Geniş Metin Koleksiyonlarından Yinelemeli Bilgi Çıkarımı

Iterative Information Extraction from Large Text Collections

Gürkan Şahin[1], M. Fatih Amasyalı[1]
[1]Elektrik Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
gurkan.sahin@hotmail.com , mfatih@ce.yildiz.edu.tr,

***Özetçe*— Geniş metinlerden bilgi çıkarımı konusunda çeşitli yöntemler bulunmaktadır. Bunlardan bir tanesi de şablonlar yöntemidir. Bu çalışmada şablonlar yöntemini kullanarak aralarında belli anlamsal ilişki bulunan ikililerin elde edilmesini sağlayan otomatik bir sistem geliştirilmiştir. Çalışma kapsamında morfolojik olarak çözümlenmiş ve çözümlenmemiş veri setleri üzerinde ayrı ayrı çalışılmıştır. Morfolojik olarak çözümlenmiş veri setinden daha iyi yapıda şablonlar elde edilmiştir. Yapılan denemeler sonucunda sürekli artan sayıda şablon kullanıldığı taktirde üretilen ikililerin doğruluklarının azaldığı görülmüştür. Sabit sayıda daha güvenilir şablonlardan büyüyen veri seti üzerinde daha iyi sonuçlar elde edilmiştir.**

Anahtar Kelimeler*— *Doğal Dil İşleme, Bilgi Çıkarımı, Şablonlar Yöntemi, Morfolojik Analiz, Anlamsal İlişki

***Abstract*— There are various methods about information extraction from large texts. One of them is method of templates. We developed an automatic system that aims to produce pairs which have semantic relation between them using templates. We worked with morphological resolved and unresolved datasets. We obtained better templates from morphological resolved dataset. In our experiments, we observed that if too many templates were used for producing pairs, accuracy of produced pairs decreased. Also, we obtain better results for fixed and more reliable templates with using growing datasets.**

Keywords*— *Natural Language Processing, Information Extraction, Templates Method, Morphological Analysis, Semantic Relation

1. GİRİŞ

Bilgi çıkarımı, bir bilgi kaynağı içerisinden çeşitli doğal dil işleme algoritmaları kullanılarak istenilen türdeki bilgilerin çıkarılması işlemidir. Bilgi kaynağı olarak hazır metin kütüphaneleri kullanılabileceği gibi web sayfalarının içerikleri de kullanılabilmektedir. Buradaki amaç sisteme dışarıdan insan müdahalesini en aza indirerek yüksek başarımlı bilgi çıkarabilen bir sistem gerçekleştirmektir.

WordNet [1] aralarında çeşitli türden ilişkilerin bulunduğu kelimelerden oluşan bir yapıdır. Günümüzde doğal dil işleme projelerinde WordNet'in kullanımı büyük bir önem teşkil etmektedir. Bu nedenle WordNet'in otomatik olarak oluşturulması fikri doğmuştur. Literatürde WordNet' in oluşturulmasıyla ilgili çeşitli çalışmalar ve yöntemler bulunmaktadır. Türkçe WordNet' in otomatik olarak oluşturulması amacıvla Balkanet projesi başlatılmış, bu proje sonucunda 11.628 eş küme ve bunlar arası 17.550 ilişki içeren bir anlamsal veritabanı oluşturulmuştur [2].

Geniş metinlerden bilgi çıkarımı konusunda çeşitli yöntemler bulunmaktadır. Şablonlar yöntemi ve öğelerine ayrılmış metinlerin kullanılması bu yöntemlere örnek olarak verilebilir [3].

Bu çalışmada aralarında çeşitli türden anlamsal ilişkilerin bulunduğu ikililerin geniş metin koleksiyonları içerisinden otomatik olarak çıkarılmasını sağlayan bir sistem geliştirilmiştir. Şablon bilgisi kullanılarak geniş metin koleksiyonları içerisinden yinelemeli bilgi çıkarımı yapılması amaçlanmıştır. 2. bölümde yöntemin ayrıntıları verilmiş ve sınırları incelenmiştir. 3. bölümde literatürdeki benzer çalışmalar hakkında bilgi verilmiş, 4. bölümde kullanılan veri kümeleri anlatılmıştır. 5. bölümde deneysel sonuçlar verilmiş, son bölümde ise elde edilen sonuçlar yorumlanmıştır.

2. ŞABLONLAR YÖNTEMİ

Bu çalışmada "şablon", aralarında anlamsal bir ilişki olan iki kelime arasında kalan, kelime ya da kelime grupları anlamında kullanılmıştır. Örneğin; elma-meyve kelimelerini üst sınıf ilişkisine (Is-A) sahip bir ikili olarak ele alalım. Burada elma varlığı, meyve ise varlığa ait üst sınıfı temsil etmektedir. Bu ikili arasında geçebilecek kelimeler ise bu ilişkiye ait şablondur. Örneğin; "… elma ve benzeri meyveler …" cümlesindeki elma ve meyve kelimeleri arasında kalan 've benzeri' kelime grubu üst sınıf ilişkisine sahip bir şablondur. Yaygın olarak kullanılan bir başka anlamsal ilişki ise aynı üst sınıfa ait ikililerden oluşan ilişki (kardeş ilişki) türüdür. Örneğin; köpek-kedi ikilisinde hem köpek hem de kedi varlıkları hayvan üst sınıfına aittir. "… köpek ya da kedi …" cümlesindeki köpek ve kedi kelimeleri arasında kalan 'ya da' kelime grubu kardeş ilişkiye ait şablondur.

Çalışma kapsamında yaygın olarak kullanılan, kolaylıkla ayırt edilebilecek şablonların üretilebileceği anlamsal ilişkiler kullanılmıştır. Bunlardan bir tanesi üst sınıf ilişkisi diğeri ise aynı üst sınıfa ait olan ikililerin oluşturduğu ilişkidir.

İlk önce her iki ilişki türü için doğruluğundan emin olduğumuz ikililer (pozitif ikililer) belirlenmiştir. Bu işlem bir kez yapılmıştır. Daha sonra bu ilişki türüne sahip olmayan ikililer (negatif ikililer) belirlenmiştir. Pozitif ve negatif ikilileri belirlememizdeki amaç pozitif ikililer için bulunan şablonların negatif ikililer için bulunan şablonlardan farklı olmasının sağlanmasıdır. Yani pozitif ikililer için bulunan bir şablon negatif ikililer için de bulunmuşsa aslında bu şablonun hiçbir önemi yoktur ve yeni ikililerin üretilmesinde kullanılmamalıdır.

Üzerinde çalıştığımız veri seti çok büyük olduğundan özellikle arama işlemlerinin hızlı bir şekilde yapılabilmesi için Apache Lucene [4, 5] kütüphanesi kullanılmıştır. Hazırladığımız pozitif ve negatif ikililer, üzerinde çalışacağımız veri setimizde Lucene ile aratılarak bu ikililerin birlikte geçtikleri cümleler bulunmuştur. Daha sonra bu cümleler içinde ikililerin arasında kalan kelime ya da kelime grupları (en fazla 2 kelime) bulunarak şablon olarak kaydedilmiştir. Pozitif ve negatif ikililere ait şablonlar bulunmuş ve ortak bulunan şablonlar pozitif şablonlardan çıkartılmıştır. Böylece sadece pozitif ikililere ait olan şablonlar elde edilmiştir.

Şablonlar bulunduktan sonra bu şablonlara ait ikili frekans değerleri hesaplanmıştır. Buradaki ikili frekans değeri ilgili şablonun başlangıç ikililerinden kaç tanesi için bulunduğunu gösteren bir sayı değeridir. Bir şablonun ikili frekans değeri ne kadar yüksek ise o şablonun daha fazla ikili için bulunduğu anlaşılmaktadır. Başka bir deyişle bir şablonun ikili frekans değeri ne kadar yüksek ise o şablon o kadar güvenilirdir denebilir.

3. Benzer Çalışmalar

Literatürde şablon bilgisini kullanarak bilgi çıkarımının yapıldığı bir çok çalışma bulunmaktadır. Marti A. Hearst 1998 yılında yaptığı çalışmasında WordNet ilişkilerinin belirlenmesinde şablon bilgisini kullanmıştır [6]. Benzer şekilde 1998 yılında Brin, şablonlar yöntemini kullanarak kitap-yazar ikililerini bulan bir çalışma yapmıştır [7]. Güncel ve en popüler çalışmaya örnek olarak ise NELL [8] verilebilir. NELL Ocak 2010 yılında geliştirilen ve web sayfaları içerisinden yinelemeli bilgi çıkarımı yapmayı hedefleyen bir sistemdir. NELL'in geliştirilmesinin amacı web sayfaları üzerinde sürekli çalışan, yeni bilgiler öğrenen ve geçmişte öğrendiği bilgileri gelecekte yeni bilgiler üretmede kullanan bir sistem gerçekleştirmektir [9]. Bu sistemin en belirgin özelliği bilgi çıkarımında şablon bilgilerini kullanması ve çok büyük miktardaki bir veri kümesi üzerinde çalışmasıdır. Bu sistemde ilk olarak doğruluğundan emin olunan ikililerden şablonlar üretilmektedir. Daha sonra ise üretilen bu şablonlar kullanılarak aynı anlamsal ilişkiyi sağlayan yeni ikililer üretilmektedir. Üretilen bu yeni ikililer tekrar sisteme giriş bilgileri olarak verilmekte, ikililerden şablonlar ve şablonlardan da yeni ikililer üretmektedir. Sonuçta bu sistem sayesinde web sayfaları üzerinden yinelemeli bir şekilde bilgi çıkarımı yapılmaktadır.

4. Kullanılan Veri Kümeleri

Bilindiği üzere bilgi çıkarımı işlemi için büyük çapta veri setine ihtiyaç vardır. Gerçekleştirdiğimiz sistemde kendi oluşturduğumuz metin kütüphaneleri kullanılmıştır. Bu kısımda bilgi çıkarımında kullandığımız veri setinin nasıl elde edildiği üzerinde durulmuştur.

Veri kütüphanelerimizin oluşturulması için farklı birçok kaynak kullanılmıştır. Elimizde bulunan bazı hazır metin dosyalarının yanı sıra bilim teknik dergisinin 45 yıllık arşivindeki pdf dosyalarından ve bilim teknik çocuk dergisinin pdf dosyalarından da çeşitli metinler elde edilmiştir. Pdf dosyalarından metin bilgilerini elde etmek için Apache Tika [10] kütüphanesi kullanılmıştır. Daha sonra elde edilen bu düzensiz yapıdaki metinler '.', '!', '?' gibi noktalama karakterlerinden bölünerek cümleler elde edilmiştir. Elde edilen bu cümleler her satıra bir cümle gelecek şekilde veri kütüphanelerimize kaydedilmiştir.

Sonuç olarak çeşitli kaynaklardan elde edilmiş yaklaşık 3.5 milyon adet cümleden oluşan 4 adet veri seti elde edilmiştir. En büyük çaptaki 'A' veri setimiz çoğunluğu haber metinleri [11] olan cümleleri içermektedir. Bu veri setimiz iyi yapıda bulunan cümlelerden oluşmaktadır. 'B' veri setimiz bilim teknik dergisinin 45 yıllık arşivinden oluşan pdf dosyalarından elde ettiğimiz cümlelerden oluşmaktadır. 'C' veri setimiz hikaye ve roman gibi kitap içeriklerinden oluşan ve genel olarak iyi yapıdaki cümlelerden oluşmaktadır. 'D' veri setimiz ise bilim teknik çocuk dergisi pdf dosyalarından elde ettiğimiz cümlelerden oluşmaktadır. 'B' ve 'D' veri setlerimiz pdf dosyalarından elde edildikleri için diğer veri setlerine göre biraz daha kötü yapıdadır ve yanlış, hatalı kelimeler içermektedir. Bu veri setleri içerisinde bulunan cümle sayıları bilgileri Tablo 1'de verilmiştir.

TABLO 1. VERİ SETİ İÇERİKLERİ VE CÜMLE SAYILARI

Veri Seti ID	Veri seti adı	İçeriği	İçerdiği cümle sayısı
A	Sentences.txt	Haber metinleri	2.090.162
B	Bilim teknik dergisi.txt	Bilimsel metinler	701.523
C	Kitap.txt	Hikaye, roman vb. kitap metinleri	518.414
D	Bilim teknik çocuk dergisi.txt	Bilimsel metinler	181.285

Üst sınıf ilişkisi için şablonlar üretilmesi amacıyla kullanılan 100 adet pozitif başlangıç ikililerinden bir kısmı Tablo 2'de verilmiştir. Üst sınıf ilişkisi için şablonlar üretilirken negatif başlangıç ikilileri olarak Tablo 3'deki kardeş ilişkisine ait ikililer kullanılmıştır.

TABLO 2. ÜST SINIF (IS-A) İLİŞKİSİ İÇİN ŞABLON ÜRETMEDE KULLANILAN POZİTİF BAŞLANGIÇ İKİLİLERİNDEN ÖRNEKLER

dolar para	türkiye devlet	boya malzeme	güneş yıldız
amerika batı	protein bileşen	ekmek besin	kanser hastalık
iran ülke	amca akraba	bakan siyasi	çatlak hasar
esnaf meslek	bütçe kaynak	kapkaç suç	salon yer
taş madde	tüfek silah	incir meyve	bez malzeme
kan doku	ehliyet belge	antibiyotik ilaç	ilaç ürün

Kardeş ilişkisi için şablonlar üretilmesi amacıyla kullanılan 100 adet pozitif başlangıç ikililerinden bir kısmı Tablo 3'de verilmiştir. Kardeş ilişkisi için şablonlar üretilirken, negatif başlangıç ikilileri olarak Tablo 2'deki üst sınıf ilişkisine ait ikililer kullanılmıştır.

TABLO 3. KARDEŞ İLİŞKİSİ İÇİN ŞABLON ÜRETMEDE KULLANILAN POZİTİF BAŞLANGIÇ İKİLİLERİNDEN ÖRNEKLER

mavi siyah	yaprak meyve	kart nakit	eksik yanlış
kandil meşale	otel pansiyon	metal plastik	emekli ssk
basit karmaşık	kestane erguvan	meyve et	anne baba
sıvı gaz	bitki hayvan	sağ sol	çocuk arkadaş
kovuk mağara	tuz şeker	bitki hayvan	kardeş arkadaş
büyü küçül	ses video	kitap kalem	amir patron

Tablo 2 ve 3'teki ikililerin tamamı Ek 1 ve Ek 2'de verilmiştir.

5. DENEYSEL SONUÇLAR

Bu kısımda morfolojik olarak çözümlenmiş ve çözümlenmemiş olan veri setleri üzerinde ayrı ayrı çalışılmış ve sonuçlar gözlemlenmiştir.

5.1. MORFOLOJİK OLARAK ÇÖZÜMLENMEMİŞ VERİ SETİ İLE ÇALIŞMA

Bu bölümde morfolojik [12] olarak çözümlenmemiş (eklerine ayrışmamış kelimelerden oluşan) olan veri setimizden şablonların üretilmesi ve üretilen şablonlardan yeni ikililerin bulunması anlatılmış, çeşitli değerlendirmeler yapılmıştır.

Eklerine ayrışmamış kelimelerden oluşan veri seti üzerinde bilgi çıkarımı yapmanın bazı olumsuz yönleri bulunmaktadır. Veri setimiz içinde aradığımız bir kelime kök haliyle bulunabilirken ekleriyle birlikte alındığında bulunamayabilmektedir. Örneğin; "… elmaların ve benzeri meyvelerin …" cümlesinin veri setimizde olduğunu düşünecek olursak ve başlangıç ikililerimizden bir tanesi olarak elma-meyve ikilisini sisteme verirsek elma ve meyve kelimeleri kök halleriyle bu cümlede bulunamayacağı için 've benzeri' şablonu da üretilemeyecektir. Oysa elma-elmaların ya da meyve-meyvelerin kelimeleri aynı nesneyi ifade etmektedir.

Bu kapsamda öncelikle pozitif başlangıç ikililerinden şablonlar elde edilmiş sonra bu şablonlardan yeni ikililer üretilmiştir. Bunun için elde edilen şablonlar morfolojik olarak çözümlenmemiş veri setimiz içerisinde aratılarak geçtiği cümleler belirlenmiştir. Daha sonra şablonların geçtiği bu cümlelerde şablonun sağında ve solunda kalan kelimeler ilgili ilişkiyi sağladıkları varsayılarak yeni ikili olarak kaydedilmiştir. Sonuçta pozitif ikililerden elde edilen şablonlar kullanılarak aynı ilişki türünü sağladığı düşünülen birçok yeni ikili üretilmiştir. Daha sonra üretilen bu yeni ikililerin şablon frekans değerleri hesaplanılmıştır.

Şablon frekans değeri, ilgili ikilinin üretilen şablonlardan kaç tanesinde geçtiğini gösteren bir sayı değeridir. Bir ikilinin şablon frekans değeri ne kadar yüksek ise o ikilinin daha fazla şablon için bulunduğu anlaşılmaktadır. Başka bir deyişle bir ikilinin şablon frekans değeri ne kadar yüksek ise o ikili o kadar güvenilirdir denebilir.

Yeni ikililer üretildikten sonar, bu ikililer de sisteme pozitif ikililer olarak negatif ikililerle birlikte verilmiş, bunlardan yeni şablonlar üretilmiş, üretilen şablonlardan gerekli elemeler yapılmış ve kalan şablonlardan yeni ikililer bulunmuştur. Böylece yinelemeli bir şekilde her iki ilişki türüne ait anlamsal ikililer otomatik olarak üretilmiş ve sonuçlar gözlemlenmiştir. Bu işlem yapılırken veri setimizin tümü (A+B+C+D) kullanılmıştır.

Üst sınıf ilişkisi için başlangıç ikililerinden üretilen şablonlar, bu şablonların ikili frekans değerleri ve bu şablonların başlangıç ikililerinden hangileri için bulunduğu bilgileri Tablo 4'de verilmiştir.

TABLO 4. ÜST SINIF (IS-A) İLİŞKİSİ İÇİN BULUNAN ŞABLONLAR

Şablon	**Şablonun ikili frekansı**	**Şablonun geçtiği başlangıç ikilileri**
ve diğer	14	[esnaf-meslek], [kan-doku], [pasaport-kimlik], [işkence-kötü], [gayrimenkul-mal], [silah-malzeme], [benzin-enerji], [baklava-tatlı], …
ve benzeri	11	[tatlı-gıda],[klor-kimyasal], [otomobil-araç], [karides-kabuklu], [faks-iletişim], [süt-ürün], [portakal-narenciye], [kiraz-meyve], …
gibi bir	2	[ehliyet-belge], [çöl-yer]
benzeri	2	[klor-kimyasal], [güneş-yıldız]
ve çeşitli	1	[polis-güvenlik]
gibi doğal	1	[sel-afet]
dışındaki	1	[silah-malzeme]
gibi çeşitli	1	[klor-kimyasal]
ve değerli	1	[çanta-eşya]
veya diğer	1	[kuş-hayvan]
gibi çok	1	[kaya-sert]
dışında diğer	1	[istanbul-il]

Tablo 4'te verilen şablonların kullanılmasıyla üretilen yeni ikililerden bazıları, bu ikililerin şablon frekansı değerleri ve hangi şablonlar için bulunduğu bilgileri Tablo 5'de verilmiştir. Bilindiği gibi morfolojik analizi yapılmamış veri seti üzerinde çalıştığımız için üretilen ikililer ekli halleriyle elde edilmiştir. Bu ikililerin ekleri parantez içinde belirtilmiştir. "*" sembolü bu ikilinin sisteme başlangıç olarak verdiğimiz ikililerden bir tanesi olduğunu, başında sembol bulunmayan ikililer ise

başlangıç ikililerinden farklı olarak üretilen ikilileri temsil etmektedir.

TABLO 5. ÜST SINIF (IS-A) İLİŞKİSİ İÇİN ÜRETİLEN İKİLİLER

Üretilen ikili	İkilinin şablon frekansı	İkilinin geçtiği şablonlar
*dolar-para	2	[ve diğer], [gibi bir]
çay-sıcak	2	[ve diğer], [ve benzeri]
tavuk-kanatlı	2	[ve diğer], [ve benzeri]
*pasaport-kimlik	2	[ve diğer], [gibi bir]
türkiye-ülke(lere)	1	[ve diğer], [ve benzeri]
güve(ler)-böcek(leri)	1	[ve benzeri]
vatoz(u)-balık (ları)	1	[ve diğer]
kükürt-kokulu	1	[ve diğer]
nane-baharat(lar)	1	[ve diğer]
kurtçuk-omurgasız(larla)	1	[ve diğer]
böcek(lerle)-küçük	1	[ve diğer]
tüsiad-kuruluş(lar)	1	[ve diğer]
tıp-alan(lar)	1	[ve diğer]
bakan(ı)-yetkili(lerle)	1	[ve diğer]
abla(sı)-akraba(lar)	1	[ve diğer]
dede(m)-akraba(larımız)	1	[ve diğer]
kitap(lara)-belge(lere)	1	[ve diğer]
kılıç(lar)-silah(lar)	1	[ve diğer]
paşa-yetkili(lere)	1	[ve diğer]
ayasofya-cami(ler)	1	[ve diğer]
üniversite(lere)-kurum(lara)	1	[ve diğer]

Kardeş ikililer için başlangıç ikililerinden üretilen şablonlar, bu şablonların ikili frekans değerleri ve bu şablonların başlangıç ikililerinden hangileri için bulunduğu bilgileri Tablo 6'da verilmiştir.

TABLO 6. KARDEŞ İLİŞKİSİ İÇİN BULUNAN ŞABLONLAR

Şablon	Şablonun ikili frekansı	Şablonun geçtiği başlangıç ikilileri
ya da	43	[basit-karmaşık], [sıvı-gaz], [balık-balina], [kız-erkek], …
hem de	8	[kız-erkek], [petrol-gaz], [özel-kamu], [sağ-sol], …
ne de	6	[kuş-böcek], [kadın-erkek], [eksik-yanlış], [iyi-kötü], …
değil	5	[iki-üç], [su-hava], [ekmek-meyve], [türk-ermeni], [sağ-sol]
ve bir	3	[kız-erkek], [bilgisayar-telefon], [kadın-erkek]
ile bir	2	[kız-erkek], [kadın-erkek]
yerine	2	[iki-üç], [kadın-erkek]
bir de	2	[kadın-erkek], [iyi-kötü]

Tablo 6'da verilen şablonların kullanılmasıyla üretilen yeni ikililerden bazıları, bu ikililerin şablon frekansı değerleri ve hangi şablonlar için bulunduğu bilgileri Tablo 7'de verilmiştir.

TABLO 7. KARDEŞ ILIŞKISI IÇIN ÜRETILEN IKILI BILGILERI

Üretilen ikili	İkilinin şablon frekansı	İkilinin geçtiği şablonlar
*kadın-erkek	6	[ya da], [hem de], [ne de], [ve bir], [ile bir], [bir de]
büyük-küçük	5	[ya da], [hem de], [ne de], [ve bir], [bir de]
yerli-yabancı	4	[ya da], [hem de], [ne de], [bir de]
gece-gündüz	4	[ya da], [hem de], [ne de], [ve bir]
sözlü-yazılı	3	[ya da], [hem de], [ne de]
kadın-çocuk	3	[ya da], [ve bir], [ile bir]
uzun-kısa	3	[ya da], [hem de], [ne de]
iktidar-muhalefet	3	[ya da], [hem de], [ne de]
ulusal-uluslararası	3	[ya da], [hem de], [ne de]
sivil-askeri	3	[ya da], [hem de], [ve bir]
sıcak-soğuk	3	[ya da], [hem de], [ne de]
*sünni-şii	3	[ya da], [hem de], [ne de]
erkek-dişi	3	[ya da], [hem de], [ve bir]
siyasi-ekonomik	3	[ya da], [hem de], [ne de]
dost-düşman	3	[ya da], [hem de], [ne de]
bireysel-kurumsal	3	[ya da], [hem de], [bir de]
türk-yunan	3	[ya da], [hem de], [ne de]
beraberlik-galibiyet	3	[ya da], [ve bir], [bir de]
bakan-başbakan	3	[ya da], [ve bir], [bir de]
altın-gümüş	2	[ya da], [ve bir]
sürücü-yolcu	2	[ya da], [ve bir]

Yapılan bu işlemler sonucunda üretilen ikililerin ekli yapıda olduğu görülmüştür. Bu durum yeni şablonlar ve ikililer üretmede verimli bir yöntem olmadığından veri setindeki kelimelerin eklerine ayrıştırılarak üzerinde işlem yapılması ve elde edilen sonuçların birbirleriyle karşılaştırılarak bir sonuç çıkartılması amaçlanmıştır.

5.2. MORFOLOJİK OLARAK ÇÖZÜMLENMİŞ VERİ SETİ İLE ÇALIŞMA

Veri setimiz içerisindeki kelimelerin morfolojik olarak çözümlenmesi için Zemberek 2 kütüphanesi kullanılmıştır. Zemberek, açık kaynak kodlu Türkçe Doğal dil işleme kütüphanesidir [13]. Tamamen Java ile geliştirilen kütüphane, yazım denetimi, hatalı kelimeler için öneri, heceleme, çözümleme gibi çeşitli işlevlere sahiptir. Zemberek' in yapabildiği işlemleri gösteren demo uygulamaya http://zemberek-web.appspot.com/ adresinden erişilebilir. Proje kapsamında kelimelerin morfolojik olarak çözümlenmesi için

Zemberek 2' nin "Çözümle" fonksiyonu kullanılarak kelimeler eklerine ayrıştırılmıştır.

Zemberek kelime çözümleme işlemini gerçekleştirirken birden fazla sonuç döndürebilmektedir. Bu sonuçlar genel olarak kelimenin kullanımı itibariyle doğruluğu yüksek olan çözümlemeden doğruluğu düşük olan çözümlemeye doğru sıralanmaktadır. Bu kapsamda bir kelime için birden fazla çözümleme olması durumunda bunlardan ilk olan çözümleme kullanılmıştır.

Zemberek çözümleme işlemini gerçekleştirirken kelimenin bütün eklerini (yapım ekleri ve çekim ekleri) ayrıştırmaktadır. Bilindiği üzere çekim ekleri eklendikleri kelimelerin anlamlarında bir değişiklik meydana getirmezken yapım ekleri eklendikleri kelimelerin anlamlarında ve hatta türlerinde (isim, sıfat, fiil vb…) bile değişiklikler meydana getirebilmektedir. Bu nedenle yapım ekleri büyük önem arz etmektedir ve kelimenin kökünden ayrıştırılmaması gerekmektedir. Çalışmada istenilen, kelimenin anlam ve tür bakımından hiçbir değişikliğe uğramaması için sadece çekim eklerinin ayrıştırılması yapım eklerine dokunulmamasıdır. Bu problemin düzeltilebilmesi için Türkçede sık kullanılan yapım ekleri ve bu eklerin eklendiği kelimenin türünde değişiklik yapıp yapmadığı (fiilden sıfat: yap-yap[an] , isimden isim: kitap-kitap[çı], …) belirlenmiştir. Daha sonar çözümlenmiş kelimenin eklerinden bir tanesi bizim belirlediğimiz sık kullanılan yapım eklerinden bir tanesi ise bu yapım eki ya da ekleri kelimenin köküne eklenmiştir. Yapılan bu işlem sayesinde sık kullanılan yapım eklerinin kelimeden ayrıştırılması durumunun önüne geçilmiş ve veri setimizde sadece çekim ekleri ayrışmış olan kelimeler elde edilmiştir. Ayrıca yapılan bu ekleme işlemleri sonucunda gövde haline gelen kelimenin tür bilgisi de eklenen yapım eklerinin yapısı göz önüne alınarak güncellenmiştir. Proje kapsamında belirlediğimiz Türkçede en çok kullanılan yapım eklerinden bazıları aşağıda verilmiştir.

-ISIM_BULUNMA_LIK	-ISIM_ANDIRMA_IMSI
-ISIM_BULUNMA_LI	-FIIL_TANIMLAMA_ICI
-ISIM_YOKLUK_SIZ	-ISIM_KUCULTME_CIK
-ISIM_ILGI_CI	-SAYI_SIRA_INCI

Eklerine ayrışmış kelimelerden oluşan veri seti üzerinde çalışmak sayesinde daha önce bahsettiğimiz kelimenin eklerinden kaynaklanan problemler aşılmaktadır. Örneğin; "… elmaların ve benzeri meyvelerin …" cümlesi morfolojik olarak çözümlendikten sonra "… elma +isim+ ISIM_COGUL_LER + ISIM_TAMLAMA_IN ve+isim benzer+isim+ ISIM_TAMLAMA_I meyve+isim+ ISIM_COGUL_LER + ISIM_TAMLAMA_IN …" şekline gelmektedir. Böylece başlangıçta verilen elma-meyve ikilisi kolaylıkla veri seti içinde bulunacak ve şablon bilgisi olarak da "ISIM_COGUL_LER+ISIM_TAMLAMA_IN ve+isim benzer+isim+ISIM_TAMLAMA_I" elde edilmiş olunacaktır.

Ancak morfolojik olarak çözümlemenin bazı olumsuz yanları da bulunmaktadır. Üzerinde çalıştığımız veri setinin bir kısmı pdf dosyalarından elde edilen verilerden oluşmaktadır. Pdf dosyalarından bilgi çekerken verilerin düzensiz bir yapıda elde edildiği görülmüştür. Bu veri setindeki kelimeler Zemberek ile çözümlenirken önce kelimenin doğru yazılıp yazılmadığı kontrol edilmektedir. Eğer kelime yanlış yazılmışsa yani anlamsız bir kelime ise Zemberek bu kelimeyi çözümleyememektedir. Dolayısıyla veri setimiz içindeki yanlış kelimeler elenmektedir. Yanlış ve anlamsız kelimelerden kurtulmak bir avantaj gibi görünse de aslında değildir. Çözümlenemeyip atılan bu yanlış veya hatalı kelimeler nedeniyle cümlelerin anlam yapıları bozulmaktadır. Bu da doğru şablon elde edilecekken yanlış bir şablonun elde edilmesine ya da doğru ikililer üretilecekken yanlış ikililerin üretilmesine neden olmaktadır.

Zemberek ile çözümleme yaptıktan sonra cümle sayılarında azalma olmuştur. Bilindiği gibi yabancı kelimeler ve pdf dosyalarından elde edilen bazı yanlış, bozuk kelimeler Zemberek tarafından çözümlenememektedir. Bu nedenle bazı cümlelerin içerdikleri kelime sayıları azalmaktadır. Bu cümlelerden de 3 kelimeden az kelime içerenler işimize yaramadığı için (min. cümle uzunluğu = [kelime+şablon+kelime]) veri seti içerisinden atılmıştır. Veri setimiz içerisinde Zemberek tarafından çözümlenip çözümlenemeyen kelime sayıları ve oranları Tablo 8'de verilmiştir.

TABLO 8. ZEMBEREK ILE ÇÖZÜMLENEN VE ÇÖZÜMLENEMEYEN KELIME SAYILARI VE ORANLARI

Veri seti ID	Başlangıçtaki cümle sayısı	Çözümlendikten sonraki cümle sayısı	Çözümlenen kelime sayısı	Çözümlenemeyen kelime sayısı	Çözümlenen kelime oranı (%)
A	2.090.162	2.037.827	27.578.775	3.106.492	% 89
B	701.523	644.100	9.177.010	2.169.712	% 80
C	518.414	499.799	4.833.143	449.871	% 91
D	181.285	173.412	1.902.072	396.916	% 82

Bu kapsamda öncelikle pozitif başlangıç ikililerinden şablonlar elde edilmiş sonra bu şablonlardan yeni ikililer üretilmiştir. Bunun için elde edilen şablonlar çözümlenmiş kelimelerden oluşan veri setimiz içerisinde aratılarak geçtiği cümleler belirlenmiştir. Daha sonra şablonların geçtiği bu cümlelerde şablonun sağında ve solunda kalan kelimeler ilgili ilişkiyi sağladıkları varsayılarak yeni ikili olarak kaydedilmiştir. Bu işlem yapılırken, her iki kelimenin tür bilgilerinin isim-isim, isim-sıfat, sıfat-sıfat, fiil-fiil, zaman-zaman gibi aynı türde olanları alınmıştır. Örneğin sistem "masa" ve "mavi" kelimelerini ikili olarak bulursa, bu kelimelerin türleri aynı olmadığından bulunan ikililer listesinden silinmektedir. Böylece anlamsız ikililerinin üretilmesi ve ilerde bu anlamsız ikililerden anlamsız şablonların üretilmesi engellenmeye çalışılmıştır. Sonrasında üretilen yeni ikililerin şablon frekans değerleri hesaplanılmıştır. Yeni ikililer üretildikten sonra bunlar sisteme pozitif ikililer olarak negatif ikililerle birlikte verilmiş, bunlardan tekrar şablonlar üretilmiş, şablonlardan da ikililer üretilerek yinelemeli bir şekilde bilgi çıkarımı yapılmış ve sonuçlar gözlemlenmiştir. Bu işlem yapılırken veri setimizin tümü (A+B+C+D) kullanılmıştır.

Üst sınıf ilişkisi için başlangıç ikililerinden üretilen şablonlar, bu şablonların ikili frekans değerleri ve bu şablonların başlangıç ikililerinden hangileri için bulunduğu bilgileri Tablo 9'da verilmiştir.

TABLO 9. IS-A ILIŞKISI IÇIN ŞABLON BILGILERI

Şablon	Şablonun ikili frekansı	Şablonun geçtiği başlangıç ikilileri
isim_çoğul_ler+ isim_belırtme_ı ve+isim diğer+isim	5	[rapor-belge], [fotoğraf-veri], [protein-bileşen], [amca-akraba], [kaya-sert]
gibi+edat bir+sayı	5	[ehliyet-belge], [çöl-yer], [güneş-yıldız], [kanser-hastalık], [dünya-gezegen]
isim_tamlama_ın ve+isim diğer+isim	4	[rapor-belge], [polis-güvenlik], [kumar-kötü], [dünya-gezegen]
gibi+edat küçük+isim	4	[köpek-hayvan], [hamsi-balık], [kuş-hayvan], [dünya-gezegen]
gibi+edat bazı+sıfat	4	[iran-ülke], [kan-doku], [köpek-hayvan], [kanser-hastalık]
gibi+edat çeşitli+isim	3	[iran-ülke], [klor-kimyasal], [kobalt-metal]
isim_tamlama_ı ve+isim benzer+isim+ isim_tamlama_ı	3	[senet-belge], [spor-faaliyet], [bez-malzeme]
gibi+edat büyük+sıfat	2	[istanbul-il], [kamyon-araç]
isim_yonelme_e yakın+fiil ol+fiil+ fiil_donusum_en	2	[istanbul-il], [güneş-yıldız]
isim_çoğul_ler ve+isim benzer+isim+ isim_tamlama_ı	2	[ceket-eşya], [banka-kurum]
isim_çoğul_ler ve+isim diğer+isim	2	[banka-işveren], [kuş-hayvan]
isim_çoğul_ler gibi+edat birçok+isim	2	[antibiyotik-ilaç], [kuş-hayvan]
isim_çoğul_ler veya+isim diğer+isim	2	[sel-felaket], [robot-mekanik]
isim_tamlama_ın ve+isim öteki+isim	2	[güneş-yıldız], [dünya-gezegen]
isim_yonelme_e ve+isim diğer+isim	2	[güneş-yıldız], [kanser-hastalık]
isim_tamlama_ı ve+isim_gibi+edat	2	[çanta-eşya], [banka-kurum]

Tablo 9'da verilen şablonların kullanılmasıyla üretilen yeni ikililerden bazıları, bu ikililerin şablon frekansı değerleri ve hangi şablonlar için bulunduğu bilgileri Tablo 10'da verilmiştir.

TABLO 10. IS-A ILIŞKISI IÇIN ÜRETILEN IKILI BILGILERI

Üretilen ikili	İkilinin şablon frekansı	İkilinin geçtiği şablonlar
*güneş-yıldız	3	[gibi+edat bir+sayı], [isim_yonelme_e yakın+fiil ol+fiil+fiil_donusum_en], [isim_yonelme_e ve+isim diğer+isim]
*kuş-hayvan	3	[gibi+edat küçük+isim], [isim_çoğul_ler ve+isim diğer+ isim], [isim_çoğul_ler gibi+edat birçok+ isim]
protein-molekül	2	[isim_çoğul_ler+isim_belırtme_ı ve+isim diğer+isim], [gibi+edat büyük+sıfat]
hormon-molekül	2	[isim_tamlama_ın ve+isim diğer+isim], [isim_çoğul_ler ve+isim diğer+isim]
böcek-omurgasız	2	[isim_çoğul_ler+isim_belırtme_ı ve+isim diğer+isim], [isim_çoğul_ler ve+isim diğer+isim]
satranç-spor	2	[gibi+edat bir+sayı], [gibi+edat bazı+sıfat]
kimya-bilim	2	[gibi+edat bir+sayı], [gibi+edat çeşitli+isim]
tavşan-hayvan	2	[gibi+edat küçük+isim], [gibi+edat bazı+sıfat]
sincap-hayvan	2	[gibi+edat bazı+sıfat], [isim_çoğul_ler ve+isim diğer+isim]
güve-böcek	2	[gibi+edat bazı+sıfat], [isim_çoğul_ler ve+isim diğer+ isim]
kızamık-hastalık	2	[gibi+edat bazı+sıfat], [gibi+edat çeşitli+isim]
*kanser-hastalık	2	[gibi+edat bazı+sıfat], [isim_yonelme_e ve+isim diğer+ isim]
grip-hastalık	2	[gibi+edat bazı+sıfat], [isim_tamlama_ı ve+isim benzer+ isim+isim_tamlama_ı]
gasp-suç	2	[gibi+edat bazı+sıfat], [gibi+edat çeşitli+isim]
akciğer-organ	2	[gibi+edat çeşitli+isim], [isim_çoğul_ler ve+isim diğer+ isim]
toplantı-etkinlik	2	[gibi+edat çeşitli+isim], [isim_çoğul_ler gibi+edat birçok+ isim]
eşya-mal	2	[gibi+edat çeşitli+isim], [isim_çoğul_ler gibi+edat birçok+ isim]
deprem-felaket	2	[gibi+edat büyük+sıfat], [isim_çoğul_ler ve+isim diğer+ isim]
hücre-yapı	2	[gibi+edat büyük+sıfat], [isim_çoğul_ler ve+isim diğer+ isim]
deprem-afet	2	[gibi+edat büyük+sıfat], [isim_yonelme_e ve+isim diğer+ isim]
*banka-kurum	2	[isim_çoğul_ler ve+isim benzer+ isim+isim_tamlama_ı], [isim_tamlama_ı ve+isim gibi+ edat]

İkili üretmede kullanılan pozitif ve negatif ikili sayıları, pozitif ikililerden üretilen şablon sayıları, bu şablonlardan ikili üretmek için kullanılan şablonların sayıları (en az 2 ikili için bulunan şablonlar) ve üretilen ikili sayıları hakkında bilgiler Tablo 11'de verilmiştir. Ayrıca üretilen ikililer şablon frekansı değerlerine göre incelenmiş ve doğruluk oranları hesaplanarak Tablo 12'de gösterilmiştir. Bu işlem üretilen çıkış ikililerinin tekrar sisteme giriş ikilileri olarak verilmesi suretiyle 3 iterasyon boyunca devam ettirilmiştir. Üretilen ikililerden şablon fekansına göre büyükten küçüğe doğru sıralanmış ilk 200 tanesi incelenerek bir doğruluk oranı çıkartılmıştır.

TABLO 11. IS-A ILIŞKISI IÇIN ITERASYONA GÖRE SONUÇLAR

Iterasyon no	Pozitif ikili sayısı	Negatif ikili sayısı	Üretilen şablon sayısı	Kullanılan şablon sayısı	Üretilen ikili sayısı
1	100	100	739	17	3.771
2	132	100	1.359	38	7.909
3	429	100	4.823	138	33.400

TABLO 12. IS-A ILIŞKISI IÇIN ŞABLON FREKANSINA GÖRE SONUÇLAR

Şablon frekansı >X olan ikililer	Ikili sayısı	Ikililerin doğruluk oranı
1.iterasyon		
X=1	38	% 68
X=2	3	% 100
X=3	1	% 100
X=4	1	% 100
X=5	0	___
2.iterasyon		
X=1	338	% 30
X=2	34	% 70
X=3	12	% 91
X=4	5	% 100
X=5	2	% 100
3.iterasyon		
X=1	5.026	% 25
X=2	969	% 25
X=3	294	% 25
X=4	136	% 30
X=5	61	% 33

Aynı üst sınıfa ait kardeş ikililer için başlangıç ikililerinden üretilen şablonlar, bu şablonların ikili frekans değerleri ve bu şablonların başlangıç ikililerinden hangileri için bulunduğu bilgileri Tablo 13'de verilmiştir.

TABLO 13. KARDEŞ ILIŞKISI IÇIN ŞABLON BILGILERI

Şablon	Şablonun ikili frekansı	Şablonun geçtiği başlangıç ikilileri
ya+isim da+isim	56	[balık-balina], [kız-erkek], [kedi-köpek], [kalp-böbrek], [ağaç-çalı], [kuş-böcek], [otel-pansiyon], …
isim_tamlama_ı ya+isim da+isim	6	[kedi-köpek], [televizyon-radyo], [telefon-radyo] [anne-baba], [kart-nakit], [tedavi-aşı]
isim_çoğul_ler ya+ isim da+isim	5	[orangutan-papağan], [bitki hayvan], [harabe-eski], [türk-ermeni], [il-ilçe]
mı+soru yoksa+isim	4	[sıvı-gaz], [sağ-sol], [kadın erkek], [iyi-kötü]
bir+sayı de+fiil	4	[sağ-sol], [kadın-erkek], [anne-baba], [iyi-kötü]
değil+isim de+fiil	3	[basit-karmaşık], [tek-çok], [iki-üç]
isim_çoğul_ler+ isim_kalma_de hem+isim de+fiil	3	[kız-erkek], [bitki-hayvan], [kadın-erkek]
isim_cıkma_den ya+isim da+isim	3	[plastik-kil], [anne-baba], [kardeş-akraba]
isim_yonelme_e mi+soru yoksa+isim	2	[mavi-yeşil], [iyi-kötü]
isim_tamlama_ı ne+isim de+fiil	2	[televizyon-radyo], [kadın-erkek]
fiil_yetersızlık_e ya+isim da+isim	2	[kız-erkek], [sağ-sol]

Tablo 13'te verilen şablonların kullanılmasıyla üretilen yeni ikililerden bazıları, bu ikililerin şablon frekansı değerleri ve hangi şablonlar için bulunduğu bilgileri Tablo 14'de verilmiştir.

TABLO 14. KARDEŞ ILIŞKISI IÇIN ÜRETILEN IKILI BILGILERI

Üretilen ikili	İkilinin şablon frekansı	İkilinin geçtiği şablonlar
*kadın-erkek	14	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], [isim_tamlama_ı ya+isim da+isim], [isim_yonelme_e ya+isim da+isim], …
hayvan-insan	11	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], [isim_tamlama_ı ya+isim da+ isim], [isim_çoğul_ler ya+isim da+isim], …
yukarı-aşağı	8	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], [isim_yonelme_e ya+isim da+ isim], [isim_tamlama_ın ya+isim da+ isim], …
atom-molekül	7	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+ isim], [isim_yonelme_e ya+isim da+ isim], [isim_çoğul_ler ya+isim da+isim], …
*sıvı-gaz	6	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], [isim_çoğul_ler+isim_tamlama_ın ya+ isim da+isim], …
robot-insan	6	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], [isim_çoğul_ler ya+isim da+isim], …
yıldız-gezegen	6	[ya+isim da+isim], [ne+isim de+fiil], [isim_çoğul_ler ya+isim da+isim], …
kasaba-köy	5	[ya+isim da+isim], [isim_yonelme_e ya+isim da+ isim], [isim_çoğul_ler ya+isim da+isim], …
deniz-okyanus	5	[ya+isim da+isim], [isim_çoğul_ler ya+isim da+isim], [isim_tamlama_ın ya+isim da+isim], …
çocuk-yetişkin	5	[ya+isim da+isim], [isim_çoğul_ler ya+isim da+isim], [isim_tamlama_ın ya+isim da+isim], …
kara-deniz	5	[ya+isim da+isim], [hem+isim de+fiil], [mı+soru yoksa+isim], …
iktidar-muhalefet	4	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], …
demir-tahta	4	[ya+isim da+isim], [isim_yonelme_e ya+isim da+isim], [isim_tamlama_ın ya+isim da+isim], …
mercek-ayna	4	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+isim], [isim_çoğul_ler ya+isim da+isim], …
hasta-doktor	4	[ya+isim da+isim], [hem+isim de+fiil], [ne+isim de+fiil], …
balkon-pencere	3	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+isim], [isim_kalma_de

		ya+isim da+isim]
cumhurbaşkanı-başbakan	3	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+isim], [mı+soru yoksa+isim]
motor-piston	3	[ya+isim da+isim], [isim_çoğul_ler ya+isim da+isim], [isim_çoğul_ler+isim_belırtme_ı ya+ isim da+isim]
lise-üniversite	3	[ya+isim da+isim], [hem+isim de+fiil], [isim_tamlama_ı ya+isim da+isim]
dürbün-teleskop	3	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+isim], [isim_tamlama_ın ya+isim da+isim]
vakıf-dernek	3	[ya+isim da+isim], [isim_tamlama_ı ya+isim da+isim], [isim_yonelme_e ya+isim da+isim]

Ikili üretmede kullanılan pozitif ve negatif ikili sayıları, pozitif ikililerden üretilen şablon sayıları, bu şablonlardan ikili üretmek için kullanılan şablonların sayıları (en az 2 ikili için bulunan şablonlar) ve üretilen ikili sayıları hakkında bilgiler Tablo 15'de verilmiştir. Ayrıca üretilen ikililer şablon frekansı değerlerine göre incelenmiş ve doğruluk oranları hesaplanarak Tablo 16'da gösterilmiştir. Bu işlemler üretilen çıkış ikililerinin tekrar sisteme giriş ikilileri olarak verilmesi suretiyle 3 iterasyon boyunca devam ettirilmiştir. Şablon fekansına göre büyükten küçüğe doğru sıralanmış ikililerden ilk 200 tanesi incelenerek bir doğruluk oranı çıkartılmıştır.

TABLO 15. KARDEŞ ILIŞKISI IÇIN ITERASYONA GÖRE SONUÇLAR

Pozitif ikili sayısı	Negatif ikili sayısı	Üretilen şablon sayısı	Kullanılan şablon sayısı	Üretilen ikili sayısı
100	100	1.059	37	9.492
161	100	3.492	97	20.601
224	100	11.443	173	32.350

TABLO 16. KARDEŞ ILIŞKISI IÇIN ŞABLON FREKANSINA GÖRE SONUÇLAR

Şablon frekansı >X olan ikililer	Ikili sayısı	Ikililerin doğruluk oranı
1.iterasyon		
X=1	434	% 59
X=2	73	% 80
X=3	25	% 92
X=4	12	% 100
X=5	9	% 100
2.iterasyon		
X=1	1.637	% 55
X=2	293	% 55
X=3	114	% 62
X=4	54	% 75
X=5	32	% 81
3.iterasyon		
X=1	6.272	% 45
X=2	1.357	% 45
X=3	544	% 45
X=4	302	% 45
X=5	171	% 50

Yapılan bu işlemler sonucunda iterasyon saysı arttıkça üretilen ikililerin doğruluklarında azalma olduğu gözlemlenmiştir. Bunun nedeninin şablon sayısının sürekli artmasından kaynaklandığı düşünülmüştür. Herhangi bir anlamsal ilişki için elde edilebilecek şablon sayısı belli bir sayıdadır. Yani çok sayıda şablon üretilmesi bunların herbirinin iyi şablonlar olduğu anlamına gelmemektedir. Bu nedenle iyi yapıda şablonları elde ettikten sonra şablon sayısını daha arttırmadan mevcut şablonları kullanmak daha iyi sonuçlar üretebilir. Ayrıca bu işlemler sonucunda ikililerin şablon frekansı değeri arttıkça yani ikilinin geçtiği şablonların sayısı arttıkça bu ikililerin doğruluğunda artma olduğu ancak üretilen ikili sayısnın az olduğu görülmüştür. Aynı şekilde ikililerin şablon frekansı değeri azaldıkça yani ikilin geçtiği şablon sayısı azaldıkça da bu ikililerin doğruluğunda azalma olduğu ancak üretilen ikili sayısnın daha fazla olduğu görülmüştür.

5.3. SABIT SAYIDA ŞABLONLAR ILE BÜYÜYEN VERI SETI ÜZERINDE ÇALIŞMA

Bu bölümde şablon sayısının sürekli artması yerine belli bir değerde sabitlenmesi sonucu elde edilen sonuçlar gözlemlenmiştir. Yeni ikililer üretmek için kullanılan bu şablonlar tüm veri setinin kullanılmasıyla 1. iterasyon sonunda elde edilen, ikili frekansı 1'den büyük olan şablonlardır. Üst sınıf ilişkisi için şablon sayısı 17, aynı üst sınıfa ait kardeş ilişkisi için ise bu şablon sayısı 37 olarak belirlenmiştir. Elde edilen bu az sayıdaki iyi şablonlar yeni ikililer üretmek amacıyla kullanılmıştır. Bu işlem yapılırken öncekinin aksine veri setinin tamamı kullanılmamış, veri seti parça parça kullanılmıştır. Bunun yapılmasındaki amaç iyi yapıda bulunan sabit sayıdaki şablonlar ile büyüyen veri seti üzerinde çalışmanın etkilerini gözlemleyebilmek ve hangi tür veri setinde nasıl sonuçlar elde edildiği görebilmektir. Şablon frekansına göre büyükten küçüğe doğru sıralanmış ikililerden ilk 200 tanesi incelenerek bir doğruluk oranı çıkartılmıştır. Her iki ilişki türü için morfolojik olarak çözümlenmiş büyüyen veri seti üzerinde elde edilen karşılaştırmalı sonuçlar Tablo 17'de verilmiştir.

TABLO 17. SABIT ŞABLON SAYISI ILE BÜYÜYEN VERI SETI IÇIN SONUÇLAR

Kullanılan veri setleri ID	Içerdiği cümle sayısı	Üretilen ikili sayısı	Ikili doğru-luk(%)	10^5*Üretilen ilişki sayısı / cümle sayısı oranı
Üst sınıf ilişkisi için sonuçlar				
A	2.037.827	1.183	% 43	6
B	644.100	917	% 43	14
C	499.799	289	% 20	6
D	173.412	210	% 38	12
C+D	673.211	498	% 40	7
B+C+D	1.317.311	1.402	% 49	11
A+B+C+D	3.355.138	2.565	% 50	8
Kardeş ilişkisi için sonuçlar				
A	2.037.827	4.064	% 46	20

B	644.100	3.452	% 44	54
C	499.799	1.415	% 40	28
D	173.412	825	% 47	48
C+D	673.211	2.229	% 49	33
B+C+D	1.317.311	5.575	% 55	42
A+B+C+D	3.355.138	9.492	% 57	28

Sınırlı sayıda ve daha güvenilir şablonlar kullanılarak elde edilen ikililerin sayısının az fakat doğruluk oranlarının daha yüksek olduğu gözlemlenmiştir. Ayrıca sınırlı sayıdaki iyi şablonları kullanarak büyüyen veri seti üzerinde çalışırken üretilen ikililerin doğruluklarında artma olduğu gözlemlenmiştir.

Yukarıdaliği verilen tablo incelendiğinde üst sınıf ilişkisi için 'A' ve 'B' veri setlerinden elde edilen ikililer için yaklaşık olarak %43 ile en yüksek doğruluk oranı elde edilmiştir. Yine 'D' veri setinin tek başına kullanılmasıyla %38, 'B' veri setinin tek başına kullanılmasıyla ise %20 başarı oranı elde edilmiştir. Buradan hareketle üst sınıf ilişkisi için haber metinlerinden oluşan 'A' ve bilimsel içeriklerden oluşan 'B' veri setlerinin daha iyi içerikler olduğu görülmüştür. Aynı şekilde bilimsel içerikli cümlelerden oluşan 'D' veri setinden de yine 'A' ve 'B' veri setlerine yakın bir değer elde edilmiştir. En düşük doğruluk oranı ise hikaye, roman vb. cümlelerin bulunduğu 'C' veri setinden elde edilmiş, bu içeriğin üst sınıf ilişkisi için iyi bir veri seti olmadığı görülmüştür.

Aynı üst sınıfa ait kardeş ilişkisi için en yüksek başarı oranı %47 ile bilimsel içerikli cümlelerden oluşan 'D' veri setinden elde edilmiştir. 'A' ve 'B' veri setlerinden de yine buna yakın değerler elde edilmiştir. 'C' veri setinden ise %40 ile en düşük başarı oranı elde edilmiştir. Bu ilişki türü için de haber ve bilimsel içeriklerden oluşan veri setlerinin uygun ikililerin bulunmasında daha verimli oldukları görülmüştür.

Üretilen ikili sayılarının veri kümelerindeki cümle oranlarına bakıldığında (son sütun) en üretken veri kümelerinin bilimsel içeriğe sahip 'B' ve 'D' oldukları görülmüştür.

6. SONUÇ

Bu çalışmada aralarında belli bir anlamsal ilişki bulunan ikililerin geniş metin koleksiyonları içerisinden yinelemeli bir şekilde çıkartılmasını sağlayan bir sistem geliştirilmiştir. Çalışmamızın katkıları aşağıdaki şekilde özetlenebilir:

-Yinelemeli bilgi çıkarımı için Türkçede ilk çalışma yapılmıştır.

-Bilgi çıkarımında kullanılan metin kaynakların türlerinin sistemin verimliliğine ve doğruluğuna etkileri incelenmiştir. Bilimsel metinlerin, haber ve edebiyat metinlerine göre daha doğru ve verimli sonuçlar ürettiği görülmüştür.

-Önce şablonların bulunması, ardından bu şablonlara uygun yeni ikililerin bulunması, bu ikililerin başlangıç ikililerine dahil edilerek bu sürece devam edilmesinin, bulunan şablon sayısını ve ikili sayısını çok fazla arttırdığı ancak, doğruluk oranlarını düşürdüğü görülmüştür. Bu sonuç, literatürdeki sonuçlarla uyumludur.

-Bu probleme çözüm olarak az sayıda ve güvenilir şablonlar kullanılarak kaynak boyutunun ve çeşitliliğinin arttırılması önerilmiş ve bu sayede hem bulunan ikili sayısını hem de doğruluk oranını arttığı gözlenlenmiştir. Veri boyutu arttığında bulunan ikili sayısının artması beklenen bir sonuçtur. Ancak doğruluk oranının da artışı çalışmamızın literatüre katkısıdır.

Çalışmamızın başarısını iki yönden değerlendirmek gerekir. İlki sistemin verimliliği diğer bir ifadeyle kullandığı verilerin boyutuna göre üretebildiği ikililerin oranıdır. Sistemimizin verimliliği bu alandaki en bilinen ve en gelişmiş sistem olan NELL ile karşılaştırılmıştır. NELL'in veri kaynağı 500 milyon web sitesidir. Bir web sitesinin yaklaşık 500 kelime içermektedir [14]. Buna göre NELL'in veri kaynağında yaklaşık $500*500*10^6 = 25*10^{10}$ kelime bulunmaktadır. NELL 2010 yılında öalıştığı ilk 63 günde 242 bin, günümüze kadar geçen 858 günde yaklaşık 2.5 milyon ikili bulmuştur. Veri kaynağındaki kelime sayısı ile bulunan ikili sayısı oranlanırsa $(25*10^{10}) / (25*10^5) = 10^5$ bulunur. Diğer bir ifadeyle veri kaynağındaki her 100 bin kelime için 1 adet ikili bulmuştur. Bizim sistemimiz (yaklaşık 45 milyon kelimeden 2500 üst sınıf, 9500 kardeş toplamda 12 bin ikili bulunmuştur) için aynı oran hesaplanırsa $(45*10^6) / (12*10^3)$ = yaklaşık $4*10^3$ bulunur. Diğer bir ifadeyle veri kaynağındaki her 4000 kelime için 1 ikili bulmuştur. NELL'in ve bizim sistemimiz için hesaplanan iki oran karşılaştırıldığında sistemimizin verimliliğinin NELL'den yüksek olduğu açıktır. Bu farklılık üzerinde çalışılan metin kütüphanelerinin türü ile açıklanabilir. Tablo 17'de göstermiş olduğumuz üzere farklı türdeki veri kaynaklarının verimlilikleri birbirinden farklıdır. Buna gore NELL'in üzerinde çalıştığı rastgele seçilmiş Web sayfalarının verimliliği bizim veri kaynaklarımızın verimliliğinden daha düşüktür.

Sistemin başarısını değerlendirmek için kullandığımız diğer kriter ikililerin doğruluk oranıdır. Karşılaştırma için yine NELL kullanılmıştır. Yayınlanan çalışmalara göre NELL'in ikilileri %74 oranında doğruluğa sahiptir. Ancak bu oran NELL'in çalıştığı 123 adet ilişki türünün ortalamasıdır. Her bir ilişki türü için değerler ayrı ayrı verilmemiştir. Bizim sistemimiz sadece 2 ilişki türü için çalıştırılmıştır. Elde edilen en iyi doğruluk oranları üst sınıf ilişkisi için %50, kardeş ilişkisi içinse %57'dir. Doğruluk oranındaki bu farkın temel sebebi sistemin kavramların tek kelimeden oluştuğu varsayımıdır. Sistemin yanlış olarak bulduğu ikililerde bu sorun açıkça görülmektedir. Örneğin sistem, "hemşireler ve diğer sağlık çalışanları" ifadesinden "Xler ve diğer Y" şablonuna göre "sağlık" ve "hemşire" arasında üst kavram ilişkisi bulmaktadır. Eğer önce kelime öbekleri bulunsaydı ilişki "sağlık çalışanı" ve "hemşire" arasında bulunabilecektir. Buna çözüm için, şablonlar bulunmadan önce metinler üzerinde varlık isim tanımlama çalışıtırıp (name entity recognition), tek bir kavramı ifade eden kelime öbeklerini (Taksim Meydanı, İstiklal Caddesi, Ege Bölgesi vb.) belirlemek gerekmektedir.

Mevcut çalışma kapsamında ayırt edici şablonlar elde edilebilmesi amacıyla yaygın olarak kullanılan iki anlamsal ilişki kullanılmış ve mevcut örnekler bu anlamsal ilişkiler üzerinden verilmiştir. Verilen bu iki örnek anlamsal ilişki haricinde de herhangi bir anlamsal ilişki için (zıt anlam, eş

anlam vb…) bu çalışmalar kolaylıkla gerçekleştirilebilir. Yapılması gereken sadece istenilen anlamsal ilişkiyi sağlayan pozitif ve negatif başlangıç ikililerinin belirlenmesi ve bunların sisteme giriş olarak verilmesidir.

Gelecek çalışma olarak, sistemin bulduğu iliklileri arttırmak için web sayfaları üzerinde sürekli büyüyen bir veri kaynağıyla çalışılması amaçlanmaktadır. Şablon sayısının sürekli artması yerine yine belli bir değerde sabitlenmesi ve bu şablonların kullanılması düşünülmektedir. Ayrıca tek kelime yerine kelime gruplarından (isim, sıfat tamlaması vb…) oluşan ikililerin de bulunması ve farklı ilişki türleri için de ikililerin çıkarılması amaçlanmaktadır.

KAYNAKÇA

[1] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K., "Introduction to WordNet: An On-line Lexical Database", 1993.

[2] Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions", Emre Yazıcı, M.Fatih Amasyalı, EMO Bilimsel Dergi, Vol. 1, No. 1, pp. 1-13, 2011

[3] Amasyalı M. F., "Türkçe Wordnet'in Otomatik Olarak Oluşturulması", *SIU 2005*, 2005.

[4] http://lucene.apache.org/core/

[5] http://tr.wikipedia.org/wiki/Lucene

[6] Hearst, M., ``Automated Discovery of WordNet Relations," in WordNet: An Electronic Lexical Database, Christiane Fellbaum (ed.), MIT Press, 1998.

[7] htpp://maya.cs.depaul.edu/~classes/etc584/papers/brin.pdf

[8] http://rtw.ml.cmu.edu/rtw/

[9] Andrew Carlson1, Justin Betteridge1, Bryan Kisiel1, Burr Settles1, Estevam R. Hruschka Jr.2, and Tom M. Mitchell., "Toward an Architecture for Never-Ending Language Learning"

[10] http://tika.apache.org/

[11] http://www.kemik.yildiz.edu.tr/?id=28

[12] http://tr.wikipedia.org/wiki/Morfoloji

[13] http://tr.wikipedia.org/wiki/Zemberek_%28yaz%C4%B1l%C4%B1m%29

[14] Levering, R., ve M. Cutler, "The Portrait of a Common HTML Web Page,", DocEng 2006, pp.198-204, 2006

EK 1. Üst sınıf (is-a) ilişkisi için şablon üretmede kullanılan pozitif başlangıç ikilileri

dolar para	türkiye devlet	boya malzeme	güneş yıldız
amerika batı	protein bileşen	ekmek besin	kanser hastalık
iran ülke	amca akraba	bakan siyasi	çatlak hasar
esnaf meslek	bütçe kaynak	kapkaç suç	salon yer
taş madde	tüfek silah	incir meyve	bez malzeme
kan doku	ehliyet belge	antibiyotik ilaç	ilaç ürün
pasaport kimlik	sel afet	sel felaket	çanta eşya
saldırı suç	sandalye mobilya	benzin enerji	süt ürün
gıda ihtiyaç	otobüs araç	plastik malzeme	banka kurum
dershane kurs	tatlı gıda	karides kabuklu	kafeterya yer
rapor belge	çavdar tahıl	bit küçük	tavuk hayvan
polis güvenlik	jimnastik spor	aspirin ilaç	portakal narenciye
muayene işlem	battaniye malzeme	faks iletişim	fon bütçe
köpek hayvan	çöl yer	plastik madde	kiraz meyve
işkence kötü	salon mekan	hamsi balık	kurt vahşi
silah alet	silah malzeme	naftalin madde	robot mekanik
banka işveren	yelken spor	baklava tatlı	kuş hayvan
gayrimenkul mal	yoğurt madde	petrol sıvı	peynir ürün
ilaç tıbbi	klor kimyasal	laboratuvar ortam	nane baharat
fotokopi belge	kumar kötü	brokoli yeşil	kaya sert
fotoğraf veri	kamyonet araç	kamyon araç	psikoloji bilim
taş nesne	ceket eşya	kobalt metal	meyve besin
istanbul il	senet belge	parazit canlı	bakır maden
sorgu işlem	spor faaliyet	jeoloji alan	su madde
rehberlik hizmet	otomobil araç	tank araç	dünya gezegen

EK 2. Kardeş ilişkisi için şablon üretmede kullanılan pozitif başlangıç ikilileri

mavi siyah	yaprak meyve	görüntü fotoğraf	eksik yanlış
kandil meşale	otel pansiyon	metal plastik	emekli ssk
basit karmaşık	orangutan papağan	meyve et	anne baba
sıvı gaz	bitki hayvan	sağ sol	çocuk arkadaş
kovuk mağara	tuz şeker	bitki hayvan	kardeş arkadaş
büyü küçül	ses video	kitap kalem	amir patron
balık balina	yürü koş	kalem kağıt	kart nakit
sperm kromozom	cam tavan	masa sandalye	sünni şii
kız erkek	kestane erguvan	televizyon radyo	ordu emniyet
kedi köpek	televizyon bilgisayar	otobüs taksi	türk ermeni
televizyon radyo	petrol gaz	bilgisayar telefon	yerli ithal
hücre kan	ev otel	telefon radyo	enerjik yorgun
mavi yeşil	kum çakıl	rüya düş	spor tiyatro
bisküvi meyve	ayna gözlük	kadın erkek	bal pekmez
tek çok	su hava	amca dayı	mülk toprak
iki üç	kaçak kayıp	fizik kimya	mart nisan
kalp böbrek	plastik cam	ekmek su	kardeş akraba
pembe mor	saz bambu	ekmek meyve	tedavi aşı
plastik kil	kum çamur	aslan kaplan	üniversite hastane
ağaç çalı	akşam sabah	aslan kartal	fazla eksik
mürekkep boya	harabe eski	asker polis	silah bomba
çadır sığınak	müze tarihi	polis memur	kişi grup
dip yüzey	özel kamu	vatan yurt	il ilçe
bal şeker	kimyasal nükleer	ömür yaşam	iyi kötü
kuş böcek	deterjan sabun	cep çanta	can mal