Hayat Bilgisi Veritabanı Kullanılarak Otomatik Cümle Üretimi

C.Berkin ÖZDEMİR[1] , M.Fatih AMASYALI[1]

[1] Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü,İstanbul

canberkozdemir@yahoo.com, mfatih@ce.yildiz.edu.tr

Özet: Bilgisayarlar ve bilgisayar ağları günümüzde yaygın olarak kullanılmakta; insanlar internet, yazılımlar sayesinde aradıkları problem çözümlerine rahatlıkla ulaşabilmektedir. Örneğin o günün akşamındaki bir konser internet üzerinden aratıldığında konser ile ilgili detaylı tüm bilgiler kolaylıkla kişinin önüne gelebilmektedir. Fakat bu bilgisayar sistemleri geliştirilerek insanların gündelik bilgilerini yorumlayabilecek, hayat ile ilgili gerçekliklerle çıkarım yapabilecek şekilde "zeki"leştirilirlerse şu senaryonun yaşanması olası hale gelebilecektir: Kullanıcı sisteme canının sıkıldığını girdiğinde sistem kullanıcısına günlük yapılabilecekler listesini -belki de o günün akşamındaki konser ile birlikte- sunabilme yetisine erişebilecektir. Burada dikkat edilmesi gereken nokta hayattaki gerçeklikleri (insanlar sıkıldıklarında neler yaparlar? vb.) semantik (anlamsal) bağlar ile bilgisayara tanıtabildiğimizde bilgisayarların bu başarımı sağlayabilecek olmasıdır. Bu çalışmada ise, bu amaca ulaşmak için atılan bir adım sunulmuştur. Çalışmada, var olan bir Türkçe hayat bilgisi veritabanı (CSdb) kullanılarak, sisteme girilen bir cümleden yeni cümlelerin üretimi gerçekleştirilmiştir. Örneğin sistem “Ali babasını düşündü.” cümlesinden “Ali ailesini gözünde canlandırdı.” cümlesini üretilebilmektedir. Bu sayede bilgi parçaları arasındaki anlamsal ilişkilerin tutulduğu bir veri tabanının, birbiriyle kelime bazında benzerliği olmayan ancak anlamca birbirine yakın olan metinlerin bulunmasında kullanılabilirliği gösterilmiştir. Bu uygulama, otomatik hikaye/metin üretimi, anlamsal metin özetleme gibi sistemlerde faydalı olabilecektir.

Anahtar Sözcükler: Doğal Dil İşleme, Hayat Bilgisi Veritabanı, Anlamsal Ağ, Makine Öğrenmesi.

Automatic Sentence Generation Using Common Sense Databases

Abstract: Computers and computer networks are widely used in our decade; people can easily reach solutions of their problems that they search thanks to internet and software. For instance, when a concert, which will happen evening of that day, is searched; whole information about the concert can be retrieved to users easily. However, if those computer systems are developed to make them intelligent in a way that they can interpret daily lives of people and make decisions about realities of life, that scenario will be possible: when user enter input as “I am bored”, system will have the ability to show the list of daily activities, maybe the concert is included in it. The main point here is that as we can express the realities of the life with semantic relations to the computers, they will achieve the goal. In this study, the main aim is production of sentences which may have relation with the sentence which is entered to the system by using an existed Turkish common sense database. The proposed system can generate “Ali visualized his family.” sentence from the “Ali thought about his father.” user input sentence. This study verify the usability of the commonsense databases at finding similar meaning texts even if they have different words. This application can be very useful at automatic story/text generation and semantic text summarization systems.

Keywords: Natural Language Processing, Common Sense Database, Semantic Web, Machine Learning.

1. Giriş

Günümüzde yapay zekâ, makine öğrenmesi, doğal dil işleme teknikleri ile bilgisayarlar insanları satrançta yenebilir, insanlarla bir şekilde sohbet edebilir hale gelmişlerdir. Kullanıcılar internet sayesinde aradıkları bilgilere rahatlıkla ulaşabilmektedir. Arama motorları insanların aramalarını sentaktik ve semantik olarak inceleyip sıralama mekanizmalarıyla kişilere sunabilmektedir [1]. Ancak hala bilgisayarlarımızın, biz kullanıcılarının yaşadıkları hayatla ilgili bilgileri bulunmamaktadır. Örneğin cep telefonlarımız insanların geceleri uyuduğunu ve insanların uyurken önemli şeyler haricinde rahatsız edilmek istemediklerini bilselerdi gecenin 3'ünde telefon şirketimizden gelen bir mesaj için bizleri uyandırmaz ve sabahı beklerdi. Ya da eğer saat farkı kavramını bilselerdi, Türkiye'den Amerika'yı aramaya kalktığımızda bize orada şu an saatin kaç olduğunu hatırlatıp "emin misiniz?" diye sorabilirlerdi. Bu örnekler daha da çoğaltılabilir tabii ki, ama buradaki önemli nokta, bu bilgilere sahip bilgisayarların hayatımıza katkılarının şimdikinden daha fazla olabileceğidir. Böyle bir sistem için 2 bileşene ihtiyaç vardır. İlki bu bilgilerin tutulacağı bir veritabanı. İkincisi ise bu bilgileri bir amaç için kullanacak çıkarım mekanizmasıdır. Bu çalışmada ilk bileşen olarak ilk Türkçe hayat bilgisi veritabanı olan CSdb(Common Sense Database) [2] kullanılmıştır. İkinci bileşen ise bir eş/ benzer anlamlı cümle üreticisi olarak tasarlanmıştır.

Çalışmada hayat bilgisi veritabanını kelime ve kelime öbekleri ile kullanabileceğimizden ötürü kullanıcıdan alınan cümle, kelime ve kelime grupları halinde alt parçalara böldükten sonra ilk olarak sentaktik açıdan incelenmiştir. Sentaktik inceleme Zemberek kelime çözümleyicisi [3] kullanılarak, kelimelerin kök, ek ve türlerine ayrımı ile gerçekleştirilmiştir.

Bildirinin devamında ise sistemin çalışma mantığı, tasarımı, Zemberek kelime çözümleyicisini ve hayat bilgisi veritabanını nasıl kullandığı, sistem arayüzü ve gelecekte yapılabilecek çalışmalara değinilmiştir.

2. Sistem ve Sistemde Kullanılan Alt Sistemler

Eş/benzer anlamlı yeni cümlelerin üretimini gerçekleştirebilmek için hayat bilgisi veritabanına nesne olarak kelime ve kelime öbeklerini taşıyıp buradan bağlantılarla yeni nesnelere ulaşmamız gerekmektedir. Bu noktada hayat bilgisi veritabanının yapısı, nasıl oluşturulduğu, döndürdüğü sonuçların tutarlılığının nasıl ölçülebileceğine değinmek gerekir.

2.1. Hayat Bilgisi Veritabanı

Hayat bilgisi veri tabanları, kelimeler arasında yaşamdaki gerçeklikleri basit ilişki yapıları ile birbirlerine bağlar ve bu sayede bilgisayarların bu gerçeklikleri yorumlayabilmesini kolaylaştırır [4].

İlk Türkçe hayat bilgisi veritabanı olan CSdb, temelinde nesnelerin farklı ilişki kalıplarıyla bağlantılı olduğu nesnelerle ilişkilendiren bir veritabanıdır. Veritabanı "bu nerede bulunur, bunun üst kavramı nedir, bu ne gerektirir, bu neyden yapılmıştır, bunun özellikleri nelerdir" gibi 40 adet ilişki ve bunların tersi ilişkilere sahiptir ve nesneler bu ilişkilerle birbirine bağlanmaktadır. Veritabanında ayrıca her ilişkinin doğruluk oranı 0-5 arasında bir değerle tutulmaktadır. Nesneleri birbirine bağlayan bu ilişkileri ve doğruluklarını, kişiler Kemik Oyun [5] adı verilen oyunla veritabanına gönderirler. Nesnelerin ilişkilerini belirleyen kullanıcılar, ilişkilerin ilgili nesneler arasındaki tutarlılıklarını 0-5 puan arasında puanları seçerek göndermektedirler. Birçok kişinin verdiği puanların ortalaması alınarak

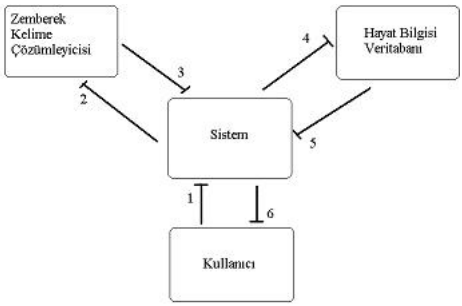
nesneler arası ilişkilerin doğruluk oranları arttırılmaktadır. Ancak ilk Türkçe hayat bilgisi veritabanının henüz istenilen olgunluğa erişmemiş olması buradan alacağımız sonuçların tutarlılık oranını düşürebilmektedir, bu dezavantaj çalışma sürecinde göz önünde tutulmuştur. Ancak oyun oynandıkça veritabanın içerdiği ilişkilerin doğruluğu artacağından bu tür uygulamalarda kullanımı artacak ve Türkçe semantik alanındaki çalışmalara büyük katkı sağlayacaktır. Eylül 2010 tarihi itibariyle CSdb'deki toplam kullanıcı sayısı 128, toplam nesne sayısı 641.183, nesneler arası ilişki sayısı ise 1.106.621'dir.

2.2. Zemberek Kelime Çözümleyicisi

Cümle üretimini sağlayabilmek için hayat bilgisi veritabanına, cümlenin kelime ve kelime öbekleri gönderilmektedir. Cümlenin kelime ve kelime öbeklerine ayrılması işlemi gerçeklendikten sonra da kelimenin veritabanına gönderilirken kök haline getirilmiş şekli, ekleri ve hayat bilgisi veritabanında sorguları gönderirken isim ya da fiil türünde olduğuna göre sorgu oluşturacağımızdan dolayı morfolojik bir kelime çözümleyicisine ihtiyaç duyulmuştur. Bu yüzden çalışmada Java programlama dili ile kodlanmış Türkçe doğal dil işleme kütüphanesi olan Zemberek uygulaması kullanılmıştır. Zemberek sisteme kelimelerin kök, ek ve türlerini sorunsuz bir şekilde döndürmektedir. Sistemde Zembereğin birden fazla çözümlemesi olan kelimeler için ürettiği sonuçlardan ilki kullanılmıştır.

3. Sistemin Tasarımı

Bu bölümde çalışmanın tasarımı sunulmuştur. Php web programlama dili ve MySql veritabanı araçlarını kullandığımız sistemin alt ve yan sistemleri ile sahip olduğu ilişkileri ve veri alışverişleri (Şekil 1) bulunmaktadır.

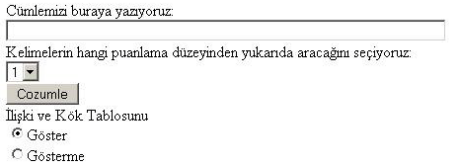


1- Kullanıcı girdisi alımı
2- Kelime çözümleme için Zemberek kullanımı
3- Kök-ek-türlerin XML olarak alınması
4- İlişki sorguları için köklerin hayat bilgisi veritabanına gönderilmesi
5- Veritabanından sorgularının cevaplarının alınması
6- Arayüzle sonuçları kullanıcıya çıktı verme

Şekil 1. Sistemdeki Veri Akışı

3.1. Sistemin Diğer Sistemlerle Etkileşimi

Sistemin kullanıcısından eş/benzer anlamlısı üretilmek istenen cümleyi, yeni cümleler üretilirken kullanılacak olan ilişkilerin CSdb'de en az sahip olacakları doğruluk puanı (Bu sayede kullanıcı istediği doğruluk oranındaki ilişkileri ve dolayısıyla cümleleri sisteme ürettirebilmektedir.) girdilerinin alınması ve üretilen cümlelerin gösterilmesi için bir web arayüzü hazırlanmıştır. Bu arayüz Şekil 2'de gösterilmiştir.



Şekil 2. Sistem Arayüzü

Sistemin ilişkide olduğu bir diğer sistem Zemberek kelime çözümleyicisidir. Burada cümleler Zemberek tarafından alınmakta ve XML olarak kelime çözümlemeleri sisteme geri döndürülmektedir. Sistem ise Zemberek'ten dönen bu verileri hayat bilgisi veritabanına, sorgulanması sistem tarafında belirli olan ve en az ortalama puanın kullanıcı tarafından belirlenmiş

olduğu sorgular gönderilmektedir. Hayat bilgisi veritabanının döndürdüğü ilişkili kelimeler, ilişki türü ve ortalama puanları kullanıcıya aktarılmak üzere sistem tarafından alınır.

3.2. Sistemin Kullanıcıya Üretilen Cümleleri Vermesi

Sistem kelime öbeklerindeki her bir kelime için verilen kelimenin ilişkili olduğu kelime sayısının bir fazlasının çarpımları sonucu kadar cümle üretebilmektedir. Bu cümlelerin hiçbiri bir diğeriyle aynı olmayacak şekilde arayüzde kullanıcıya aktarılmıştır. Örneğin dört kelimeden oluşan bir cümlede ilk kelimenin 2, son kelimenin de 1 adet ilişkilisi bulunduğunda üretilen cümle sayısı (2+1)*(0+1)*(0+1)*(1+1) = 6 olacaktır. Bu ilişki Eşitlik 1'de verilmiştir.

$$\ddot{u}cs = \prod_{k=1}^{n} (is_k + 1) \qquad (1)$$

Eşitlik 1'de: n : cümledeki toplam kelime (öbeği) sayısını, is_k : k. kelime (öbeği)nin ilişki sayısını ve $\ddot{u}cs$: üretilen cümle sayısını göstermektedir.

Kullanıcıya döndürülen sonuçlarda oluşturulan her bir cümle için tutarlılık yüzdesi hesaplanmaktadır. Bu üretilen bir cümlenin, kullanıcı tarafından girilen cümleye göre ne kadar tutarlı olabildiğini hesaplayıp bilgilendirme amacıyla yapılmıştır. Bu tutarlılık hesaplanırken bir cümlenin içindeki hayat bilgisi veritabanından gelen her bir ilişkili kelimenin ortalama puanı ve o kelimenin ilişkisinin ilişki yüzdesi çarpılarak kullanılmaktadır. Örneğin sistem tarafından oluşturulmuş üç kelimeli bir cümlede birinci kelime değişmemiş olsun, ikinci kelimenin ilişkisi "üst kavramıdır" ilişkisinin yüzdesi 80 ve ortalama puanı 5 üzerinden 4 olsun ve üçüncü kelimenin de ilişkisi "benzer anlam" ilişkisi ve yüzdesi 70, ortalama puanı 5 üzerinden 3,2 olsun. Cümlemizin tutarlılık yüzdesi şu şekilde hesaplanacaktır: (1)x(4/5x0,8)x(3,2/5x0,7) = % 28,67

Sistem tarafından üretilen cümlelerin tutarlılığı Eşitlik 2'de verilmiştir.

$$cyt = \prod_{k=1}^{n} \frac{(ity_k * iop_k)}{5} \qquad (2)$$

Eşitlik 2'de: n : cümledeki toplam kelime(öbeği) sayısını, cyt : cümlenin yüzde tutarlılığını, iop_k : k. kelime (öbeği)nin sahip olduğu ilişkinin doğruluk puanı / 5, ve ity_k : k. kelime (öbeği)nin sahip olduğu ilişki türünün tutarlılık yüzdesini göstermektedir.

Şekil 3'te kullanıcının girdiği cümledeki kelime(öbeği)ler için CSdb'den bulunan kelime(öbeği)leri ve doğruluk oranlarının verildiği ekran gösterilmiştir.


Cümlemizi buraya yazıyoruz:
Bildiklerini yazdı.
Kelimelerin hangi puanlama düzeyinden yukarıda aracağını seçiyoruz:
5
Cozumle
İlişki ve Kök Tablosunu
Göster
Gösterme

Kelime 1	İlişki	Kelime 2	Ortalama Puan
bildik	Eşanlamlı	tanıdık	5
yaz	Benzer Fiiller	yazışmak	5
yaz	Benzer Fiiller	okumak	5
yaz	Benzer Fiiller	kaleme almak	5
yaz	Benzer Fiiller	kağıda dökmek	5
yaz	Benzer Fiiller	iki satır yazmak	5
yaz	Benzer Fiiller	yayımlamak	5
yaz	Benzer Fiiller	bestelemek	5

Şekil 3. Sistemde Bulunan Kelime İlişkisi Örnekleri

Tablo 1'de ise kullanıcı cümleleri ve bunlara karşılık sistem tarafından üretilen cümlelerden örnekler ve üretim açıklamaları verilmiştir.

Giriş Cümlesi	Üretilen Cümle	Üretimde Kullanılan ilişkiler
Ali odada uyur.	Ali binada yatar.	1.ilişki: Oda - Bütünün Bölümü - bina 2.ilişki: uyumak - Bu hangi olayın parçasıdır? –yatmak
Ali limon yedi.	Ali meyve yedi.	1.ilişki: Limon - üst Kavramıdır – meyve

Gözüne vurdu.	Yüzüne saldırdı.	1.ilişki: Göz - Bunun parçaları nelerdir? - yüz 2.ilişki: vurmak - Benzer Fiiller -saldırmak
Hayatını sona erdirdi.	Yaşamını yıktı.	1.ilişki: Hayat - Eşanlamlı - yaşam, 2.ilişki: Sona erdirmek - Benzer Fiiller - yıkmak

Tablo 1. Üretilen Cümle Örnekleri

4. Sonuç

Sunulan çalışmada, ilk Türkçe hayat bilgisi veritabanı (CSdb) kullanılarak bir bilgisayar sisteminin girilen bir cümleden çıkarım yaparak girilenle benzer/aynı anlamda yeni cümleler ve doğruluk oranları üretmesi sağlanmıştır. Çalışmanın olası uygulama alanları olarak, otomatik hikaye/metin üretimi sistemleri, anlamsal metin özetleme uygulamaları, anlamsal metin sınıflandırma / kümeleme çalışmaları ve anlamsal bilgiye erişim sistemleri sayılabilir. Bu çalışma hayat bilgisi veritabanlarının, doğal dil işleme kütüphane ve uygulamalarının, gelecekteki bilgisayar sistemlerini ve interneti yönlendireceği ve bilgisayarların gündelik hayatımızı semantik çalışmalar sayesinde daha fazla kolaylaştıracağı aşikardır [6].

5. Gelecek Çalışmalar

Sistemin performansının / veriminin arttırılması için yapılabilecek çalışmalar ve yeni uygulamalar bu bölümde maddeler halinde sunulmuştur.

- Sistemin performansı kullandığı CSdb'ye çok bağımlıdır. Bu nedenle CSdb'nin içeriğinin zenginleştirilmesi ve kalitesinin arttırılması otomatik olarak uygulamamıza yansıyacaktır.
- Zemberek çözümlemelerinden sadece ilkinin kullanımından doğan hatalar, bir kelime anlamı durulaştırma işlemiyle çözümlenebilir.
- Kök ek ayrımından sonra üretilen yeni kelime köklerini tekrar eklerle birleştirecek bir sınıfın yazılması sistemin daha kullanıcı dostu olmasını sağlayacaktır.
- Üretilen cümlelerin tutarlılığını kullanıcıdan geri besleme alarak arttırabilmesi mümkündür.
- Yeni cümle üretilirken, değişim yapılan öğe türüne göre bu işlem gerçekleştirilebilir. Örneğin yüklemi sonuç ile bağlı olduğu bir kavramla değiştirirken üretilen yeni cümlenin yükleminin kipine olasılık eklenebilir. Bu sayede "Ali top oynayacak." cümlesinden "Ali yorulacak." cümlesi yerine "Ali yorulabilir." cümlesini üretmek mümkün olabilecektir.
- CSdb'de bir şeyin nerelerde bulunduğu, nerelerde yapıldığı gibi ilişkiler mevcut olduğundan kullanıcının girdiği cümlede yer almasa bile üretilen cümlelerde bu bilgiler yer alabilir. Örneğin "Ali futbol oynadı." cümlesinden "Ali stadyumda topla futbol oynadı." cümlesi üretilebilir.

Kaynaklar

[1] S. Cohen, J. Mamou, Y. Kanza, Y. Sagiv: XSEarch: A Semantic Search Engine for XML. VLDB (2003).

[2] "Türkçe Hayat Bilgisi Veri Tabanının Oluşturulması", M.Fatih Amasyalı, Bahar İnak, M.Zeki Ersen, AB 2010, Muğla, Türkiye

[3] http://code.google.com/p/zemberek/

[4] Chklovski, T. 2003. Learner: a system for acquiring commonsense knowledge by analogy. In K-CAP '03: Proceedings of the 2nd International Conference on Knowledge Capture, 4–12. New York, NY, USA: ACM Press.

[5] http://www.kemikoyun.yildiz.edu.tr/commonsense/

[6] Berners-Lee, T. Hendler, J. Lassila, O. : The Semantic Web. Scientific American, 2001