

Çeviri Yazılımlarında Sözcüklerin Bağlam İçindeki Anlamını Algılamaya Yönelik Bir Öneri

Gökhan Silahtaroğlu¹ Fırat Demircan²

¹Beykent Üniversitesi, Bilgisayar Mühendisliği Bölümü

²Beykent Üniversitesi, Fen Bilimleri Enstitüsü

Özet: Türkçe den başka bir dile ya da başka bir dilden Türkçe'ye çeviri yapan yazılımların en büyük sorunlarından bir tanesi iki dil arasında, birçok sözcüğün iki, üç hatta dört farklı şeyi ifade etmesidir. Yazılımların uygun bir çeviri sağlayabilmesi için cümle içinde geçen sözcüğün o cümlede ya da o bağlamda hangi anlamda kullanıldığının yine yazılım tarafından belirlenmesi gerekir. Bunun içinse çeviri yazılımının uygun bir anlamsal çözümleme işlevine sahip olması gerekir.

Bu çalışmada Türkçe sözcüklerin bir birleriyle olan ilişkilerini, hangi sözcüğün hangi sözcüklerle birlikte geçtiğinin haritası bir bakıma anlamsal sözcük grafını çıkartılması ve buna bağlı olarak İngilizce – Türkçe ve Türkçe – İngilizce Sözlüğün bu anlamsal graf'a göre yeniden düzenlenmesi ve uygun bir algoritmayla sözcüklerin bağlam içindeki gerçek anlamlarının ortaya çıkarılması için bir uygulama ve öneri sunulmaktadır.

1. Giriş

Makine çevirilerinde ortaya çıkan bir çok sorun henüz tam anlamıyla çözülmüş değildir. Türkçe'den İngilizce'ye ya da İngilizce'den Türkçe'ye yapılan makine çevirileri henüz tam olarak güvenilir sonuçlar vermemektedirler. Çevirelerde, dil bilgisi, zaman dönüşümü, deyimler, atasözleri gibi hususların yanı sıra her iki dilde de bulunan eş sesli, çok anlamlı sözcüklerin varlığı da makine çevirilerinde üstesinden gelinmesi gereken bir sorundur. Örneğin, Türkçe'de kullanılan *pas* sözcüğü, futbol dilinde *topu takım arkadaşına kazandırma, verme* anlamını taşıırken, başka bir anlamı da *metallerin oksitlenmesi sonucu ortaya çıkan bir maddedir*. Cümle içinde bu sözcüğün hangi anlamda kullanıldığını anlamak için bağlamına ya da gerek cümle gerekse de aynı paragraf içinde birlikte geçtiği diğer sözcüklere bakmak gerekecektir. “*Kaleci Demir geri pası tuttu*” ya da “*Demir pas tuttu*” cümlelerinde geçen pas sözcüğünün gerçek anlamı, yanında, ötesinde berisinde geçen sözcüklere bakılarak bulunabilir.

Bu çalışmada, Türkçe'de anlam kargaşasına yol açan sözcüklerin, birlikte geçtiği sözcükler belirlenerek oluşturulan grafa dayanarak yapılacak bir Türkçe – İngilizce sözlük yardımıyla çok anlamlılıktan kayanaklanan makine çeviri sorunlarına bir çözüm önerilmektedir.

2. Metin Madenciliği

Metin madenciliği veri madenciliğinin bir yan uygulama alanıdır. Bilinen şekliyle veri madenciliği sınıflandırma, kümeleme ve birliktelik analizi modellerini inceler [1]. Birliktelik analizi incelenirken veri setinde kullanıcı tarafından belirtilmiş nesnelerin birbirleriyle ilişkisi ya da birlikte olma sıklıkları ortaya çıkartılır[2]. Veri madenciliğinin bu modeli en çok pazar sepeti analizinde uygulama alanı bulmuştur. Pazar sepeti analizinde müşterinin aldığı ürünlerin bir arada bulunma olasılıkları belirlenir ve bu birliktelik algoritma tarafından hesaplanan güven ve destek seviyeleriyle birlikte değerlendirilir[1],[2].

Metin madenciliğinde ise müşterinin aldığı ürünlerin bir arada bulunma sıklığı yerine sözcüklerin belli koşullar altında bir birini izleme sıklığı ortaya çıkartılır. Bu uygulamada veri seti bir düz dosyadaki metindir. İnternetin yaygınlaşmasıyla birlikte elektronik ortamdaki metinlerin birçoğu artık web sayfalarında yer almaktadır [3]. Metin madenciliğini besleyen metinlerde bu yüzden çoğunlukla web sayfalarından sağlanmaktadır. BU yüzden metin madenciliği zaman zaman İngilizce'de Web Mining (Web Madenciliği) ismiyle de anılmaktadır. Ancak işin özü ve kullanılan algoritmalar değişmemektedir. Metin madenciliğinde de birliktelik analizleri dışında sınıflandırma, kümeleme gibi model uygulamaları yapılabilir.

Herhangi bir kavramın metinde bir kez ya da belirli sayıda geçme olasılığının belirlenmesi bununla ilgili kurallar türetilme çalışmaları metin madenciliğinde sınıflandırma çalışmalarına tipik bir örnektir. Örneği biraz daha açarsak, *bulut* sözcüğünün bir metinde üç ve üzerinde geçmesi için gerekli kurallar çeşitli analizlerin yapılması ve metinler üzerinde algoritmaların koşturulmasından sonra şu veya benzer bir şekilde oluşacak ya da algoritma tarafından üretilip kullanıcıya sunulacaktır.

IF **title** INCLUDE *bulut* AND **body** INCLUDE *bilişim*(> 3) AND **body** INCLUDE *veri* (>0) THEN TRUE (Support %3, Confidence %78, Lift 236).

Yukarıdaki sonuç cümlesi şu şekilde açıklanabilir: *Bulut* sözcüğünün bir metinde üç defa ve üzerinde geçmesi için ilk şartın başlıkta *bulut* sözcüğünün geçmesidir, bunu yanı sıra metin içinde *bilişim* sözcüğünün en az dört defa ve *veri* sözcüğünün en az bir defa geçmesi gerekmektedir. Bu kural tüm metinler göz önüne alındığında veri setinde %3 sıklıkla rastlanılmış ve her bir rastlantıda *bulut* sözcüğünün üç ve üzerinde gerçekleşme olasılığı % 78'dir. Bu kural ağacın tümü göz önüne alındığında verilere ulaşma açısından 2.36 kat daha etkindir. Metin madenciliğinde kullanılabilir bir başka veri madenciliği modeliye kümelemedir.

Metin madenciliğinde kümeleme metinlerin içinde barındırdıkları tüm sözcüklere göre kaç ayrı kümeye ayrılması gerektiği konusuyla ilgilenir[1],[4]. Örneğin İnternetteki tüm metinler metin madenciliği ile doğal kümeler ayrılıp daha sonra otaya çıkartılan her bir kümenin, yapısal ve anlamsal özellikleri belirlenebilir. Daha sonra da bununla ilgili kurallar türetilerek, öğrenme evreni dışından karşımıza çıkacak her hangi bir metnin hangi küme, yapısal ya da anlamsal olarak ait olduğu söylenebilir.

Bu çalışmanın da ana konusunu oluşturan üçüncü modelse birliktelik kuralları analizdir[3]. Metin madenciliğinde bu model, hangi şartlarda ve konularda örneğin, hukuk, spor, magazin vs hangi sözcüklerin hangi sözcüklerle birlikte geçtiği ve bu birlikteliğin mutlak ve göreceli gücünü ortaya koyan çalışmadır [5]. Bu yöntemle ortaya çıkartılan kurallar ve örüntüler web semantiği alanında temel kaynak veya kullanılacak olan algoritmanın temel bileşeni olarak kullanılabilir. Bağlantı analiz için kullanılan algoritmalarından en yaygın olanları APRIORI, AIS ve SETM algoritmalarıdır [1],[3],[5]. Üç algoritma da en kısa süre içinde nesne kümeleri yaratarak birliktelik kurallarını ortaya çıkarır.

3. Anlam ve Bağlam

Anlam sözcüklerin ilk okunduğu ya da duyulduğu anda zihinde meydana getirdikleri olgudur[6]. Konuya başlamadan önce bir sözcüğün ilk duyulduğu anda bizde ne hissettirdiğini anlamamız gerekmektedir. İnsan beynini bir bilgisayar olarak düşündüğümüz zaman sözcüğü gördüğü ya da duyduğu anda o sözcüğü daha önce bağdaştırdığı bir sürü anlam gelir aklına ama cümleye veya konuya göre o anlamlardan bir tanesini seçer ve o anlama göre karşısındaki kişiye cevap verir [7]. Şimdi düşünecek olursak biz İngilizce bir sözcüğü duyduğumuzda ne yaparız? Arkadaşlarımız ya da yakınlarımızdan birine şu sözcüğün anlamı ne diye sorarız bize cevap verirken birbirinden farklı bir kaç anlam söylediklerini ya da bunları söylemeden önce bize o sözcüğü nereden duyduğumuzu veya nasıl bir cümle içinde duyduğumuzu sorduklarını görürüz. Peki karşımızdaki kişi bunu niye yapar? Anadili İngilizce olmadığı için mi? Hayır, sözcüğün anlamını bilmediği ve bunu cümleden yararlanarak bulmak istediği için mi? Hayır, peki niye? Çünkü sözcüğü duyduğu anda beynine birçok anlam gelir ve bunların hangisinin doğru olduğunu bilemez işini şansa bırakmamak için ve karşısındaki kişiye en iyi şekilde yardımcı olabilmek için sorar. Sözcüğün o cümle içinde kullanılan başka bir deyişle o bağlam içindeki anlamını karşısındaki kişiye verir. Türkçe'de ve diğer bütün dillerde de bu böyledir. Eş sesli kelimeler, eş anlamlılar, anlam genişlemeleri ve anlam daralmaları kelimenin birden çok anlam ifade etmesini veya birden fazla kelimenin sadece bir anlam ifade etmesini sağlarlar[8]. Bu da çevirilerde işimizi bir hayli zorlaştırır. Bunu aynı dili konuşan kişiler fark etmeyebilir ya da önemsemeyebilirler ama çeviri gibi hassa konularda çok önemlidirler.

Konuşma dili yazım dilinden farklı olduğu için yazarak karşımızdaki hissettirdiğimiz şeyi konuşurken farklı şekilde açıklarız[9]. Çünkü yazı dilinde vurgu yoktur ve bunun için çoğu insan yazarken duygularını tam olarak ifade etmediğini söyler. Yazım dilinde kurallar, kullanılacak sözcükler ve vurgular konuşma diline göre çok daha önemlidir ve daha büyük bir yer kaplar. Eş anlamlıların çokluğu zaman kavramları bütünlük derinlik ilişkileri gerçek anlamda önemlidir. Bunun için insanlar okudukları zaman dili daha iyi kullandıklarını görürler. Çünkü birçok kavramı takip etmesi gerekir. Önce sözcüğün gerçek anlamını, daha sonra deyimleri, daha sonra cümleyi ve belki paragrafın tümünü hesaba katarak o sözcüğün gerçekte ne anlam ifade ettiğini görmeye çalışır. Bizim üzerinde çalışacağımız konu işte kelimelerin gerçekte ya da bulunduğu bağlam içinde ne anlam ifade ettiğini bu özelliklerden birkaç tanesini kullanarak bilgisayara öğretmek ve çevirilerin daha rahat, daha sağlıklı olmasını sağlamaktır.

4. Uygulama Adımları ve Öneri

4.1 Amaç

Makine çevirilerinde karşılaşılan temel güçlüklerden bir tanesi iki dil arasında oluşan sözcükler arası çok anlamlılıktır. Örneğin, İngilizce’de **strong** sözcüğü güçlü kuvvetli anlamına geldiği gibi ‘**strong tea**’ olarak kullanıldığında ‘*demli*’ anlamını taşımaktadır. Bu nedenle iki dil arasında da makine çevirisi yaparken bu anlam farklılıkları, sözcüğün kullanıldığı bağlam da göz önünde bulundurularak değerlendirilmeli ve çeviri işlemi makine tarafından bu şekilde yürütülmelidir.

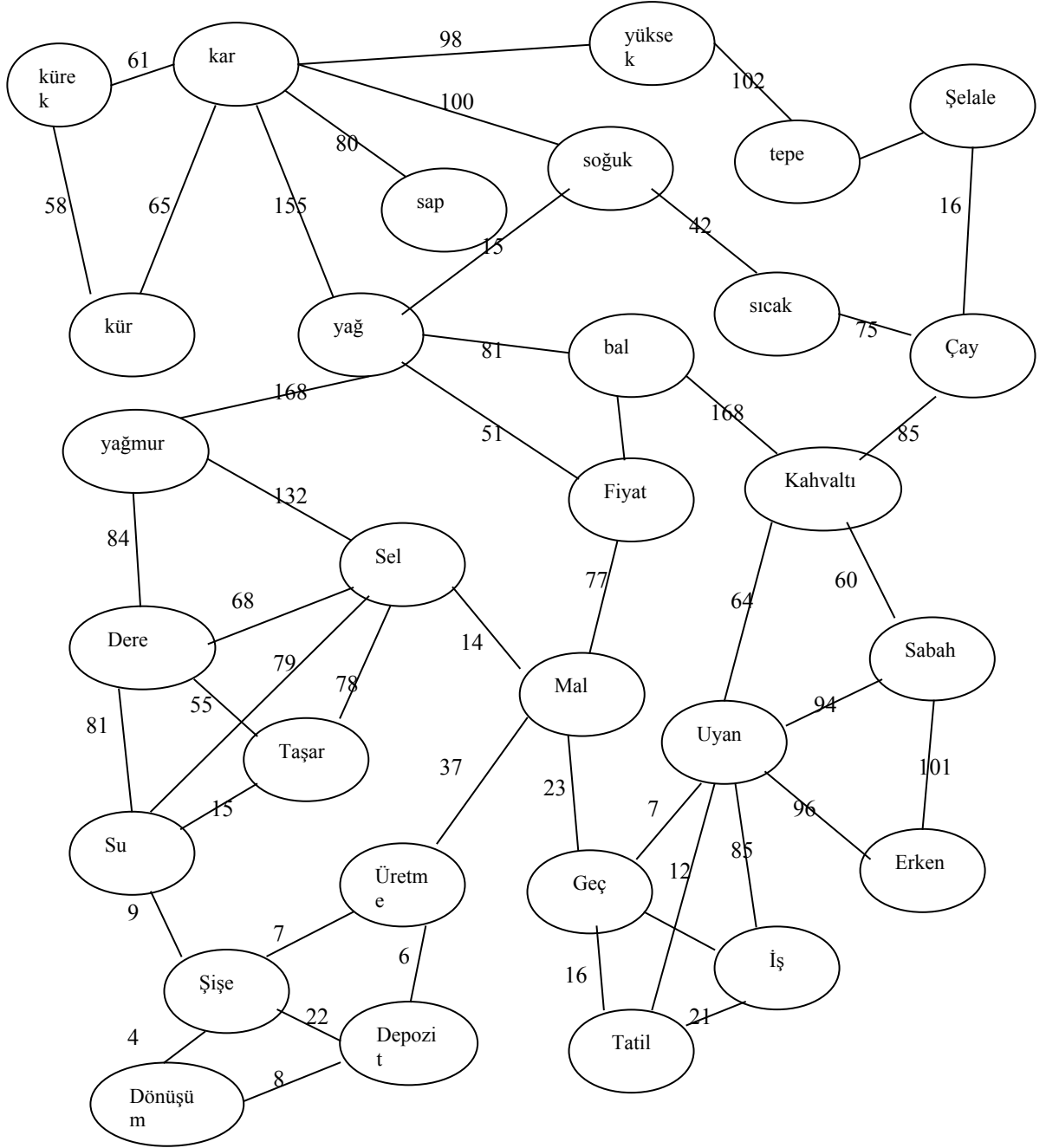
Bu araştırmadaki temel çözüm önerisi Türkçe – İngilizce çevirilerde hedef dil İngilizce olduğu durumlarda , iki dil arasındaki sözcükler arasındaki çok anlamlılık ya da anlam kargaşasını gidermektir.

Bunun başarılabilmesi için aşağıdaki hususların yerine getirilmesi gerekmektedir:

1. Tercihen her iki dilin de ama en azından hedef dilin anlam haritasının çıkartılması.
2. Bu anlam haritasına uygun Türkçe – İngilizce yeni bir sözcük hazırlanması
3. Makine çeviri programının yeni sözlüğü okuyacak şekilde düzenlenmesi.

4.2 Uygulama

Birinci aşamayı başarmak için veri kaynağı internet olmak üzere, toplam 13500 paragraf Türkçe metinlerden taranarak Türkçe deki sözcüklerin bir birleriyle olan bağları, bir graf şeklinde ortaya konmuştur. Başka bir deyişle sözcüklerin anlam haritası çıkartılmıştır. Bu anlam haritası bize kelimelerin birbirleriyle ne kadar sıklıkla bir arada geçtiklerini göstermektedir. Aşağıdaki Şekil 1 ‘de ortaya çıkartılan grafın temsili görüntüsü görülmektedir.



Şekil 1. Oluşturulan grafın temsili görüntüsü

Burada şekilde görüldüğü gibi paragraflarda geçen kelimelerin birbiriyle ne kadar sıklıkla geçtiğini, birbirleriyle nasıl bağlandıklarını görmekteyiz. Burada sözcüklerin bir arada hangi sıklıkla kullanıldığını rahatlıkla görmekteyiz. Gördüğünüz gibi bazı yerlerde üçgenler, dörtgenler oluşmuştur. Bu gibi durumlarda o üç veya dört kelimenin birbirleriyle aynı paragrafta çok geçtiğini, birbirleriyle çok fazla kullanıldığı görülmektedir. Bu tabii ki bu kelimelerin anlamlarını araştırırken, kelimenin tam anlamını çok rahat bir şekilde verebilmemizi sağlayacaktır. Burada tam olarak gösteremediğimiz ve aşağıda tabloda daha rahat bir şekilde göstereceğimiz şey ise bunların birbirleriyle bağlandıklarında aralarında bu bağın gücünü gösteren sayıların olması, bunu daha çok birbirleriyle geçen kelimeleri bağlayarak gösterdik ama tam olarak anlayabilmek için aşağıdaki tabloya dikkat etmemiz gerekir.

İkinci aşamada, hazırlanmış olan anlam haritası veri tabanı kullanılarak yeni bir Türkçe – İngilizce sözlük hazırlanacaktır. Bu sözlük hazırlanırken, hem manuel hem de bir program aracılığı ile otomatik yapılandırma sağlanmalıdır. Hazırlanacak olan sözlüğün prototipi aşağıdaki tabloda verilmiştir.

Tablo 1. Önerilen Türkçe – İngilizce sözlükten bir kesit.

Türkçe	İngilizce	100	90	80	70	60
Kar (i)	Snow	Soğuk	Yağ	Tut	beyaz	Yüksek
Yağ (mak) (f.)	Snow	Soğuk	Kış	Tut	Tatil	Beyaz
Kar (i)	Profit	Yüksek	Para	Muhasebe	Talep	Ürün
Yüz	Face	Göz	Gözlük	Havlu	Bak	Vur
Yüz	Swim	Deniz	Su	Boğul	Havuz	Havlu
Yüz	Hundred	Para	Matematik	Banknot	Bakiye	Kişi

Prototipte görüldüğü gibi öncelikle Türkçe kelimeler hemen yanında İngilizce anlamları ve yanlarında her bir sözcük için o sözcüğün anlam haritasından yararlanarak bulunmuş onunla birlikte geçen diğer sözcükler görülmektedir. Karşılaşılan büyük sorunlardan bir tanesi de aynı sözcüğün eş sesli bir başka kelimenin farklı anlamlarında geçme ihtimalinin olmasıdır. İşte bu sorunu çözmek için kelimelerin kullanımlarını derecelendirdik ve bu derecelendirme sayesinde o kelimenin eş sesli sözcüğün anlamını ne kadar etkilediğini rahatlıkla görebiliriz.

Örneğin yukarıda ki tabloda “yüksek” sözcüğü hem kar (yağış) hem de kar (kazanç) anlamındaki kelimelerin ikisiyle de bağlantılı olduğu gözükmektedir. Ama yüksek kelimesi kar (kazanç) anlamındaki kelimelerle daha fazla geçtiği için derecesi 100 ancak bir yağış türü olan kar anlamındaki kelimeyle daha az geçtiği için derecesi ise 60 olduğu görülmüştür.

Daha önce yapılan programların ve internet sayfalarının bu konuda neden yetersiz olduğuna gelince; bu gibi programlar eş sesli veya anlam daralması yaşamış sözcükleri göz önüne almadan, sözcüğün en bilindik anlamını çevirmeye çalışırlar. Yazıdaki kelimeleri birbirleriyle bağlantılarına ya da bağlam içinde kullandıkları anlamlarına bakmadan doğrudan, kelimelere odaklanarak çeviri işlemini yürütürler. Örneğin, son zamanların en çok kullanılan çeviri sitesi Google Translate uygulamasında cümle içinde “*kara saplanmak*” tamlamasını çevirmeye çalıştığımızda bize **clung** ve **land** ya da **clung** ve **black** sözcükleriyle çeviriyi yapmaktadır. Ama burada **land** *kara parçası*, **black** *siyah* ve **clung** ise *karaya oturmak* anlamındadır. Başka bir örnekte ise; “suda yüzmek” ifadesi “face in the water” şeklinde çevirmekte burada face sözcüğü *sima* anlamında kullanılmış ama aslında yüzmek fiili anlamına gelen swim kullanılmalıydı. Burada eş sesli sözcükler kullanarak örnekleri çoğaltabiliriz.

Şimdi ise programın çalışma mantığına gelelim. Anlam haritasını kullanarak otomatik ve manuel olarak bir sözlük prototipi ürettik. Bu sözlükte kelimelerimiz, İngilizce karşılıkları ve onlarla birlikte en çok kullanılan sözcükler bulunmaktadır. Bu sözcükleri ise kullanılma sıklıklarına göre derecelendirdik. Programın yapacağı şey

öncelikle İngilizceye çevrilecek sözcüğün içinde bulunduğu paragrafı bulup, algoritma yardımıyla o paragrafta en çok geçen sözcükleri sıralayıp, daha sonra önceden oluşturduğumuz sözlüğümüzde o kelimenin farklı anlamlarındaki onlarla birlikte geçmiş (daha önce derecelendirdiğimiz) kelimelere bakarak hangi anlamlarındaki sözcükler ile incelediğimiz paragraftaki sözcükler birbirine daha çok uyduğuna bakarak sözcüğün gerçek anlamını bulacaktır.

5. Sonuç

Çeviri; farklı dilleri konuşan insanlığın birbirleriyle daha rahat iletişim kurabilmesi için gerekli olup gün geçtikçe gelişen, üzerinde çok fazla araştırma yapılan konuların başında gelir. İnternetin tüm dünyada yaygınlaşmasından sonra, farklı ülkelerdeki farklı dilleri konuşan insanların birbirleriyle daha fazla iletişim kurma istekleri, farklı dillerdeki kaynaklardan yararlanma istekleri çeviri yapan programların yaygınlaşmasını ve yeni çözüm önerileri çıkmasını sağlamıştır. Biz de en önemli konulardan biri olan eş sesli ve anlam daralması yaşamış kelimeler sorununu çözmeye çalıştık. Daha önce yapılan araştırmaları inceleyerek ve onların veri ve sonuçlarından yararlanarak bu sorun için bir algoritma geliştirdik. Öncelikle kelimelerin paragraflar içindeki birbirlerine yakınlıklarından yararlanıp yeni bir tür Türkçe İngilizce sözlük üretip bu sözlük ile kelimenin kullanıldığı cümle, paragraf ya da başka bir deyişle kullanıldığı bağlam içindeki anlamını bulmaya çalıştık.

Bildiğiniz gibi çeviri gibi hassas konular, çok farklı dillerde çok farklı yapılar olduğundan dolayı tüm dünyada aynı hızla ilerlememektedir. Ancak, farklı ülkelerde kendi dillerinin yapısını çözümlmek için birçok araştırma yürütülmektedir. Kendi dilimizin yapısını çözdüğümüz ve dilimizin yapısı hakkında yeteri kadar fikir sahibi olduğumuz durumlarda çeviri gibi hassas konular daha hızlı ilerleme gösterebilir. Gelişen dünyada yeni teknikler sayesinde gelecekte insanların birbirlerinin dillerini öğrenmek zorunda kalmadan birbirleriyle çok rahat bir şekilde iletişim kurabilecekleri görebiliyoruz. Bu süreci kısaltmak insanların kendi anadillerine verdikleri özenden ve kendi anadillerinin yapısını ne derece iyi bilmelerinden kaynaklandığı bir gerçektir..

6. Kaynaklar.

[1] SİLAHTAROĞLU, Gökhan, **Kavram ve Algoritmalarıyla Temel Veri Madenciliği**, İstanbul, Papatya Yayıncılık, 2008.

[2] KANTARDZIC, Mehmed, **Data Mining Concepts, Models, Methods, and Algorithms**, A John Wiley & Sons, Inc., Publication, United States of America, 2003.

[3] ÇELİKİYAY, E. Kübra, **Metin Madenciliği Yöntemiyle Türkçe’de En Sık Kullanılan Ve Birbirini Takip Eden Harflerin Analizi Ve Birliktelik Kuralları**, Beykent Üniversitesi Fen Bilimleri Enstitüsü, 2010, (Yayımlanmamış Yüksek Lisans Tezi).

[4] CABENA, Peter- HADJINIA, Pablo -Rolf- VERHEES, Jaap – ZANASI, Alessandro – HALL, Prentice, **Discovering Data Mining: From Concept To Implementation**, USA, International Business Machines Corporation, 1998.

[5] BAYER, HARUN, **Veri madenciliğinde bir metin madenciliği uygulaması**, Beykent Üniversitesi Fen Bilimleri Enstitüsü, 2011, (Yayımlanmamış Yüksek Lisans Tezi).

[6] Türk Dil kurumu, **Anlam Sözcüğünün Türkçe Karşılığı**

[7] Condon, John (1998), **Sözcüklerin Büyüklü Dünyası (Anlambilim ve İletişim)-**, İstanbul, İnsan Yayınları,

[8] Muzaffer BARIN, **Dinleme ve Konuşma Becerilerinin Önemi**, Elazığ, 1997

[9] Feridun KARAARSLAN, **Konuşma Ve Yazma Eğitiminde Beyin Fırtınası Tekniğinin Etkinliği**, Sakarya 2010