Derin Öğrenme ile Anahtar Kelime ve Anahtar İfade Çıkarımı Üzerine bir İnceleme

A Survey on Keyword and Key Phrase Extraction with Deep Learning

Özlem Ünlü (Kılıç)
Bilgi İşlem D. B.
Aksaray Üniversitesi
Aksaray, Türkiye
ozlemhumaunlu@yahoo.com

Aydın Çetin
Computer Engineering Department
Gazi Üniversitesi,Faculty of Technology
Ankara, Türkiye
acetin@gazi.edu.tr

***Özet* – Teknolojik gelişmelerle birlikte yüksek miktarda veri üretilmeye başlanmıştır. Daha önceden insan gücü ile kayıt altına alınan terabaytlarca veri kişisel bilgisayarların kullanımı ile dijitalleştirilmiştir. Sonuçta hızla çoğalan veri yığınları oluşmuş, anlamlandırılmamış bu verilerin arasından bilgiyi bulmak zorlaşmıştır. Bu verileri anlamlandırma gerekliliği önceden tanımlanmış istatistiki yöntemleri daha önemli kılmıştır. Tek bir dokümandan veya doküman yığınlarından metin anlamlandırma yöntemleri ile ihtiyaç duyulan bilgilere ulaşmak mümkündür. Önceleri daha çok istatistiki yöntemler ya da Doğal Dil İşleme (Natural Language Processing - NLP) teknikleri ile çözülen bu problem, Makine öğrenmesi algoritmaları ve yapay sinir ağları ile çözülmeye başlanmıştır. Son yıllarda yapay sinir ağlarının özelleşmiş bir çalışma alanı olan derin öğrenmenin birçok problemde mevcut istatistiki ve NLP yöntemlerden daha iyi sonuç vermesi makine çevirisi, anahtar kelime çıkarımı, özetleme gibi problemlerde bu yöntemlerin uygulanmasını sağlamıştır. Bu çalışmada anahtar kelime ve anahtar ifade çıkarımında kullanılmış olan derin öğrenme yöntemleri incelenmiştir.**

Anahtar Kelimeler – anahtar kelime, anahtar ifade, inceleme, derin öğrenme, RNN

***Abstract* – With the technological developments, a large amount of data has been produced. Tera bytes of data previously recorded by manpower were digitized with the use of personal computers. As a result, rapidly growing data stacks were formed, making it difficult to find information among these unanticipated data. The need to make sense of this data has made predefined statistical methods more important. It is possible to access the required information from a single document or from the document stacks by means of text mining methods. This problem, which was previously solved mostly by statistical methods or Natural Language Processing (NLP) techniques, has been started to be solved by machine learning algorithms and artificial neural networks. In recent years, deep learning, which is a specialized study area of artificial neural networks, gives better results than the current statistical and NLP methods in many problems and has provided the application of these methods in problems such as machine translation, keyword extraction and summarizing. In this study, deep learning methods used in the extraction of keywords and key phrases are examined.**

Keywords – keyword, key phrase, review, deep learning, RNN

I. GİRİŞ

İnternetin gelişmesi ile birlikte bilgisayarlar ve bilgisayar sistemleri yerel çalışabilme kısıtından kurtulmuş, bu sayede savunma sistemleri, elektronik cihazlar, kara, hava ve deniz taşıtları, akıllı ev sistemleri gibi çok geniş bir yelpazede kullanım alanı bulmuştur. Bununla birlikte birçok kaynaktan derlenen çoğunlukla metin biçiminde üretilen verilerin işlenmesi ve anlamlandırılması problemi ortaya çıkmıştır. Bu problem önceden tanımlanmış istatistiki yöntemleri daha önemli hale getirmiştir. Makine çevirisi, anahtar kelime ve anahtar ifade çıkarımı, dokümanların ilişkilendirilmesi, başlık takibi, metin özetleme gibi problemler metin madenciliği alanında çözülmektedir.

Metin anlamlandırma konusunda yapılan birçok çalışma vardır. Örneğin Sosyal ağlar metin açısından zengin içeriklere sahiptir. Sosyal ağlar, anahtar kelime arama, sınıflandırma ve kümeleme gibi çok çeşitli uygulamalar için metin incelemesi algoritmaları gerektirir. Arama ve sınıflandırma çok çeşitli senaryolar için iyi bilinen uygulamalar olsa da, sosyal ağlar hem metin hem de bağlantılar açısından daha zengin bir yapıya sahiptir. Alandaki çalışmaların çoğu ya tamamen metin içeriğini ya da yalnızca bağlantı yapısını kullanır [1].

Biyomedikal literatürün büyümesi, alan uzmanlarının karşılaştığı aşırı bilgi yükünü ele almak için metin anlamlandırma tekniklerinin biyomedikal metne uygulanmasına olan ilginin artmasına neden olmuştur [2].

Günlük yayınlanan haber etiketlerinin oluşturulması, hızla artan sayıda bilimsel makalelere anahtar kelime atanması, dokümanların özetlerinin çıkarılması, farklı iki dilin birbirine çevrilmesi, dokümanlar arasında bağlantı kurulması gibi insan gücü ile yüksek maliyetle yapılan birçok iş otomatik sistemler ile az bir maliyete gerçekleştirilebilmektedir.

Bu çalışma sırasıyla anahtar kelime/anahtar ifade çıkarımı/atanması problemi, probleminin çözümü için geliştirilmiş istatistiki, NLP ve grafik tabanlı yöntemler, makine öğrenmesi yöntemleri, derin öğrenme yöntemleri ve Yinelenen Sinir Ağlar (Recurrent Neural Network - RNN) , sonuçlar bölümlerini içermektedir.

978-1-7281-3789-6/19/$31.00 ©2019 IEEE

II. ANAHTAR KELİME/ANAHTAR İFADE ÇIKARIMI/ATANMASI

Her gün binlerce kitabın ve yayının üretildiği dijital dünyada aradığı bilgiye ulaşabilmek problem haline gelmiştir. Bu sebeple metinlerden bilgi çıkarımı ve özet çıkarımı işlemleri önemli hale gelmiştir [3].

Metin madenciliği kullanılan teknolojiler 8 kategoride gruplandırılmıştır:

i. Bilgi Çıkarımı (BÇ- Information Extraction)
ii. Başlık Takip Etme (Topic Tracking)
iii. Özet Çıkartma (Summarization)
iv. Kategorizasyon (Categorization)
v. Kümeleme (Clustering)
vi. Kavram Bağlantılama (Concept Linkage)
vii. Bilgi Görselleştirme (Information Visualization)
viii. Soru Cevaplama Sistemleri (Question Answering-Chatbot) [4]

Bilgi çıkarma yazılımı, anahtar ifadeleri ve metin içindeki ilişkileri tanımlar. Konu izleme sistemi, kullanıcı profillerini koruyarak çalışır ve kullanıcının görüntülediği belgelere dayanarak, kullanıcının ilgisini çeken diğer belgeleri tahmin eder. Anahtar Kelime çıkarımı konu izleme başlığı altında çalışılmaktadır. Özetleme işlemi uzun dokümanlardan daha kısa sürede konunun temel prensiplerini anlatan kısa yazılar oluşturur. Kategorizasyon, önceden tanımlı üst başlıklara dokümanları kategorize eder. Kümelemede ise kategorizasyondan farklı olarak üst başlıklar önceden tanımlı değildir. Kavram Bağlantılama, geleneksel bağlantılama yöntemlerinden farklı olarak dokümanlar arasında kavram temelli ilişkiler kurar. Bilgi görselleştirme, verilerin görsel olarak sunumudur. Soru Cevaplama Sistemleri, insan yerine akıllı sistemler tarafından kullanıcıya cevap veren chatbotlardır.

Anahtar kelimeler ve anahtar ifadeler bütün metini tanımlayan en küçük birimler olduğundan bilgi çıkarımı, başlık takip etme, özet çıkarma, kategorizasyon, kümeleme, bilgi görselleştirme, kavram bağlantılama işlemlerinde kullanılabilmektedir.

Uluslararası Bilgi ve Kütüphane Bilimi Ansiklopedisi [5], anahtar kelimeyi "Konuyu veya konunun genel çerçevesini özlü ve doğru şekilde tanımlayan bir kelime" olarak tanımlamaktadır. Hem tek kelimeler (anahtar kelime) hem de ifadeler (anahtar ifade) anahtar kelimeler olarak adlandırılabilir. Manning ve Schutze, İstatistiki Doğal Dil İşlemenin Temelleri adlı kitabındaki ifadeler hakkında şunları söylemiştir: Kelimeler sadece eski bir sırayla gerçekleşmez. Dillerin kelime sırası ile ilgili kısıtlamaları vardır. Fakat bir cümle içindeki kelimeler, bir kolye üzerindeki boncuklar gibi sadece konuşma dizileri olarak dizilmemektedir. Aksine, kelimeler sözcük grupları halinde düzenlenir. Temel fikirlerden biri, belirli kelime gruplarının bileşen olarak davranmasıdır [6].

Başka bir çalışmada anahtar kelimeler, bir makalede içeriğinin okurlara yüksek düzeyde açıklanmasını sağlayan önemli kelimeler grubu olarak tanımlanmıştır [7]. Örneğin çok sayıda çevrimiçi haber verisinden anahtar kelimeler belirlemek, haber makalelerinin kısa bir özetini verebilmesi için çok yararlıdır. Anahtar kelimeler, akademik makalelerde okuyucuya makalenin içeriği hakkında fikir vermek için kullanılır. Bir ders kitabında, okuyucular için belirli bir bölümle ilgili aklındaki ana noktaları tanımlamaları ve muhafaza etmeleri yararlıdır. Anahtar kelimeler bir metnin ana temasını temsil ettiğinden, metin kümelemesi için bir benzerlik ölçüsü olarak kullanılabilirler [8]. Kaur ve arkadaşları anahtar kelimeyi en önemli verileri gösteren kelimeler olarak tanımlarken [3], Chen ve arkadaşları kısa metin parçacıkları olarak tanımlamıştır [9]. Lott anahtar kelimelerin, bilgileri bulmak ve iki test parçasının birbiriyle ilişkili olup olmadığını belirlemek için arama motorlarında ve doküman veri tabanlarında yaygın olarak kullanıldığını [10], Meng anahtar ifadelerin belgeyi anlama, kategorilendirme için kullanılabilecek bilgi sağladığını belirtmiştir [11].

Anahtar kelime bir metinin veya metin belgesinin içeriği hakkında bilgi veren kelimelerdir. Anahtar ifadeler ise tamlamalar gibi birden fazla kelimeden oluşan ve metin veya belge içeriği hakkında kullanıcıya bilgi veren kelime öbekleridir.

Otomatik anahtar kelime ve anahtar ifade üretimi iki yaklaşımla yapılabilir. Daha önceden oluşturulmuş kontrollü bir sözlükten anahtar kelime atanması yöntemi, basit ve tutarlı olması açısından avantajlıdır. Fakat doğru bir kontrollü sözlük oluşturmak zor ve zaman alıcı bir işlemdir. Sözlükte olmayan yeni kelimeler hiçbir zaman bulunamaz. Dokümanda bulunan kelimelerin en önemli olanları anahtar kelime olarak belirlenen ikinci yöntem, sözlük oluşturulmaması açısından avantajlı olsa da kararlı olamaz [8].

III. İSTATİSTİKİ, NLP VE GRAFİK TABANLI YÖNTEMLER

Literatürde yapılan anahtar kelime çıkarımı uygulamalarına bakıldığında çok eski çalışmalarda Terim Frekansı (TF), Ters Terim Frekansı gibi bilgiler kullanılarak sadece tek bir dokümana ait istatistiksel ölçümle anahtar kelime belirlendiği görülmektedir. 1972 yılında terim frekansının ölçülü o korpusta bulunan diğer dokümanlarla ilişkilendirilerek Terim Frekansı- Ters Doküman Frekansı (kısaca TF-TDF veya TF-IDF) çıkarılarak kullanılmaya başlanmıştır.

$$tftdf(t, d, D) = tf(t, d) \; x \; tdf(t, D) \quad (1)$$

olarak hesaplanmaktadır. Eşitlik (1)' de *tf(t,d)* terim frekansı istatistiksel yöntemi olup $tf(t,d) = f_{t,d}$ ($f_{t,d}$ t teriminin d dokümanında geçme frekansı)' dır.

tdf(t, D) Ters Doküman Frekansı istatistiği ise

$$tdf(t, D) = log \frac{N}{} \quad (2)$$

olarak hesaplanmıştır.

Eşitlik (2)' de N: korpustaki |D| dokümanlarının sayısı, dt : t teriminin görüldüğü doküman sayısını ifade etmektedir.

TF-TDF yöntemi Salton ve Buckley tarafından anahtar kelime çıkarımı için kullanılmıştır.

Matsuo ve Ishizuka tarafından yapılan bir çalışmada anahtar ifade çıkarımı için birlikte görülme eğilimi terim önemi olarak kullanılmıştır. Terim frekansı düşük ise eğilim güvenilir değildir. Örneğin t_1 teriminin a terimi ile birlikte bir sefer göründüğünü düşünüldüğünde (1.0 olasılığı ile) ve t_2' nin 100 defa a ile birlikte göründüğünü düşünüldüğünde (1.0 olasılığı ile); sezgisel olarak, t_2 daha güvenilir şekilde eğilimli görünmektedir. Çalışmada eğilimlerin istatistiksel önemini değerlendirmek için X^2 testi kullanılmıştır. X^2 testi beklenen frekanslar ve gözlenen frekanslar arasındaki eğilimlerin değerlendirilmesinde çok yaygındır. Beklenen olasılık p_g, sık görülen $g \in G$ teriminin mutlak olasılığını ifade eder, ve t ifadesinin sık görülen G ifadeleri ile birlikte görülme sayısını

n_t olarak belirtiriz. t ve g ifadelerinin birlikte oluşma sıklığı, *frek (t, g)* olarak yazılmıştır. X^2'nin istatistiksel değeri;

$$X^2(t) = \sum_{g \in G} \frac{(frek(t,g) - n_t p_g)^2}{n_t p_g} \quad (3)$$

olarak hesaplanmaktadır [12]. Algoritma, bir korpus kullanılmasına gerek kalmadan çalışabildiği için tftdf ile yarışabilir ve yüksek performanslıdır. Daha fazla elektronik doküman bulunan durumlarda, özellikle alan bağımsız anahtar kelimelerin çıkarılması için kullanılabilmektedir.

NLP' de, bazı varlık türlerini çıkarmak için kullanılan sözcük türlerini (yerler, kişiler vb.) bulmak mümkündür. Geliştirilen tekniklerin yelpazesi, dizgelerin basit manipülasyonundan doğal dil sorgularının otomatik olarak işlenmesine kadar uzanır. Ayrıca, dilbilimsel analiz teknikleri, metnin işlenmesi için diğer tekniklerin yanı sıra kullanılır [13].

Stapley ve arkadaşları bir Medline koleksiyonundan, her Saccharomyces cerevisiae geninin birlikte görülme sayısını çıkarmıştır. Yaptıkları sorgu otomatik olarak gen takma adları da içerecek şekilde oluşturulmuştur. Ek olarak, alınan doküman seti kullanıcı tarafından bir MeSH terimi ile filtrelenebilir. Bu birlikte-görülme verilerinden, ortak ve bireysel görülme istatistiklerini temel alarak, her bir gen çiftinin farklılık ölçümleri yapılmaktadır [13].

Iratxeta ve arkadaşları yaptıkları çalışmada aynı özette bulunan kelimeler (isimler) arasındaki ilişkiler, doğal dil ilişkilerinin modellenmesi için diğer yaklaşımlardan daha uygun olan iki bulanık ikili ilişki kullanılarak açıklanmıştır [14].

Alana özgü bir yaklaşım geliştirilerek ve kural tabanlı dilsel yaklaşımlar (NLP) kullanarak Radyoloji raporlarını anahtar kelimeler ile ilişkilendiren bir sistem başka bir çalışmada önerilmiştir [15].

Arafat Awajan istatistiksel analiz ve dilbilimsel bilgiyi birlikte kullanan Arapça belgelerden anahtar kelimeleri çıkaran bir yaklaşım sunmaktadır. Bu yaklaşım denetimsizdir ve iki aşamadan oluşmaktadır. İlk aşama, anahtar kelimeler olarak kabul edilebilecek tüm N gramlarını çıkarır. İkinci aşamada, N gramları, türemiş kelimelerin kökleri olan basit kelimelerin kök formlarıyla değiştirmek için morfolojik olarak değerlendirir. Aynı köke sahip olanlar yeniden sayılır. Son olarak görülme sıklığına göre anahtar kelimeler çıkarılmaktadır [16].

Liang ve arkadaşları tarafından yapılan çalışmada Çince haber makalelerindeki anahtar ifadeleri çıkarmak için, grafik tabanlı bir öğrenme algoritması olan TextRank kullanılmıştır [17]. Bu çalışmada sorgu günlüğü bilgisini kullanarak Çin haber makaleleri için anahtar sözcükler çıkarmaya yönelik pratik bir yaklaşım önerilmiştir. Ek olarak, iki öğretici özellik, ifadelerin uzunlukları ve konumları, TextRank modeline daha iyi sonuçlar elde edebilmek için dâhil edilmiştir.

Litvak ve Last tarafından anahtar kelime çıkarımı için denetimli ve denetimsiz olmak üzere iki yeni yöntem önerilmiştir [18]. Önerilen yaklaşımlar belgenin yapısal özelliklerini de dikkate alarak geleneksel vektör uzay modelini geliştiren metin ve web belgelerinin grafik tabanlı söz dizimsel gösterimini temel almaktadır. Denetimli yaklaşımda özetler üzerinden eğitim gerçekleştirildikten sonra belgelerin dilden bağımsız grafik gösterimi yapılır. Bu gösterim sayesinde, grafik yapım prosedüründe herhangi bir değişiklik yapmadan yöntem çeşitli dillere ve etki alanlarına uygulanabilir. Denetimsiz yaklaşımda ise belgelerin söz dizimsel gösterimi çıkarıldıktan sonra HITS sıralama algoritması kullanarak anahtar kelime çıkarılır.

Zhai ve arkadaşları tarafından gerçekleştirilen başka bir çalışmada Çince ve Vietnamca haber yapan iki dilli haber belgelerini temsil etmek için hipegrafi grafiksel gösterimi kullanılmıştır [19]. Çalışmada farklı dilde olan iki belge aynı belge alanına eşitlenmeye çalışılmaktadır. Aynı belge alanına eşlenene kelimeler için yönlendirilmiş difüzyon yöntemi kullanılarak anahtar kelime ağırlığı yeniden hesaplanmaktadır.

Anahtar kelime çıkarımının söz dizesi etiketleme işlemi olarak yapıldığı grafiksel model Koşullu Rastgele Alan (Conditional Random Field - CRF) modeli Zhang tarafından kullanılmıştır [20].

IV. MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Makine öğrenme algoritmaları denetimli ve denetimsiz olarak iki grupta toplanmaktadır. Anahtar kelime/ifade üretimi problemine özgü olarak denetimli öğrenme modelleri kullanılmaktadır.

Yasin Uzun bir metindeki anahtar kelimelerin ayırt edici özelliklerini belirlemek ve bu bilgileri kullanarak metinden anahtar kelimeleri çıkarmak için makine öğrenme yöntemlerinden biri olan Baiyes yöntemini kullanmıştır [21]. Çıkartılmış anahtar kelimeler içeren bir dizi eğitim belgesi olduğunu varsayarak, denetimli öğrenme yöntemi kullanılmıştır.

Collier ve arkadaşları Hidden Markov Modeli kullanarak MEDLINE üzerinde yayınlanmış 100 özet korpusunda anahtar kelime çıkarımı gerçekleştirmiştir [22]. Alan uzmanları tarafından çıkarılan kelimeler değerlendirildiğinde 0.78 f-ölçüsü elde edilmiştir.

Turney bir karar ağacı algoritması olan C4.5 algoritmasını anahtar kelime üretimi için kullanılmıştır [23]. İfadeleri pozitif veya negatif anahtar kelimelerin örnekleri olarak sınıflandırarak, bir ifadenin aday anahtar olup olmadığına karar vermiştir. Aynı çalışmada hibrit bir genetik algoritma türü olan GenEx algoritması da anahtar ifade çıkarımı için test edilmiştir. GenEx algoritmasının Genitor genetik algoritması ve Extractor anahtar kelime çıkarma algoritması olmak üzere iki bileşeni vardır.

Çeşitli varyasyonları bulunan KEA algoritması üç aşamadan oluşmaktadır. Birinci algoritma aşamada dilsel özelliklerle aday anahtar kelimeleri bulur. İkinci aşamada her bir aday için özellik değerlerini hesaplar ve üçüncü aşamada e Naïve Bayes modeli kullanılarak sınıflandırma yapılmaktadır. Özellik değerleri hesaplanırken tftdf ölçüsü kullanılır [24].

En küçük kareler destek vektör makinesini temel alan bir anahtar sözcük çıkarma algoritmasında, yalnızca tf ve tdf değil, aynı zamanda cümlenin yapısal özellikleri de kullanılmıştır [25].

Destek Vektör Makineleri (Support Vector Machines - SVM)' ne dayalı bir sınıflandırma problemi çözümü ile anahtar kelime/ifade çıkarımı yapan başka bir algoritmada 3 farklı sınıf kullanılmıştır. Bunlar "iyi anahtar kelime", "önemsiz anahtar kelime" ve "kötü anahtar kelime"' dir [26]. Aday anahtar kelimeler üçlü ifade olarak seçilir ve özellikleri tanımlanır.

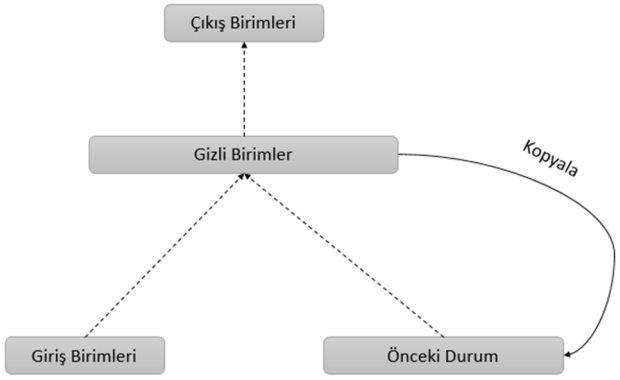
Qingguo ve arkadaşları tarafından önerilen K en yakın komşu (K Nearest Neighbor - KNN) modelini kullanan başka bir algoritmada verimli ve çok boyutlu bir indeks yapısı oluşturulmuştur [27]. Test belgesindeki aday anahtar

kelimeler, bu belgedeki en yakın K anahtar kelimesinden oluşmaktadır. Aday anahtar kelimelerin ağırlığı, test dokümanı ile aday anahtar kelimeyi içeren en yakın komşular arasındaki mesafenin toplamı ile hesaplanmaktadır. Aday anahtar kelimeler ağırlıklarına göre sıralandıktan sonra n aday anahtar kelime, anahtar kelimeler olarak kabul edilir.

V. Derin Öğrenme Yöntemleri ve RNN

Yapay sinir ağları belirli bir boyutta girdi alıp, belirli bir boyutta çıktı üretmektedir. Doğal dil işleme gibi birbirine bağlı ve önceki girdilerin hafızada tutulması gerektiği durumlarda derin öğrenme yöntemleri kullanılmaktadır. Sarkar ve arkadaşları tarafından önerilen bir algoritmada anahtar ifade çıkarımı problemi, sınıflandırma problemi gibi değil bir sıralama problemi gibi düşünülerek çok katmanlı algılayıcı (multilayer perception - MLP) modeli kullanılarak çıkarılmıştır. tftdf, yer özelliklerine ilaveten, ifade uzunluğu, kelime uzunluğu, ifadeler arasındaki bağlantı özellikleri de kullanılmıştır [28]. Başka bir çalışmada Evrişimli Sinir Ağı (Convolutional Neural Network - CNN) tabanlı CopyCNN yöntemi önerilmiştir [29]. Bu yöntemde gizli anahtar ifadeleri çıkarımı için, metinde mevcut anahtar kelimeler çıkarıldıktan sonra kopyalama mekanizması kullanılmıştır. Daha başarılı sonuç elde etmek için modele önem mekanizması ve yer bilgisi eklenmiştir.

CNN, MLP gibi derin öğrenme yöntemlerinden bir diğeri olan RNN doğal dil işlemede kullanılmaktadır. Bir sinir ağı; giriş katmanı, gizli katmanlar ve çıkış katmanından oluşmaktadır. RNN içerisinde yinelenen ve ağın kendisini besleyen bir yapı bulunmaktadır. Yani sinir ağı içerisindeki nöronlar zaman içerisinde birbirine bağlanmaktadır. Bu şekilde çalışma zamanının öncesinde çalışmış bir ağdan bilgi alabilmektedir. Hafıza akışı bu şekilde sağlanmaktadır. Hafıza akışı ile bilgi alabilmesi anahtar kelime/ifade çıkarımında RNN' in yaygın kullanılmasını sağlamıştır.



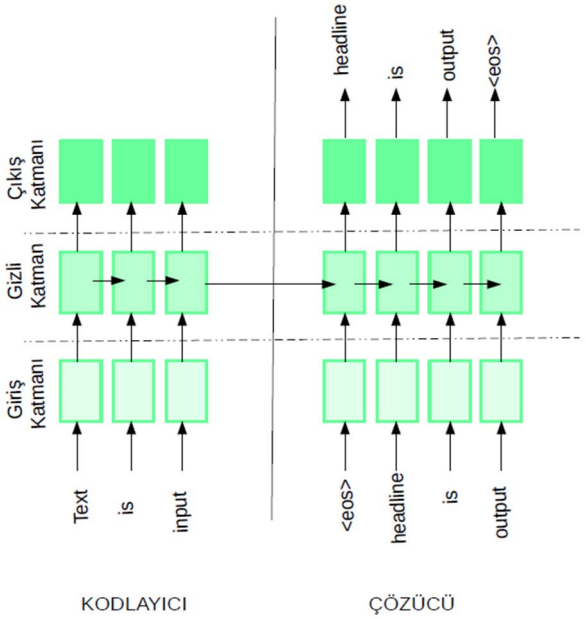
Şekil 1. Basit bir RNN Yapısı [30]

Basit bir sinir ağı hem bir önceki ileri yayılımdan gelen aktivasyon ile birlikte mevcut durum girişlerinden beslenmektedir. Gizli birimlerden gelen bilgiler RNN' e önceki durum hakkında bilgi vermektedir. RNN fonksiyon olarak :

$$h_t = fw(h_{t-1} + x_t) \quad (4)$$

şeklinde ifade edilebilir. Eşitlik (4) 'te *fw :* w parametreli fonksiyon, h_t : yeni durum, h_{t-1} : önceki durum, x_t : girdi vektörü olarak tanımlanmaktadır.

Makine çevirisi, haber başlığı çıkarımı, anahtar ifade çıkarımı gibi birçok problemde diziden-diziye (sequence-to-sequence seq-to-seq) modeli kullanılmaktadır. Seq-to-seq modeli birden fazla girişten birden fazla çıkış üretileceği zaman kullanılan RNN modelidir. Bunun dışında yazılan bir yorumun olumlu mu, olumsuz mu olduğuna dair yapılan ikili sınıflandırma problemlerinde diziden-bire (sequence-to-one), resim başlığı gibi bir girdiye birden fazla çıktı üretilmesi probleminde birden-diziye (one-to-sequence) modelleri kullanılmaktadır. Şekil 2' de seq-to-seq RNN modeli görülmektedir [31]. Seq-to-seq modeli iki RNN ağıdan oluşmaktadır. Bunlar Kodlayıcı ve Çözücü olarak isimlendirilir.

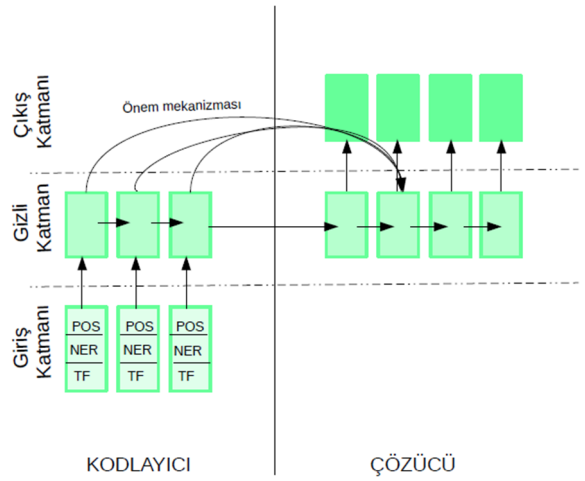


Şekil 2- seq-to-seq RNN [31]

RNN' in geri yayılımı sırasında oluşan ve hafıza aktarımından kaynaklanan kaybolan ve patlayan eğim problemlerine çözüm olarak önerilen Uzun Kısa Dönemli Hafıza (Long Short-Term Memory - LSTM) ve Geçitli Tekrarlayan Birim (Gated Recurrent Unit - GRU) versiyonları bulunmaktadır. Doğal dil işleme problemlerinde bu iki model sıkça kullanılmaktadır. GRU çalışma zamanı ve hafıza açısından performans göstermektedir. LSTM ise daha eski bir yöntemdir ve parametrelerini ayarlamak daha kolaydır.

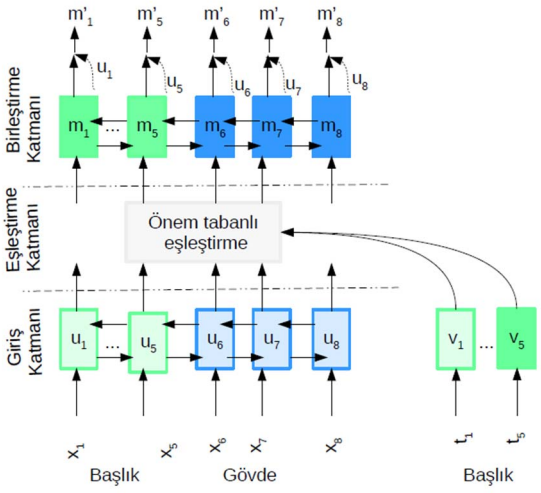
Meng ve arkadaşları tarafından önerilen CopyRNN modelinde seq-to-seq RNN kodlayıcı ve çözücü RNN ağları sırası ile iki yönlü GRU ve öne doğru GRU ile değiştirilmiştir [11]. Çözücü ağına bir önem mekanizması eklenerek performans arttırılmıştır. Gizli anahtar ifadelerinin üretilebilmesi için kopyalama mekanizması kullanılmıştır. Bu mekanizma kelimelerin yer bilgisini, kelimenin önem vektörünü hesaplarken kullanmıştır.

Nallapati ve arkadaşları tarafından metin özetlemede kullanılmak üzere önerilmiş başka bir model tek yönlü GRU dan oluşan kodlayıcı, önem mekanizmalı GRU çözücü içeren bir yapıya sahiptir [32]. Kodlayıcı tf, tdf, cümlenin öğesi(Part of Speech-POS), isimlendirilmiş varlık tanıma (Named Entity Recognition - NER) bilgilerini giriş katmanında vektör olarak almaktadır. Girişte kullanılan bilginin çeşitliliği sayesinde model Zengin Özellikli Kodlayıcı (words large vocabulay trick - words-lvt) olarak isimlendirilmiştir. Şekil 3' te Zengin Özellikli Kodlayıcı modeli görülmektedir.



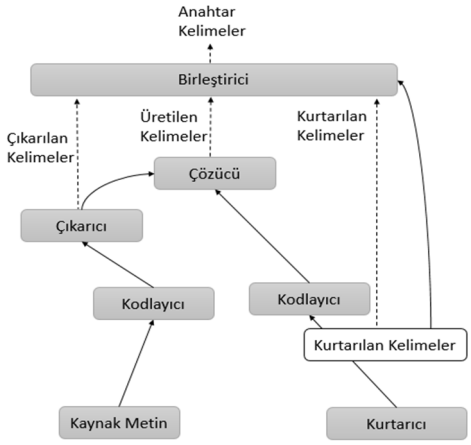
Şekil 3. Zengin Özellikli Kodlayıcı [32]

Başlık Güdümlü Ağ (Title Guided Network - TG-Net) olarak isimlendirilen Chen ve arkadaşları tarafından önerilen model CopyRNN modelinin bir uzantısıdır [33]. CopyRNN ve CopyCNN' de başlık ve özet bilgisinin eşit ağırlıkta giriş metni olarak alınır. Kodlayıcı bu metinin tamamını kelime girdi vektörleri ile alıp işlemektedir. Bu da başlığın bir metinin içeriğini belirtmesi durumunu göz ardı etmektedir. Probleme çözüm olarak TG-Net' de başlık ve gövde metni için iki yönlü ayrı GRU kodlayıcı tanımlanmaktadır. Şekil 4' te başlık güdümlü kodlayıcı görülmektedir. Çözücü olarak önem tabanlı, kopyalama mekanizmasına sahip model kullanılmaktadır.



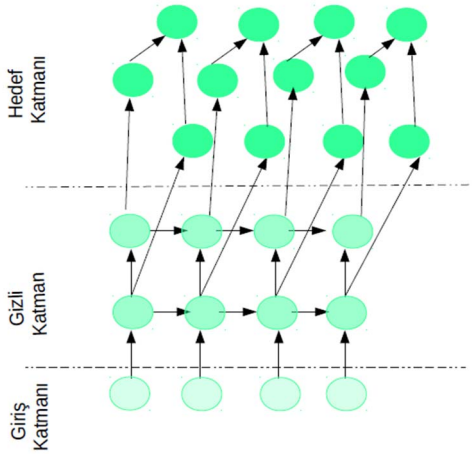
Şekil 4. Başlık Güdümlü Kodlayıcı [33]

Anahtar İfade Çıkarımı (Keyphrase Generation - KG) kütüphanesi iki kodlayıcı, bir çıkarıcı, bir kurtarıcı ve bir birleştirme modülünden oluşmaktadır [9]. Kütüphanenin temel amacı hem metinden çıkarılabilen anahtar ifadeleri (görünen ifadeler) hem de metinden üretilebilen anahtar ifadeleri (gizli ifadeler) bulmaktır. Kurtarıcı modül ilk K anahtar ifadeyi (sık kullanılan terimler hariç) Jaccard benzerliğine göre çıkarır. Kodlayıcılar iki yönlü GRU' dan oluşmaktadır. Çözücü ise modül ileri GRU tabanlıdır. Şekil 5' te KG modeli bulunmaktadır.

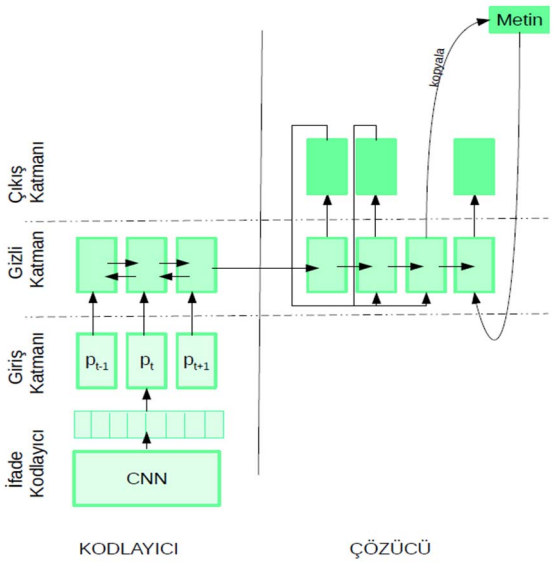


Şekil 5. KG Kütüphanesi [9]

Birleşik-katman RNN (joint-layer RNN) olarak isimlendirilen yöntemde (Şekil 6) iki çıkış katmanı birleştirilerek hedef katman tanımlanmıştır [34]. İki tane gizli katman bulunan modelde son gizli katman ve bir önceki gizli katmandan üretilen çıktılar doğrusal çakıştırma fonksiyonu ile birleştirilmekte ve hedef katmanı oluşturmaktadır.



Şekil 6. Birleşik Katman RNN [34]



Şekil 7. LSTM CNN Hibrit Model [35]

Şekil 7'de verilen iki derin öğrenme ağının kodlayıcıda birlikte kullanıldığı LSTM-CNN tabanlı hibrid modelde, kelimeler ifade çözücü olarak isimlendirilen CNN ağından geçtikten sonra üretilen ifade vektörü LSTM' in giriş katmanına gönderilmektedir [35].

VI. SONUÇLAR

Bir doküman hakkında en özet bilgiyi hızlı bir şekilde edinebileceğimiz anahtar kelimeler, veri miktarının katlanarak arttığı günümüzde çözülmesi gereken önemli bir problemdir. Anahtar kelime/ifade çıkarımı birçok metin anlamlandırma problemin çözümünde kullanılmaktadır. Bilgi çıkarımı, başlık takibi, özet çıkarma problemlerinde bir aşama olması ile birlikte haber metinleri ve bilimsel makalelere anahtar kelime atanması problemlerinde çözülmesi gereken temel durum olarak karşımıza çıkmaktadır.

Bu çalışmada, anahtar kelime/ifade için literatürde yapılan çalışmalar incelenmiş, anahtar ifade çıkarımı için kullanılan Derin öğrenme yöntemleri kategorik olarak sınıflandırılmıştır. Yöntemlerin daha iyi anlaşılabilmesi için literatürde yapılan çalışmalar 3 ayrı başlık altında toplanmış, derin öğrenme tabanlı anahtar kelime ve anahtar ifade üretimi yöntemlerinin mimari detayları incelenmiştir.

REFERANSLAR

[1] Ananiadou, Sophia, and John McNaught. Text mining for biology and biomedicine. London: Artech House, 2006.

[2] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." Ldv Forum. Vol. 20. No. 1. 2005.

[3] Kaur, Jasmeen, and Vishal Gupta. "Effective approaches for extraction of keywords." International Journal of Computer Science Issues (IJCSI) 7.6 (2010): 144.

[4] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence. 2009 Aug 1;1(1):60-76.

[5] Feather, J. and S. P., International encyclopedia of information and library science. London & New York: Routledge, 1996.

[6] Manning, C., and Schütze, H., 1999. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA.

[7] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence. 2009 Aug 1;1(1):60-76.

[8] Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications. 2015 Jan 1;109(2).

[9] Chen, Wang, et al. "An integrated approach for keyphrase generation via exploring the power of retrieval and extraction." arXiv preprint arXiv:1904.03454 (2019).

[10] Lott, Brian. "Survey of keyword extraction techniques." UNM Education 50 (2012): 1-11.

[11] Meng, Rui, et al. "Deep keyphrase generation." arXiv preprint arXiv:1704.06879 (2017).

[12] Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools. 2004 Mar;13(01):157-69.

[13] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. InBiocomputing 2000 1999 (pp. 529-540).

[14] Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. Trends in biochemical sciences. 2001 Sep 1;26(9):573-5.

[15] Sosyal E, Çiçekli NB, Baykal N. Radyoloji Raporları için Türkçe Bilgi Çıkarım Sistemi. VI. Ulusal Tıp Bilişimi Kongresi, Antalya, Türkiye. 2009:304-12.

[16] Awajan AA. Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents. InProceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014) 2014 (pp. 175-184).

[17] Liang W, Huang CN, Li M, Lu BL. Extracting keyphrases from chinese news articles using textrank and query log knowledge. InProceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2 2009 (Vol. 2).

[18] Litvak M, Last M. Graph-based keyword extraction for single-document summarization. InProceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization 2008 Aug 23 (pp. 17-24). Association for Computational Linguistics.

[19] Liang W, Huang CN, Li M, Lu BL. Extracting keyphrases from chinese news articles using textrank and query log knowledge. InProceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2 2009 (Vol. 2).

[20] Zhang, Chengzhi. "Automatic keyword extraction from documents using conditional random fields." Journal of Computational Information Systems 4.3 (2008): 1169-1180.

[21] Uzun Y. Keyword extraction using naive bayes. InBilkent University, Department of Computer Science, Turkey www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/Yasin_Uzun. Pdf 2005.

[22] Collier N, Nobata C, Tsujii JI. Extracting the names of genes and gene products with a hidden Markov model. InProceedings of the 18th conference on Computational linguistics-Volume 1 2000 Jul 31 (pp. 201-207). Association for Computational Linguistics.

[23] Turney, Peter D. "Learning algorithms for keyphrase extraction." Information retrieval 2.4 (2000): 303-336.

[24] Witten, Ian H., et al. "Kea: Practical automated keyphrase extraction." Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. IGI Global, 2005. 129-152.

[25] Wang, Jiabing, and Hong Peng. "Keyphrases extraction from web document by the least squares support vector machine." The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). IEEE, 2005.

[26] Zhang, Kuo, et al. "Keyword extraction using support vector machine." international conference on web-age information management. Springer, Berlin, Heidelberg, 2006.

[27] Qingguo, Zhang, and Zhang Chengzhi. "Automatic chinese keyword extraction based on KNN for implicit subject extraction." 2008 International Symposium on Knowledge Acquisition and Modeling. IEEE, 2008.

[28] Sarkar, Kamal, Mita Nasipuri, and Suranjan Ghose. "A new approach to keyphrase extraction using neural networks." arXiv preprint arXiv:1004.3274 (2010).

[29] Zhang, Yong, Yang Fang, and Xiao Weidong. "Deep keyphrase generation with a convolutional sequence to sequence model." 2017 4th International Conference on Systems and Informatics (ICSAI). IEEE, 2017.

[30] Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.

[31] Lopyrev, Konstantin. "Generating news headlines with recurrent neural networks." arXiv preprint arXiv:1512.01712 (2015).

[32] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[33] Chen, Wang, et al. "Title-Guided Encoding for Keyphrase Generation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[34] Zhang, Qi, et al. "Keyphrase extraction using deep recurrent neural networks on twitter." Proceedings of the 2016 conference on empirical methods in natural language processing. 2016.

[35] Song, Shengli, Haitao Huang, and Tongxiao Ruan. "Abstractive text summarization using LSTM-CNN based deep learning." Multimedia Tools and Applications 78.1 (2019): 857-875.