

# Soru Cevaplama

## Question Answering - QA

Prof.Dr. Banu Diri

FAQ

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



# İÇERİK

- Soru Cevaplama Nedir?
- Soru Cevaplama Sistemlerine Genel Bir Bakış
- TREC Yarışması
- Önceki Sistemler
- YTU-Tasarlanan Sistemler



# Soru Cevaplama Sistemleri

Arama Motorları = Site Arama Motorları

Soru Cevaplama Sistemleri = Cevap Arama Motorları



# Ad-Hoc Sorgular

- İlk çalışmalar çoğunlukla ad-hoc bir sorgunun cevabı ile ilgilenirdi :

## *General Clark*

- Sadece birkaç kelime; Sorunun tamamı değil
- İstenilen konu hakkında genel bir bilgi

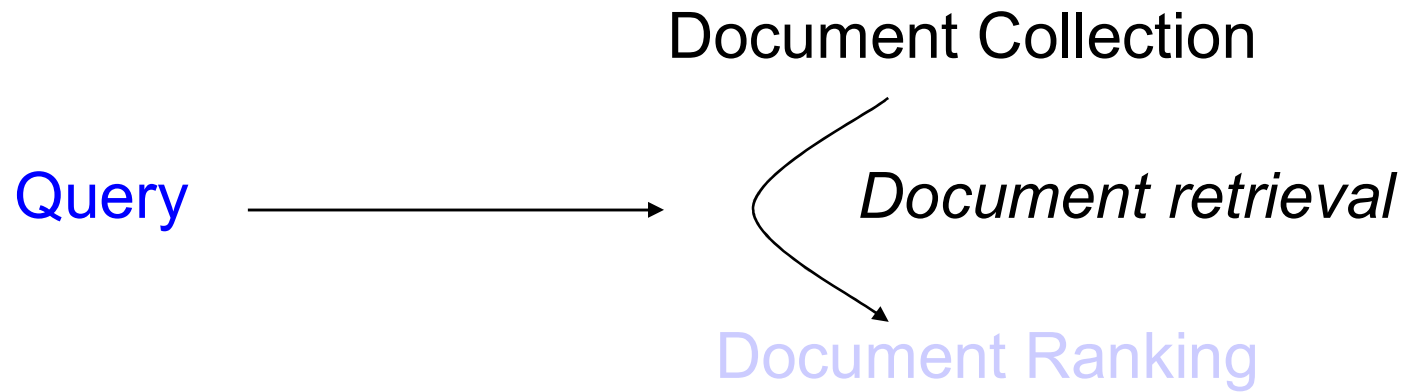


# Ad-Hoc Sorguların Cevapları

- Bilgiye Erişimdeki ilk çözümler:
  - Vector-based yöntemler
  - SVD, sorgu genişletme, dil modelleri
  - Cevap olarak döndürülen sayfa
- Sonuç
  - Google, Altavista, Yahoo



# Geleneksel IR



# Bütün sorgular Ad-Hoc değildir!

## *How old is General Clark?*

- Bu soru Ad-hoc paradigmasına uymaz...
  - “How” ve “is” kelimeleri sorunun cevabı için bir ipuçtu olsa bile geleneksel ad-hoc IR sistemler bu kelimeler ile ilgilenmezler



## Geleneksel IR sistemler sayfa döndürür

- Aranan sorgu kelimesi hakkında genel bir bilgidir
- Aranılan cevap

*How old is General Clark? 58*

*How long did Clark serve in the military? 36 years*

*Will Clark run for president? Maybe*





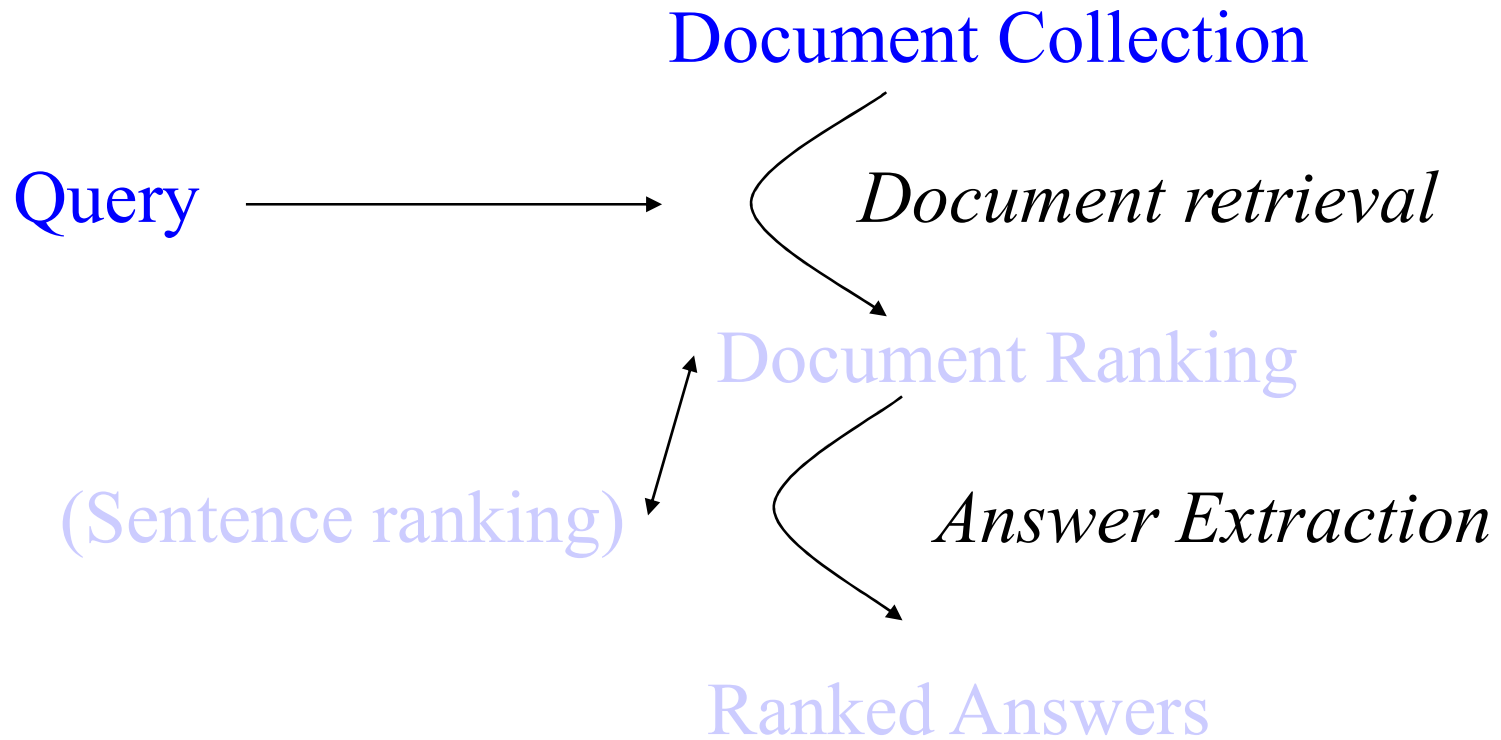
# Geri getirim olarak Soru Cevap

Verilen bir **doküman topluluğu** ve **soru**:

Soru Cevaplama sistemi, sorunun cevabı içinde olacak kısa özet bir text getirir.



# Soru Cevaplama



- Mükemmel Recall için, sorunun cevabı doküman koleksiyonu içerisinde bir kez görülmelidir
- Soru Cevaplama sistemleri aslında bir dokümanı anlamayı da zorunlu kılmaktadır
- Soru Cevaplama sistemleri sadece bilgi getirimi yapmazlar, dokümanı da anlamaya çalışır



# QA basamak taşları

- Soru Cevaplama sistemleri sorunun cevabına doğrudan odaklanır
  - Soruların cevapları doküman içerisinde aynen gözükebilir

*How old is General Clark?*

- Doküman içerisinde cevabı tam olarak görünmeyen sorulara nasıl cevap vereceğiz?

*How long has Clark been in the military?*

*Will Clark run for President?*



# Soru Cevaplama Yöntemleri

İki temel yöntem vardır

- Answer Preference Matching
- Answer Context Matching



# 1. Answer Preferences

- Question Analysis
- Type identification
- Learning Answering Typing

# 2. Answer Context

## Learning Context Similarity

- Alignment
- Surface Text Patterns



## Cevap tipinin belirlenmesi

Sorunun kendisinden olası cevabın türünü çıkarmak

- *How old is General Clark?*
  - How Old →
- *When did Clark retire?*
  - When →
- *Who is the NBC war correspondent?*
  - Who →



# 1. Answer Preferences

- *How old is General Clark?*
  - How Old → Age
- *When did Clark retire?*
  - When → Date
- *Who is the NBC war correspondent?*
  - Correspondent → Person





# Wh- Words

Who	Person, Organization, Location
When	Date, Year
Where	Location
In What	Location
What	??



## 1. Answer Preferences

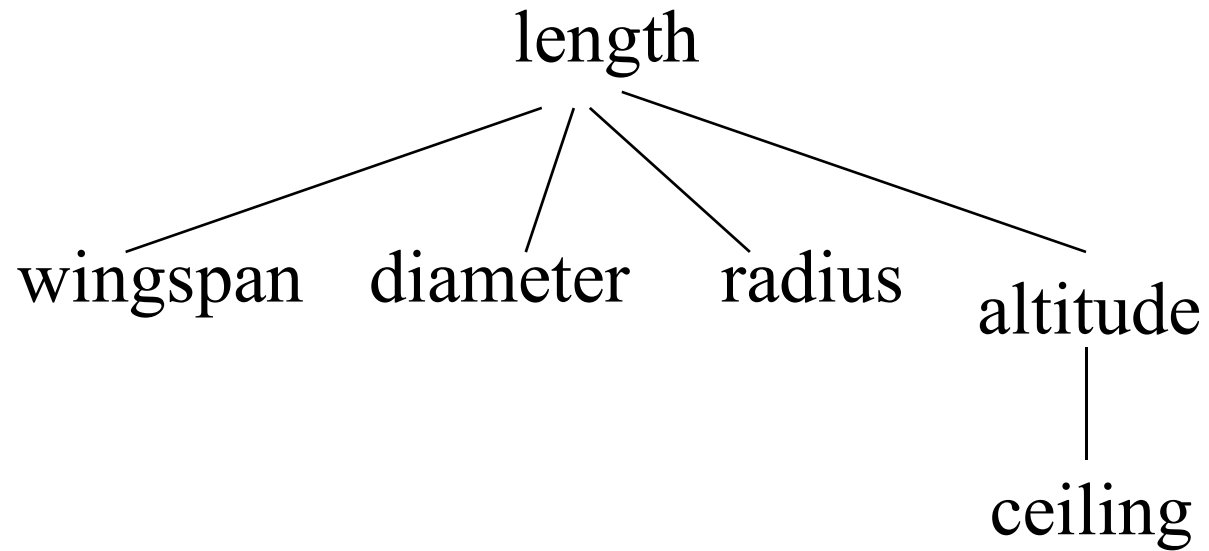
Tüm olasılıkları çıkarmak zordur...

*What is the service ceiling for a PAC750?*

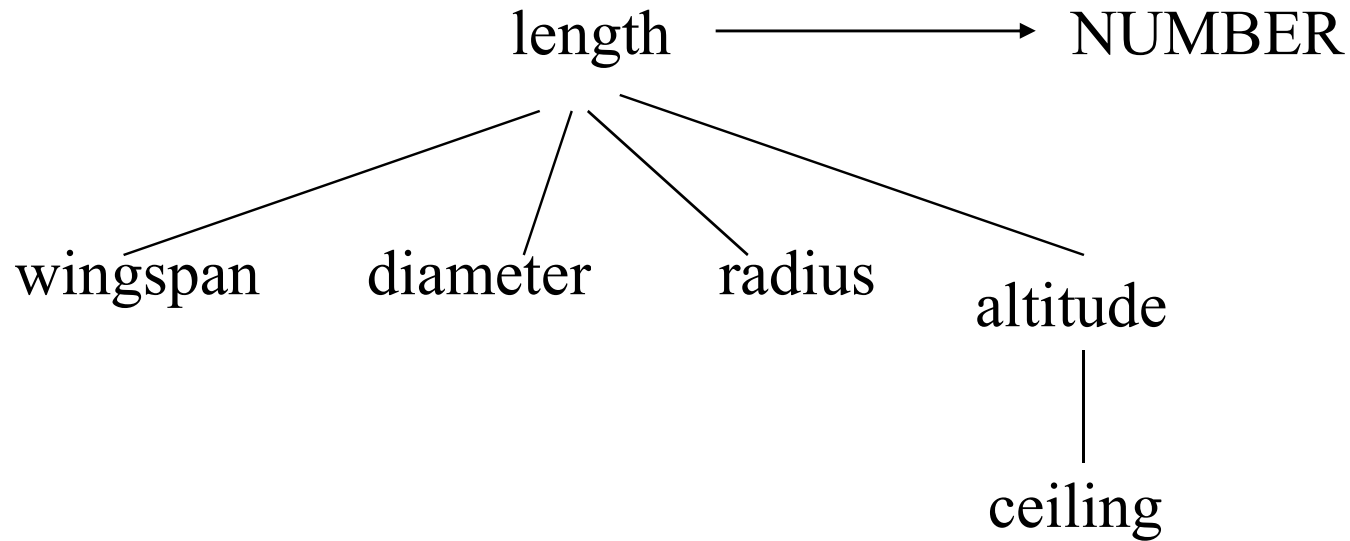


# 1. Answer Preferences

## WordNet



## Cevabın türü için Word Net yardımcı olur



*What is the service **ceiling** for a PAC750?*

# Cevabın tipi bize ön bilgi verecektir...

- Soru hakkında sahip olunan ön bilgi bize cevabın tipini verecektir
- Bu ön bilgiyi kullanabilmek için de potansiyel aday cevapları çıkarmalıyız



# 1. Answer Preferences

- Number  $\rightarrow$   $[0-9]^+$
- Date  $\rightarrow$  (\$month) (\$day) (\$year)
- Age  $\rightarrow$  0 – 100

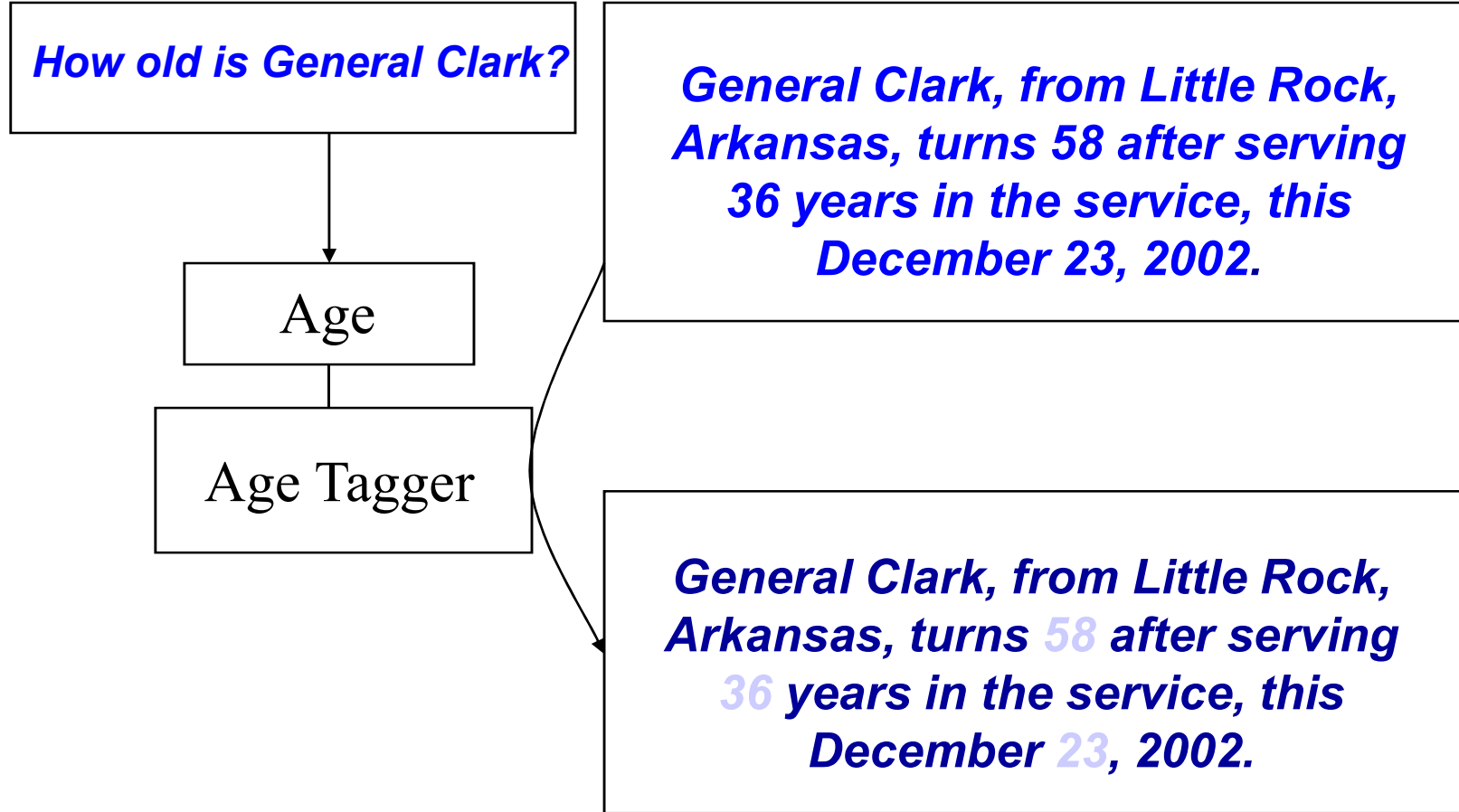


## ... Diğer Karışık Durumlar

- ***Who shot Martin Luther King?***
  - Person ön bilgi olarak verilmiş
  - Named Entity Identifier ihtiyaç vardır
- ***Who saved Chrysler from bankruptcy?***
  - Cevap sadece *person* ile sınırlı değildir
  - Uygun adayları bulmak için etiketlenmiş cümleler gerekir



# Basit Cevap Çıkarımı





## Cevabın türünü nasıl öğreniriz?

- Cevabın türünü tahmin eden bir model kurulur  
 $P(\text{type}|\text{question})$ 
  - Who →
    - Person (“Who shot Kennedy?” Oswald)
    - Organization (“Who rescued Chrysler from bankruptcy?” The Government)
    - Location (“Who won the Superbowl?” New England)



# Eđitim verisi nedir?

- Etiketlenmiř sorular
  - “*Who shot Kennedy*” [PERSON]
- Problems :
  - Etiketleme oldukça pahalıdır



## MI Model (Mutual Information)

- Etiketler kullanılarak MI model kurulabilir
  - Soru içerisindeki head-Word den tahmin yapılabilir
    - MI(Question Head Word, Type Tag)

$$= P(\text{Type Tag} | \text{QuestionHeadWord})$$

-----

$$P(\text{Type Tag})$$



# MaxEnt (Maximum Entropy) Model

- Eğitim için sadece head word leri kullanmak yerine MaxEnt model kullanılarak cümle içerisindeki diğer özelliklerde birleştirilerek kullanılabilir
  - Doküman içerisindeki kelimelerin kullanım sıklıklarına bakarken, türlerinin ve özelliklerinin de ne olduğuna bakılmalı
  - Cevap için uygun türü belirlerken bu birlikte geçme istatistikleri de kullanılmalıdır

*“What was the year in which Castro was born?”*



## Multinomial Logistic Regression = MaxEnt

Rasgele bir deęişken olan  $x$  örneęinin hangi sınıfa dahil olacaęını tahmin eder  $C$  ( $c_1, c_2, c_3, \dots, c_n$ )

$$p(c|x) = \frac{1}{Z} \exp \sum_i w_i f_i$$

$Z$  normalizasyon parametresidir

$$Z = \sum_C p(c|x) = \sum_{c' \in C} \exp \left( \sum_{i=0}^N w_{c'i} f_i \right)$$

$$p(c|x) = \frac{\exp \left( \sum_{i=0}^N w_{ci} f_i \right)}{\sum_{c' \in C} \exp \left( \sum_{i=0}^N w_{c'i} f_i \right)}$$

$f$  deęeri örnek

$w$  ise ilgili örneęin aęırlık deęeridir

MaxEnt Genelleştirilmiş hali

$$p(c|x) = \frac{\exp \left( \sum_{i=0}^N w_{ci} f_i(c, x) \right)}{\sum_{c' \in C} \exp \left( \sum_{i=0}^N w_{c'i} f_i(c', x) \right)}$$



## Örnek:

Secretariat/NNP is/BEZ expected/VBN to/TO **race**/?? tomorrow/

↑  
current word

$$f_1(c, x) = \begin{cases} 1 & \text{if } word_i = \text{"race"} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(c, x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

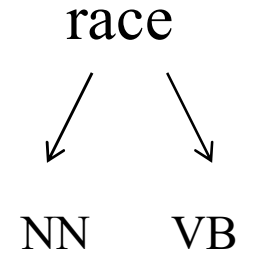
$$f_3(c, x) = \begin{cases} 1 & \text{if } \text{suffix}(word_i) = \text{"ing"} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(c, x) = \begin{cases} 1 & \text{if } \text{is\_lower\_case}(word_i) \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_5(c, x) = \begin{cases} 1 & \text{if } word_i = \text{"race"} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_6(c, x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

Her bir  $f$  özelliği için  $w$  ağırlık değeri verilir



VB etiket değeri ile geçen **race** kelimesinin ağırlık değeri  $w_1(c, x)$  olsun. Kendisinden önce gelen kelimenin **TO** olması durumunda kelime sınıfının da **VB** olması halinde ağırlık değeri  $w_2(c, x)$  olup,  $w_1$ 'den daha kuvvetlidir.



		f1	f2	f3	f4	f5	f6
VB	f	0	1	0	1	1	0
VB	w		.8		.01	.1	
NN	f	1	0	0	0	0	1
NN	w	.8					-1.3

Değerleri eşitlikte yerine koyduğumuzda

$$P(NN|x) = \frac{e^{.8} e^{-1.3}}{e^{.8} e^{-1.3} + e^{.8} e^{.01} e^{.1}} = .20$$

$$P(VB|x) = \frac{e^{.8} e^{.01} e^{.1}}{e^{.8} e^{-1.3} + e^{.8} e^{.01} e^{.1}} = .80$$

Maximum değeri aldığımızda *race* kelimesinin POS tag değeri **VB** dir

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|x)$$



## Cevabın türü yeterli midir?

- Doğru cümleyi bulunp, cevabın türü de bilinse cevap için halen belirsizlikler devam edebilir
- Denemeler göstermiştir ki, her sorunun cevabı olabilecek 2/3 oranında uygun türde cevap cümlesi olabilir
- Yüksek başarımlı isteyen sistemler için bu kabul edilebilir bir durum değildir





*Who shot Martin Luther King?*



Answer Preference



Answer Context



- Çoğu sistem uygun görünen bağlam içerisinde doğru türdeki cevabı arar.
  - Çakışan kelime sayısı
  - Genişletilmiş sorgu sayısı



## Bir diğer alternatif...

- Sorumuz “*Who shot Kennedy*” ise
- Şablon ile eşleşen tüm cevaplar aranır
  - “X shot Kennedy”
- veya
  - “Kennedy was shot by X”



# Local Alignment

Who shot Kennedy?

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.



Tipe uygun 3 adet potansiyel cevap



Question

Head  
word → *Who shot Kennedy?*

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

### Matching Context



*Who shot Kennedy?*

Anchor word

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.



*Who shot Kennedy?*

Potential alignments

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.



*Who shot Kennedy?*

One Alignment

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

Three Alignment Features :





*Who shot Kennedy?*

One Alignment

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

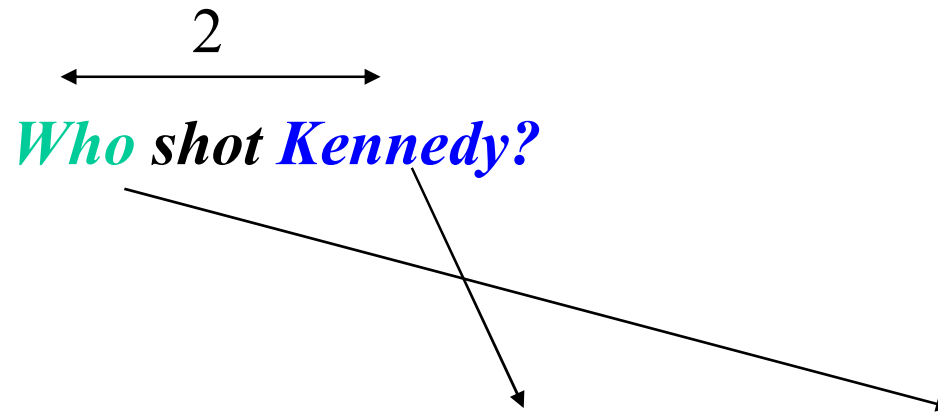


1

Three Alignment Features :

1. **Dws** : Distance between Question Head word and Anchor in the sentence





Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

Three Alignment Features :

2. **Dwq** Distance between Question Head word and Anchor  
In the question



*Who shot Kennedy?*

← Headword position flipped

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

Three Alignment Features :

3. **R** : Has the Head Word changed position?



- $\Pr(\text{answer} \mid \text{question, sentence}) =$   
 $\Pr(\text{Dws} \mid \text{answer, question, sentence})$   
 $*\Pr(\text{Dwq} \mid \text{answer, question, sentence})$   
 $*\Pr(\text{R} \mid \text{answer, question, sentence})$
- Yeni terimler eklenerek model zenginleştirilebilir

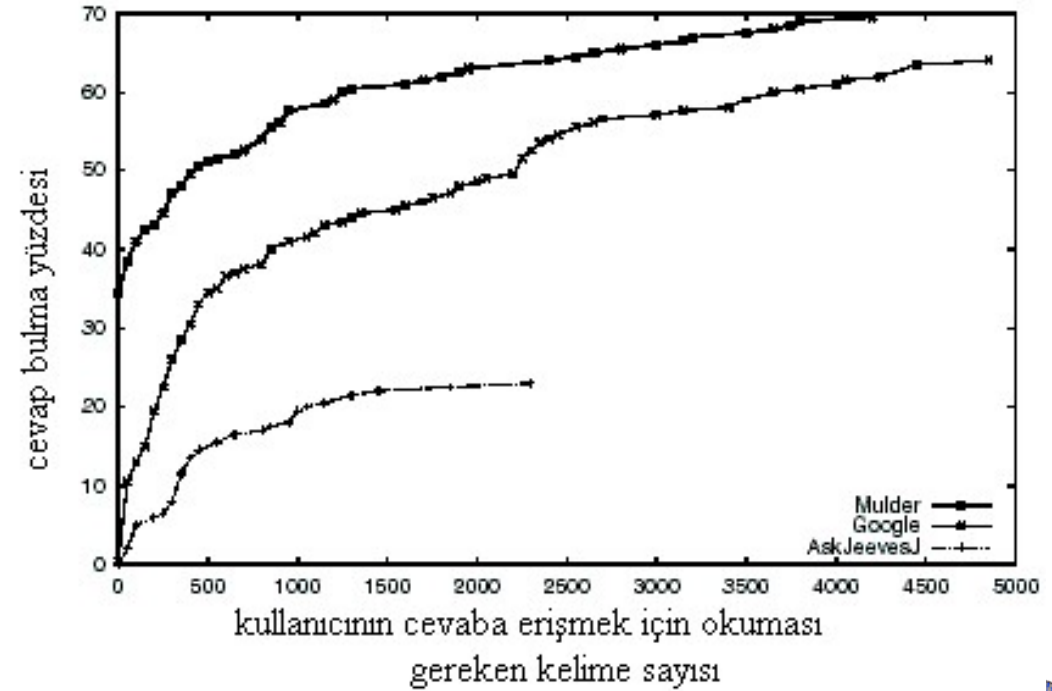
*Local Alignment Model*, cevabı içeren cümleden doğru cevabı ayırmak için sorunun bağlamını kullanan güçlü bir yöntemdir.



# Soru Cevaplama Sistemlerinin Avantajları

- Doğal Dil kullanıcılar
- Cevabı içeren sitelerin listesi yerine cevabın kendisini vermesi

Sonuç olarak;  
Daha az kullanıcı emeği



# TREC (Text REtrieval Conference) Konferansları

- Büyük dokümanlar (1 milyon gazete makalesi) içerisinde cevap arayan sistemlerin yarıştığı bir konferans
- Her yıl yaklaşık 30 sistem katılıyor
- Her sisteme aynı sorular sorulup, puan veriliyor



# Yarışmanın Kuralları

- Cevaplar dokümanlardan elde edilmeli
- Sistemlerden sıralanmış 5 adet cevaptan oluşan bir liste isteniyor



# Sistemlerin Puanlanması (Mean Reciprocal Rank)

- Sistemlerin verdiği sıralı cevap listesinden kaçının doğru cevabı içerdiğine göre sistemler puanlanır.

Sistemin  $i$ . sırada verdiği doğru cevap için aldığı puan

$$RR_i = \frac{1}{r_i}$$

Sistemin  $n$  adet sorudan aldığı puanların ortalaması (MRR)

$$MRR = \frac{1}{n} \sum_{i=1}^n RR_i$$





# TREC Soru ve Cevap Örnekleri

<b>Soru 1</b>	How far is it from Denver to Aspen?
Cevap 1	The Aspen/Snowmass area is about 200 miles southwest of Denver.
<b>Soru 2</b>	What country is Modesto, California in?
Cevap 2	Modesto in Stanislaus Country.
<b>Soru 3</b>	Who was Galileo?
Cevap 3	Galileo Galileo, the astronomer.
<b>Soru 4</b>	What is an atom?
Cevap 4	Atoms, long considered to be the smallest units of matter.
<b>Soru 5</b>	What was the name of the first Russian astronaut to do a spacewalk?
Cevap 5	Aleksei A. Leonov.



# TREC ve Internet Veritabanları Arasındaki Farklar

	TREC	INTERNET
Doküman sayısı	<1,000,000	<3,000,000,000
Doküman türü	Gazete makaleleri	Her tür
İşlem	Internet'ten bağımsız	Internet'e bağımlı
Cevap formatı	50-250 bytes	Her türlü
Yanlış bilgi	Nadir	Çok fazla
Doğru cevabı içeren doküman sayısı	Az	Çok fazla



# Önceki Sistemler

- **WEBCLOPEDIA**

Southern California Üniversitesi,1999; Soru türü kütüphanesi; Cevap şablonlarının otomatik olarak toplanması; Tanımlar için WORDNET

## QuASM (Question Answering System)

Amerika'daki istatistiksel veritabanlarını sorgulamak için geliştirilmiş bir sistemdir; Sistem [www.fedstats.gov](http://www.fedstats.gov) adresindeki yaklaşık 170,000 web sitesinde arama yapmaktadır; Kullanıcının girdiği soru incelenerek türü belirlenir. Bu işlem Naive Bayes metodu kullanılarak yapılmıştır. Soru ve türlerinden oluşan bir veritabanında kelimelerin hangi tür sorularda bulduklarının olasılıkları  $P(w_i|C)$  önceden çıkarılmıştır; Kullanıcı sorusunun hangi sınıfa ait olduğu ise aşağıdaki formülle bulunmuştur.

- **AnswerBus**

$$P(Q|C) = P(w_1|C) * P(w_2|C) * \dots * P(w_n|C)$$

Altı farklı dili destekler. İngilizce sorguları doğrudan ararken, Almanca-İtalyanca-Portekizce-Fransızca-İspanyolca soruları BabelFish ile İngilizceye çevirip arama yapar

- **Start**

İngilizce dilini destekler



# Otomatik Şablon Toplama-1

- Örnek soru-cevap ikilileri alınmış (Mozart-1756),(Gandhi-1869) ve arama motoruna sorgu olarak gönderilir
- İlk 1000 sonuç sayfasındaki bu ikililerin beraber geçtiği cevap cümlelerinde en uzun eşleşen parçalar bulunmuştur.
- Cevap cümlelerindeki en uzun eşleşen parçadaki soru kelimesi; <Name>, cevap kelimesi ise <Answer> ile etiketlenmiştir.



Mozart (1756-1791).
Mozart was born on January 27, 1756
Mozart was born in 1756 in Salzburg, Austria
Mozart was born in Salzburg, Austria, Jan. 27, 1756, the son of Leopold Mozart
Mozart, born in 1756
Mozart was born there in 1756



# Otomatik Şablon Toplama-2

- Doğum yılını soran yukarıdaki ikilliler için
  - 1.00 <NAME> ( <Answer> - )
  - 0.87 <NAME> was born on < Answer > , in
  - 0.87 <NAME> was born on < Answer >
  - 0.86 <NAME> was born on < Answer > in
  - 0.86 <NAME> was born on < Answer >

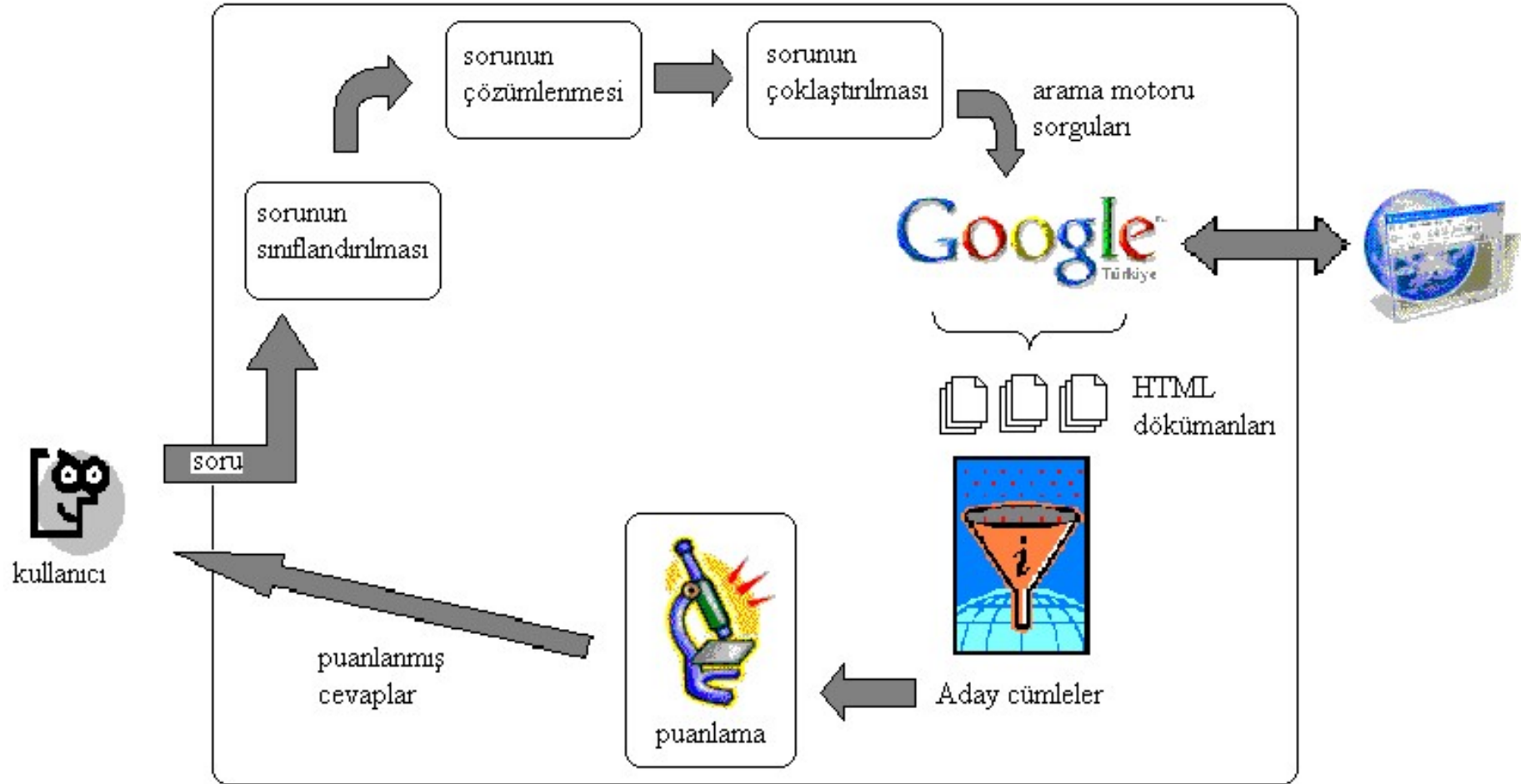
şeklinde birçok şablon elde edilmiştir.

- Şablonların başlarındaki ağırlık değerleri ise; şablonun geçtiği doğru cevabı içeren cümle sayısının, şablonun geçtiği cevap cümlesi sayısına bölümünden elde edilmiştir.



# YTÜ-Baybilmiş

## Soru Cevaplama Sisteminin Mimarisi\*



\* Hemen hemen tüm QA sistemleri benzer bir mimariye sahiptir.

# Sistemin Akışı

- Sorunun Tipinin Belirlenmesi
- Kullanıcı Sorusunun Çözümlemesi
- Kullanıcı Sorusunun Çoklaştırılması
- Arama Motorundan Sonuçların Alınması
- Sonuç Sayfalarından Aday Cümlelerin Seçilmesi
- Aday Cümlelerin Puanlanması



# Sorunun Tipinin Belirlenmesi

- Sorgudaki soru kelimeleri incelenerek, elle oluşturulan bir tablo kullanılmıştır.

sorukelime (head word)	sorutürü	Silinecek kelime
yüzde kaç	oran	kaç
yıl nedir?	yıl	nedir?
yıl kaçtır?	yıl	kaçtır?
yapıtaşları nelerdir?	bileşen	nelerdir?
tarih?	zaman	tarih?
tarih nedir?	zaman	tarih nedir?
tarifi nedir?	nedir	nedir?
şehri hangisidir?	İl	hangisidir?
sebebi nedir?	neden	nedir?
sayısı nedir?	sayı	nedir?
nüfusu nedir?	nüfus	nedir?
nüfusu ne kadardır?	nüfus	ne kadardır?
nüfusu kaçtır?	nüfus	kaçtır?





# Kullanıcı Sorusunun Çözümlemesi

- Sistemde, ilk olarak kelimenin olası tüm gövdeleri bulunmaktadır. Örneğin; “ağacı” kelimesi için “ağ”, “ağa” ve “ağaç” kelimeleri gövde olarak bulunmaktadır.
- Çekim eklerinin varlığı araştırılmaktadır. Eğer gövde isim soylu ise kelimenin içerdiği isim çekim ekleri, şayet fiil soylu ise çekim ekleri ve fiilimsi ekleri araştırılır.
- Türkçe’deki eklerin birbirini içermesinden dolayı (örnek: -m, -miz gibi), bu işlemler sonucunda birden fazla muhtemel çözüm bulunmaktadır. Bu çözümlerden bir kısmı, kelimenin tamamını içermediğinden dolayı zaten elenir.



- Geriye kalan olası çözümler içerisinde doğru çözümü bulurken, kelimenin cümle içerisindeki durumuna bakılmaktadır. Örneğin, “**evi**” kelimesi için “**ev**” gövde olarak ve “**-i**” eki ise ya *hal eki* ya da *iyelik eki* olarak bulunur. Eğer cümlede “**Ayşe’nin evi**” benzeri bir tamlama varsa bunun *iyelik eki* olduğu anlaşılır. “**Şu evi beğendim**” cümlesinde ise “**-i**” ekinin *hal eki* olduğu anlaşılır.



# Çözüm Örnekleri

Cümle	Orijinal kelime	Eklerine ayrılmış kelime	Gövde ve eklerin tanımları	Kelimenin görevi
13 Aralık'ta okullar açıldı.	13	13	i[i]	sayı
	Aralık'ta	aralık*da	i[i]*bul	zaman
	okullar	okul*lar	i[i]*ç	
	açıldı	açıl*dı	f[f]*dgz	
29 Mayıs 1453'te okullar açıldı.	29	29	i[i]	sayı
	Mayıs	mayıs	i[i]	zaman
	1453'te	1453*te	i[i]*bulmayr	zaman
	okullar	okul*lar	i[i]*ç	
	açıldı	açıl*dı	f[f]*dgz	
29 km yürüdüm.	29	29	i[i]	sayı
	km	km	i[i]	uz
	yürüdüm	yürü*dü*tm	f[f]*dgz*setl	
29 km'lik yolu 3 saate aldık.	29	29	i[i]	sayı
	km'lik	km*lik	i[i]*bulmayr	uz
	yolu	yol*lu	i[i]*bli	
	3	3	i[i]	sayı
	saate	saat*te	i[i]*yon	zaman
	aldık	al*dı*ık	f[f]*dgz*seçl	
Ali üniversiteyi 3 yılda bitirdi.	Ali	ali	ö[i]	özel isim
	üniversiteyi	üniversite*y*ı	i[i]*kh*bli	
	3	3	i[i]	sayı
	yılda	yıl*da	i[i]*bul	zaman
	bitirdi	bitir*di	f[f]*dgz	
Ankara ilini gördüm.	Ankara	ankara	ö[i]	özel isim
	ilini	il*i*ni*	i[i]*iet3*kh*bli	yer
	gördüm	gör*dü*tm	f[f]*dgz*setl	
Bir fil 2500 kilodur.	Bir	bir	sf[sf]	
	fil	fil	i[i]	
	2500	2500	i[i]	sayı
	kilodur	kilo*dur	i[i]*ef_gez	ağ

# Kelimelerin etiketlemedesinde kullanılan anahtar kelimelere örnekler

[yer]	[zaman]	[ag]	[uz]
ili	dakika	gram	metre
barajı	saat	kilo	km
dağı	gün	kg	kilometre
gölü	hafta	kilogram	santim
şehri	ay	ton	cm
yöresi	yıl	litre	santimetre
bölgesi	asır		mil
kenti			



# Kullanıcı Sorusunun Çoklaştırılması

- Kullanıcının sorduğu sorudan iki adet arama sorgusu elde edilmiştir.
  - 1- Soru kelimeleri atılarak elde edilen
  - 2- Çekim ekleri de atılarak elde edilen

Kullanıcının Sorusu	1. sorgu Soru kelimeleri atılmış hali	2. Sorgu Eklerinden arındırılmış hali
Asya ile Avrupa'yı ayıran boğazın adı nedir?	Asya ile Avrupa ayıran boğazın adı	Asya ile Avrupa ayır boğaz ad
enzimlerin vücuttaki görevi nedir?	enzimlerin vücuttaki görevi	enzim vücut görev
Deli Dana hastalığının belirtileri nelerdir?	Deli Dana hastalığının belirtileri	Deli Dana hastalık belirti
dünyanın en uzun boylu insanı kimdir?	dünyanın en uzun boylu insanı	dünya en uzun boy insan
su hangi atomlardan oluşur?	su atomlardan oluşur	su atom oluş



# Google

Google'da Ara: Öğrenme algoritmaları - Microsoft Internet Explorer

Dosya Düzen Görünüm Sık Kullanılanlar Araçlar Yardım

Geri - - - - - Ara Sık Kullanılanlar Ortam

Adres <http://www.google.com.tr/search?hl=tr&ie=UTF-8&oe=UTF-8&q=%C3%96%C4%9Frenme+algoritmalar>

Google™ Öğrenme algoritmaları Google'da Ara

Web'de Ara Türkçe sayfalarda ara

Web Grafikler Gruplar Dizin

Web Öğrenme algoritmaları için arandı.

[Ethem Alpaydın'a ödül](#)  
... Ödül alma nedenim Yapay **Öğrenme**, özellikle de yapay sinir ağlarındaki **öğrenme algoritmaları** konusunda yaptığım çalışmalar. ...  
[www.boun.edu.tr/~pubrel/news/ocak2003/6.htm](http://www.boun.edu.tr/~pubrel/news/ocak2003/6.htm) - 7k - [Önbellek](#) - [Benzer sayfalar](#)

[Derslerin Listesi](#)  
... Sınırlanmamış optimizasyon ve **öğrenme algoritmaları**. Lineer, quadratik programlama ve lineer tümlenmeli problemler için sinir ağları. ...  
[www.fbe.ktu.edu.tr/anabilimdallari/bilgisayar/bilgisidet.html](http://www.fbe.ktu.edu.tr/anabilimdallari/bilgisayar/bilgisidet.html) - 24k - [Önbellek](#) - [Benzer sayfalar](#)

[İstanbul Tabip Odası](#)  
... Bölümü □ Bilgisayar mühendisliği alanında yapay **öğrenme** ve özellikle yapay sinir ağlarında **öğrenme algoritmaları** konusundaki uluslararası ...  
[www.istabip.org.tr/genel/tubitak2002.asp](http://www.istabip.org.tr/genel/tubitak2002.asp) - 18k - [Önbellek](#) - [Benzer sayfalar](#)

[BAŞKENT ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ ...](#)  
... Sinirsel çağrışimli yardımcı bellekler, şekil sınıflayıcıları, fonksiyon yaklaşıtıncılar ve **öğrenme algoritmaları**. ...  
[www.cs.baskent.edu.tr/ylderslericerik.html](http://www.cs.baskent.edu.tr/ylderslericerik.html) - 23k - [Önbellek](#) - [Benzer sayfalar](#)

- En büyük içerik: Milyarlarca web sayfası
- En hızlı arama
- En iyi puanlandırma tekniği
- Sayfa özetleri
- Önbelleğe sahip



# Arama Motorundan Sonuçların Alınması

- Cevap cümlelerinin elde edilebileceği iki kaynak bulunmaktadır.
  - Arama motorunun sonuç sayfasındaki sayfa özetlerinden
  - Sonuç sayfasındaki bağlantıların gösterdiği sayfalardan
- Her iki yaklaşımın avantaj ve dezavantajları aşağıda verilmiştir.

Cevabın nerede aranacağı	Avantajları	Dezavantajları
Sadece sayfa özetlerinde arama	hızlı	Cümleler kesik ve anlamsız dolayısıyla daha az doğru sonuçlar (%43.9)
Sayfaların kendilerinde arama	Cümleler tam ve anlamlı ve dolayısıyla daha doğru cevaplar (%43.9)+%16.6	Çok yavaş



# Google'ın Ön Belleği

- Sayfaların kendilerinden cevap üretilirken sayfaların orijinal hallerinin yerine Google'ın ön belleğindeki halleri kullanılmıştır. Bunun iki önemli avantajı vardır.
  - Sayfalardaki bilgilerin indirilmesi daha az zaman almaktadır
  - Sayfanın arama yapılırkenki (belki değiştirilmiş ya da yayından kaldırılmış) içeriği yerine Google'da indekslendiği hali kullanılarak daha doğru bilgiye erişim sağlanmaktadır.





# Sonuç sayfalarından aday cümlelerin seçilmesi

- Bir cümlenin cevap cümlelerinden olabilmesi için aşağıdaki şartları sağlaması gerekmektedir.
  - İçinde soru kelimelerinin sayısının en az yarısı kadar sorgu kelimesi içermeli
  - Sorgudaki özel isim olarak etiketlenmiş en son kelimeyi içermeli
  - Soru türü, cevapta sayısal bir değer beklenen türde ise (yıl, mesafe, oran, ağırlık, vb.) içinde sayı bulundurmalıdır
- Puanlama işlemi zaman alıcı bir işlem olduğundan, yukarıdaki filtreleme sayesinde sistemin cevap verme süresi kısaltılmıştır



# Aday cümlelerin puanlanması

- Filtreden geçebilen cümleler 3 farklı şekilde puanlanmaktadır:
  - Eşleştirme puanı
  - Sıralama puanı
  - Şablon puanı



# Eşleştirme puanının belirlenmesi

- Cümlelerin içerdiği sorgu kelimesi sayısı ile bulunur
- Ör: “**Boğaz Köprüsü’nün yüksekliği ne kadardır?**” sorusu için

Cevap cümlesi	Eşleştirme puanı
Boğaz Köprüsü'nün yüksekliğinin yaklaşık 3 katına eşdeğerdir .	3
64 metre yüksekliğindeki Boğaz Köprüsü'yle, Saipem 7000 .	3
Boğaz Köprüsü'nün denizden yüksekliği ise 60 metre.	3
Mesela Boğaziçi Köprüsü'nün deniz seviyesinden yüksekliği sadece 64 .	3
boyu 165 metre olan Birinci Boğaz Köprüsü'nün .	2



# Sıralama puanının hesaplanması-1

- Cümle içindeki sorgu kelimelerinin birbirine yakınlığı, cevabın o cümlede olma olasılığını arttırmaktadır. Bu yüzden sorgudaki kelimelerin, cümledeki sorgu kelimelerinin sayısına ve yakınlığına bağlı olarak bir formül yardımıyla 0-10 arasında bir sıralama puanı hesaplanır.

## Parametreler:

$$SP = \left[ \frac{10 * C_k * (C_k - 1)}{Tp * S_k} \right]$$

$C_k$  cümlenin içerdiği sorgu kelimesi sayısı

$S_k$  sorgudaki kelime sayısı

$Tp$  cümledeki sorgu kelimelerinin birbirlerine kelime bazındaki uzaklıklarının toplamı



## Sıralama puanının hesaplanması-2

- Ör: “**A B C D**” sorgu cümlesi için  
“**A E F G H B C**” cevap cümlesinin sıralama puanının hesaplanması:

Cevap cümlesindeki sorgu kelimelerinin cümledeki kaçınıcı kelime oldukları bulunur. (1 – 6 – 7)

***Tp*** ifadesi; bu dizideki her rakamın bir sonrakinden farkı bulunduktan sonra, bu farkların mutlak değerleri toplanarak elde edilir.

$$T_p = |(1-6)| + |(6-7)| = 6$$

Örneğin sıralama puanı

$$**SP** = [10 * 3 * (3-1)] / (6 * 4) = 2.5$$



# Sıralama puanının hesaplanması-3

- Ör: “Boğaz Köprüsü’nün yüksekliği ne kadardır?” sorusu için

Cevap cümlesi	Sıralama puanı
64 metre <b>yüksekliğindeki Boğaz Köprüsü</b> ’yle, Saipem 7000 arasında yaklaşık 8 metre mesafe kalacak.	10
Fatih Sultan Mehmet <b>Köprüsü</b> ’nün güvenli geçiş <b>yüksekliği</b> 64 metre olduğu için bu köprünün altından geçişte bir sorun görülüyor.	2
Gemi, denizden 64 metre <b>yüksekliğinde</b> olan <b>Boğaziçi Köprüsü</b> ’nün altından yaklaşık 10 dakikada geçişini tamamladı.	6



# Şablon puanının belirlenmesi -1

- Kullanıcının girdiđi soru türüne ait şablonlar veritabanından çekildikten sonra cümlede bu şablonların olup olmadığı araştırılır. Eğer varsa, o var olan şablona ait ağırlık değeri cümlenin şablon puanı olarak belirlenir.



## Şablon puanının belirlenmesi -2

- Örnek olarak “**plastik neden elde edilir?**” sorusunun türü “**bileşen**” olarak belirlendikten sonra; şablonlar veritabanından “**bileşen**” sorusuna cevap olabilecek şablonlar seçilir.

Şablonlar		
sorutipi	Cevaptipi	ağırlık
bileşen	orj=ana%maddesi	5
bileşen	orj=elde%edil	6
bileşen	orj=hammadde	5
bileşen	orj=içerir	5
bileşen	orj=içinde	5
bileşen	orj=kullanılarak	5
bileşen	orj=yapılır	7
bileşen	tnm=ayr#orj=elde%edil	9
bileşen	tnm=ayr#orj=oluş	7
bileşen	tnm=ayr#orj=üretilir	8
bileşen	tnm=ayr#orj=yapılır	9





# Şablonlar Veritabanınının Oluşturulması-1

- Şablonlar iki farklı yolla elde edilmiştir:
  - “Nedir” şablonlarının oluşturulması
  - “Nedir haricindeki” şablonların oluşturulması



# Şablonlar Veritabanının Oluşturulması-2

- “Nedir” şablonlarının oluşturulması:
  - “Nedir” soruları bir kelimenin tanımını sorduğu için Internet'teki bir sözlükten yararlanılmıştır. Yaklaşık 4000 Türkçe kelimedenden oluşan listedeki kelimeler Türk Dil Kurumu'nun Internet sitesindeki sözlüğe (<http://www.tdk.gov.tr/tdksozluk/sozara.htm>) gönderilip sonuç sayfalarındaki kelime anlamları bir dosyaya kaydedilmiştir.
  - Bir sözlükteki en çok tekrar eden kelimeler Türkçe'deki tanım ifadelerini içereceğinden; kelime anlamlarını içeren dosyadaki kelimelerden en çok tekrarlanan 50 kelime (Türkçe'de en sık tekrarlanan kelimeler çıkarıldıktan sonra) “nedir sorularının” şablonları olarak belirlenmiştir.



# Şablonlar Veritabanının Oluşturulması-3

## “Nedir” sorularının şablonları

orj=:	orj=gösteren	orj=olarak%bilin
orj=ad	orj=herhangi	orj=olarak%tanım
orj=adı%verilir	orj=için%	orj=olmak
orj=alan	orj=için%kullan	orj=olmayan
orj=alınan	orj=ifade%eder	orj=oluşan
orj=amacıyla	orj=ilgili	orj=oluşturan
orj=anlamında	orj=ismidir	orj=sağlamak
orj=araç	orj=kalan	orj=sağlayan
orj=aracı	orj=karşılık	orj=tanım
orj=bakımından	orj=kimse	orj=tanımlamak%için%kullan
orj=biçimde	orj=kısaltılmış%halidir	orj=tarafından
orj=biçiminde	orj=kısaltma	orj=türlü
orj=bulunan	orj=kısaltması	orj=veren
orj=denir	orj=kişi	orj=verilen
orj=dir%.	orj=konulan	orj=verilen%addır
orj=duruma	orj=kullanılan	orj=yani
orj=eden	orj=kullanılır	orj=yapan
orj=edilen	orj=madde	orj=yapıl
orj=etmek	orj=olan	orj=yarayan
orj=genellikle	orj=yanlış	orj=yasayan



# Şablonlar Veritabanınının Oluşturulması-4

- Nedir haricindeki soruların şablonlarının oluşturulması
  - Her soru türü için soru cevap çiftleri belirlenmiş
  - İnternet'te soru ve cevabın birlikte geçtiği cümleler bulunmuştur
  - Bu cümleler incelenerek, her soru türü için sıkça kullanılan kelimeler ve yapılar elle belirlenmiştir.



# Şablonlar Veritabanının Oluşturulması-5

Zaman şablonlarının elde edilmesinde kullanılan cümlelerden örnekler

X (Y -	ERZURUM KONGRESİ (23 Temmuz -
Y X	23 TEMMUZ ERZURUM KONGRESİ
X Y de	Erzurum Kongresi 23 Temmuz 1919'da
Y günü ... X	4 Eylül 1919 günü toplanan Sivas Kongresi
Y de ... X	4 Eylül 1919'da toplanan Sivas Kongresi'nin
X, Y günü	Sivas Kongresi, 4 Eylül 1919 günü
Y tarihinde ... X	17 Ağustos 1999 tarihinde saat 03.02 te meydana gelen
Y X (Y)	13 Kasım. Napolyon Viyana'ya girdi (1805).
Y tarihli X	3 Kasım 1839 tarihli Gülhane Hattı Hümayunu'nun
Y: X	19 Temmuz: Paris'te metro hizmete girdi.
Y yılında ... X	25 Haziran 1950 yılında çıkan Kore Savaşı

**X' ler soruyu, Y' ler ise cevabı temsil etmektedir.**



# Cevapların kullanıcıya verilmesi

- 3 farklı şekilde puanlandırılan cümlelerin aldıkları puanlar toplanarak o cümlenin puanı elde edilmiştir.
- Puanlanmış cümlelerden, birbirinden farklı olan ve en yüksek puanı almış ilk 5 cümle kullanıcıya cevap olarak iletilmektedir.



# Uygulama Örnekleri-1

soru		Cevaplar
<b>TBMM ne zaman açıldı?</b>	1	İşte hemen sonra da Ankara'da 23 Nisan 1920 tarihinde TBMM açıldı.
	2	Sonuçta 23 Nisan 1920 Tarihinde Ankara'da TBMM açıldı ve Türk .
	3	23 Nisan 1920'de geniş bir katılımıla TBMM açıldı.
	4	Teklif, bugün (24 Şubat 2000) koalisyon partileri genel başkanlarının düzenledikleri zirveden hemen sonra TBMM'de imzaya açıldı.
	5	1 Kasım 1927, TBMM'nin üçüncü dönemi açıldı.
<b>Gaziantep'in kurtuluş günü ne zamandır?</b>	1	Şöyle ki, bugün, 25 Aralık 1999, Gaziantep İlimizin kurtuluş günüdür.
	2	25 Aralık'ta Gaziantep'in kurtuluş gününde açacağımız sergiye gelecek insanlardan böyle bir grup çıkarabileceğimizi ve gelecek sene beş .
	3	-
	4	-
	5	-
<b>Tarkan ne zaman doğdu?</b>	1	Tarkan Süalp 1968 yılında İstanbul'da doğdu.
	2	TARKAN GÖZÜBÜYÜK Tarkan Gözübüyük Bass Gitar & Vokal 1970 yılında Erzurum'da doğdu.
	3	Tarkan Gözübüyük: Bass, Vocal 1970 yılında Erzurum'da doğdu.
	4	-
	5	-



# Uygulama Örnekleri-2

<b>Türkiye İnternet ile ne zaman tanıştı?</b>	1	12 Nisan 1993'de Ankara-Washington 64K'lık bağlantısı ile Türkiye İnternet'le tanıştı.
	2	Türkiye İnternet ile 12 nisan 1993'de tanıştı ve İnternetimiz 6.
	3	12 Nisan 1993'te Ankara-Washington arasında kiralık hatla kurulan bağlantı ile Türkiye internetle tanıştı.
	4	Türkiye İnternet'le 12 nisan 1993'de ODTÜ'den gerçekleştirilen Ankara-Washington bağlantısı ile tanıştı.
	5	-
<b>Boğaz Köprüsü'nün yüksekliği ne kadardır?</b>	1	Boğaz Köprüsü'nün yüksekliğinin yaklaşık 3 katına eşdeğerdir .
	2	64 metre yüksekliğindeki Boğaz Köprüsü'yle, Saipem 7000 .
	3	Boğaz Köprüsü'nün denizden yüksekliği ise 60 metre.
	4	Mesela Boğaziçi Köprüsü'nün deniz seviyesinden yüksekliği sadece 64 .
	5	boyu 165 metre olan Birinci Boğaz Köprüsü'nün .
<b>Ay'ın Dünya'dan uzaklığı ne kadardır?</b>	1	Pembe-mavi benekli bu uydunun Neptün'e uzaklığı, yaklaşık olarak Dünya-Ay arası uzaklığı kadardır (354.200 Km).
	2	Ay'ın Yer yüzeyine uzaklığı 350000 km ile 400000 km .
	3	Ay'ın Dünya'ya olan ortalama uzaklığı 384000 km dir.
	4	Yer çevresinde yaklaşık olarak her 48 saatte bir dönmekte ve bu hareketi sırasında kimi zaman Yer'e Ay uzaklığı kadar yaklaşmakta, kimi zaman da Ay .
	5	Bu uzaklık, Ay-Dünya arasındaki uzaklığın (6) katıdır.





Internet'teki bilginin güvenilirliđi ???

'Atatürk öldüğünde kaç yaşındaydı?' sorusu için  
**cevap:**

15 yıllık Cumhurbaşkanlığı  
olan Atatürk, öldüğünde 54 yaşındaydı.



# Sistemin Deęerlendirilmesi-1

- Soru cevaplama sistemlerinin deęerlendirilmesi olduka g bir iřlemdir. TREC yarışmalarında bile sistemlerin deęerlendirilmesi elle yapılmaktadır.
- Birbirinden farklı sorularla test edilen iki sistemin birbirleriyle karşılaştırılması pek bilimsel bir yöntem değildir.
- Bu yüzden geliştirilen sistemin dięer sistemlerle karşılaştırılabilmesi için aynı sorularla olmasa bile aynı tür sorularla test edilebilmesi için öncelikle TREC-9 ve TREC-10 soruları yarışmanın sitesinden toplanmış ve daha sonra Türke'ye evrilmiştir.



## Sistemin Deęerlendirilmesi-2

- Türke web sitelerinde cevabı olması mümkün gzkmeyen bazı zel sorular elendikten sonra 4 kiřiye bu soruları baři trl nasıl sorabilecekleri sorulmuřtur.
- Kiřilerden gelen sorular birleřtirilerek 524 adet test sorusu elde edilmiřtir.
- Cevapları sadece Google'ın sonu sayfasında arayan sisteme, 524 soru verilmiř ve cevapları aldıkları puanlarla birlikte bir veritabanına kaydedilmiřtir.



# Sistemin Değerlendirilmesi-3

**Cevapların Değerlendirilmesi**

Kriter =  Soru Göster

Soru No =  Soru Türü =

Soru =

**CEVAPLAR :**

Eyleme	<input type="text" value="Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde *,&lt;br/&gt;4'ün Today (Bugün) adlı programında Amerika'nın bağımsızlık .&lt;br/&gt;Eylül 2001 Çarşamba Bilindiği üzere 11 Eylül 2001 Salı günü Amerika'da çok .&lt;br/&gt;Amerika'nın Bağımsızlık Günü'nden banane; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günler"/>	Cevap - 1 <input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5	Sonraki Soru
Eyleme - Sıralama	<input type="text" value="4'ün Today (Bugün) adlı programında Amerika'nın bağımsızlık .&lt;br/&gt;Eylül 2001 Çarşamba Bilindiği üzere 11 Eylül 2001 Salı günü Amerika'da çok .&lt;br/&gt;Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde *,&lt;br/&gt;Amerika'nın Bağımsızlık Günü'nden banane; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günler"/>	Cevap - 2 <input type="radio"/> 0 <input checked="" type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5	
Eyleme - Şablon	<input type="text" value="Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde *,&lt;br/&gt;Eylül 2001 Çarşamba Bilindiği üzere 11 Eylül 2001 Salı günü Amerika'da çok .&lt;br/&gt;Amerika'nın Bağımsızlık Günü'nden banane; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günler&lt;br/&gt;4'ün Today (Bugün) adlı programında Amerika'nın bağımsızlık ."/>	Cevap - 3 <input type="radio"/> 0 <input type="radio"/> 3 <input checked="" type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5	
Sıralama - Şablon	<input type="text" value="Eylül 2001 Çarşamba Bilindiği üzere 11 Eylül 2001 Salı günü Amerika'da çok .&lt;br/&gt;Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde *,&lt;br/&gt;4'ün Today (Bugün) adlı programında Amerika'nın bağımsızlık .&lt;br/&gt;Amerika'nın Bağımsızlık Günü'nden banane; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günler"/>	Cevap - 4 <input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5	
Eyleme - Sıralama - Şablon	<input type="text" value="Eylül 2001 Çarşamba Bilindiği üzere 11 Eylül 2001 Salı günü Amerika'da çok .&lt;br/&gt;Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde *,&lt;br/&gt;4'ün Today (Bugün) adlı programında Amerika'nın bağımsızlık .&lt;br/&gt;Amerika'nın Bağımsızlık Günü'nden banane; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günler"/>	Cevap - 5 <input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input checked="" type="radio"/> 5	Soru Puan Kaydet

Start | Cev... | tezi... | İara... | Fw... | MSN... | den... | Fw... | 11... | Micr... | Delp... | Prof... | NTR | 9:39 PM

# Sistemin Deęerlendirilmesi-4

- Sistemin verdięi cevaplar 5 farklı puanlama teknięine gore sıralanmıřtır
  - 1. eřleme
  - 2. eřleme+sıralama
  - 3. eřleme+řablon
  - 4. sıralama+řablon
  - 5. eřleme+sıralama+řablon
- Doęru cevabın kaıncı sırada olduęu bir tabloya kaydedilmiřtir.



# Sistemin Deęerlendirilmesi-5

Soru ve doęru cevap sayılarının soru türlerine göre dağılımını gösteren tablodan bir kesit

Soru türü	Soru Sayisi	Eşleme	Eşleme+Sıralama	Eşleme+Şablon	Sıralama + Şablon	Eşleme+Sıralama+Şablon
neden	6	2	1	3	1	1
nedir	117	33	33	40	39	40
nelerdir	13	2	2	2	2	2
nerededir	20	12	12	13	13	13
nesne	11	2	2	2	2	2
nüfus	10	7	6	6	7	5
olumzaman	3	3	3	3	3	3
oran	11	8	7	8	7	7
para	4	2	1	2	2	2
renk	7	1	1	1	1	1
sayı	26	15	15	14	14	14
şehir	3	1	1	1	1	1
sembol	1	0	1	1	1	1
süre	7	1	1	1	1	1
süreyıl	1	1	1	1	1	1
yl	11	3	3	3	3	3
zaman	50	22	21	21	22	22
<b>toplam</b>	<b>524</b>	<b>224</b>	<b>210</b>	<b>230</b>	<b>218</b>	<b>220</b>
		42.70%	40.00%	43.90%	41.60%	41.98%



# Sistemin Deęerlendirilmesi-6

Soru turlerine gre MRR puanlarının daęılımını gsteren tablodan bir kesit

Soru tr	Soru Sayısı	Eşleme	Eşleme+ Sıralama	Eşleme+ Şablon	Sıralama+ Şablon	Eşleme+ Sıralama+ Şablon
neden	6	0.25	0.083333333	0.305555556	0.083333333	0.083333333
nedir	117	0.178917379	0.197578348	0.212393162	0.235327635	0.242735043
nelerdir	13	0.115384615	0.115384615	0.115384615	0.115384615	0.115384615
nerededir	20	0.268333333	0.276666667	0.435	0.405833333	0.433333333
nesne	11	0.181818182	0.090909091	0.181818182	0.075757576	0.090909091
nfus	10	0.483333333	0.375	0.458333333	0.475	0.35
olumzaman	3	0.777777778	0.777777778	0.777777778	0.777777778	0.777777778
oran	11	0.545454545	0.590909091	0.545454545	0.454545455	0.5
para	4	0.375	0.25	0.375	0.333333333	0.333333333
renk	7	0.028571429	0.028571429	0.028571429	0.028571429	0.028571429
Sayı	26	0.509615385	0.477564103	0.490384615	0.448717949	0.467948718
şehir	3	0.333333333	0.333333333	0.333333333	0.333333333	0.333333333
sembol	1	0	1	0.333333333	1	1
Sure	7	0.142857143	0.071428571	0.142857143	0.071428571	0.071428571
Sreyil	1	1	1	0.5	0.5	0.5
Yıl	11	0.151515152	0.2	0.227272727	0.212121212	0.212121212
Zaman	50	0.362333333	0.361666667	0.373333333	0.385	0.395



# Sonuç-1

- En fazla soruya doğru cevap veren puanlama türü  
– eşleme+şablon (%43.9)
- En yüksek MRR puanına sahip puanlama türü  
– eşleme+sıralama+şablon (0.313)





# Sonuç-2

TREC yarışmasının en iyileri (MRR puanları)

	TREC-8	TREC-9	TREC-10
<b>En iyi</b>	0.66	0.58	0.68
<b>En iyi ikinci</b>	0.56	0.32	0.59

Sistem örnekleri (MRR puanları)

Sistem	MRR
Quantum	0.223
Shapaqa	0.545
Webslopedia	0.435
QuASM	0.253
AskMSR	0.507

Geliştirilen Sistemin Puanlama türlerine göre MRR puanları

	Soru Sayısı	Eşleme	Eşleme+ Sıralama	Eşleme+ Şablon	Sıralama+ Şablon	Eşleme+ Sıralama+ Şablon
<b>MRR</b>	524	0,298	0,289	0,308	0,301	<b>0,313</b>



# Gelecek Çalışmalar İçin Öneriler

- Sistemde kullanılan **şablonlar veritabanı** daha da geliştirildiğinde sistemin performansı artacaktır.
- Sisteme **yeni kelime etiketlerinin (Name Entity)** eklenmesi, şablonların ve filtrelerin dolayısıyla tüm sistemin performansını arttıracaktır.
- Aday cevapların puanlanmasında ve kullanıcı sorusunun arama motoru sorgularına dönüştürülmesinde, **eşanlımlı kelimelerin ya da yapıların** kullanılması sistemin başarısına en büyük katkıyı sağlayacaktır.



# Aşamalı Sorgulama

- Soru: *What is the occupation of Bill Clinton's wife?*
  - Cevabı içeren doküman yoksa
- Aşamalı sorgu: soruyu ikiye böl
  - Önce *Who is Bill Clinton's wife?* = X'i bul.
  - Sonra *What is the occupation of X?* 'i ara.



- Kabul edilebilirlik
  - *İstanbul Ankara arası 12 metredir.*
  - *Haliç Köprüsünün yüksekliği 23 cm'dir.*
- Wordnet:
  - Cevap türü şehir olan bir sorunun olası cevaplarını Wordnet'ten kontrol edebiliriz.

WordNet: Helsinki, Tampere, Luanda... şehir isimleri olarak yer almaktalar.



# TURQUAS: TURkish QUestion Answering System

Bu çalışmada klasik soru-cevaplama sistemlerinin dışına çıkılarak farklı bir model geliştirilmiştir. Alınan sonuçlar çok başarılı değildir ancak model üzerindeki çalışmalar devam etmektedir.

- Önerilen model içerisinde kelimelerin Word2Vector değerlerine ihtiyaç vardır.
- Verilen bir corpus içerisine yer alan her farklı kelime w2v değerini hesaplayan bir modül yazılmıştır. Ancak, süreç çok uzun sürdüğünden dolayı önceden çıkarılmış w2v değerleri kullanılmıştır.
- Sistemi eğitirken Keras kütüphanesindeki 3 çeşit RNN katmanından biri olan LSTM (Long Short-Term Memory) kullanılmıştır. (Zaman sıralı gelen verilerde kullanılır.)



- Kullanıcıdan gelen soru cümlesi LSTM verilerek soru cümlesine karşılık sayısal bir değer elde edilir.
- LSTM'e cümle içerisindeki kelimelerin w2v değerlerini nereden alacağı, cümle içerisindeki stop-word lerin kullanılıp kullanılmayacağı bilgisi, çıkışta oluşacak olan giriş soru cümlesinin vektör değerinin boyut bilgisi parametre olarak verilmektedir. Sonucunda soru cümlesi için sayısal bir değer elde edilir.
- Kullanıcıdan alınan soru olduğu gibi arama motoruna verilerek dönen sayfalardan içerisinde min 5 max 20 kelime olan ilk 5 sayfa çekilir.



- Sayfa içerisindeki cümleler aday cevap cümlesi olarak seçilir. Bu seçim sırasında soru cümleleri ve Türkçe kelimelerden oluşmayan cümleler elenir.
- Elde edilen bütün aday cevap cümleleri sırası ile LSTM verilir. Her biri için aynı boyutta sayısal değerler elde edilir.
- Kullanıcı sorusunun vektör değeri ile aday cevap cümlelerinin vektör değerleri arasındaki benzerlik ölçülür.
- Benzerlik değerleri en yakından en uzağa doğru sıralanarak kullanıcıya sunulur.



- Tabloda sistemin soru türlerine göre MRR ve İlk 5 te cevabın bulunup bulunmadığına göre başarısı gösterilmektedir.

<u>Soru Kalıbı</u>	<u>MRR (%)</u>		<u>İlk 5 Cevapta Var Mı (%)</u>	
	<u>LSTM</u>	<u>Average</u>	<u>LSTM</u>	<u>Average</u>
<i>5'er Soru İle Test Edildi</i>				
<u>yüzde kaç</u>	80	60	100	100
<u>yıl nedir</u>	51.6	41.6	80	80
<u>yıl kaçtır</u>	55	25.6	80	80
<u>tarih nedir</u>	49	30	80	40
<u>başkenti neresidir</u>	90	84	100	100
<u>kaç tane</u>	60	35	100	60
<u>uzunluğu nedir</u>	70.6	70	100	80
<u>hangisidir</u>	45	66.6	80	100
<i>10'ar Soru İle Test Edildi</i>				
<u>ne</u>	55.3	41.6	90	60
<u>ne zaman</u>	62.5	50.8	80	80
<u>nerede</u>	67	71.6	100	100
<u>kim</u>	73.3	67.5	90	100
<u>neden</u>	35.8	29.9	60	70
<u>nasıl</u>	44.1	41.1	90	80





- Sistemin performansını artırmak için neler yapılabilir?
  - Soru cümlesinde geçen kelimelerin yarı sayısındaki kelimelerden oluşan cümleler seçilebilir
  - Soru cümlesi içinde yer alan head-word cevabı olabilecek türün aday cevap cümle içerisinde olmasına dikkat edilebilir
  - ???



# Kaynaklar

1. Bir Soru Cevaplama Sistemi: Baybilmiş, M.Fatih Amasyalı, Banu Diri, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2005
2. Open-Domain Question Answering, Eric Nyberg
3. Questioning and Answering, Alison Huettner
4. Processing of large document collections, Helena Ahonen-Myka
5. An Introduction to Information Retrieval and Question Answering , Jimmy Lin



# Teşekkürler

