Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Hikaye Bağlantı Algılama Görevinin Başarımına Etkisi

Proje No: 111K030

Doç.Dr. İrem Soydal
Doç.Dr. Umut Al

Ocak 2014
Ankara

ÖNSÖZ

TÜBİTAK tarafından desteklenen bu proje ile Konu Tespit ve Takip (Topic Detection and Tracking) programında tanımlı Hikaye Bağlantı Algılama (Story Link Detection) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesi ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunmasını amaçlanmaktadır. Projede Bilkent Üniversitesi tarafından hazırlanmış olan BilCOL-2005 derlemi kulanılmıştır. Haberlerde geçen ve temel olarak kim (who), nerede (where) ve ne zaman (when) sorularına yanıt verecek etiketlerle işaretleme yapılmıştır.

Proje süresince yapılan çalışmalar ve elde edilen bulgular uluslararası kamuoyu ile paylaşılmıştır (bkz. Ek). 4-6 Eylül 2013 tarihleri arasında Limerick Teknoloji Enstitüsü tarafından düzenlenen *4th International Symposium on Information Management in a Changing World* adlı toplantıda “Supervised news classification based on a large-scale news corpus”; 17-20 Kasım 2013 tarihleri arasında IEEE tarafından düzenlenen *International Conference on Web Intelligence* toplantısında ise “Story link detection in Turkish Corpus” başlıklı bildiriler sunulmuştur.

Projeye çok sayıda kişinin emeği geçmiştir. Derlem etiketleme işini yapan Bilgi ve Belge Yönetimi Bölümü öğrencilerinin dışında projenin bursiyerleri olarak Güven Köse, Hamid Ahmadlouei ve İpek Şencan çalışmaya katkı sağlamışlardır. Ayrıca projenin öneri aşamasında destekte bulunan Yaşar Tonta’ya teşekkürü bir borç biliriz.

İÇİNDEKİLER

ÖZ

Bu proje, Konu Tespit ve Takip (Topic Detection and Tracking - TDT) programında tanımlı Hikaye Bağlantı Algılama (Story Link Detection - SLD) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesini ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunmasını amaçlamaktadır. Bu kapsamda, TDT içerisinde başarımları kanıtlanmış olan Vektör Uzayı Modeli (Vector Space Model) ve Dil Modeli (Language Model) temel yöntemler olarak kabul edilmiş ve bu yöntemlerle birlikte Varlık İsimlerinin (Named Entity) kullanılmasının başarım üzerindeki etkileri değerlendirilmiştir. Projede tanımlanan yöntemlerin test edilebilmesi için BilCOL-2005 derlemi haberlerde geçen ve temel olarak kim (who), nerede (where) ve ne zaman (when) sorularına yanıt verecek etiketlerle işaretlenmiş ve sistem testleri bu etiketli veriler kullanılarak gerçekleştirilmiştir.

Bu raporda proje önerisinde hedeflenen durum açıkça ortaya konulacak, belirlenen bu hedefleri gerçekleştirmek için yapılan çalışmalar anlatılacak ve projenin kazanımları açıklanacaktır.

ABSTRACT

This project aims to test the performance of the Story Link Detection (SLD) task as part of the Topic Detection and Tracking (TDT) program using different retrieval algorithms and combinations thereof on a Turkish corpus, and find the one that provides the best precision/recall values. To do this, Vector Space Model (VSM) and Relevance Model (RM) are used as the main methods since their performance is proven in TDT studies, and evaluate the impact of Named Entities on performance. In order to test the performance of these methods, the BilCOL-2005 corpus is used, after tagging the news items, so that who, where and when type questions can be answered. This report describes the methodology, presents the results and explains the achievements of the project that was targetted in the project proposal.

TABLOLAR LİSTESİ

1. BÖLÜM: GİRİŞ

Geleneksel Bilgi Erişim Sistemleri üzerindeki akademik çalışmalar son yıllarda ağırlıklı olarak Konu Tespit ve Takip (Topic Detection and Tracking - TDT) programı üzerinde yoğunlaşmıştır. TDT çalışmalarının amacı; gazete, radyo ya da televizyon haberleri ile ilgili hikâyelerin organize edilmesi, belirlenen bazı hikayelerin tespit edilmesi ve zaman içerisinde bunların takip edilebilmesini sağlayacak teknolojilerin geliştirilmesini sağlamaktır (Allan, 2002). Belirlenen bu hedefi gerçekleştirmek için TDT çalışmaları, sisteme ulaşan haber yayınlarını her biri bağımsız bir olayı tartışacak şekilde ayırmayı amaçlayan "*Hikaye Bölümleme (Story Segmentation)"*, sisteme ulaşan haberin daha önce karşılaşılmamış yeni bir hikaye olduğunu belirlemeyi amaçlayan "*İlk Hikaye Algılama (First Story Detection)"*, sisteme ulaşan haberin hangi konu kümesine ait olduğunu belirlemeyi amaçlayan "*Küme Algılama (Cluster Detection)"*, belirlenen bir haberin sistem tarafından takip edilmesini amaçlayan "*Hikaye İzleme (Topic Tracking)"* ve sisteme ulaşan iki bağımsız haberin aynı konuyu tartışıp tartışmadıklarını anlamayı amaçlayan "*Hikaye Bağlantı Algılama (Story Link Detection)"* isimleri altında beş temel göreve bölünmüştür.

"Hikaye Bağlantı Algılama" görevinin, TDT çalışmalarında kritik bir öneme sahip olduğu belirtilmiştir (Allan ve diğerleri, 1998; Allan, 2002; Lavrenko ve diğerleri, 2002). Buna göre, sisteme verilen iki bağımsız hikayenin aynı haber konusunu tartışıp tartışmadığını anlamayı hedefleyen Hikaye Bağlantı Algılama görevinin başarıyla gerçekleştirilmesi halinde, TDT için pek çok problemin de beraberinde çözülebileceği öngörülmektedir (Allan ve diğerleri, 1998; Allan, 2002).

Hikaye Bağlantı Algılama görevi, geleneksel bilgi erişim sistemlerinde bir sorgu ile derlemde bulunan belgelerin eşleştirilmesine çok benzemekle birlikte TDT içerisinde sorgunun yerini belge almakta ve iki farklı dokümanın aynı konuyu tartışıp tartışmadığı belirlenmeye çalışılmaktadır. Bu kapsamda geleneksel bilgi erişim sitemlerinde ilgililik kestirmeleri için kullanılan pek çok yöntemin TDT içerisinde de kulanıldığı görülmektedir. Bu yöntemler; Boole modeli, vektör uzayı modeli, olasılıksal modeller, dil modeli ve ilgi modeli (Salton, Wong ve Yang, 1975; Robertson, 1977; Maron, 1988; Maron ve Kuhns, 1960; Sparck Jones, Walker ve Robertson, 2000; Ponte ve Croft, 1998; Lavrenko ve Croft, 2001) olarak karşımıza çıkmaktadır.

TDT alanında gerçekleştirilen akademik çalışmalar son yıllarda özellikle erişim fonksiyonu bacağında farklı yöntemler birlikte kullanılarak erişim başarımının artırılıp artılamayacağı konusunda yoğunlaşmıştır (Can ve diğerleri, 2010; Yang ve diğerleri, 2002; Hatzivassiloglou, Gravano ve Maganti, 2000; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005). Farklı yöntemlerin birleştirilmesi konusunda yapılan çalışmalar genellikle sistemin anma (recall) değerlerini artırırken aynı zamanda ilgisiz pek çok belgenin de getirilmesini sağlamakta ve sistemin duyarlık (precision) değerinin dolayısıyla başarımın düşmesine neden olmaktadır.

Bu nedenle, bu tür farklı erişim fonksiyonlarının birlikte kullanılacağı çalışmalarda sistem başarımını en üst seviyeye çıkarabilmek için anma ve duyarlık arasındaki dengeyi gözetecek modellerin geliştirilmesi son derece önemlidir. Kısaca bu tür sistemlerin, ideal olarak, derlemdeki tüm ilgili belgelere erişim sağlamasını aynı zamanda da ilgisizlerin dışarıda bırakılmasını sağlayacak şekilde uygun stratejileri desteklemesi gerekmektedir.

Bu çerçevede gerçekleştiriken bu projenin amacı; Hikaye Bağlantı Algılama (Story Link Detection) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesini ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunmasını sağlamaktır. Bu kapsamda, TDT içerisinde başarımları kanıtlanmış olan Vektör Uzayı Modeli (Vector Space Model) ve Dil Modeli (Language Model) temel yöntemler olarak kabul edilmiş ve bu yöntemlerle birlikte Varlık İsimlerinin (Named Entity) kullanılmasının başarım üzerindeki etkileri değerlendirilmiştir.

Deneysel çalışmaların gerçekleştirilebilmesi amacıyla, Bilkent Üniversitesi'nde geliştirilen ve benzer makale çalışmalarında kullanılan BilCol-2005 (Can ve diğerleri, 2010) haber derlemi varlık isimlerinin doküman benzerliklerinin belirlenmesindeki etkilerini belirleyebilmek için etiketlenerek kullanılmıştır.

2. BÖLÜM: LİTERATÜR DEĞERLENDİRMESİ

Bilgi erişim sistemleri, farklı ortamlarda bulunan belgeler içerisindeki bilginin bulunarak onunla ilgilenen kullanıcılara sunulmasını amaçlayan sistemlerdir (Meadow, 1992). Bir bilgi erişim sistemi: belgelerin bulunduğu derlem, kullanıcı sorguları ve kullanıcıların sorgu cümlelerinde yer alan terimlerle derlemdeki belgelere verilen terimleri karşılaştırarak ilgili belgeleri belirlemek için kullanılan bir erişim fonksiyonundan oluşur. Bu noktada bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel derlemdeki ilgili (relevant) belgelerin tümüne erişmek, ilgili olmayanları da ayıklamaktır (Tonta, Bitirim ve Sever, 2002).

Geleneksel bilgi erişim sistemlerinden farklı olarak TDT programında kullanıcı sorgularının yerini derlemdeki belgelerle ilgili olup olmadığı bilinmeyen yeni belgeler almaktadır. Bu kapsamda hikaye bağlantı algılama görevinin gerçekleştirilmesinde erişim fonksiyonu sorgu-belge yerine belge-belge eşleştirmesi yapmak zorundadır. Bu eşleştirmeler için kullanılan erişim fonksiyonları geleneksel bilgi erişim sitemlerinde kullanılan yöntemlerle benzerlikler göstermektedir. Bu yöntemlerden bazıları; Boole modeli (Robertson, 1977), vektör uzayı modeli (Salton, Wong ve Yang, 1975), olasılıksal modeller (Robertson, 1977; Maron, 1988; Maron ve Kuhns, 1960; Sparck Jones, Walker ve Robertson, 2000), dil modeli (Ponte ve Croft, 1998) ve ilgi modeli (Lavrenko ve Croft, 2001) olarak karşımıza çıkmaktadır.

Önerilen proje kapsamında kullanılacak olan vektör uzayı modeli (vector space model) ve ilgi modeli (relevance model) yöntemlerine kısaca bakmakta yarar vardır. Vektör uzayı modeli, klasik bilgi erişim sistemleri tarafından erişim fonksiyonu olarak sıkça kullanılan ve 1960'ların sonlarında geliştirilmiş olan ve günümüzde de hâlâ yoğun olarak kullanılan oldukça popüler bir yaklaşımdır (Salton, Wong ve Yang, 1975; Salton, 1989; Frakes ve Baeza Yates, 1992; Schultz ve Liberman, 1999). Bu yöntemi kullanan bilgi erişim sistemlerinde, sorgular ve belge koleksiyonunda bulunan her bir belge, koleksiyonda bulunan *t1,t2...tn* gibi *n* adet tekil kelimeden oluşan bir vektör gibi gösterilir. Belgenin vektör biçiminde gösterilmesinde kullanılan *t1,t2...tn* katsayılarının değerleri, ilgili koleksiyon kelimesinin *(t_i),* belge veya sorgu içerisinde bulunup bulunmamasına ya da kaç kez bulunduğuna göre belirlenir. Vektör uzayı modelinde, terim ağırlıkları *idf-ağırlıklı kosinüs katsayısı* olarak tanımlanır ve *tf.idf (term frequency X inverse document frequency)* olarak gösterilir (Salton ve McGill, 1983). TDT çalışmalarında karşılaştırılması gereken iki belge olduğu için burada her bir belge için birer doküman vektörü oluşturulur ve belgeler arasındaki benzerlik aşağıdaki eşitlikte olduğu gibi hesaplanır. Eşitlikte kullanılan $tf_a(w)$, *w* kelimesinin *a* belgesi içerisindeki sıklığı, $tf_b(w)$, *w* kelimesinin *b* belgesi içerisindeki sıklığı ve *idf(w)*' de *w* belgesinin derlem içerisindeki sıklığını ifade etmektedir.

$$sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w).idf(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2(w)}.\sqrt{\sum_{w=1}^{n} tf_b^2(w)}}$$

Geleneksel bilgi erişim sistemlerinde yoğun olarak kullanılan dil modelinin (language model) gelişmiş bir versiyonu olan ilgi modeli (relevance model) TDT programında hikaye bağlantı algılama görevinin gerçekleştirilmesinde yoğun olarak kullanılmıştır (Berger ve Lafferty, 1999; Miller, Leek ve Schwartz, 1999; Song ve Croft, 1999; Lavrenko ve Croft, 2001). İlgi modeli, dil modelinin uygulanması için gerekli olan eğitim verilerinin bulunmadığı ortamlarda, olasılıkların kestirilmesi için yeni bir yaklaşım sunmaktadır. Lavrenko ve Croft (2001), ilgi modelini, *"Bir sorgu ile ilgili bir belge içerisinde, w kelimesinin bulunma olasılığını ifade eden ve R'nin sorguyla ilgili belgelerin kümesini gösterdiği bir evrende, P(w|R) koşullu olasılığının kestirilmesini sağlayan mekanizma"* olarak tanımlamışlardır. Buna göre *P(w|R),* kelimenin koleksiyon içerisinde bulunma olasılığı kullanılarak, doğrusal aradeğerleme yapılan maksimum benzerlik (maximum likelihood) kestirmesi ile aşağıdaki eşitlikte olduğu gibi kestirilebilir.

$$P(w \mid D) = \lambda P_{ml}(w \mid D) + (1-\lambda) P_{bg}(w) = \lambda \frac{tf_{w,D}}{|D|} + (1-\lambda) \frac{cf_w}{coll.size}$$

Karşılaştırılacak her bir belge için yukarıdaki eşitlik kullanılarak konu modelleri (topic model) oluşturulur. Bu aşamadan sonra iki olasılık dağılımı olarak elimizde bulunan konu modelleri Kullback-Leibler yöntemi kullanılarak karşılaştırılır ve belgelerin ne kadar benzer oldukları belirlenir (Lavrenko ve diğerleri, 2002; Lavrenko ve Croft, 2001).

Hikaye bağlantı algılama görevinin gerçekleştirilmesinde kullanılan pek çok yöntem, karşılaştırılan iki hikaye arasında ne kadar fazla sayıda kelimenin örtüştüğünü araştırır. Karşılaştırılan iki hikaye arasında ne kadar fazla sayıda örtüşen kelime varsa, bu iki hikayenin aynı konuyu tartışma olasılığının da o kadar yüksek olduğu kabul edilir. Bu yaklaşım, vektör uzayı modellerinden (Frakes ve Baeza Yates,1992; Allan, Lavrenko ve Swan, 2002; Schultz ve Liberman, 1999; Schultz ve Liberman, 2002; Xu ve Croft, 2000; Ponte ve Croft, 1997) başlayıp, istatistiksel dil modellerine kadar (Berger ve Lafferty, 1999; Miller, Leek ve Schwartz, 1999; Song ve Croft, 1999; Ponte ve Croft, 1998; Lavrenko ve Croft, 2001) geliştirilen bütün yöntemlerin temelini oluşturmuştur. Pek çok bilgi erişim sisteminde olduğu gibi, çoğu araştırmacı, hangi kelimelerin seçileceği, bu kelimelerin nasıl

ağırlıklandırılacağı ve ağırlıklandırılmış olan bu kelimelerin en etkili biçimde nasıl karşılaştırılacakları konularına odaklanmışlardır.

Doküman gösterimleri (document representation) hem geleneksel bilgi erişim sistemleri hem de TDT görevleri için son derece önemli bir aşamadır. Çalışılan alanlara bağlı olmak koşulu ile doküman gösterimi için kelime tabanlı yöntemler (Salton, 1989), dil modelleri (Ponte ve Croft, 1998) ve çizge (graph) tabanlı yöntemler (Thompson ve Callan, 2005) kullanılmaktadır. Doküman gösterimi ile ilgili olarak kullanılan yöntemlerden bazıları konudan bağımsız olarak geniş bir kullanım alanı bulurken diğer bazı yöntemler sadece sınırlı alanlarda kullanılabilmiştir. TDT çalışmaları da doğası gereği doküman gösteriminin kritik bir öneme sahip olduğu alan olarak karşımıza çıkmaktadır. TDT haber metinleri içerisinde ifade edilen olaylar (events) ile doğrudan ilgilidir ve bu program içerisinde bir olay; özel bir mekanda, belirli kişi ya da organizasyonların katılımı ile belirli bir zaman diliminde gerçekleşen eylemler olarak tarif edilmektedir (Shah, Croft ve Jensen, 2006). Bu kapsamda TDT içerisinde bir haber metninin gösteriminde varlık isimlerinin (named entity) kullanılması, program içerisindeki olay (event) kavramının tanımı ile eşleşmesi açısından bir zorunluluk gibi görünmektedir.

Shah, Croft ve Jensen (2006) çalışmalarında; TDT içerisinde tanımlı olan Story Link Detection (SLD) görevinin gerçekleştirilmesi amacıyla haber benzerliklerinin belirlenmesinde varlık isimlerinden yararlanmışlardır. Çalışmada tf.idf ağırlıklandırma yöntemi baz olarak kabul edilmiş ve bu yöntemin başarımı varlık ismi tabanlı tf.idf, ağırlıklandırılmamış varlık ismi genişletme yöntemi ve ağırlıklandırılmış varlık ismi genişletme yöntemleri ile karşılaştırılmıştır. Testler esnasında TDT3 ve TDT4 derlemleri kullanılmıştır. Bu çalışmada varlık isimleri kullanılarak uygulanan ilk yöntemde (tf.idf on entities) BBN's Identifier (Bikel, Schwartz ve Weischedel, 1999) kullanılarak varlıklar otomatik olarak tespit edilmiş ve haber metinlerinde geçen diğer kelimeler (isimlendirilmiş varlıklar dışındakiler) atılmıştır. Sonraki aşamada, her bir doküman için belirlenen varlık isimleri kullanılarak doküman vektörleri oluşturulmuştur. Doküman benzerliklerinin belirlenmesinde vektör uzayı modeli kullanılmıştır. Bu yöntemde en büyük problem, bazı dokümanların sağlıklı bir karşılaştırma yapacak kadar varlık ismine sahip olmamasıdır. Bu problemi gidermek için varlıklar arasındaki ilişkileri gösteren çizgeler oluşturulmuş ve aynı haberde 1 kez birlikte geçen varlık isimleri ilişkili olarak kabul edilmiştir. Bu yaklaşımda, doküman vektörleri oluşturulurken sadece doküman içinde geçen varlıklar değil bunlarla ilişkili diğer varlıklar da kulanılmıştır (unweighted expansion). Uygulanan son yöntemde ise çizge üzerinde birbiri ile ilişkili varlık isimlerine ilişki derecelerine göre bazı ağırlıklar verilmiş ve yeni doküman vektörleri bu ağırlıklar göz önüne alınarak oluşturulmuştur. Testler sonucu elde edilen veriler SLD görevinde haber benzerlikleri belirlenirken varlık isimlerinin kullanılmasının sistem başarımı üzerinde anlamlı bir maliyet düşüşü sağladığını göstermiştir (Shah, Croft ve Jensen, 2006).

Varlık isimlerinin TDT programında “New Event Detection – NED” görevi için kullanıldığı diğer önemli bir çalışma da Kumaran ve Allan (2004) tarafından gerçekleştirilmiştir. Bu çalışmadan elde edilen sonuçlar, NED görevinin gerçekleştirilmesinde varlık isimlerinin kullanılmasının, belirli konularda başarım üzerinde olumlu etkisi olduğunu göstermektedir (Kumaran ve Allan, 2004).

Bu çalışmanın devamında Can ve arkadaşları (2010) Türkçe bir derlem üzerinde NED görevinin gerçekleştirilmesinde varlık isimlerinin sistem başarımı üzerindeki etkilerini araştırmıştır. Araştırmada doküman vektörleri oluşturulurken dört farklı yöntem kullanılmıştır. Bu yöntemler; 1) varlık ismi dışındaki tüm kelimelerin alınması 2) sadece varlık isimlerinin alınması 3) tüm kelimelerin alınması ve 4) Kumaran, Allen ve McCallum (2004) tarafından önerilen üçgenleme (triangulation) yaklaşımıdır. Buna göre, dokümanlar içerisindeki tüm kelimelerin kullanıldığı vektör gösterimi yaklaşımı en başarılı yöntem olarak rapor edilmiştir (Can ve arkadaşları, 2010). Bu çalışmada dokümanlar içerisindeki varlık isimlerinin belirlenmesinde otomatik çıkarsama yöntemleri kullanılmıştır.

Geleneksel bilgi erişim sistemlerinde kulanılan doküman gösterme yöntemlerinin aslında TDT için yetersiz kaldığı ve olay tabanlı bu alanda destekleyici farklı yöntemlerin kullanılması gereğine literatürde sıkça vurgu yapılmıştır (Allan, Lavrenko ve Jin, 2000; Makkonen, Ahonen-myka ve Salmenkivi, 2003; Makkonen, Ahonen-myka ve Salmenkivi, 2002; Qiu, Liao ve Dong, 2008; Qiu ve Liao, 2008; Mori, Miura ve Shioya, 2006; Jin ve diğerleri, 2005; Kim ve Myaeng, 2004). Bu bakış açısı ile TDT içerisindeki dokümanları klasik terim vektörleri ile ifade etmek yerine hikâyeler içerisindeki isimleri, yerleri, zamanı ve konuyu adresleyen olay vektörlerinin (event vector) kullanılmasının daha anlamlı olacağı fikri destek görmüştür (Makkonen, Ahonen-myka ve Salmenkivi, 2003). Buna göre bir olay vektörü; olaya katılan aktörleri ifade eden kişiler (who), olayın gerçekleştiği zamanı ifade eden zaman (when), olayın gerçekleştiği mekânı ifade eden konum (where) ve olayın eylemini ifade eden konu (what) vektörlerinden oluşacak biçimde ifade edilebilir.

Kumaran ve Allan (2005) tarafından yine NED ile ilgili olarak gerçekleştirilen sonraki bir çalışmada yine varlık isimleri kullanılarak iki farklı hikâyenin karşılaştırılması için isimler, konular ve tam metinler dikkate alınarak bazı deneyler yapılmıştır. Kumaran ve Allan (2004) tarafından daha önce gerçekleştirilen çalışmada vektörler, varlıkların türüne bakılmaksızın belirlenen tüm varlık isimleri kullanılarak oluşturulmuştu. Yeni çalışmada ise (Kumaran ve Allan, 2005), TDT içerisindeki olay (event) tanımından yola çıkarak bir hikâyenin kişiler (who), yerler (where), zaman (when) ve eylemi belirleyen (what) kelimeler kullanılarak ifade edilebileceğini söylemiştir. Bu kabûle göre; eğer iki farklı hikâye aynı konuda ise bu hikâyelerin aynı varlık isimlerini ve konu terimlerini paylaşmaları gerekir. Diğer taraftan, eğer iki hikâye birbirine yakın ancak farklı konularda ise varlık isimleri ya da konu terimleri arasında bir eşleşme olsa da muhtemelen her ikisi birden eşleşmeyecektir (Kumaran ve

Allan, 2005). Kumaran ve Allan (2005) bu çalışmada, varlık isimleri kullanılarak gerçekleştirilen sınıflandırma yöntemlerinin vektör uzayı modeli baz alınarak gerçekleştirilen temel sınıflandırma modelinden anlamlı olarak daha başarılı sonuçlar elde edildiğini rapor etmişlerdir.

Benzer bir yaklaşım daha önceleri Makkonen, Ahonen-myka ve Salmenkivi'nin (2002) çalışmalarında da kullanılmıştır. Araştırmacılar, haberlerde geçen isim, yer ve zaman bilgilerini ayrı ayrı vektörlerle ifade etmişlerdir. Bu çalışmada isim, yer ve zaman gibi varlık isimleri otomatik çıkarsama yöntemleri ile elde edilmiş ve doküman içerisinde bunlar dışındaki terimlerin haberin konusunu (what) ifade edeceği belirtilmiştir. Yazarlar, varlık isimlerinin kullanılmasının yeni haber tespit etme probleminde önemli bir başarım artışı sağladığını raporlamışlardır (Makkonen, Ahonen-myka ve Salmenkivi, 2002). Araştırmacılar yine aynı konuda takip eden çalışmalarında (Makkonen, Ahonen-myka ve Salmenkivi, 2003) TDT için sadece doküman terimleri kullanılarak gerçekleştirilen doküman gösterimlerinin yeterli olmadığını ve etkili bir sistem için varlık isimleri kullanılması gerektiğini vurgulamışlardır. Araştırmacılar her iki çalışmalarında da özellikle yer ve zaman karşılaştırmaları için kesişime dayanan benzerlik metrikleri önermişlerdir (Makkonen, Ahonen-myka ve Salmenkivi, 2002; Makkonen, Ahonen-myka ve Salmenkivi, 2003).

TDT görevlerinin gerçekleştirilmesinde varlık isimlerinin kullanılmasının literatürde genellikle başarım üzerindeki olumlu etkilerinden bahsedilmekle birlikte bunun tersinin savunulduğu çalışmalarda vardır. Kim ve Myaeng (2004), Korece haberlerden oluşturulmuş olan derlem üzerinde gerçekleştirdikleri çalışmalarında zaman (when) bilgisinin konu takibi (topic tracking) için gerçekleştirilen deneylerde başarımı anlamlı bir oranda artırmadığını ifade etmişlerdir. Bu çalışma, TDT içerisinde varlık isimleri kullanımı ile ilgili genellikle başarım üzerinde anlamlı artışların rapor edildiği literatürde ilgi çekici görünmektedir.

TDT programı içerisinde doküman gösterimi (Salton, 1989; Ponte ve Croft, 1998; Thompson ve Callan, 2005; Shah, Croft ve Jensen, 2006; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005; Can ve diğerleri, 2010; Allan, Lavrenko ve Jin, 2000; Makkonen, Ahonen-myka ve Salmenkivi, 2003; Makkonen, Ahonen-myka ve Salmenkivi, 2002; Qiu, Liao ve Dong, 2008; Qiu ve Liao, 2008; Mori, Miura ve Shioya, 2006; Jin ve diğerleri, 2005; Kim ve Myaeng, 2004) ve farklı erişim fonksiyonlarının sonuçlarının birleştirilmesi (Can ve diğerleri, 2010; Yang ve diğerleri, 2002; Hatzivassiloglou, Gravano ve Maganti, 2000; Kumaran ve Allan, 2004; Kumaran ve Allan, 2005) ile ilgili çalışmalar geçmişten günümüze aktif olarak araştırılmış ve günümüzde hâla popülerliğini korumaktadır.

TDT programında Hikaye Bağlantı Algılama (Story Link Detection - SLD) görevinin gerçekleştirilmesinde farklı doküman gösterim yöntemlerinin ve farklı erişim fonksiyonlarının kullanılması ve elde edilen sonuçların farklı kombinasyonlarının test edilmesi konusu

literatürde çalışılan bir konu olmasına rağmen seçilen ve özellikle Türkçe bir derlem üzerinde uygulanacak yöntemler açısından özgün değer taşımaktadır.

Sistem testleri esnasında baz yöntemler olarak seçilen vektör uzayı ve dil modelleri geçmişten günümüze, bilgi erişim sistemleri ile ilgili çalışmalarda erişim fonksiyonu olarak genellikle tek başlarına kullanılmıştır. Pek çok çalışma, bu alanda uygulanan bir yöntemi diğerine göre daha başarılı olarak gösterirken, yöntemler arasındaki başarım farkının nerelerden kaynaklandığı konusunda ayrıntılı bir çalışma gerçekleştirilmemiştir. Özellikle bu çalışmanın kapsamı içerisinde yer alan vektör uzayı ve ilgi modeli, erişim fonksiyonu olarak TDT çalışmalarında yoğun olarak kullanılmıştır (Lavrenko ve diğerleri, 2002; Allan ve diğerleri, 1998; Allan, 2002; Leek, Schwartz ve Sista, 2002). Bu çalışmalarda ilgi modeli kullanılarak hem dil modeli hem de vektör uzayı modelinden daha başarılı sonuçlar alındığı gösterilmesine rağmen, farklılığı yaratan etkenler üzerinde herhangi bir yorum bulunmamaktadır (Lavrenko ve diğerleri, 2002).

Vektör uzayı ve ilgi modellerinin arkasında yatan felsefeye bakıldığında erişim fonksiyonu olarak her iki yöntemin de farklı temeller üzerine kurulduğunu söylemek yanlış olmayacaktır. Vektör uzayı yöntemi karşılaştırılan belgelerdeki terim çakışmalarına göre benzerlik değerlerini hesaplarken, ilgi modeli genişletilmiş konu modelleri oluşturmakta ve bu modelleri doğrudan karşılaştırmaktadır. Bu bağlamda, doküman benzerliklerinin belirlenmesinde vektör uzayı modelinin ilgi modeline göre daha seçici olduğunu söylemek yanlış olmayacaktır. Bu çıkarsama aynı zamanda vektör uzayı yönteminin kaçırdığı konuyla ilgili belgelerin ilgi modeli tarafından yakalanma olasılığının da yüksek olduğunu göstermektedir. Bu bağlamda, SLD görevinin gerçekleştirilmesinde vektör uzayı ve ilgi modelinin vereceği bağımsız kararların OR (VEYA) mantıksal operatörü ile birleştirilmesi sonucu sistemin anma (recall) değerinin oldukça yüksek çıkması, diğer bir deyişle ilgili belgelerin büyük bir çoğunluğuna erişilmesi sağlanacaktır. Diğer taraftan bu tür bir birleştirme muhtemelen ilgili belgelerin yanında ilgisizleri de getireceği için duyarlık (precision) düşecektir. Bununla birlikte vektör uzayı ve ilgi modelinin vereceği bağımsız kararların AND (VE) mantıksal operatörü ile birleştirilmesi ile elde edilecek sonuç yöntemlerin birlikte verdikleri ilgililik kararlarının yorumlanması, diğer bir deyişle bir yöntemin diğerinden farklı olarak verdiği doğru kararların belirlenmesi açısından açıklayıcı olacaktır.

Tüm bu özgün çalışmaların yanında, proje içerisinde SLD içerisinde haber benzerliklerinin belirlenmesinde varlık isimlerinin (named entity) kullanılacak olması Türkçe derlemler üzerinde bu kapsamdaki çalışmaların çok sınırlı olması nedeni ile projenin özgün içeriğini oldukça kuvvetlendirmektedir. Türkçe için gerçekleştirilen benzer çalışmalar ağırlıklı olarak metinlerden varlık isimlerinin (isim, yer, zaman, organizasyon v.b.) otomatik olarak çıkarılmasını sağlayan makine öğrenme yöntemleri üzerine yoğunlaşmıştır (Dalkılıç, Gelişli ve Diri, 2010; Tür, Hakkani-Tür ve Oflazer, 2003; Bayraktar ve Taşkaya-Temizel, 2008;

Küçük ve Yazıcı, 2009a; Küçük ve Yazıcı, 2009b; Küçük ve Yazıcı, 2010). Bilgi erişimin bir parçası olarak varlık isimlerinin erişim fonksiyonu ya da bunu destekler nitelikte kullanıldığı çalışmalar ise oldukça sınırlıdır (Can ve diğerleri, 2010; Uyar, 2009). Can ve diğerleri (2010) yaptıkları çalışmada Türkçe derlemler üzerinde varlık isimlerinin kullanılması ile elde edilecek erişim etkinliği konusunda sınırlı çalışmalara dikkat çekmişler ve bu konuda daha derinlemesine çalışmalar yapılması gerektiğini vurgulamışlardır.

Bu bağlamda who (kim), where (nerede) ve when (ne zaman) etiketleri ile işaretlenmiş bir derlemde bu varlık isimlerinin gerek teker teker gerekse tümü bir arada değerlendirilerek haber benzerlikleri üzerindeki etkileri açık bir şekilde ortaya konulabilecektir. Proje kapsamındaki varlık isimleri ile ilgili çalışmaların iki boyutta incelenmesi planlanmaktadır. Bunlardan birincisinde; haberler içerisindeki varlık isimleri gerek teker teker (who, where ve when ayrı ayrı) gerekse birlikte kullanılarak haber benzerlikleri üzerindeki etkileri araştırılacaktır. İkincisinde ise; haberler içerisindeki varlık isimleri gerek teker teker gerekse birlikte kullanılarak, iki haberin farklı konularda olup olmadıklarını belirlemede ne kadar etkili olduklarına bakılacaktır. Projenin bu yönü literatürde daha önce bu tür bir araştırma hiç yapılmamış olmasından dolayı oldukça yenilikçidir. Bu kapsama yakın bir çalışma Köse (2004) tarafından TDT derlemi üzerinde gerçekleştirilmiş ve sadece haberlerdeki asıl aktörlere (who) bakılarak iki haberin aynı konuda olmadığı ile ilgili güçlü bir karar verilebileceği tezi savunulmuştur. Önerilen projede Türkçe bir derlem üzerinde daha ayrıntılı deneyler yapılarak (who, where, when ve what terimleri incelenerek) ilginç sonuçlar elde edilebileceği düşünülmektedir. Bununla birlikte, eğer varlık isimleri haber farklılıkları konusunda anlamlı sonuçlar üretirse, vektör uzayı ve ilgi modellerinin birleştirilmesi ile ortaya çıkan düşük duyarlık (precision) problemine de bir çözüm bulunmuş olacaktır.

Sonuç olarak, bu proje kapsamında uygulanacak olan yöntemler mükemmel bir bilgi erişim sistemine ulaşmak için ihtiyaç duyulan *"ilgili belgelerin tamamına erişim sağlama ilgisizleri ise dışarda bırakma"* prensibine bizleri yaklaştıracak bazı sonuçlar üretecektir.

3. BÖLÜM: AMAÇLAR

Bu projenin temel amacı; Konu Tespit ve Takip (Topic Detection and Tracking - TDT) programında tanımlı Hikaye Bağlantı Algılama[1] (Story Link Detection - SLD) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesini ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunmasını sağlamaktır. Bu amacı gerçekleştirmek için proje çalışmaları, ilgili derlemin etiketlenmesi, test senaryolarının oluşturulması ve gerekli yazılımların geliştirilmesi ile testlerin uygulanması olarak üç adımda yürütülmesi hedeflenmiştir.

Bu kapsamda birinci aşamada; Can ve diğerleri (2010) tarafından geliştirilmiş olan BilCol – 2005 derlemindeki haberlerden seçilen ve hangi konuya ait olduğu bilinen haberlerdeki varlık isimlerinin etiketlenmesi sağlanarak deneyler için gerekli derlem oluşturulması, ikinci aşamada sistem testlerinin gerçekleştirilebilmesi için gerekli senaryoların oluşturularak yazılımların geliştirilmesi ve son aşamada önerilen yöntemlerin başarımlarının sınanacağı testlerin gerçekleştirilmesi amaçlanmıştır. Bu bağlamda test edilecek senaryoların aşağıdaki gibi oluşturulması planlanmıştır.

- Vektör uzayı yöntemi (VUM) kullanılarak haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- Dil (ilgi) modeli (DM) yöntemi kullanılarak haber benzerliklerindeki anma duyarlık ve f-ölçü değerlerinin belirlenmesi,
- VUM ve DM, OR mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- VUM ve DM, AND mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- Haberlerdeki tüm varlık isimleri kullanılarak (who, where, when) haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- Haberlerdeki tüm varlık isimleri kullanılarak (who, where, when) haber farklılıklarının belirlenmesindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- VUM, DM ve varlık isimleri yöntemlerinin sonuçları OR mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,
- VUM ve DM yöntemlerinin sonuçları OR mantıksal operatörü ile birleştirilirken varlık isimleri tarafından tespit edilemeyen haberlerin ilgisiz olarak işaretlenmesi ile haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi,

[1] Hikaye Bağlantı Algılama: Sisteme verilen iki bağımsız haberin aynı konuda olup olmadıklarını belirlemek için TDT içerisinde tanımlanmış olan görevdir.

4. BÖLÜM: YÖNTEM

Bu bölümde raporun amaçlar bölümünde üç grupta sınıflandırılan hedeflerin gerçekleştirilmesi için yapılan çalışmalar ve kullanılan yöntemler anlatılacaktır.

4.1. Test Derleminin Oluşturulması

Projede hedeflenen amaçların ve deneysel çalışmaların gerçekleştirilebilmesi amacıyla, Bilkent Üniversitesi'nde geliştirilen ve benzer makale çalışmalarında kullanılan BilCol-2005 (Can ve diğerleri, 2010) haber derleminin kullanılması planlanmıştır. BilCol-2005 haber derlemi TDT çalışmalarından esinlenerek hazırlanmıştır. Bu derlem 209.296 gazete haberinden oluşan dokümanlardan oluşturulmuştur. Ancak bu derlem içerisinde geçen haberlerden sadece 5.872 tanesinin önceden belirlenmiş olan 80 konu başlığı (ya da haber) ile ilgili olduğu bilinmekte olup, bu çalışmayı gerçekleştiren araştırmacılar kalan tüm haberlerin bu konu başlıkları ile ilgisiz olduğunu kabul etmiştir. Bu araştırma da bu kabuller temel alınarak gerçekleştirilmiştir. Bu proje, varlık isimlerinin otomatik yöntemlerle çıkarılmasını sağlayan makine öğrenme yöntemleri yerine varlık isimlerinin doküman benzerliklerinin belirlenmesindeki etkilerine odaklandığı için BilCol-2005 derleminin bu proje kapsamında kullanılabilmesi için etiketlenmesine karar verilmiştir.

Bir haber içeriğini oluşturan metinde geçen kelimelerin, nitelediği veya cevapladığı soru zamirlerine göre işaretlenmesi işlemi, "*etiketleme*" olarak adlandırılmaktadır. Bu işlem sırasında, BilCol-2005 derlemi içerisinden alınmış ve konuları net olarak belirlenmiş olan 5.872 haberin okunması ve haber metni içerisindeki kelimelerin özenle seçilerek doğru bir şekilde işaretlenmesi gerekmiştir.

Bu etiketleme çalışmasını gerçekleştirmek amacıyla Java tabanlı bir web uygulaması geliştirilmiş ve proje çalışmalarına katılan Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü 3. sınıf ve yüksek lisans öğrencilerinin bu uygulamayı kullanarak hızlı ve etkin bir biçimde etiketleme yapabilmeleri sağlanmıştır.

Proje kapsamında haberler içerisindeki varlık isimleri belirlenirken "kim (who)", "ne zaman (when)", "nerede (where)" ve "ne (what)" sorularına cevap verecek etiketlemelerin yapılması planlanmış, ancak literatürde farklı çalışmalarda daha ayrıntılı etiketlemeler yapıldığı gözlenmiştir (Shah, Croft ve Jensen,2006; Bikel, Schwartz ve Weischedel, 1999; Kumaran ve Allan, 2004; Dalkılıç, Gelişli ve Diri, 2010; Tür, Hakkani-Tür ve Oflazer, 2003; Bayraktar ve Taşkaya-Temizel, 2008; Küçük ve Yazıcı, 2009a; Küçük ve Yazıcı, 2009b; Küçük ve Yazıcı, 2010). Bu kapsamda proje önerisinde belirtilen ve yukarıda sıraladığımız etiketler genişletilerek varlık isimlerinin *"kurum (organization)", "kişi (person)", "konum (location)", "tarih (date)", "zaman (time)", "yüzde (percentag*e)", "para (money") ve *"bilinmeyen (unknown)"* olarak etiketlenmesine karar verilmiştir. Bu sayede hem bu proje kapsamında belirlenen yöntemler test edilebilecek hem de oluşturulan etiketlenmiş derlemin çok daha

geniş bir akademik çevre tarafından kullanılabilmesi sağlanacaktır. Bu kapsamda etiketleme ile ilgili olarak belirlenen bazı ön kurallar aşağıda sunulmuştur.

- Etiketlenecek ifadeler mümkün olduğunca en küçük parçaya bölünerek etiketlenecektir. Bu sayede varlık isimleri arasındaki kesişme olasılıkları artırılarak başarımın yükseltilmesi hedeflenmektedir. Örneğin: "İzmir Atatürk Stadı" benzeri ifadeler bölünecek ("İzmir": *Location*, "Atatürk": *Person*).
- Aynı haber içinde açık adı ve kısaltması birlikte verilen kurum adları ayrı ayrı etiketlenecektir. Örneğin: "BM": *Organization*, "Birleşmiş Milletler": *Organization*.
- Kurum isimleri (örneğin üniversite adları) bölünmeyecek, tamamı *Organization* olarak etiketlenecektir. Örneğin: "İstanbul Medeniyet Üniversitesi": *Organization*.
- *Organization* etiketi yalnızca resmi niteliği olan kurumlar için kullanılacaktır.
- Herhangi bir şekilde kişi adı geçiyorsa *Person* etiketi kullanılacaktır, kişi kast edilerek kullanılan mahlas ya da unvanlar (örneğin "Başbakan", "Doç.Dr.") etiketlenmeyecektir.
- Ülke kısaltmaları *Location* olarak etiketlenecektir (Örneğin: TC, ABD, UK, vb.).
- Ülke, eyalet, bölge, il, ilçe, semt, köy adları *Location* olarak etiketlenecektir.
- Mahalle, stat, spor salonu, vb. yer isimleri etiketlenmeyecektir.
- Doğrudan "yüzde" yazıyorsa ya da "%" işareti kullanılmışsa *Percentage* şeklinde etiketlenecektir.
- Gün, ay, yıl belirtilen ifadelerin her biri ayrı ayrı olmak kaydıyla *Date* olarak işaretlenecektir.
- Irk belirten ifadeler etiketlenmeyecektir.
- İsim olduğu bilinen ancak belirlenen kategorilere atanamayan ifadeler *Unknown* olarak işaretlenecektir.

Yukarıda sunulan genel kurallara uyularak BilCol-2005 derleminde konu başlıkları bilinen 5881 haber, etiketleme çalışmasında görev yapan öğrencilere paylaştırılmış ve öğrencilerin ilgili yazılım üzerinde hızlı ve etkin bir biçimde varlık isimlerini belirlemeleri sağlanmıştır. Bu ilk etiketleme bittikten sonra haberler çaprazlama olarak öğrencilere tekrar dağıtılmış ve etiketlenen varlık isimlerinin ikinci bir kontrolden geçirilmesi sağlanarak derleme son hali verilmiştir. Bu kapsamda etiketlenen derlemde son durum itibari ile 45.201 person, 35.255 locaition, 29.059 organization, 10.622 date, 1.118 time, 2.708 money, 2.608 percentage ve 10.258 unknown etiketleri bulunan varlık isimleri oluşturulmuştur. Tüm kontrol ve düzenleme çalışmaları bitirildikten sonra her bir haber için bir XML dosyası oluşturulmuş ve haberler, içeriklerinde varlık isimleri ilgili etiketler de gösterilerek sistem testleri için hazı hale getirilmiştir.

4.2. Test Senaryolarının Oluşturulması

Bu bölümde projede hedeflenen amaçların gerçekleştirilebilmesi için sınanacak test senaryolarının nasıl oluşturulduğu anlatılacaktır. Bu kapsamda aşağıda ayrıntıları sunulan test senaryoları temel olarak vektör uzayı modeli, dil modeli ve varlık isimleri yöntemleri kullanılarak uygulanacak olan testlerin nasıl gerçekleştirildiğini ayrıntılı bir biçimde ortaya koymaktadır. Bunun yanında proje önerisinde değinilmemiş olmasına rağmen TDT içerisinde tanımlı diğer bir görev olan Konu İzleme[2] (Topic Tracking - TT) görevi için de sistem testleri gerçekleştirilmiş olup bu yaklaşıma ait test senaryosu da takip eden bölümde sunulmuştur.

4.2.1. SLD İçin Vektör Uzayı ve Dil Modeli Test Senaryosu

- Testler ilgililik değerlendirmesi yapılmış olan haberler üzerinden yapılacaktır.
- Öncelikle her bir konu eğitim ve test belgeleri olmak üzere iki kısma ayrılacaktır.
    - Her bir konuda var olan belge sayısının üçte biri eğitim üçte ikisi de test belgesi olarak kabul edilecektir.
    - Eğitim belgeleri seçilirken, tarih sırasına göre derlemdeki ilk N belge seçilecektir, kalan belgeler test belgesi olarak kullanılacaktır.
- Her bir konu ile ilgili olarak eğitim belgeleri belirlendikten sonra ilgili belgeleri belirlemek için gerekli olan uygun eşik değerinin seçilmesi işlemi şu şekilde gerçekleştirilecektir;
    - Öncelikle, derlemde bulunan 209.296 belgenin üçte biri (1.961 + 67.804) eğitim belgesi olarak belirlenecek ve eğitim verisi olarak dizinlenecektir.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belgenin üçte biri olan 1.961 belge eğitim için sorgu olarak kabul edilecektir.
    - Derleme gönderilecek her bir sorgu için dil modeli ve vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlenecektir.
    - Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarılacak ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edilecektir.
    - Bu başlangıç eşiğine göre her bir konu için anma/duyarlık ve f-ölçüsü değerleri hesaplanacaktır.
    - Sonraki aşamada eşik değeri belirli oranda azaltılıp-artırılarak anma/duyarlık değerleri her bir eşik için tekrar hesaplanacaktır.
    - Anma ve duyarlığın birlikte en yüksek oldukları (ya da birbirlerine en yakın oldukları) değer sistemin kesin eşik değeri olarak hesaplanacak ve sistem testleri bu değere göre gerçekleştirilecektir.

[2] Konu İzleme: Sisteme verilen bir haberin konusunun tespit edilerek yeni gelen haberlerin bu konu ile ilgili olup olmadığının tespit edilmesini amaçlamaktadır.

- Kesin eşik değeri belirlendikten sonra test derlemi üzerindeki değerlendirmeler aşağıdaki gibi gerçekleştirilecektir;
    - Derlemde bulunan 209.296 belgenin üçte ikisi (3.922 + 135.609) test belgesi olarak belirlenecek ve test verisi olarak dizinlenecektir.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belgenin üçte ikisi olan 3.922 belge test için sorgu olarak kabul edilecektir.
    - Derleme gönderilecek her bir sorgu için dil modeli ve vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlenecektir.
    - Her bir sorgu sonucu için ikili sınıflandırma tablosu yaratılacaktır.
    - Tüm sorgular tamamlandıktan sonra mikro ortalama yöntemi kullanılarak tüm testler için ortak bir ikili sınıflama tablosu oluşturulacaktır.
    - Bu ikili sınıflama tablosuna göre anma ve duyarlık değerleri hesaplanarak sistemin başarımı belirlenecektir.
- Vektör Uzayı Modeli ve Dil Modeli için yukarıda belirlenen her bir aşama, belgeleri ifade etmek için seçilecek terim sayısına göre (1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 275, 300, 400, 500 ve 1000 terim için) tekrarlanacaktır.
- Vektör Uzayı ve Dil Modeli Yöntemlerinin sonuçları AND ve OR mantıksal operatörleri ile birleştirilecek ve oluşan yeni erişim çıktısı için anma ve duyarlık değerleri belirlenerek başarım hesaplanacaktır.

4.2.2. SLD İçin Varlık İsimlerinin Kullanıldığı Test Senaryoları

Bu alt bölümde Hikaye Bağlantı Algılamaya yönelik dokuz farklı (Varlık isimleri olmadan vektör uzayı modeli senaryosu, tüm varlık isimlerinin tek vektörle ifade edildiği senaryo, vb.) senaryo ve proje kapsamında yapılanlar ile ilgili bilgi verilmektedir.

4.2.2.1. Varlık İsimleri Olmadan Vektör Uzayı Modeli Senaryosu

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Testler her konu için eğitim ve test belgeleri ayırımı yapmadan tüm belgeler (5.872 adet) üzerinde yapılacaktır.
- Vektör Uzayı Modelinin haber benzerliklerini tespit etmede başarısını belirlemek için aşağıdaki adımlar gerçekleştirilecektir;
    - Haberler üzerinde Zemberek Kütüphanesi kullanılarak gövdeleme işlemi yapılacaktır.
    - Apache Lucene kütüphanesi kullanılarak haberlerin vektör modelleri oluşturulacaktır.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belgenin hepsi sorgu olarak kabul edilecektir.

- Her bir sorgu için vektör uzayı modeli kullanılarak üretilen sorgu-belge eşleşme skorları belirlenecektir.
- Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarılacak ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edilecektir.
- Bu eşik değeri için sistemin ikili sınıflandırma tablosu oluşturularak anma, duyarlık ve f-ölçü değerleri hesaplanacaktır.
- Sonraki aşamada seçilen başlangıç eşik değeri küçük oranlarda artırılıp azaltılarak her defasında anma, duyarlık ve f-ölçü değerleri tekrar hesaplanacaktır.
- Anma ve duyarlılığın en yüksek olduğu eşik değeri sistem testlerinde kullanılacak olan eşik değeri olarak belirlenecektir.

4.2.2.2. Tüm Varlık İsimlerinin Tek Vektörle İfade Edildiği Senaryo

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Sonraki aşamada sistem başarımını belirlemek için aşağıdaki işlemler gerçekleştirilecektir;
  - Haberlerde etiketlenen tüm varlık isimleri belirlenecektir.
  - Varlık isimleri üzerinde gövdeleme yapılacaktır.
  - Her bir haber için belirlenen varlık isimleri kullanılarak haber vektörleri oluşturulacaktır.
  - İlgililik değerlendirmesi yapılmış olan 5.872 belgenin hepsi sorgu olarak kabul edilecektir.
  - Her bir sorgu derlemde bulunan diğer tüm haber varlık vektörleri ile karşılaştırılacaktır.
  - Her bir eşleşme için benzerlik skor değerleri üretilecektir.
  - Belirlenen tüm bu skor değerleri içerisinden, sorgunun ilgili olduğu bilinen belgeler için üretilen skor değerleri çıkarılacak ve ilgili sorgu-belge eşleşmeleri için ortalama skor değeri başlangıç eşiği olarak kabul edilecektir.
  - Bu eşik değeri için sistemin ikili sınıflandırma tablosu oluşturularak anma, duyarlık ve f-ölçü değerleri hesaplanacaktır.
  - Sonraki aşamada seçilen başlangıç eşik değeri küçük oranlarda artırılıp azaltılarak her defasında anma, duyarlık ve f-ölçü değerleri tekrar hesaplanacaktır.
  - Anma ve duyarlılığın en yüksek olduğu eşik değeri sistem testlerinde kullanılacak olan eşik değeri olarak belirlenecektir.

- Son aşamada belirlenen bu eşik değerine göre tüm sorgular tekrar yürütülerek sistemin son başarım değeri hesaplanacaktır.

4.2.2.3. Farklı Varlık İsimlerinin Farklı Vektörlerle İfade Edildiği Senaryo

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Sonraki aşamada başarımını belirlemek için aşağıdaki işlemler gerçekleştirilecektir;
  - Haberlerde etiketlenen tüm varlık isimleri türlerine göre (Person", "Location", "Organization", "Date", "Time", "Money", "Percentage", "Unknown") ayrıştırılacaktır.
  - Varlık isimleri üzerinde gövdeleme yapılacaktır.
  - Her bir haber için farklı varlık ismi türlerine göre vektörler oluşturulacaktır.
  - İlgililik değerlendirmesi yapılmış olan 5.872 belge sorgu olarak kabul edilecektir.
  - Her bir varlık ismi türü için, oluşturulan vektörler kullanılarak ortalama eşik değeri belirleme yöntemi ile ortalama eşik değerleri başlangıç eşiği olarak belirlenecektir.
  - İkili sınıflandırma tabloları yaratılarak anma ve duyarlık değerleri hesaplanacaktır.
  - Başlangıç eşiği küçük oranlarda değiştirilerek her bir eşik için yeni ikili sınıflandırma tabloları, anma ve duyarlık değerleri hesaplanacaktır.
  - Anma ve dyarlığın en yüksek olduğu noktalar her bir varlık türü için eşik değeri olarak kabul edilecek ve sistem testleri bu eşik değerleri baz alınarak gerçekleştirilecektir.
- Her bir varlık türü için belirlenen eşik değerlerine göre aşağıdaki testler vektör uzayı modeli kullanılarak gerçekleştirilecektir;
  - Haberlerde geçen "Person" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Location" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Organization" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Date" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Time" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Money" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Percentage" varlık isim vektörleri sorgu olarak kullanılacaktır.
  - Haberlerde geçen "Unknown" varlık isim vektörleri sorgu olarak kullanılacaktır.

- Her bir varlık türü için gerçekleştirilen testlere göre sistem başarımını hesaplamak için ikili sınıflandırma tabloları oluşturularak anma, duyarlık ve f-ölçü değerleri belirlenecektir.

4.2.2.4. Varlık İsimlerinin Eşleştirilmesinde Kesişime Bakılan Senaryo

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Sonraki aşamada başarımını belirlemek için aşağıdaki işlemler gerçekleştirilecektir;
    - Haberlerde etiketlenen tüm varlık isimleri türlerine göre (Person", "Location", "Organization", "Date", "Time", "Money", "Percentage", "Unknown") ayrıştırılacaktır.
    - Varlık isimleri üzerinde gövdeleme yapılacaktır.
    - Her bir haber için farklı varklık ismi türlerine göre karşılaştırma (eşleştirme) tabloları oluşturulacaktır.
    - İlgililik değerlendirmesi yapılmış olan 5.872 belge sorgu olarak kabul edilecektir,
    - Varlık ismi türlerine göre farklı haberlerde geçen varlık isimlerinin eşleşip eşleşmediği kontrol edilerek sistem testleri gerçekleştirilecektir.
    - İki farklı haber karşılaştırılırken, aynı varlık ismi türüne ait kelimelerden bir tanesi bile eşleşse (örneğin person varlık ismi türü için haberlerde geçen isimlerin aynı olması gibi) haberler aynı kabul edilecektir.
- Her bir varlık türü için varlık ismi kesişme yöntemi kullanılarak aşağıdaki testler gerçekleştirilecektir;
    - Haberlerde geçen "Person" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Location" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Organization" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Date" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Time" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Money" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
    - Haberlerde geçen "Percentage" varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.

- Haberlerde geçen “Unknown” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.

- Her bir varlık türü için gerçekleştirilen testlere göre, sistem başarımını hesaplamak için, ikili sınıflandırma tabloları oluşturularak anma, duyarlık ve f-ölçü değerleri belirlenecektir.

4.2.2.5. Varlık İsimlerinin Eşleştirilmesinde Fonksiyona Bakılan Senaryo

- Varlık isimlerinin kesişim testi gerçekleştirdikten sonra Benzerlik= (D1 ∩ D2) / (D1 U D2) fonksiyonu (D1 ve D2 haberler kesişen varlık isimlerinin sayısı / D1 ve D2 haberlerde geçen tüm varlık isimlerinin sayısı) benzerlik fonksiyonu kullanılarak haber benzerliklerinin belirlenmesi için sistem testleri oluşturulacaktır.
- Sonraki aşamada başarımını belirlemek için aşağıdaki işlemler gerçekleştirilecektir;
  - Haberlerde etiketlenen tüm varlık isimleri türlerine göre (Person”, ”Location”, ”Organization”, ”Date”, ”Time”, ”Money”, ”Percentage”, “Unknown”) ayrıştırılacaktır.
  - Varlık isimleri üzerinde gövdeleme yapılacaktır.
  - Her bir haber için farklı varklık ismi türlerine göre karşılaştırma (eşleştirme) tabloları oluşturulacaktır.
  - İlgililik değerlendirmesi yapılmış olan 5.872 belge sorgu olarak kabul edilecektir.
  - Varlık ismi türlerine göre farklı haberlerde geçen varlık isimleri benzerlik fonksiyonu kullanılarak eşleştirilip her bir eşleşme için skor değerleri hesaplanacaktır.
  - Her bir varlık ismi türü için, elde edilen skorlar kullanılarak ortalama eşik değeri belirleme yöntemi ile ortalama eşik değerleri başlangıç eşiği olarak belirlenecektir.
  - İkili sınıflandırma tabloları yaratılarak anma ve duyarlık değerleri hesaplanacaktır.
  - Başlangıç eşiği küçük oranlarda değiştirilerek her bir eşik için yeni ikili sınıflandırma tabloları, anma ve duyarlık değerleri hesaplanacaktır.
  - Anma ve duyarlığın en yüksek olduğu noktalar her bir varlık türü için eşik değeri olarak kabul edilecek ve sistem testleri bu eşik değerleri baz alınarak gerçekleştirilecektir.
- Her bir varlık türü aşağıdaki testler benzerlik fonksiyonu yöntemi kullanılarak gerçekleştirilecektir;
  - Haberlerde geçen “Person” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.

- Haberlerde geçen “Location” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Organization” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Date” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Time” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Money” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Percentage” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.
- Haberlerde geçen “Unknown” varlık isim tablolarındaki veri sorgu olarak kullanılacaktır.

- Her bir varlık türü için gerçekleştirilen testlere göre sistem başarımını hesaplamak için ikili sınıflandırma tabloları oluşturularak anma, duyarlık ve f-ölçü değerleri belirlenecektir.

4.2.2.6. Varlık İsimlerinin Eşleştirilmesinde Birlikte Geçme Durumlarına Bakılan Senaryo

- Bu senaryoda varlık isimlerinin ikişer ve üçerli kombinasyonları baz alınarak haber benzerlikleri hesaplanmıştır. Buna göre bir haber içerisinde geçen varlık ismi türlerinden ikisinin ya da üçünün aynı anda eşleşme durumları baz alınarak haberlerin benzer ya da farklı oldukları belirlenmeye çalışılacaktır.
- Sonraki aşamada başarımını belirlemek için aşağıdaki işlemler gerçekleştirilecektir;
  - Haberlerde etiketlenen tüm varlık isimleri türlerine göre (“Person", "Location", "Organization", "Date", "Time", "Money", "Percentage", “Unknown”) ayrıştırılacaktır.
  - Varlık isimleri üzerinde gövdeleme yapılacaktır.
  - Her bir haber için farklı varlık ismi türlerine göre karşılaştırma (eşleştirme) tabloları oluşturulacaktır.
  - İlgililik değerlendirmesi yapılmış olan 5.872 belge sorgu olarak kabul edilecektir.
  - Varlık ismi türlerine göre farklı haberlerde geçen varlık isimlerinin eşleşip eşleşmediği kontrol edilerek sistem testleri gerçekleştirilecektir.
  - İki farklı haber üzerinde aynı varlık ismi türlerinin ikili ve üçlü kombinasyonlarına ait kelimelerden birer tanesi bile eşleşse (örneğin, person-

location kombinasyonunda iki farklı haberdeki birer person ve birer location eşleşmesi gibi) haberler aynı kabul edilecektir.

- Haberler için varlık isimleri kullanarak aşağıda sunulan farklı kombinasyon birleşim tabloları oluşturulacaktır;
    - "Location" ve "Time" varlık isimlerinin kesişmesi
    - "Location" ve "Date" varlık isimlerinin kesişmesi
    - "Location", "Time" ve "Date" varlık isimlerinin kesişmesi
    - "Person" ve "Time" varlık isimlerinin kesişmesi
    - "Person" ve "Date" varlık isimlerinin kesişmesi
    - "Person", "Time" ve "Date" varlık isimlerinin kesişmesi
    - "Organization" ve "Time" varlık isimlerinin kesişmesi
    - "Organization" ve "Date" varlık isimlerinin kesişmesi
    - "Organization" , "Time" ve "Date" varlık isimlerinin kesişmesi
    - "Person" ve "Location" varlık isimlerinin kesişmesi
    - "Person", "Location" ve "Time" varlık isimlerinin kesişmesi
    - "Person", "Location" ve "Date" varlık isimlerinin kesişmesi
    - "Organization" ve "Location" varlık isimlerinin kesişmesi
    - "Organization" , "Location" ve "Time" varlık isimlerinin kesişmesi
    - "Organization" , "Location" ve "Date" varlık isimlerinin kesişmesi
    - "Person", "Organization" ve "Time" varlık isimlerinin kesişmesi
    - "Person" , "Organization" ve "Date" varlık isimlerinin kesişmesi
    - "Person" , "Organization" ve "Location" varlık isimlerinin kesişmesi
- Her bir sorgu için derleme gönderildiğinde varlık isim birleşim tablo kesişim modeli kullanılarak üretilen sorgu-belge kesişim sonuçları çıkarılacaktır.
- Tüm bu sonuçlar içerisinden, sorgunun ilgili olduğu bilinen belgeler için anma/duyarlık ve f-ölçü skor değerleri hesaplanacaktır.

4.2.2.7. Vektör Uzayı Modeli OR Varlık İsim Vektörü Birleşim Senaryosu

Bu aşamada daha önce gerçekleşen iki yöntemin OR mantıksal birleşimiyle elde edilen sonuçlarının başarım testlerinin yapılması hedeflenmektedir. Bunun için daha önce yapılan vektör uzayı modeli ile varlık isimlerinin tamamı ve her bir varlık ismi için bağımsız olarak oluşturulan vektörler kullanılarak gerçekleştirilen eşleşmelerin OR mantıksal bağlacı kullanılarak birleştirilmesi sonucu oluşturulan testler gerçekleştirilecektir.

4.2.2.8. Vektör Uzay Modeli OR Varlık İsim Kesişim Modeli Senaryosu

Bu testte vektör uzayı modelinden elde edilen sonuçlar ve varlık isimlerinin kesişimi yöntemi ile elde edilen sonuçlar OR mantıksal operatörü ile birleştirilerek başarım hesaplanacaktır.

4.2.2.9. Vektör Uzay Modeli OR Varlık İsimlerinin Birlikte Geçme Durumları Senaryosu

Bu testte vektör uzayı modelinden elde edilen sonuçlar varlık isimlerinin birlikte geçme durumlarına göre benzerlik tespiti yöntemi ile elde edilen sonuçlar ile birleştirilerek başarım hesaplaması yapılacaktır.

4.2.3. TT İçin Test Senaryoları

4.2.3.1. Kümeleme Test Senaryosu

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Öncelikle her bir konu eğitim ve test belgeleri olmak üzere iki kısma ayrılacaktır;
    - Her bir konuda var olan belge sayısının üçte biri eğitim (1.931 adet), üçte ikisi de test belgesi (3.941 adet) olarak kabul edilecektir.
    - Eğitim belgeleri seçilirken, tarih sırasına göre derlemdeki ilk N belge seçilecektir, kalan belgeler test belgesi olarak kullanılacaktır.
    - Her bir konu kümesini yaratmak için tarih sırasına göre ilk dört belge kullanılacaktır.
- Her bir konu ile ilgili olarak eğitim belgeleri belirlendikten sonra konu modellerini oluşturmak ve gerekli parametreleri belirlemek için aşağıdaki adımlar gerçekleştirilecektir;
    - Metin kümeleme işleminde k-means algoritması kullanılacaktır.
    - K-means algoritmasının yürütülmesi için gerekli başlangıç merkez vektörleri (centroid), Canopy algoritması kullanılarak belirlenecektir.
    - Başlangıç noktalarına göre k-means algoritması yürütülerek her bir konu için konu kümeleri yaratılacaktır.
    - Her bir konu kümesi için küme centroidleri ve her bir kümeye ait olan doküman vektörleri belirlenecektir.
    - Her bir konu merkezi vektörü ile o konuya ait olan doküman vektörleri arasındaki uzaklıklar CosineSimilarity yöntemi ile belirlenecektir.

- Eşik değeri belirlenirken, eğitim kümesindeki dokümanlar sorgu olarak, en yüksek anma/duyarlık değerindeki eşik değeri sistem eşiği olarak kullanılacaktır.

- Konular için eşik değerleri ve konu kümeleri belirlendikten sonra test derlemi üzerindeki değerlendirmeler aşağıdaki gibi gerçekleştirilecektir;
  - Derlemde bulunan ve ilgililik değerlendirmesi yapılmış olan 3.941 belge test belgesi olarak kullanılacaktır.
  - Test belgesi olarak belirlenen bu 3.941 belge test için sorgu olarak kabul edilecektir.
  - Her bir sorgu daha önce konu modelleri oluşturulmuş olan kümelerle CosineSimilarity yöntemi kullanılarak karşılaştırılacaktır.
  - Bu karşılaştırmalar sonucunda elde edilen uzaklık eğitim aşamasında belirlenen eşik değerden düşük ya da eşitse belge konuyla ilgili kabul edilecektir; aksi durumda belge ilgisizdir.
  - Her bir eşleştirme sonucu için ikili sınıflandırma tablosu yaratılacaktır.
  - Tüm sorgular tamamlandıktan sonra mikro ortalama yöntemi kullanılarak tüm testler için ortak bir ikili sınıflama tablosu oluşturulacaktır.
  - Bu ikili sınıflama tablosuna göre anma ve duyarlık değerleri hesaplanarak sistemin başarımı belirlenecektir.

4.2.3.2. Vektör Uzayı Modeli Test Senaryosu

- Testler ilgililik değerlendirmesi yapılmış olan belgeler (5.872 adet) üzerinden yapılacaktır.
- Öncelikle her bir konu eğitim ve test belgeleri olmak üzere iki kısma ayrılacaktır;
  - Her bir konuda var olan belge sayısının üçte biri eğitim (1.931 adet), üçte ikisi de test belgesi (3.941 adet) olarak kabul edilecektir.
  - Eğitim belgeleri seçilirken, tarih sırasına göre derlemdeki ilk N (seçilen test derleminin üçte birine denk gelecek haber sayısını ifade etmektedir) belge seçilecektir, kalan belgeler test belgesi olarak kullanılacaktır.
  - Her bir konu kümesini yaratmak için tarih sırasına göre ilk dört belge kullanılacakır.
- Derlemde bulunan ve ilgililik değerlendirmesi yapılmış olan 3.941 belge test belgesi olarak kullanılacaktır.

- Test belgesi olarak belirlenen bu 3.941 belge test için sorgu olarak kabul edilecektir.
- Her bir sorgu ilgili konuyu temsil eden dört eğitim belgesi ile Vektör Uzayı yöntemi kullanılarak karşılaştırılacaktır.
- Bu karşılaştırmalar sonucunda elde edilen uzaklık eğitim aşamasında belirlenen eşik değerden düşük ya da eşitse belge konuyla ilgili kabul edilecektir; aksi durumda belge ilgisizdir.
- Her bir eşleştirme sonucu için ikili sınıflandırma tablosu yaratılacaktır.
- Tüm sorgular tamamlandıktan sonra mikro ortalama yöntemi kullanılarak tüm testler için ortak bir ikili sınıflama tablosu oluşturulacaktır.
- Bu ikili sınıflama tablosuna göre anma ve duyarlık değerleri hesaplanarak sistemin başarımı belirlenecektir.

4.3. Testlerin Gerçekleştirilmesi

Oluşturulan test senaryolarına göre her bir senaryonun uygulanması için Java programlama dili kullanılarak ayrı ayrı uygulamalar geliştirilmiştir. Kullanılan yöntemler yeniden kodlanmamış vektör uzayı modeli için Lucene (http://lucene.apache.org/core/), dil modeli için Lemur (http://www.lemurproject.org/) ve kümeleme yöntemleri için Mahout (http://mahout.apache.org/) açık kaynak kodlu kütüphanelerinden yararlanılmıştır.

5. BÖLÜM: BULGULAR VE YORUM

Bu bölümde proje çalışmalarında belirlenen amaçlara ulaşabilmek için oluşturulan test senaryolarına göre gerçekleştirilen sistem testlerinin sonuçları sunularak bu sonuçlar üzerinde elde edilen bulgular yorumlanacaktır.

5.1. SLD İçin Vektör Uzayı ve Dil Modeli Test Sonuçları

Yukarıdaki senaryoya göre vektör uzayı modeli için gerçekleştirilen eğitim ve test sonuçları için elde edilen tablolar aşağıda sunulmuştur. Her bir tabloda bir haberi ifade etmek için seçilen terim sayısı (her bir haber içerisinde tf.idf değeri en yüksek olan ilk N kelime seçilmiştir), testleri gerçekleştirirken ilgililik kararının verildiği eşik değeri (eşik değerin altında kalan benzerlik değerlerinde haberler ilgisiz, üzerinde kalan benzerlik değerlerinde haberler ilgili kabul edilmiştir), ilgili belgelere hangi oranda erişildiğini gösteren anma, erişilen belgelerin ne kadar ilgili olduğunu gösteren duyarlık ve anma-duyarlık değerlerinin harmonik ortalamasını veren f-ölçü değerleri bulunmaktadır.

Tablo 1'de, vektör uzay modelinin eğitim sonuçları sunulmuştur. Gerçekleştirilen eğitim testleri sonucunda en yüksek başarım 0,3558'lik f-ölçü değeri ile, 30 terim için 0,0262 eşik değeri için elde edilmiştir.

Tablo 1. VUM Eğitim Test Sonuçları

VUM Eğitim Sonuçları				
Terim	Eşik	Anma	Duyarlık	F-Ölçü
1	0,5000	0,1535	0,1538	0,1536
2	0,2330	0,2422	0,2429	0,2425
3	0,1810	0,2911	0,2914	0,2912
4	0,1410	0,3107	0,3104	0,3105
5	0,1140	0,3239	0,3239	0,3239
10	0,0600	0,3428	0,3425	0,3426
15	0,0420	0,3503	0,3508	0,3505
20	0,0340	0,3550	0,3547	0,3546
25	0,0293	0,3550	0,3559	0,3554
30	0,0262	0,3556	0,3560	0,3558
35	0,0243	0,3546	0,3542	0,3544
40	0,0231	0,3507	0,3519	0,3513
45	0,0222	0,3473	0,3477	0,3475
50	0,0215	0,3456	0,3443	0,3449
55	0,0210	0,3430	0,3414	0,3422
60	0,0207	0,3387	0,3398	0,3392
65	0,0204	0,3352	0,3363	0,3357
70	0,0201	0,3339	0,3327	0,3333
75	0,0199	0,3317	0,3305	0,3311
80	0,0198	0,3293	0,3293	0,3293
85	0,0197	0,3282	0,3275	0,3278
90	0,0196	0,3274	0,3262	0,3268
95	0,0195	0,3256	0,3243	0,3249
100	0,0195	0,3237	0,3237	0,3237
125	0,0194	0,3186	0,3194	0,3190
150	0,0193	0,3172	0,3169	0,3170
175	0,0192	0,3166	0,3154	0,3160
200	0,0192	0,3163	0,3157	0,3160
225	0,0192	0,3166	0,3159	0,3162
250	0,0192	0,3173	0,3161	0,3167
275	0,0193	0,3165	0,3177	0,3171
300	0,0193	0,3168	0,3178	0,3173
400	0,0193	0,3180	0,3177	0,3177
500	0,0193	0,3180	0,3176	0,3178
1000	0,0194	0,3170	0,3187	0,3178

Tablo 2. VUM Test Sonuçları

VUM Test Sonuçları				
Terim	Eşik	Anma	Duyarlık	F-Ölçü
1	0,5000	0,1385	0,1523	0,1451
2	0,2330	0,1922	0,2222	0,2061
3	0,1810	0,2242	0,2720	0,2458
4	0,1410	0,2442	0,2916	0,2658
5	0,1140	0,2525	0,3006	0,2744
10	0,0600	0,2644	0,3238	0,2911
15	0,0420	0,2665	0,3286	0,2943
20	0,0340	0,2641	0,3378	0,2964
25	0,0293	0,2623	0,3406	0,2964
30	0,0262	0,2642	0,3393	0,2970
35	0,0243	0,2628	0,3365	0,2951
40	0,0231	0,2595	0,3320	0,2913
45	0,0222	0,2577	0,3267	0,2882
50	0,0215	0,2564	0,3208	0,2850
55	0,0210	0,2540	0,3153	0,2814
60	0,0207	0,2516	0,3107	0,2781
65	0,0204	0,2505	0,3067	0,2758
70	0,0201	0,2500	0,3022	0,2737
75	0,0199	0,2489	0,2989	0,2716
80	0,0198	0,2471	0,2963	0,2694
85	0,0197	0,2457	0,2937	0,2675
90	0,0196	0,2453	0,2916	0,2664
95	0,0195	0,2446	0,2893	0,2651
100	0,0195	0,2437	0,2887	0,2643
125	0,0194	0,2406	0,2846	0,2607
150	0,0193	0,2406	0,2822	0,2597
175	0,0192	0,2410	0,2795	0,2589
200	0,0192	0,2412	0,2788	0,2587
225	0,0192	0,2417	0,2787	0,2589
250	0,0192	0,2420	0,2786	0,2590
275	0,0193	0,2411	0,2799	0,2591
300	0,0193	0,2414	0,2803	0,2594
400	0,0193	0,2417	0,2806	0,2597
500	0,0193	0,2420	0,2811	0,2601
1000	0,0194	0,2409	0,2831	0,2603

Tablo 2'de, vektör uzayı modeli için test sonuçları sunulmuştur. Gerçekleştirilen testler sonucunda en yüksek başarım 0,2970'lık f-ölçü değeri ile (anma: 0,2642 ve duyarlık: 0,3393) 30 terim için elde edilmiştir.

Yukarıdaki senaryoya göre dil modeli için gerçekleştirilen eğitim ve test sonuçları ile ilgili tablolar aşağıda sunulmuştur.

Tablo 3. Dil Modeli Eğitim Sonuçları

Dil Modeli (DM) Eğitim Sonuçları				
Terim	Eşik	Anma	Duyarlık	F-Ölçü
1	-5,9200	0,1567	0,1575	0,1571
2	-6,6100	0,2204	0,2200	0,2202
3	-6,7000	0,2162	0,2156	0,2159
4	-6,7800	0,2140	0,2131	0,2135
5	-6,8000	0,2186	0,2177	0,2181
10	-6,7600	0,2064	0,2060	0,2062
15	-6,6400	0,1886	0,1886	0,1886
20	-6,5000	0,1794	0,1781	0,1787
25	-6,3000	0,1652	0,1679	0,1665
30	-6,1300	0,1562	0,1588	0,1575
35	-5,9200	0,1453	0,1453	0,1453
40	-5,7800	0,1413	0,1415	0,1414
45	-5,6600	0,1391	0,1402	0,1396
50	-5,5600	0,1377	0,1389	0,1383
55	-5,5000	0,1376	0,1355	0,1365
60	-5,4400	0,1377	0,1380	0,1378
65	-5,3900	0,1361	0,1383	0,1372
70	-5,3600	0,1383	0,1406	0,1394
75	-5,3300	0,1401	0,1396	0,1398
80	-5,2800	0,1382	0,1394	0,1388
85	-5,2500	0,1383	0,1387	0,1385
90	-5,2200	0,1394	0,1363	0,1378
95	-5,1900	0,1389	0,1405	0,1397
100	-5,1700	0,1405	0,1399	0,1402
125	-5,0400	0,1413	0,1428	0,1420
150	-4,9500	0,1421	0,1404	0,1412
175	-4,8800	0,1409	0,1416	0,1412
200	-4,8100	0,1389	0,1414	0,1401
225	-4,7500	0,1372	0,1362	0,1367
250	-4,7000	0,1341	0,1387	0,1364
275	-4,6600	0,1319	0,1351	0,1335
300	-4,6200	0,1306	0,1296	0,1301
400	-4,5000	0,1181	0,1197	0,1189
500	-4,4400	0,1111	0,1125	0,1118
1000	-4,3800	0,1024	0,1032	0,1028

Tablo 3'te, dil modeli eğitim sonuçları sunulmuştur. Gerçekleştirilen eğitim testleri sonucunda bu yöntemde en yüksek başarım 0,1910'luk f-ölçü değeri ile 4 terim için -6,7800 eşik değerinde elde edilmiştir.

Tablo 4. Dil Modeli Test Sonuçları

DM Test Sonuçları				
Terim	Eşik	Anma	Duyarlık	F-Ölçü
1	-5,9200	0,1322	0,1544	0,1424
2	-6,6100	0,1734	0,2039	0,1874
3	-6,7000	0,1609	0,2179	0,1851
4	-6,7800	0,1625	0,2316	0,1910
5	-6,8000	0,1581	0,2318	0,1880
10	-6,7600	0,1425	0,2439	0,1799
15	-6,6400	0,1310	0,2316	0,1673
20	-6,5000	0,1213	0,2162	0,1554
25	-6,3000	0,1131	0,2015	0,1449
30	-6,1300	0,1079	0,1838	0,1360
35	-5,9200	0,0979	0,1809	0,1270
40	-5,7800	0,0965	0,1773	0,1249
45	-5,6600	0,0941	0,1767	0,1228
50	-5,5600	0,0930	0,1748	0,1214
55	-5,5000	0,0960	0,1623	0,1206
60	-5,4400	0,0980	0,1653	0,1231
65	-5,3900	0,1005	0,1733	0,1272
70	-5,3600	0,1069	0,1795	0,1340
75	-5,3300	0,1115	0,1803	0,1378
80	-5,2800	0,1073	0,1880	0,1366
85	-5,2500	0,1062	0,1899	0,1362
90	-5,2200	0,1057	0,1890	0,1356
95	-5,1900	0,1049	0,1878	0,1346
100	-5,1700	0,1051	0,1898	0,1353
125	-5,0400	0,0949	0,2068	0,1301
150	-4,9500	0,0923	0,1970	0,1257
175	-4,8800	0,0915	0,1864	0,1227
200	-4,8100	0,0896	0,1797	0,1196
225	-4,7500	0,0875	0,1737	0,1164
250	-4,7000	0,0852	0,1691	0,1133
275	-4,6600	0,0830	0,1690	0,1113
300	-4,6200	0,0810	0,1732	0,1103
400	-4,5000	0,0691	0,1783	0,0996
500	-4,4400	0,0646	0,1760	0,0945
1000	-4,3800	0,0563	0,1643	0,0839

Tablo 4'te, dil modeli için test sonuçları sunulmuştur. Gerçekleştirilen testler sonucunda en yüksek başarım 0,1910'luk f-measure değeri ile (anma: 0,1625 ve duyarlık: 0,2316) 4 terim için elde edilmiştir.

Vektör Uzayı ve Dil Modeli Sonuçlarının Birleştirilmesi aşamasında her iki yöntemin de erişim çıktıları incelenmiş sırasıyla OR ve AND işleçleri kullanılarak yeni bir erişim çıktısı yaratılmıştır.

Tablo 5. VUM ve DM Sonuçları Birleştirme (AND)

VUM AND DM Sonuçları			
Terim	Anma	Duyarlık	F-Ölçü
1	0,0807	0,2680	0,1240
2	**0,1587**	0,3475	0,2179
3	0,1474	0,3886	0,2137
4	0,1507	0,4183	**0,2216**
5	0,1469	0,4180	0,2174
10	0,1340	0,4259	0,2038
15	0,1213	0,4251	0,1887
20	0,1103	0,4307	0,1757
25	0,1016	0,4282	0,1642
30	0,0949	0,4255	0,1553
35	0,0860	0,4204	0,1427
40	0,0841	0,4126	0,1397
45	0,0825	0,4065	0,1371
50	0,0818	0,4075	0,1363
55	0,0827	0,4069	0,1374
60	0,0838	0,4212	0,1397
65	0,0855	0,4357	0,1429
70	0,0900	0,4452	0,1497
75	0,0926	0,4597	0,1541
80	0,0899	0,4776	0,1513
85	0,0894	0,4883	0,1512
90	0,0888	0,4966	0,1506
95	0,0875	0,5052	0,1491
100	0,0873	0,5131	0,1493
125	0,0793	0,5642	0,1390
150	0,0745	0,6113	0,1329
175	0,0704	0,6380	0,1269
200	0,0666	0,6592	0,1209
225	0,0632	0,6814	0,1157
250	0,0601	0,7011	0,1107
275	0,0574	0,7154	0,1063
300	0,0549	0,7256	0,1020
400	0,0467	0,7618	0,0880
500	0,0432	0,7795	0,0819
1000	0,0393	**0,7987**	0,0749
MAX	0,1587	0,7987	0,2216
MIN	0,0393	0,2680	0,0749
AVG	0,0890	0,5117	0,1461

Tablo 5'te Vektör Uzayı ve Dil Modeli yöntemlerinin AND birleşiminde ise en yüksek başarım 0.2216'lık f-ölçü değeri ile (anma: 0,1504 ve duyarlık: 0,4183) 4 terim için elde edilmiştir.

Tablo 6. VUM ve DM Sonuçları Birleştirme (OR) Tablosu

VUM OR DM Sonuçları			
Terim	Anma	Duyarlık	F-Ölçü
1	0,1900	0,1297	0,1542
2	0,2068	0,1643	0,1831
3	0,2377	0,2009	0,2178
4	0,2560	0,2172	0,2350
5	0,2636	0,2252	0,2429
10	0,2729	0,2513	0,2617
15	0,2762	**0,2531**	**0,2641**
20	0,2750	**0,2531**	0,2636
25	0,2738	0,2502	0,2615
30	**0,2771**	0,2425	0,2587
35	0,2747	0,2459	0,2595
40	0,2719	0,2423	0,2563
45	0,2694	0,2408	0,2543
50	0,2675	0,2367	0,2512
55	0,2674	0,2240	0,2438
60	0,2659	0,2208	0,2413
65	0,2656	0,2212	0,2413
70	0,2669	0,2187	0,2404
75	0,2677	0,2143	0,2381
80	0,2645	0,2174	0,2386
85	0,2624	0,2165	0,2372
90	0,2622	0,2146	0,2360
95	0,2621	0,2129	0,2349
100	0,2615	0,2130	0,2348
125	0,2562	0,2201	0,2368
150	0,2583	0,2155	0,2349
175	0,2621	0,2109	0,2337
200	0,2643	0,2093	0,2336
225	0,2660	0,2081	0,2335
250	0,2671	0,2076	0,2336
275	0,2667	0,2096	0,2347
300	0,2675	0,2135	0,2374
400	0,2641	0,2224	0,2415
500	0,2634	0,2246	0,2425
1000	0,2579	0,2254	0,2405
MAX	0,2771	0,2531	0,2641
MIN	0,1900	0,1297	0,1542
AVG	0,2615	0,2198	0,2387

Vektör Uzayı ve Dil Modeli Sonuçlarının OR birleşiminde en yüksek başarım 0,2641'lik f-ölçü değeri ile (anma: 0,2762 ve duyarlık: 0,2531) 15 terim için elde edilmiştir (Bkz. Tablo 6).

Tablo 1, 2, 3, 4, 5 ve 6'da yer alan verileri elde etmek için kullanılan yöntem şu şekilde özetlenebilir:

- Her bir yöntemde öncelikli olarak farklı terim sayılarına göre haberleri ifade etmek için en kıymetli kelimeler tf.idf katsayıları kullanılarak belirlenmiştir.
- Farklı terim sayılarına göre seçilen kelimeler eğitim kümesi üzerinde ilgili yöntem (VUM ve DM) kullanılarak test edilmiş ve seçilen terim sayısı için en uygun eşik değer belirlenmiştir.
- Daha sonra bu eşik değeri kullanılarak her bir terim sayısı için sistem testleri gerçekleştirilmiş ve başarım hesaplanmıştır.

Bu yöntem uygulanarak test kümesi üzerinde elde edilen sonuçlara göre başarımın en yüksek olduğu durumlar aşağıdaki gibidir;

Vektör uzayı modeli için gerçekleştirilen testler sonucu bu yöntemde en iyi başarım 0,2970'lik f-ölçü değeri ile (anma: 0,2642 ve duyarlık: 0,3393) 30 terim için elde edilmiştir. Dil modelinde ise en iyi başarım 0,1910'luk f-ölçü değeri ile (anma: 0,1625 ve duyarlık: 0,2316) 4 terim için elde edilmiştir. Bu bağlamda, sorgu oluşturmak için seçilen en uygun terim sayısı yöntemler arasında önemli ölçüde değişiklik göstermiştir. Vektör uzayı modelinde; en etkin sonuçlar terim sayısı 10-55 arasındayken elde edilirken, dil modelinde bu aralık 4-10 terim gibi görünmektedir.

Diğer taraftan yöntemlerin birleştirildiği durumlara bakıldığında beklenen sonuçların görüldüğü söylenebilir. Bu kapsamda OR ile birleştirmede en yüksek başarım 0,2641'lik f-measure değeri ile (anma: 0,2762 ve duyarlık: 0,2531) 15 terim için elde edilmiştir. Buna göre sonuçları OR ile birleştirme; VSM ile en yüksek başarımın elde edildiği 30 terimde anma değerini %1,3, DM ile en yüksek başarımın elde edildiği 4 terimde ise anma değerini %9,35 oranında artırmıştır. Diğer taraftan OR birleşiminde duyarlık değerleri 30 terim VUM için %9,7 ve 4 terim DM için %1,44 düşmüştür.

AND birleşimlerine bakıldığında ise en yüksek başarımın 0.2216'lık f-measure değeri ile (anma: 0,1504 ve duyarlık: 0,4183) 4 terim için sağlandığı görülmektedir. AND birleşiminde VUM ile en yüksek başarımın elde edildiği 30 terimde duyarlık değerini %8,62, DM ile en yüksek başarımın elde edildiği 4 terimde ise duyarlık değerini %18,67 oranında artırmıştır. Buna karşılık AND birleşiminde anma değerleri 30 terim VUM için %16,92 ve 4 terim DM için %1,18 düşmüştür.

5.2. SLD İçin Varlık İsimlerinin Kullanıldığı Test Sonuçları

SLD için varlık isimlerinin kullanıldığı durumda başarım hesaplaması yapabilmek ve sonuçları karşılaşatırabilmek amacıyla aynı test senaryoları uygulanarak varlık isimleri kullanılmadan temel VUM yöntemi ile gerçekleştirilen testlerde 0,59 (anma= 0,58 ve duyarlık= 0,61) f-ölçü değeri elde edilmiş ve bu değer diğer tüm yöntemlerle karşılaştırılırken baz yöntem olarak kabul edilmiştir. Bu bağlamda uygulanacak olan varlık ismi yöntemlerinden ede edilen başarım değerleri bu 0,59 f-ölçü değerinin üzerinde yer alırsa bilgi erişim başarımının artmış olduğu ve daha etkin bir bilgi erişim performansı elde edildiği sonucu çıkacaktır.

Haberler içerisinde geçen varlık isimlerinin vektörler ile ifade edildiği ve bu varlık isim vektörlerinin VUM yöntemi kullanılarak gerçekleştirilen testlerde ise en yüksek başarım 0,67 (anma= 0,55 ve duyarlık= 0,89) f-ölçü değeri ile “unknown” etiketli varlık isimlerinde elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %8’lik bir başarım artışı anlamına gelmektedir.

Haberler içerisinde geçen varlık isimlerinin eşleştirilmesinde bir erişim fonksiyonu kullanmak yerine doğrudan eşleşmelere bakılan testlerde ise en yüksek başarım 0,89 (anma=0,91 ve duyarlık=0,87) f-ölçü değeri ile yine “unknown” etiketli varlık isimlerinde elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %30’luk, vektörlerin kullanıldığı yönteme göre de %22’lik bir başarım artışı anlamına gelmektedir.

Varlık isimlerinin eşleştirilmesinde erişim fonksiyonunun kullanıldığı yöntemde ise en yüksek başarım 0,72 (anma= 0,59 ve duyarlık= 0,94) f-ölçü değeri ile yine “unknown” etiketli varlık isimlerinde elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %13’lük bir başarım artışı anlamına gelmektedir.

Varlık isimlerinin eşleştirilmesinde birlikte geçme durumlarına bakılan yöntemde ise en yüksek başarım 0,76 (anma= 0,80 ve duyarlık= 0,72) f-ölçü değeri ile *person-location* eşleşmesinde elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %17’lik bir başarım artışı anlamına gelmektedir.

Baz VUM yöntemi ile varlık isim vektörlerinin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0,64 (anma= 0,62 ve duyarlık= 0,65) f-ölçü değeri ile VUM OR Person ve VUM OR Organization eşleşmeleri için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %5’lik bir başarım artışı anlamına gelmektedir.

Baz olarak kabul edilen VUM yöntemi ile varlık isimlerinin kesişim modelinin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0,89 (anma= 0,87 ve duyarlık= 0,90) f-ölçü değeri ile VUM OR Tüm Varlık İsimlerinin Kesişimi için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %30’luk bir başarım artışı anlamına gelmektedir.

Baz olarak kabul edilen VUM yöntemi ile varlık isimlerinin birlikte geçme durumlarına göre benzerlik tespiti yapılan modelin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0.72 (anma= 0,70 ve duyarlık= 0,73) f-ölçü değeri ile VUM OR Person-Location birleşimi için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %13'lük bir başarım artışı anlamına gelmektedir.

Tablo 7. Varlık İsimleri Olmadan Vektör Uzayı Modeli Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik-Değer
VUM	0,61	0,58	0,59	0,05

Tablo 8. Tüm Varlık İsimlerinin Tek Vektörle İfade Edildiği ve Farklı Varlık İsimlerinin Farklı Vektörlerle İfade Edildiği Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik-Değer
Tüm Varlık İsimler Vektörü	0,65	0,56	0,60	0,02
"Person" varlık isim vektörü	0,77	0,39	0,51	0,01
"Location" varlık isim vektörü	0,51	0,51	0,51	0,25
"Organization" varlık isim vektörü	0,46	0,46	0,46	0,01
"Date" varlık isim vektörü	0,33	0,36	0,34	0,20
"Time" varlık isim vektörü	0,34	0,33	0,34	0,25
"Money" varlık isim vektörü	0,49	0,28	0,36	0,01
"Percentage" varlık isim vektörü	0,76	0,22	0,34	0,01
"Unknown" varlık isim vektörü	**0,89**	**0,55**	**0,67**	**0,01**

Tablo 9. Varlık İsimlerinin Eşleştirilmesinde Kesişime Bakılan Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü
Tüm varlık isimlerinin kesişim testi	0,23	0,98	0,37
"Person" kesişim testi	0,63	0,85	0,72
"Location" kesişim testi	0,25	0,97	0,39
"Organization" kesişim testi	0,39	0,88	0,54
"Date" kesişim testi	0,22	0,98	0,35
"Time" kesişim testi	0,29	0,57	0,38
"Money" kesişim testi	0,46	0,60	0,52
"Percentage" kesişim testi	0,77	0,55	0,64
"Unknown" kesişim testi	**0,87**	**0,91**	**0,89**

Tablo 10. Varlık İsimlerinin Eşleştirilmesinde Erişim Fonksiyonu Kullanılan Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik-Değer
Bütün varlık isimlerin benzerlik tespiti	0,98	0,13	0,22	0,02
“Person” varlık isminin benzerlik tespiti	0,95	0,32	0,48	0,01
“Location” varlık isminin benzerlik tespiti	0,59	0,62	0,61	0,01
“Organization” varlık isminin benzerlik tespiti	0,78	0,24	0,37	0,01
“Date” varlık isminin benzerlik tespiti	0,41	0,37	0,39	0,01
“Time” varlık isminin benzerlik tespiti	0,36	0,29	0,32	0,01
“Money” varlık isminin benzerlik tespiti	0,96	0,30	0,46	0,01
“Percentage” varlık isminin benzerlik tespiti	0,62	0,11	0,18	0,01
“Unknown” varlık isminin benzerlik tespiti	**0,94**	**0,59**	**0,72**	**0,01**

Tablo 11. Varlık İsimlerinin Eşleştirilmesinde Birlikte Geçme Durumlarına Bakılan Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü
Location-Time	0,69	0,17	0,27
Location-Date	0,44	0,80	0,56
Location-Time-Date	0,86	0,14	0,24
Person-Time	0,97	0,16	0,28
Person-Date	0,84	0,52	0,64
Person-Time-Date	0,95	0,14	0,25
Organization-Time	0,79	0,16	0,26
Organization-Date	0,66	0,54	0,59
Organization-Time-Date	0,94	0,13	0,22
Person-Location	**0,72**	**0,80**	**0,76**
Person-Location-Time	0,94	0,15	0,25
Person-Location-Date	0,90	0,50	0,64
Organization-Location	0,57	0,81	0,66
Organization-Location-Time	0,98	0,16	0,27
Organization-Location-Date	0,76	0,51	0,61
Person-Organization-Time	0,96	0,15	0,25
Person-Organization-Date	0,97	0,41	0,56
Person-Organization-Location	0,89	0,64	0,74

Tablo 12. Vektör Uzay Modeli OR Varlık İsim Vektörü Birleşim Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik Değer
Vektör uzay model OR tüm varlık isimler vektörü	0,61	0,58	0,59	0,05
Vektör uzay model OR Person varlık isim vektörü	**0,65**	**0,62**	**0,64**	**0,05**
Vektör uzay model OR Location varlık isim vektörü	0,62	0,59	0,60	0,05
Vektör uzay model OR Organization varlık isim vektörü	**0,65**	**0,62**	**0,64**	**0,05**
Vektör uzay model OR Date varlık isim vektörü	0,63	0,60	0,62	0,05
Vektör uzay model OR Time varlık isim vektörü	0,61	0,58	0,59	0,05
Vektör uzay model OR Money varlık isim vektörü	0,61	0,58	0,59	0,05
Vektör uzay model OR Percentage varlık isim vektörü	0,61	0,58	0,59	0,05
Vektör uzay model OR Uknown varlık isim vektörü	0,62	0,58	0,59	0,05

Tablo 13. Vektör Uzay Modeli OR Varlık İsim Kesişim Modeli Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik Değer
VUM OR tüm varlık isimler kesişimi	**0,90**	**0,87**	**0,89**	**0,05**
VUM OR Person varlık isim kesişimi	0,75	0,72	0,73	0,05
VUM OR Location varlık isim kesişimi	0,83	0,80	0,81	0,05
VUM OR Organization varlık isim kesişimi	0,77	0,74	0,76	0,05
VUM OR Date varlık isim kesişimi	0,75	0,71	0,73	0,05
VUM OR Time varlık isim kesişimi	0,71	0,68	0,69	0,05
VUM OR Money varlık isim kesişimi	0,71	0,68	0,69	0,05
VUM OR Percentage varlık isim kesişimi	0,71	0,68	0,69	0,05
VUM OR Unknown varlık isim kesişimi	0,71	0,68	0,69	0,05

Tablo 14. Vektör Uzay Modeli OR Varlık İsimlerinin Birlikte Geçme Durumlarına Göre Benzerlik Tespiti Test Sonuçları

Yöntem	Duyarlık	Anma	F-Ölçü	Eşik Değer
VUM OR Location-Time	0,71	0,68	0,69	0,05
VUM OR Location-Date	0,73	0,70	0,71	0,05
VUM OR Location-Time-Date	0,71	0,68	0,69	0,05
VUM OR Person-Time	0,71	0,68	0,70	0,05
VUM OR Person-Date	0,71	0,68	0,70	0,05
VUM OR Person-Time-Date	0,71	0,68	0,69	0,05
VUM OR Organization-Time	0,71	0,68	0,69	0,05
VUM OR Organization-Date	0,71	0,68	0,69	0,05
VUM OR Organization-Time-Date	0,71	0,68	0,69	0,05
VUM OR Person-Location	**0,73**	**0,70**	**0,72**	**0,05**
VUM OR Person-Location-Time	0,71	0,68	0,69	0,05
VUM OR Person-Location-Date	0,71	0,68	0,70	0,05
VUM OR Organization-Location	0,71	0,68	0,69	0,05
VUM OR Organization-Location-Time	0,71	0,68	0,69	0,05
VUM OR Organization-Location-Date	0,71	0,69	0,70	0,05
VUM OR Person-Oragnization-Time	0,71	0,68	0,69	0,05
VUM OR Person-Organization-Date	0,71	0,68	0,69	0,05
VUM OR Person-Organization-Location	0,71	0,69	0,70	0,05

5.3. TT İçin Test Sonuçları

Hikaye izleme (Topic Tracking) görevi kapsamında gerçekleştirilen testlerde 0,7677'lık F-ölçü değeri ile kümeleme yöntemi, 0,7335'lık F-ölçü değeri elde edilen vektör uzayı modelinden daha başarılı sonuçlar üretmiştir. Diğer taraftan, kümeleme yöntemi uygulanmadan önce eşik değerinin belirlenmesinde farklı yaklaşımlar test edilmiştir. Gerçekleştirilen testlerde anma/duyarlık değerinin en yüksek olduğu değerin eşik olarak kullanılması diğer yöntemler ile karşılaştırıldığında en başarılı yöntem olarak öne çıkmaktadır.

5.3.1. Kümeleme Test Sonuçları

Hikaye İzleme (Topic Tracking) görevi kapsamında gerçekleştirilen testlerde, her bir konu ile ilgili olarak eğitim kümesindeki ilk dört doküman kullanılarak konu kümeleri yaratılmış ve her bir yeni haberin bu konuyla ilgili olup olmadığı anlaşılmaya çalışılmıştır. Ayrıca farklı eşik belirleme yöntemlerinin başarım üzerindeki etkileri belirlenmeye çalışılmıştır. Sonuçlar Tablo 16'da gösterilmektedir ve sütunların anlamları şu şekildedir;

A: Eşik değeri, eğitim kümesinde anma/duyarlık değerlerinin en yüksek olduğu noktada seçilmiştir.

B: Eşik değeri, tüm konular için küme merkezi vektörlerine en uzak olan dokümanın mesafesi alınarak seçilmiştir.

C: Eşik değeri, her bir konu merkezi vektörüne o konuyla ilgili dokümanlardan en uzak olanlarının mesafelerinin ortalamaları alınarak seçilmiştir.

D: Eşik değeri, her bir konu için ayrı ayrı hesaplanmıştır. Her bir konunun merkezi vektörüne en uzak olan o konuyla ilgili dokümanın mesafesi eşik olarak seçilmiştir.

E: Küme merkezi vektörlerine en yakın dokünanın mesafesi eşik olarak seçilmiştir.

Tablo 15. Kümeleme Yöntemi Test Sonuçları

	EĞİTİM	TEST
Eşik	0,8660	0,8660
Anma	0,7929	0,6686
Duyarlık	0,7941	0,9012
F-Ölçü	0,7935	0,7677
Sorgu sayısı	1,9310	3,9410

Tablo 16. Farklı Eşik Belirleme Yöntemlerinin Başarım Üzerinde Etkileri

	A	B	C	D	E
Eşik	0,8660	0,9975	0,7043	Dinamik	0,4308
Anma	0,6686	1,0000	0,1355	0,8206	0,0090
Duyarlık	0,9012	0,0130	0,9963	0,2650	1,0000
F-Ölçü	0,7677	0,0257	0,2386	0,4006	0,0178
Sorgu sayısı	3,9410	3,9410	3,9410	3,9410	3,9410

5.3.2. Vektör Uzayı Modeli Test Sonuçları

Tablo 17'de vektör uzayı modeli test sonuçları sunulmaktadır.

Tablo 17. Vektör Uzayı Modeli Test Sonuçları

	EĞİTİM	TEST
Eşik	0,8754	0,8754
Anma	0,7628	0,6341
Duyarlık	0,7632	0,8698
F-Ölçü	0,7630	0,7335
Sorgu sayısı	1,9310	3,9410

6. BÖLÜM: SONUÇ VE TARTIŞMA

Bu projede TDT programında tanımlı Hikaye Bağlantı Algılama (Story Link Detection - SLD) görevinin Türkçe bir derlem üzerinde farklı erişim fonksiyonları ve bunların kombinasyonları kullanılarak başarımının test edilmesi ve optimum anma/duyarlık değerlerini sağlayacak kombinasyonun bulunması hedeflenmiştir. Bu kapsamda proje çalışmaları, ilgili derlemin etiketlenmesi, test senaryolarının oluşturulması ve gerekli yazılımların geliştirilmesi ile testlerin uygulanması olarak üç adımda yürütülmüştür.

Sistem testlerinin gerçekleştirilebilmesi için BilCol – 2005 derleminde konu başlıkları bilinen 5.872 adet haber Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü son sınıf ve yüksek lisans öğrencileri tarafından incelenerek bu haberler içerisindeki varlık isimleri etiketlenmiştir. Bu etiketleme çalışması sonunda haberler içerisinde; 45.201 Person, 35.255 Location, 29.059 Organization, 10.622 Date, 1.118 Time, 2.708 Money, 2.608 Percentage ve 10.258 Unknown etiketi oluşturulmuş ve derlem sistem testlerini yapmaya hazır hale getirilmiştir.
Sonraki aşamada proje önerisinde belirlenen hedeflere ulaşabilmek için test senaryoları oluşturulmuş ve her bir senaryoyu uygulayabilmek için Java programlama dili ve yöntemlerle ilgili açık kaynak kodlu kütüphaneler kullanılarak gerekli yazılımlar geliştirilmiştir.
Oluşturulan senaryolara uygun olarak gerçekleştirilen test çalışmaları sonunda, proje önerisindeki hedefler kapsamında, aşağıdaki sonuçlara ulaşılmıştır:

Hedef 1: Vektör uzayı yöntemi (VUM) kullanılarak haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: SLD için vektör uzayı yöntemi ile ilgili testler iki farklı senaryo kullanılarak gerçekleştirilmiştir. Bunlardan birincisinde; BilCol-2005 derleminde bulunan 209.206 haberin tamamı kullanılmış ve en iyi başarım 0,2970'lik f-ölçü değeri ile (anma: 0,2642 ve duyarlık: 0,3393) 30 terim için elde edilmiştir.
İkinci senaryoda ise derlem üzerinden varlık isimlerinin işaretlendiği 5.872 haber üzerinde testler gerçekleştirilmiş ve 0,59 (anma= 0,58 ve duyarlık= 0,61) f-ölçü değeri elde edilmiştir.
İkinci senaryo kullanıldığında yöntem değişmemiş olmasına rağmen başarımın oldukça yükseldiği görülmüştür. Bu beklenmeyen başarım artışının nedeni ikinci senaryoda kullanılmayan yaklaşık 204 bin haber olarak düşünülmektedir. Hangi konu ile ilgili olduğu bilinmeyen bu haberler testler esnasında doğal olarak hiçbir konu ile ilgili olmayan haberler olarak düşünülmüştür. Kanımızca bu haberler içerisinde ciddi miktarlarda kirli veri ve muhtemelen takip edilen konularla ilgili haberler bulunmaktadır.

Hedef 2: Dil (ilgi) modeli (DM) yöntemi kullanılarak haber benzerliklerindeki anma duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: SLD için DM ile ilgili testlerde BilCol-2005 derleminde bulunan 209.206 haberin tamamı kullanılmış ve en iyi 0,1910'luk f-ölçü değeri ile (anma: 0,1625 ve

duyarlık: 0,2316) 4 terim için elde edilmiştir. Bu sonuçlar dil modeli ile elde edilen başarımın VUM'a göre yaklaşık %10 daha düşük olduğunu göstermektedir. Literatürdeki çalışmalarda genellikle DM'nin VUM'den daha yüksek başarım sağladığı görülmesine rağmen bu çalışmada başarımın düşük kalması uygulanan yöntemin ayrıntılarında gizlidir. Benzer çalışmalarda uygulanan DM yaklaşımlarında haberler sorgu genişletme yöntemleri ile iyileştirilerek kullanılırken bu çalışmada bu tür bir ön işlem gerçekleştirilmemiştir. Bu nedenle DM'nin VUM'den daha düşük bir başarım sağlaması normal olarak karşılanmıştır.

Hedef 3: VUM ve DM, OR mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi

Elde Edilen Sonuçlar: OR ile birleştirmede en yüksek başarım 0,2641'lik f-ölçü değeri ile (anma: 0,2762 ve duyarlık: 0,2531) 15 terim için elde edilmiştir. Buna göre sonuçları OR ile birleştirme; VUM ile en yüksek başarımın elde edildiği 30 terimde anma değerini %1,3, DM ile en yüksek başarımın elde edildiği 4 terimde ise anma değerini %9,35 oranında artırmıştır. Diğer taraftan OR birleşiminde duyarlık değerleri 30 terim VUM için %9,7 ve 4 terim DM için %1,44 düşmüştür.

Hedef 4: VUM ve DM, AND mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: AND birleşimlerine bakıldığında ise en yüksek başarımın 0,2216'lık f-ölçü değeri ile (anma: 0,1504 ve duyarlık: 0,4183) 4 terim için sağlandığı görülmektedir. AND birleşiminde VUM ile en yüksek başarımın elde edildiği 30 terimde duyarlık değerini %8,62, DM ile en yüksek başarımın elde edildiği 4 terimde ise duyarlık değerini %18,67 oranında artırmıştır. Buna karşılık AND birleşiminde anma değerleri 30 terim VUM için %16,92 ve 4 terim DM için %1,18 düşmüştür.

Hedef 5: Haberlerdeki tüm varlık isimleri kullanılarak (who, where, when) haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: Haberler içerisinde geçen tüm varlık isimlerinin vektörler ile ifade edildiği ve bu varlık isim vektörlerinin VUM yöntemi kullanılarak gerçekleştirilen testlerde ise en yüksek başarım 0,67 (anma= 0,55 ve duyarlık= 0,89) değeri f-ölçü değeri ile "unknown" etiketli varlık isimlerinde elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %8'lık bir başarım artışı anlamına gelmektedir.

Hedef 5: Haberlerdeki tüm varlık isimleri kullanılarak (who, where, when) haber farklılıklarının belirlenmesindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: Haber farklılıklarının ya da birbirinden farklı haber çiftlerinin belirlenebilmesi için farklı yöntemlerden elde edilen ve anma değerinin en yüksek olduğu yöntemin tespit edilebilmesi gerekmektedir. Bu kapsamda sadece tüm varlık isimlerinin kullanıldığı yöntem yerine tüm yöntemlerin sonuçlarının değerlendirilmesinin daha uygun

olacağı düşünülmüştür. Buna göre; varlık isimlerinin vektörlerle ifade edildiği ve erişim fonksiyonu olarak VUM kullanılan testlerde en yüksek anma 0,56'lık değer ile tüm varlık isimlerinin tek vektör olarak ifade edildiği senaryoda elde edilmiştir. Haber benzerliklerinin belirlenmesinde; varlık isimlerinin kesişimine bakılan yöntemde en yüksek anma 0,98 ile tüm varlık isimlerinin kesişimine bakıldığı yöntemde, benzerlik fonksiyonu kullanılarak uygulanan yöntemde 0i62 ile Location eşleşmelerinde, varlık isimlerinin birlikte geçme durumlarına bakılan yöntemde anma 0,80 ile Location-Date ve Person-Location kombinasyonlarında, baz VUM OR Varlık İsim Vektörü birleşiminde anma 0.62 ile VUM OR Person ve VUM OR Organization kombinasyonlarında, VUM OR Varlık İsim Kesişim birleşiminde anma 0.87 ile VUM OR Tüm Varlık İsimleri Kesişiminde ve VUM OR Varlık İsimlerinin Birlikte Geçme Durumlarına Göre Benzerlik Tespiti yönteminde anma 0.70 ile VUM OR Location-Date ve VUM OR Person-Location kombinasyonlarında elde edilmiştir.

Bu sonuçlara göre anma değerinin en yüksek elde edildiği (anma= 0,98) tüm varlık isimlerinin kesişimine bakıldığı yöntem kullanılması halinde iki haberin farklı konularda olduğunun tespiti işlemi %98 başarı ile gerçekleştirilebilmektedir. Diğer taraftan sadece Location-Date veya Person-Location birlikte geçme durumlarına bakılarak da haberlerin %80 oranında farklı konularda olduklarını belirlemek mümkündür.

Hedef 6: VUM, DM ve varlık isimleri yöntemlerinin sonuçları OR mantıksal operatörü ile birleştirilerek haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: Proje önerisinde bu testler esnasında dil modelinin de kullanılması öngörülmüş olmasına rağmen varlık isimleri çıkarıldıktan sonra sadece bu bilgiler kullanılarak haberler için dil modellerinin oluşturulması bu modelin doğası gereği mümkün olmamıştır. Dil Modeli dışında diğer yöntemler uygulanarak elde edilen sonuçlar şu şekildedir;

Baz VUM yöntemi ile varlık isim vektörlerinin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0.64 (anma= 0,62 ve duyarlık= 0,65) f-ölçü değeri ile VUM OR Person ve VUM OR Organizastion eşleşmeleri için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %5'lik bir başarım artışı anlamına gelmektedir.

Baz olarak kabul edilen VUM yöntemi ile varlık isimlerinin kesişim modelinin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0,89 (anma= 0,87 ve duyarlık= 0,90) f-ölçü değeri ile VUM OR Tüm Varlık İsimlerinin Kesişimi için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %30'luk bir başarım artışı anlamına gelmektedir.

Baz olarak kabul edilen VUM yöntemi ile varlık isimlerinin birlikte geçme durumlarına göre benzerlik tespiti yapılan modelin OR mantıksal operatörü ile birleşiminin kullanıldığı yöntemde ise en yüksek başarım 0,72 (anma= 0,70 ve duyarlık= 0,73) f-ölçü değeri ile VUM OR Person-Location birleşimi için elde edilmiştir. Bu baz olarak alınan VUM yöntemine göre %13'lük bir başarım artışı anlamına gelmektedir.

Hedef 7: VUM ve DM yöntemlerinin sonuçları OR mantıksal operatörü ile birleştirilirken NE tarafından tespit edilemeyen haberlerin ilgisiz olarak işaretlenmesi ile haber benzerliklerindeki anma, duyarlık ve f-ölçü değerlerinin belirlenmesi.

Elde Edilen Sonuçlar: Bir önceki hedefte olduğu gibi bu testlerde de dil modeli kullanılamamıştır. Bu kapsamda yükarıda ifade edilmiş olan hedef aslında VUM ve varlık isimlerinin AND birleşimlerine işaret etmektedir. Bu kapsamda gerçekleştirilen testler sonucunda 0,52 (anma= 0,50 ve duyarlık= 0,53) f-ölçü değeri elde edilmiştir.

KAYNAKÇA

Allan, J. (2002). Introduction to topic detection and tracking. *Topic detection and tracking: Event-based information organization*, J. Allan, ed., Kluwer Academic Publishers, pp. 1-16.

Allan, J., Lavrenko, V. ve Jin, H., (2000). First story detection in TDT is hard. In P*roceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*, pp. 374-381.

Allan, J., Lavrenko, V. ve Swan, R. (2002). Explorations within topic tracking and detection. *Topic detection and tracking: Event-based information organization*, J. Allan, ed., Kluwer Academic Publishers, pp. 197-224.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. ve Yang, Y. (1998). Topic detection and tracking pilot study: Final report. *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 194-218.

Bayraktar, Ö. ve Taşkaya-Temizel, T. (2008). Person name extraction from Turkish financial news text using local grammar based approach. In *Proceedings of the International Symposium on Computer and Information Science (ISCIS)*.

Berger, A. ve Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, pp. 222-229.

Bikel, D.M., Schwartz, R.M. ve Weischedel, R.M. (1999). An algorithm that learns what's in a name. In *Proceedings of Machine Learning*, pp. 211-231.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C. ve Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, 61(4), 802-819.

Dalkılıç, F.E., Gelişli, S. ve Diri, B. (2010). Türkçe kural tabanlı varlık ismi tanıma. *18. Sinyal İşleme ve Uygulama Kurultayı, Diyarbakır, (22-24 Nisan) 2010*.

Frakes, W. ve Baeza Yates, R. (1992). *Information retrieval: Data structure and algorithm, clustering algorithm*. Prentice-Hall, Englewood Cliffs.

Hatzivassiloglou, V., Gravano, L. ve Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'00)* (pp. 224-231). Athens, Greece: ACM.

Jin, Y., Myaeng, S.H., Lee, M., Oh, H. ve Jang, M. (2005). Effective use of place information for event tracking. *Lecture Notes in Computer Science*, 3689(2005), 410-422.

Kim, P. ve Myaeng, S.H. (2004). Usefulness of temporal information automatically extracted from news articles for topic tracking. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4), 227-242.

Köse, G. (2004). *Konu algılama ve izleme programında olay modeli*. Yayımlanmamış Yüksek Lisans Tezi. Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Kumaran, G. ve Allan, J. (2004). Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'04)* (pp. 297-304). Sheffield, UK: ACM.

Kumaran, G., Allen, J. ve McCallum, A. (2004). Classification models for new event detection. 13 Ocak 2014 tarihinde http://works.bepress.com/cgi/viewcontent.cgi?article=1059&context=andrew_mccallum adresinden erişildi.

Kumaran, G. ve Allan, J. (2005). Using names and topics for new event detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*.

Küçük, D. ve Yazıcı, A. (2009a). Rule-based named entity recognition from Turkish texts. *International Symposium on INovations in Intelligent Systems and Applications. Trabzon, Turkey, June 29-July 1, 2009*.

Küçük, D. ve Yazıcı, A. (2009b). Named entity recognition experiments on Turkish texts. *In Proceedings of the International Conference on Flexible Query Answering Systems. Roskilde, Denmark*. T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 524-535.

Küçük, D. ve Yazıcı, A. (2010). A hybrid named entity recognizer for Turkish with applications to different text genres. In *Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS)*. London, UK. E. Gelenbe et al. (Eds.): Computer and Information Sciences, LNEE 62, pp. 113-116.

Lavrenko, V. ve Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01*. ACM, New York, NY, 120-127.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V. ve Thomas, S. (2002). Relevance models for topic detection and tracking, *Proceedings of the Human Language Technology Conference (HLT)*, 104-110.

Leek, T., Schwartz, R. ve Sista, S. (2002). Probabilistic approaches to topic detection and tracking. T*opic detection and tracking: Event-based information organization*, J. Allan, ed., Kluwer Academic Publishers, 67-83.

Makkonen, J., Ahonen-myka, H. ve Salmenkivi, M. (2002). Applying semantic classes in event detection and tracking. In *Proceedings of International Conference on Natural Language Processing (ICON'02)*, pp. 175-183.

Makkonen, J., Ahonen-myka, H. ve Salmenkivi, M. (2003). Topic detection and tracking with spatio-temporal evidence. In *Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003),* pp. 251-265.

Maron, M. E. (1988). Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing & Management*, 24(3), 249-255.

Maron, M.E. ve Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7, 216-244.

Meadow, C. T. (1992). T*ext information retrieval systems*. San Diego: Academic Press.

Miller, D., Leek, T. ve Schwartz, R. (1999). A hidden markov model information retrieval system. In *Proceedings on the 22nd Annual International ACM SIGIR Conference,* pp. 214–221.

Mori, M., Miura, T. ve Shioya, I. (2006). Topic detection and tracking for news web pages. In W*I '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 338-342.

Ponte, J. M. ve Croft, W. B. (1998). A language modeling approach to information retrieval. In P*roceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98*. ACM, New York, NY, 275-281.

Ponte, J. ve Croft, W. B. (1997). Text segmentation by topic, In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL*), pp. 113-125.

Ponte, J. ve Croft, W. B. (1998). A language modeling approach to information retrieval, In *Proceedings on the 21st Annual International ACM SIGIR Conference*, pp. 275–281.

Qiu, J. ve Liao, L.J. (2008). Add temporal information to dependency structure language model for topic detection and tracking. *Machine Learning and Cybernetics*, 1575 – 1580.

Qiu, J., Liao, L.J. ve Dong, X.J. (2008). Topic detection and tracking for Chinese news web pages. *Advanced Language Processing and Web Information Technology, ALPIT '08*, 114-120.

Robertson, S.E. (1977). Theories and models in information retrieval. *Journal of Documentation*, 33, 126-148.

Salton, G. (1989). *Automatic text processing*. Addison-Wesley. Reading, Mass.

Salton, G. ve McGill, M.J. (1983). *Introduction to modern information retrieval*. McGraw Hill Book Co., New York.

Salton, G., Wong, A. ve Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*. 18(11), 613–620.

Schultz, J.M. ve Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient, *DARPA Broadcast News Workshop Proceedings*.

Schultz, J.M. ve Liberman, M. (2002). Towards a universal dictionary for multi-language IR applications. *Topic detection and tracking: Event-based information organization*, J. Allan, ed., Kluwer Academic Publishers, pp. 225-239.

Shah, C., Croft, W. B. ve Jensen, D. (2006). Representing documents with named entities for story link detection (SLD). A poster presentation at the *ACM Fifteenth Conference on Information and Knowledge Management (CIKM) 2006, Arlington VA, November 6-11, 2006*.

Song, F. ve Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, pp. 279–280.

Sparck Jones, K., Walker, S. ve Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, 36(2), 809-840.

Thompson, K.C. ve Callan, J. (2005.) Query expansion using random walk models. In *Proceedings of the Fourteenth International Conference on Information and Knowledge Management (CIKM'05). ACM*.

Tonta, Y., Bitirim. Y. ve Sever. H. (2002). *Türkçe arama motorlarında performans değerlendirme*. Ankara: Total Bilişim Ltd. Sti.

Tür, G., Hakkani-Tür, D. ve Oflazer, K., (2003). A statistical information extraction system for Turkish. *Natural Language Engineering*. 9(2), 181-210.

Uyar, E. (2009). *Near-duplicate news detection using named entities*. Master Thesis, Computer Engineering Department, Bilkent University. 10 Ocak 2014 tarihinde http://www.cs.bilkent.edu.tr/~canf/bilir_web/theses/erkanUyarThesis.pdf adresinden erişildi.

Xu, J. ve Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1), 79-112.

Yang, Y., Carbonell, J., Brown, R., Lafferty, J., Pierce, T., & Ault, T. (2002). Multi-strategy learning for topic detection and tracking. In J. Allan (Ed.), *Topic detection and tracking: Event-based information organization,* (pp. 85-114). Norwell, MA: Kluwer Academic Publishers.

EK. PROJE KAPSAMINDAKİ ÇALIŞMALAR

Proje kapsamında bir dizi çalışma gerçekleştirimiştir ve lisansüstü tez çalışmaları da yakın zamanda tamamlanacaktır. Bu çalışmaların başlıkları aşağıda listelenmektedir.

- Köse, G., & Ahmadlouei, H. Supervised news classification based on a large-scale news corpus. *Beyond the Cloud: Information…Innovation…Collaboration…: 4th International Symposium on Information Management in a Changing World, September 4-6, 2013, Limerick, Ireland. Abstracts*. içinde s. 46-49. Ankara: Hacettepe University Department of Information Management.
- Köse, G., Tonta, Y., Polatkan, A.C. & Ahmadlouei, H. (2013). Story link detection in Turkish Corpus, The 2013 IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 17-20, 2013 Atlanta GA USA. (http://doi.ieeecomputersociety.org/10.1109/WI-IAT.2013.23)
- Doktora Çalışması (Tez Aşamasında): Güven Köse, *Sınırlı Alanlarda Konu Tespit Ve Takibi İçin Genişletilmiş Bir Mimari Yapı Önerisi*. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi A.B.D.
- Yüksek Lisans Çalışması (Tez Aşamasında): Hamid Ahmadlouei, *Türkçe Haber Benzerliklerinin Belirlenmesinde Varlık İsimlerinin Etkisi*. Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği A.B.D.
- Yüksek Lisans Çalışması (Tez Aşamasında): *İpek Şencan, Haber Metinlerinde Varlık İsimleri Ve Ontolojik Yaklaşımla Erişim Performansı Değerlendirmesi: Bilcol-2005 Örneği*. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi A.B.D.