BYOYO 2018

Boğaz'da Yapay Öğrenme
İsmail Arı Yaz Okulu 2018

ÖZGÜN PARAGRAF TABANLI ÇIKARIM TEKNİĞİ KULLANARAK OTOMATİK ÇOKLU DOKÜMAN ÖZETLEME

DR. METİN TURAN/İSTANBUL TİCARET ÜNİVERSİTESİ

RAHMETLİ SEVGİLİ HOCAM

PROF. DR. AHMET ÇOŞKUN SÖNMEZ ANISINA

1-Metin Turan, Coşkun Sönmez, IPCT, "Outlier Document Filtering Applied to the Extractive Summarization", 2014, Roma.
2-Metin Turan, Coşkun Sönmez, Murat Can Ganiz (2014), "The Benchmark of Paragraph and Sentence Extraction Summaries using Outlier Document Filtering based Multi-Document Summarizer", Information Technology and Control(SCI Indexed), Vol:43, Issue:4, pp. 433-439

1-Metin Özetleme Tanımı

Mani[1] özetlemeyi aşağıdaki gibi tanımlar:

"Bir bilgi kaynağından içerik çıkartma ve en önemli kısımlarını kullanıcı veya uygulama ihtiyaçlarına uygun bir biçimde sıkıştırılmış olarak sunmaktır. "

2-Metin Özetleme Amacı

Günümüzde yayınlanmış belge (doküman) sayısındaki korkunç artış gerçek bilgiye ulaşımı zorlaştırmıştır.

Amaç, uzman özetlerine yakın özetlerin otomatik olarak elde edilebilmesidir.

3-Çıkarıma Dayalı Özetlemede Temel Sorunlar

- Çıkarılan birimler genelde ortalamadan daha uzundurlar.
- Önemli bilgi birimler arasında yayılmıştır, özet boyu yeterince uzun değilse bunları içermez.
- Tekrarlı bilgi önlenemeyebilir.
- Adıllar (dangling anaphora) özetin tutarlılığını etkileyebilir.
- Okunurluluk oldukça düşüktür.

4-Çalışmanın Kısıtları

- İngilizce dokümanları kullanacaktır.
- Paragraf Tabanlı Çıkarım Özetleme(PTÇÖ) yapacaktır
- Bilgi getirme (Information Retrieval) uygulaması değildir. Referans kabul edilebilecek bir sorgu anahtarları içermeyecek ve otomatik özet oluşturacaktır.

Hipotez - I

- Doküman genel yapısı:
  - ✓Bir konu ve bunun alt-konularından oluşur. Doküman içindeki birimler birbiriyle ilişkilidir.
  - ✓İçeriğini oluşturan birimler kelime (kök hali terim), cümle, paragraf veya metin bloklarıdır.
  - ✓Cümle ufak bir birimdir, metin bloğu ise sınırlarının belirlenmesi açısından zor bir birimdir. Özetleme amacıyla en uygun birim paragraf gözükmektedir.

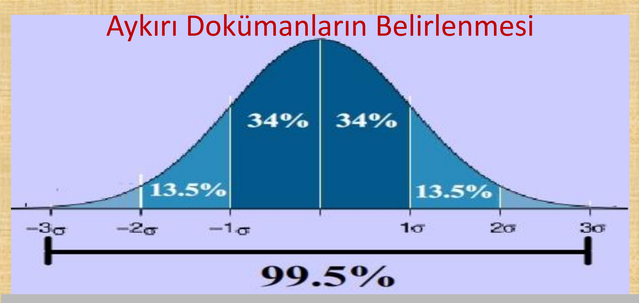
Bulgu

- Mitra ortalamada %50 civarında başarıya ulaşmış, önemli bir tespiti ise iki insan özetleyicinin tercihlerinin ancak ortalamada %46 eşleşme gösterdiğini gözlemlemiş olmasıdır.

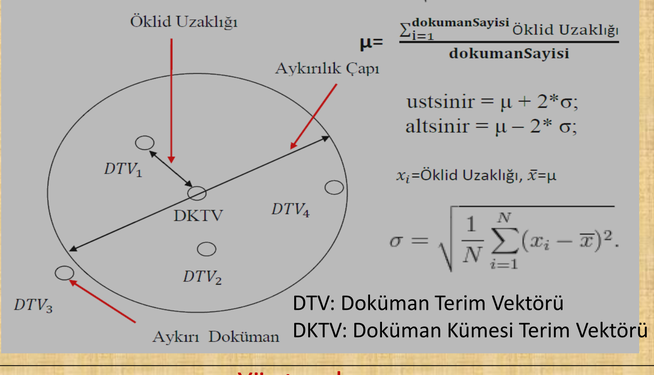
Hipotez - II

- Çoklu Doküman Özetleme tanımına göre tüm dokümanlar aynı konuya hitap etmelidir.
- Doküman kümesindeki dokümanların içeriklerinin (alt-konular) sapma göstermesi mümkündür (aykırı olanların belirlenmesi).
- Dokümanlardan elde edilen terimlerden sadece fazla bilgi taşıyanlar (belirleyici terimler) sayesinde sapmalar doğru tespit edilebilir.

Aykırı Dokümanların Belirlenmesi



$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$



$$\mu=\frac{\sum_{i=1}^{\text{dokumanSayisi}}\text{Öklid Uzaklığı}}{\text{dokumanSayisi}}$$

ustsinir = μ + 2*σ;
altsinir = μ − 2* σ;

x_i=Öklid Uzaklığı, $\bar{x}$=μ

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}.$$

DTV: Doküman Terim Vektörü
DKTV: Doküman Kümesi Terim Vektörü

Yöntemler

- Fizikteki Gestalt teoreminin Helmholtz ilkesine göre istatiksel bir değer elde edilmekte ve belli bir eşik değeri üzerinde kalan kelimeler anlamlı kelimeler olarak belirlenmektedir.
- Kelimelerin TF-IDF değerlerini esas alınmakta ve birimler arası görülme sıklığına göre önemli kelimeler belirlenmeye çalışılmaktadır.
- Çizge yaklaşımı kullanarak kelimeler arası birliktelik ilişkilerden faydalınmaktadır.

- D, tüm dokümanları içeren doküman kümesi,
- T, D içinde görülen tüm farklı terimler kümesi
- t_b, belirleyici terimler kümesi olmak üzere
- d_{nt}, n. dokümanın terimler kümesi (DTV) olmak üzere;
  - ✓ D= { $d_1, d_2, d_3, \dots, d_n$ },
  - ✓ T= { $t_1, t_2, t_3, \dots, t_z$ },
  - ✓ $d_{nt} \subseteq$ T ve $t_b \subseteq$ T 'dir.

Çalışmada belirleyici terimleri tespit etmek amacı ile, terimin kaç dokümanda görüldüğü sayısının toplam doküman sayısına oranı olarak ifade edilen Terim Yayılma Oranı (TYO) kullanılmıştır.

Hipotez - III

- Özet alt konulardan en önemli olanların belirlenmesi ve bu alt konuyu en iyi belirleyen birimlerden oluşmalıdır.
- Özet dökümanlar arası dağılımları gözönüne almalıdır.
- Uzun birimlerin öne çıkmasını önleyecek yaklaşım gereklidir.

Sıralama Kriteri (Çözüm)

- Bir PTV'nin Eşleşme Yüzdesi (EY) değeri aşağıdaki gibi bulunur.
- Birimin uzunluğunun önemi paydada uzunluk değerinin kullanılmasından dolayı önlenmiş olur.

$$EY=\frac{\{\text{eğer terim } t_i \in \text{PTV ve } t_i \text{ aynı zamanda } t_b \;\; 1 \;\; \text{değilse} \;\; 0\}}{\text{PTV 'deki toplam terim sayısı}}$$

Uygulanan tekniklerin İngilizce isimlerinin ilk harfi göz önüne alındığında geliştirilen modele
MOD (M –Matching Percent, O –Outlier Detection, D –Descriptive Terms) adı verilmiştir.

VERİ SETİ VE DENEYLER

DUC 2006 Doküman Kümesi

- Bu doküman kümesi toplam 50 adet farklı konu kümesi içermektedir.
- Bu konular Financial Times of London ve Los Angales Times gazetelerinin haberlerinden seçilmiştir.
- Her bir konu 25 adet haberden oluşmaktadır.
- Uzman ve sistem özetleri 250 kelime ile sınırlandırılmıştır.
- Her bir konu için toplam 4 uzman referans özeti çıkarılmıştır.

ROUGE Nedir?

- ROUGE iki özet arasındaki benzerlikleri bulmak üzere geliştirilmiş bir yazılımdır.

Metrik	Açıklama
ROUGE-1	1-gram tabanlı birliktelik istatistiğini verir
ROUGE-2	2-gram tabanlı birliktelik istatistiğini verir
ROUGE-3	3-gram tabanlı birliktelik istatistiğini verir
ROUGE-4	4-gram tabanlı birliktelik istatistiğini verir
ROUGE-L	En uzun ortak altdizi (EOA) tabanlı istatistiği verir.
ROUGE-W	Sıralı EOA tabanlı istatistiği verir.
ROUGE-S	2-gram hariç birliktelik istatistiğini verir.
ROUGE-SU	1-gram ve 2-gram hariç birliktelik istatistiğini verir.

PTÇÖ: Paragraf Tabanlı Çoklu Özetleme **CTÇÖ: Cümle Tabanlı Çoklu Özetleme**

MOD'un CTÇÖ ile DUC 2006 katılımcı ortalamasının metriklerinin kıyaslanması

	DUC-Katılımcıları	TYO %25	TYO %50	TYO %75
ROUGE-1				
Ortalama-H	0,371	0,641	0,635	0,637
Ortalama-T	0,386	0,540	0,540	0,535
Ortalama-F	0,377	0,583	0,581	0,579
ROUGE-2				
Ortalama-H	0,073	0,413	0,413	0,417
Ortalama-T	0,076	0,347	0,351	0,349
Ortalama-F	0,074	0,375	0,377	0,379
ROUGE-3				
Ortalama-H	0,020	0,344	0,344	0,348
Ortalama-T	0,021	0,289	0,292	0,291
Ortalama-F	0,021	0,312	0,315	0,316
ROUGE-4				
Ortalama-H	0,008	0,301	0,302	0,305
Ortalama-T	0,009	0,253	0,256	0,255
Ortalama-F	0,008	0,273	0,276	0,277
ROUGE-SU4				
Ortalama-H	0,128	0,408	0,399	0,401
Ortalama-T	0,133	0,293	0,293	0,285
Ortalama-F	0,130	0,334	0,332	0,330

- Ortalama-F değeri baz alınırsa TYO'nın %75 olduğu sınama sonuçları en iyi değerlere ulaşmaktadır.
- Bu sonuçlara göre MOD'un CTÇÖ için anlamlı terim seçiminin %50 - %75 aralığında tutulmasının başarıyı arttırdığı sonucuna varılabilir.
- TYO %50 değeri için tutarlılık özelliği daha iyiyken, TYO %75 olduğu durumda hatırlama özelliği daha iyi sonuç vermektedir.

MOD'un DUC katılımcı ortalamasına göre performansı

	TYO %25	TYO %50	TYO %75
ROUGE-1			
Ortalama-F	+%54,6	+%54,1	+%53,5
ROUGE-2			
Ortalama-F	+%506	+%509	+%512
ROUGE-3			
Ortalama-F	+%1485	+%1500	+%1504
ROUGE-4			
Ortalama-F	+%3412	+%3450	+%3462
ROUGE-L			
Ortalama-F	+%49,1	+%50,8	+%50,8
ROUGE-W			
Ortalama-F	+%49,6	+%51,1	+%51,1
ROUGE-SU4			
Ortalama-F	+%256	+%255	+%253

- En önemli sonuç, gram sayısı (n, 1 – 4 aralığında) artmasına rağmen MOD'un başarısının diğer katılımcıların elde ettiği sonuçlar gibi keskin bir biçimde düşmemesidir.
- ROUGE-L ve ROUGE-W metrikleri (en uzun benzerlik zincirleri istatistikleri) incelendiğinde özet için seçilen cümlelerin uzmanlarınkine benzerliğinin yaklaşık %60 daha iyi olduğu aşikardır.
- Bu durumu, SU4 (1-gram ve 2-gram hariç) metriği %265 daha iyi sonuç vererek desteklemektedir.

MOD'un CTÇÖ ve PTÇÖ'lerinin metriklerinin kıyaslanması

	TYO %25		TYO %50		TYO %75	
	Cümle	Paragraf	Cümle	Paragraf	Cümle	Paragraf
ROUGE-1						
Ortalama-H	0,641	0,608	0,635	0,604	0,637	0,606
Ortalama-T	0,540	0,575	0,540	0,574	0,535	0,566
Ortalama-F	0,583	0,589	0,581	0,587	0,579	0,584
ROUGE-2						
Ortalama-H	0,413	0,383	0,413	0,382	0,417	0,386
Ortalama-T	0,347	0,362	0,351	0,363	0,349	0,359
Ortalama-F	0,375	0,371	0,377	0,371	0,379	0,371
ROUGE-3						
Ortalama-H	0,344	0,316	0,344	0,300	0,348	0,320
Ortalama-T	0,289	0,298	0,292	0,307	0,291	0,298
Ortalama-F	0,312	0,306	0,315	0,307	0,316	0,308
ROUGE-4						
Ortalama-H	0,301	0,276	0,302	0,277	0,305	0,280
Ortalama-T	0,253	0,260	0,256	0,262	0,255	0,260
Ortalama-F	0,273	0,267	0,276	0,268	0,277	0,269
ROUGE-SU4						
Ortalama-H	0,408	0,362	0,399	0,359	0,401	0,362
Ortalama-T	0,293	0,324	0,293	0,325	0,285	0,316
Ortalama-F	0,334	0,339	0,332	0,337	0,330	0,334

- Genel anlamda ROUGE metriklerine göre PTÇÖ tutarlılığı arttırırken hatırlama değerini düşürmektedir. CTÇÖ ise hatırlamayı arttırmaktadır.
- Tutarlılık formülü düşünüldüğünde, özetteki doğru cümle oranı artmıştır. Bu da dağınık birimler seçmek yerine daha büyük önemli birimleri özete koymanın bir sonucudur.
- Hatırlama formülü düşünüldüğünde, özetteki doğru cümle sayısının toplam doğru cümle sayısına oranı yükselmiştir. Bu da daha küçük birimler seçerek daha fazla doğru cümle bulma şansını arttırmasındadır.

ÖRNEK UYGULAMA

Örnek bir doküman kümesinin PTV'leri

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
PTV_{11}	1	0	1	1	0	2	0	1	1	0	0	0	0	0
PTV_{12}	0	1	0	1	1	0	0	0	2	1	0	1	0	1
PTV_{13}	1	1	0	0	0	1	1	0	0	0	2	0	1	0
PTV_{21}	2	0	1	1	0	0	0	1	0	0	0	1	1	0
PTV_{22}	0	2	0	2	1	0	0	0	0	1	0	1	0	0
PTV_{31}	1	1	0	0	2	1	0	0	1	0	1	0	0	3
PTV_{32}	2	1	1	1	0	0	0	1	0	2	0	0	0	0
PTV_{33}	0	0	0	2	1	0	2	0	1	1	0	0	1	0
PTV_{34}	1	0	0	0	0	0	1	1	0	1	1	0	0	1

TYO → %75 =>MGS (Minimum Görülme Sayısı)= [TYO * dokumanSayisi] =3

Örnek PTV'ler için DTV'ler

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
DTV_1	2	2	1	2	1	3	1	1	3	1	2	1	1	1
DTV_2	2	2	1	3	1	0	0	1	0	1	0	2	1	0
DTV_3	4	2	1	3	3	1	3	2	2	4	2	0	1	4

Örnek PTV'ler için DKTV

	t_1	t_2	t_3	t_4	t_5	t_8	t_{10}	t_{13}
DKTV	4	2	1	3	3	2	4	1

- Aykırı doküman tespiti için Öklid Uzaklığı (OU) hesaplandığında aşağıdaki değerler elde edilir.
- OU_1 = 6,63 , OU_2 = 4,69 , OU_3 = 5,83, ortalama μ= 5,71 ve standart sapma σ= 1,37 bulunur
- Bu koşullar altında σ bir olarak seçilse bile hiçbir doküman aykırı olarak tespit edilemeyecektir, doküman içerikleri birbirleriyle örtüşmektedir.

- Paragrafların, PKTV ile benzerlikleri hesaplanarak sıralama yapılmalıdır.

Örnek PTV'ler için EY değerleri

	t_1	t_2	t_3	t_4	t_5	t_8	t_{10}	t_{13}	EY
PTV_{11}	1	0	1	1	0	1	0	0	4/8= 0,50
PTV_{12}	0	1	0	1	1	0	1	0	4/8= 0,50
PTV_{13}	1	1	0	0	0	0	0	1	3/8= 0,375
PTV_{21}	2	0	1	1	0	1	0	1	5/8= 0,675
PTV_{22}	0	2	0	2	1	0	1	0	4/8= 0,50
PTV_{31}	1	1	0	0	2	0	0	0	3/8= 0,375
PTV_{32}	2	1	1	1	0	1	2	0	6/8= 0,75
PTV_{33}	0	0	0	2	1	0	1	1	4/8= 0,50
PTV_{34}	1	0	0	0	0	1	1	0	3/8= 0,375
DKTV	4	2	1	3	3	2	4	1	

- Bu çizelgeden görüleceği üzere EY değerleri tahminen daha uzun paragraflara (terim sayısı fazla) üstünlük tanımamıştır.