü açık veri kümesi bildiğimiz kadarıyla bulunmamaktadır.

C. İlişki Çıkarımı Yapılan Veri Kümeleri

Açık Amerikan Ulusal Derlemi (Open American National Corpus - OANC) [29], 1990'dan beri üretilen konuşma ve yazma dilinde bulunan 22 milyon kelimeyi içeren bir Amerikan İngilizcesi metin derlemidir. Bu derlem, e-posta, tweet ve web verisi içermektedir. Bunun yanında MASC, dil içerisindeki kelimelerden yararlanılarak otomatik üretilen etiketlerin yanısıra 19 farklı Amerikan İngilizcesi içerisinde bulunan elle üretilmiş ve doğrulanmış etiketler içermektedir. MASC-NEWS [30], adlı varlıklar ve kelime anlamlarından yola çıkarak MASC derlemini otomatik olarak etiketleyen bir çalışmadır.

Zeka için Gelişmiş Soru Cevaplama (Advanced Question Answering for Intelligence - AQUAINT) [31], farklı dil, tür ve formatlardaki yüksek miktarda verinin içerisinden benzer konularda, anlamsal ilişkili, güncel bilgiyi çıkarmak için geliştirilmiştir. Bu yaklaşımda kullanıcıların sorularına cevaplar sunulmaktadır. Sisteme bir metin ve metin hakkında bir soru verilmektedir. Metne verilen cevap kesin (Metnin doğru olduğundan yola çıkarak sorunun cevabı da doğrudur) veya akla yakındır (Metnin doğruluğundan yola çıkarak sorunun cevabı doğru olabilmesine rağmen ek bilgiler cevapta değişikliklere neden olabilir). Bu sistemde etiketler, metnin kaynağı ve polaritesine bakarak oluşturulur. Kaynak, sorunun veya metnin içinde olmayan bir bilgiye ihtiyaç olmadan cevaplama yapılıp yapılamayacağını belirler. Polarite ise cevabın doğru, yanlış veya bilinmeyen şeklinde kategorize edilmesini sağlar.

V. SONUÇ

Bu çalışmada sunulan WeDGeM aracı basit ve hızlı bir şekilde Vikipedi ve DBpedia destekli herhangi bir doğal dilde ve alanda değerlendirme veri kümesinin oluşturulmasını amaçlamaktadır. Sinema çalışma alanı için bir kullanım senaryosu, Türkçe ve İngilizce dillerinde üretilen etiketlenmiş metinler ile incelenmiştir. Ayrıca üretilen veriseti, Varlık Bağlama sistemlerinin performanslarını değerlendirmek için GERBIL'e entegre edilmiştir. Babelfy ve DBpedia Spotlight da bahsedilen iki dil için GERBIL üzerinde değerlendirme yapmak amacıyla ele alınmıştır.

Gelecek çalışmalarda WeDGeM, web üzerinde çalışan bir araç olarak implemente edilecektir. Kullanıcılar, herhangi Wikipedia kategori ve dilini belirterek varlık etiketleme sistemleri için kıyaslama (benchmark) verisetlerini yaratabilecektir. Ayrıca, Varlık Bağlama sistemleri arasında daha gelişmiş karşılaştırmalar yapmak için anlam karmaşıklığı ve farklılığını içeren kalite standartları geliştirilecektir. Varlık Bağlama görevine ek olarak, kullanıcılar diğer alan bağımlı değerlendirme verisetlerini WeDGeM üzerinde üretebilecektir. Bu verisetleri, farklı türlerdeki konu tespiti ve ilişki çıkarımı gibi doğal dil işleme görevlerini kapsayacaktır.

TEŞEKKÜR

Yrd. Doç. Dr. Fatih Tekbacak bildirinin yazımı esnasında Alanya Alaaddin Keykubat Üniversitesi (ALKÜ)'nde görev yapmakta olup ALKÜ'ye bu süreçte verdiği destekten dolayı teşekkür ederiz.

KAYNAKÇA

[1] O. Hassanzadeh and M. P. Consens, "Linked Movie Data Base," in *LDOW*, 2009.

[2] P. Ernst, A. Siu, and G. Weikum, "Knowlife: A Versatile Approach for Constructing A Large Knowledge Graph for Biomedical Sciences," *BMC Bioinformatics*, vol. 16, no. 1, 157, 2015.

[3] D. Dou, H. Wang, and H. Liu, "Semantic Data Mining: A Survey of Ontology-Based Approaches," *in Semantic Computing (ICSC)*, 2015 IEEE International Conference on. IEEE, 244–251, 2015.

[4] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data," 2007.

[5] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective Annotation of Wikipedia Entities in Web Text," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 457–466, 2009.

[6] S. Singh, A. Subramanya, F. Pereira, and A. McCallum, "Wikilinks:

A Large-Scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia," *University of Massachusetts, Amherst, Tech. Rep.* UM-CS-2012-015, 2012.

[7] R. Navigli, "Babelnet and Friends: A Manifesto for Multilingual Semantic Processing,", *Intelligenza Artificiale*, vol. 7, no. 2, 165–181, 2013.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

[8] S. Strassel, M. A. Przybocki, K. Peterson, Z. Song, and K. Maeda, "Linguistic Resources and Evaluation Techniques for Evaluation of Crossdocument Automatic Content Extraction." in *LREC*, 2008.

[9] A. Moro, F. Cecconi, and R. Navigli, "Multilingual Word Sense Disambiguation and Entity Linking for Everybody," *in Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ser. ISWC-PD'14. Aachen, Germany, Germany: CEUR-WS.org, 25–28, 2014.

[10] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia Spotlight: Shedding Light on the Web of Documents," in *Proceedings of the 7th International Conference on Semantic Systems*, ser. I-Semantics'11. New York, NY, USA: ACM, 1–8, 2011.

[11] A. Kilgarriff and C. Fellbaum, "Wordnet: An Electronic Lexical Database," 2000.

[12] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Webscale Information Extraction in KnowItAll: (preliminary results)," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 100–110, 2004.

[13] S. Brin, "Extracting Patterns and Relations from the World Wide Web," in *Selected Papers from the International Workshop on The World Wide Web and Databases*, ser. WebDB '98. London, UK, UK: Springer-Verlag, 172–183, 1999.

[14] J. Ellis, J. Getman, J. Mott, X. Li, K. Griffitt, S. Strassel, and J. Wright, "Linguistic Resources for 2013 Knowledge Base Population Evaluations," *in Proceedings of the Sixth Text Analysis Conference, TAC 2013*, Gaithersburg, Maryland, USA, November 18-19, 2013.

[15] R. Navigli and S. P. Ponzetto, "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network," *Artificial Intelligence*, vol. 193, 217–250, 2012.

[16] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving Efficiency and Accuracy in Multilingual Entity Extraction," in *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[17] X. Li, S. Strassel, H. Ji, K. Griffitt, and J. Ellis, "Linguistic Resources for Entity Linking Evaluation: From Monolingual to Crosslingual," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey, May 23-25, 3098–3105, 2012.

[18] M. Cornolti, P. Ferragina, and M. Ciaramita, "A Framework for Benchmarking Entity-Annotation Systems," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, 249–260.

[19] A. Mitchell, S. Strassel, S. Huang, and R. Zakhary, "Ace 2004 Multilingual Training Corpus," *Linguistic Data Consortium*, Philadelphia, vol. 1, 1–1, 2005.

[20] E. F. Tjong Kim Sang, F. De Meulder, "Introduction to the Conll-2003 Shared Task: Language-Independent Named Entity Recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Volume 4, Association for Computational Linguistics, 142-147, 2003.

[21] V. I. Spitkovsky, A. X. Chang, "A Cross-Lingual Dictionary for English Wikipedia Concepts." in *LREC*, 3168–3175, 2012.

[22] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, L. Wesemann, "GERBIL – General Entity Annotation Benchmark Framework," in *24th WWW Conference*, 2015.

[23] T. Pedersen, S. V. Pakhomov, S. Patwardhan, C. G. Chute, "Measures of Semantic Similarity and Relatedness in the Biomedical Domain," *Journal of Biomedical Informatics*, vol. 40, no. 3, 288–299, 2007.

[24] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G. B. Melton, "Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study," in *AMIA Annual Symposium Proceedings*, vol. 2010. American Medical Informatics Association, 572, 2010.

[25] M. Arcan, M. Turchi, S. Tonelli, P. Buitelaar, "Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment," in *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, 54–68, 2014.

[26] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages," *arXiv preprint cs/0609058*, 2006.

[27] T. Pamay, U. Sulubacak, D. Torunoglu-Selamet, G. Eryigit, "The Annotation Process of the ITU Web Treebank," in *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, 95, 2015.

[28] H. Sak, T. Güngör, M. Saraçlar, "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus," in *GoTAL 2008*, ser. LNCS, vol. 5221. Springer, 417-427, 2008.

[29] N. Ide, K. Suderman, "The American National Corpus First Release," in *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, 1681–1684, 2004.

[30] A. Moro, R. Navigli, F. M. Tucci, R. J. Passonneau, "Annotating the MASC Corpus with Babelnet," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4214-4219, May 2014.

[31] D. Crouch, S. Roser, F. Abraham, "AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines," May 2005.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.