Gəlin bu bölməni daha da kompleks suallara cavab verən daha qabaqcıl araşdırma sistemlərinə

baxaraq sonlandıraq. "What is water spinach?" adlı tərif sualını nəzərdən keçirin və

gördüyünüz kimi mən suala potensial en uzun cavabı vermişəm. Bunun kimi tərif sualları üçün

cavab yarada bilərikmi? Deyək ki, onlar vebdə yaxşı bir yerdə deyillər və özümüz cavabı

yaratmaq istəyirik. Və ya tibbi bir sualı götürün. Sual belədir: " What is the efficacy of a

particular therapy for a particular disease?" Siz cavab vermək istəyirsiniz. İndi baxdığımız

suala uyğun cavab PubMed verilənlər bazasında olan xüsusi məqalədən götürülüb. Bu çox

çətindir, xüsusilə də bir və ya bir neçə sənəddən suala cavab vermək üçün qısa məzmunun

hazırlanması. Faktiki olaraq belə kompleks suallara cavab vermək üçün yarış gedir. Və mən

sizə bu tip sualların sadələşdirilmiş variantını vermişəm. Məsələn " How is compost made

and used for gardening?", " What causes train wrecks? What can be done to prevent them?",

" What's the human toll in death or injury of tropical storms in recent years?" və s. Bu tipli

suallar çoxdur və onlara cavab vermək üçün avtomatik şəkildə bir neçə sənəd oxunmalı,

onlardan informasiya götürülməli və bu informasiyanın qısa məzmunu çıxarılmalıdır. Bu çətin

məsələdir və bu tipli çox çətin suallara cavab vermək sorğu-əsaslı müxtəlif sənədlərdən qısa

məzmun çıxartmanın işidir (query-focused multiple document

summarization). Bunun üçün iki standard alqoritm vardır.

Birincisi aşağıdan yuxarıya və ya

mətn parçası stilli metoddur hansı ki, biz tək sənəddən qısa məzmun çıxartma zamanı

görmüşdük. Bu metodu yerinə yetirərkən ilk addım uyğun sənədlər toplusunun tapılmasıdır.

İkinci addım həmin sənədlərdən informativ cümlələrin çıxarılmasıdır. Sonda isə biz

cümlələr üzərində bəzi sıralamalar və dəyişikliklər edə bilərik. Başqa sözlə desək biz cümlələri

sənəddən götürürük və onları birlikdə qarışdırırıq.

Yuxarıdan-aşağıya və ya informasiya çıxardılması metodunda isə müxtəlif sual növləri üçün

xüsusi cavablar hazırlayacağıq. Biz tərif suallarına, bioqrafiya suallarına və bəzi tibbi suallara bu

suallara cavab hazırlama prosesində xüsusi informasiya çıxardaraq cavab yarada bilərik.

Və biz bunların hər ikisini bu balaca mühazirədə görəcəyik. İlk olaraq mətn parçasına əsaslanan

metoda baxaq hansı ki, aşağıdan yuxarıya sorğu-əsaslı müxtəlif sənədlərdən qısa məzmun

çıxartma metodudur. Biz yenə də müxətlif sənədlərdən cümlələrin çıxardılması ilə başlayacağıq.

Təkcə bir sənədə baxmayacağıq. Və bizim cümlələr toplumuz var. Biz onların üzərində

dəyişikliklər aparacağıq. Cümlənin sadələşdirilməsi tez-tez edilən əməliyyatdır. Bir neçə

saniyədən sonra sizə daha dəqiq göstərəcəm: Biz cümlələri götürürük və onları müxtəlif

yollarla sadələşdirəcəyik. Sonda cümlələrimizin bir neçə sadələşdirilmiş versiyası olacaq.

Yəni əlimizdə bir neçə cümlə buludlarının ailələri olacaq. Biz loqarifma ehtimal nisbəti testini

və cümlələr toplusundan yaxşı cümlələri seçə bilən digər metodları tətbiq edəcəyik. İndi bizim

əlimizdə burda da qara nöqtələrlə göstərdiyim kimi çıxardılmış cümlələr var. Biz bu çıxardılmış

cümlələri tək sənəddən qısa məzmun çıxartmada olduğu kimi sıralayacağıq. Və biz onları

müxtəlif yollarla dəyişdirib real cümlələr əldə edəcəyik. Biz tez-tez qısa məzmunun ölçüsünü

artıran vacib olmayan hissələri silməklə cümlələri sadələşdirəcəyik. Bunun üçün ən ümumi

yollardan biri budama və ya bölüşdürmənin (parsing) daxil olduğu metoddur. Biz cümlələri

hissələrə bölürük və sonra hansı modifikatorları (modifiers) budamağın daha uyğun olduğunu

təklif edən bölüşdürmə ağacına (parse tree) əsaslanan əllə yazılmış qaydalardan istifadə edirik.

Məsələn ara cümlələri (appositives) siz budaya bilərsiniz - burda " an artist who was living at

the time in Philadelphia" ara cümləsini aradan qaldıra bilərik. Və ya istinad bəndlərini aradan

qaldırmaq mümkündür, məsələn burda "International Observer said Tuesday" və s.

Bizi əsasan onların nə dedikləri maraqlandırır, ona görə cümlənin bu hissəsi -

" Rebels agreed to talks" önəmlidir. Ancaq istinad bəndini silə bilərik, çünki bizim

üçün o qədər də vacib fakt deyil. Sözönü ifadələrini xüsusilə də adlandırılmış varlıqları olmayan

sözönü ifadələrini silə bilərik və bu bizim qısa məzmunumuz daha yığcam edəcək. Daha yığcam

məzmunda "increased to a sustainable number" yerinə sadəcə "increased" istifadə edə bilər.

Və məsələn "on the other hand", "as a matter of fact" kimi başlanğıcda gələn zərfliklər

(adverbials) də aradan qaldırıla bilər. Bu tip zərfliklər böyük sənəd üçün uyğundur, ancaq

kiçik xülasə və ya qısa məzmun üçün heç də uyğun deyil. Beləliklə ən sadə metod bu tip

qaydaları yazaraq götürməkdir. Burda göstərilən məqalələr də sizə digər maraqlı qaydalar

toplusu verir ki, siz onlardan cümlələri sadələşdirmək üçün istifadə edə bilərsiniz. Biz bu

əməliyyatı bitirdikdən sonra əlimizdə geniş cümlə toplusu olacaq və onların içərisində orijinal

cümlə, ara cümlələr silimiş cümlə, sözönü ifadələri silinmiş cümlə və s. vardır. Beləliklə,

müxtəlif cümlələrin tam dəstinə malik oluruq. Orijinal cümlə və müxtəlif qısaldılmış cümlələr.

Bəs biz növbəti addım olaraq nə edirik? Yaxşı, biz artıq qalan cümlələrdən bəzilərini

qısa məzmunda yerləşdirmək üçün seçmək istəyirik. Biz iki səbəbə görə onları artıq (redundant)

adlandırırıq. Çünki, onlar eyni hadisədən danışan müxətlif sənədlərdən götürülmüşdür. Biz

əlimizdə sadələşdirilmiş cümlələr olsa da, hər cür artıqlıq onlarda müşahidə olunur. Bu

artıqlığı olan cümlələr verilməklə məzmun seçə bilən iterativ metodlardan biri Maksimal

Kənar Münasiblik - MMR (maximal marginal relevance) adlanır. İdeyası belədir, iterasiya və

ya tamahkar yolla qısa məzmunumuza daxil etmək üçün cümlələr toplusunda ən yaxşı

cümləni seçirik və ən yaxşı cümlənin iki əsas xassəni təmin etməsini istəyirik.

Biz həmin cümlənin istifadəçinin sorğusuna maksimum münasib olmasını istəyirik və əmin olmaq üçün

həmin cümlənin istifadəçinin sorğusuna yüksək kosinus oxşarlığının olmasını təyin etməliyik.

Və biz onun yeni və ya öyrənilməmiş olmasını istəyirik, belə ki, biz onun qısa məzmunda

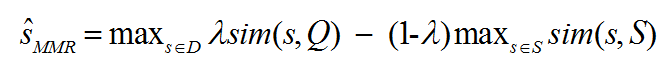
minimum yer tutmasını arzulayırıq. Əgər biz cümlə daxil ediriksə, daxil edilmiş cümlələrlə

demək olar ki, eyni olan cümləni daxil etmək istəmirik. Və biz bunu qiymətləndirmək istəyiriksə

yalnızca cümlənin qısa məzmuna kosinus oxşarlığını ölçürük və seçmə zamanı məzmunla

kosinus oxşarlığı az olanlardan istifadə edirik.

MMR-ın bir versiyası bu cürdür, bu iki faktoru ölçmək üçün bizim lambda kəmiyyətimiz olacaq.



Biz sorğuya yüksək oxşarlığı olan cümlə istəyirik. Oxşarlığı az olanlar üçün biz cümlənin

oxşarlığını qısa məzmunda olan ən oxşar olan cümlənin oxşarlığından çıxacağıq. Beləliklə,

biz qısa məzmunda olan cümlələrə oxşar olmayan cümlələri seçəcəyik, ancaq seçdiyimiz

cümlələr sorğuya çox bənzəyəcək və iterativ şəkildə cümlələri əlavə edəcəyik və hər hansı

bir meyara çatdığımız zaman və ya istədyimiz uzunluğu əldə edən zaman dayancağıq.

Biz informativ cümlələri seçən loqarifma ehtimal nisbətinin və artıqlığı olmayan

cümlələri seçən MMR-ın intuisiyalarını birləşdirəcəyik. Bu intuisiyaları birləşdirə biləcəyimiz

çoxlu yollardan birini də başlatdığımız zaman, ilk olaraq loqarifma ehtimal nisbətini

əsas götürüb gözlədiyimizdən çox baş verən sözlərdən və ya sorğuda olan sözlərdən

istifadə etməklə hər cümlənin qiymətini hesablayacağıq. İndi ən yüksək loqarifma ehtimal

nibəti dəyəri olan cümlələri qısa məzmuna daxil etməklə başlayaq. Daha sonra iterativ

şəkildə digər cümlələri qısa məzmundakı cümlələrlə artıqlıq yaratmayacaq şəkildə əlave edəcəyik.

Əgər biz tək sənədin qısa məzmununu hazırlayırıqsa qısa məzmundakı cümlələrin sırası

sənəddəki sıralamaya əsaslanacaq. Bunu etmək bir neçə sənəddən qısa məzmun çıxartdığımız

zaman çətinləşir. Bu məsələni həll etmək üçün müxtəlif şeylərdən istifadə edə bilərik.

Əgər xəbərlərin qısa məzmunun çıxardırıqsa xronloji ardıcıllıqdan istifadə edə bilərik,

yəni sənədlərin yaranma vaxtına baxıb həmin sıraya görə edə bilərik.

Uzlaşma əsaslı sıralama də etmək mümkündür. Biz cümlələri eyni mənalara malikdirlərsə

yan-yana qoya bilərik. İki cümlə arasındakı kosinusa baxarıq və yüksək kosinuslu cümlələri

yan-yana qoya bilərik. Və ya cümlələrin danışdıqları varlıqlara baxa bilərik və əgər iki cümlə

eyni varlıqdan danışırsa o zaman biz onları yan-yana qoya bilərik.

Biz daha qəribə metoddan istifadə edə bilərik, harda ki, biz mənbə sənədlərə baxar və

sənədlərdə baş verən semantik mövzuları nəzərdən keçirər və sonda sıralamanı təyin edib

çıxış cümlələrinə tətbiq edə bilərik. Gördüyünüz kimi, müxtəlif yollar informasiya sıralamasını

yerinə yetirə bilər. Bu sorğu-əsaslı müxtəlif sənədlərdən qısa məzmun çıxardılması üçün

aşağıdan yuxarıya və ya mətn parçası əsaslı yanaşma idi. Biz tam cümlələr toplusunu

götürür, onları LLR və MMR-la dərəcəsini təyin edir, qısa məzmunu həmin dərəcələndirməyə

uyğun hazırlayır və ən sonda hər hansı metod ilə sıralayırıq.

Alternativ metod isə informasiya çıxardılması metodudur və bu halda bizim spesifik sualımız

var və həmin suala cavabda nələr ola biləcəyini əvvəlcədən bilirik. Məsələn, bilirik ki, yaxşı

bioqrafiyada doğum və ölüm tarixləri qeyd olunur və ya onlar niyə görə məşhurdur, təhsili

və milliyəti və s. vurğulanır.

Ancaq yaxşı tərifdə qenus adlana yüksək ranklı cümlə olur. Məsələn, " The Hajj is, what is it?"

Bu dini ayinin bir növüdür (It's a type of ritual).

Və ya dərmanın istifadəsinə aid tibbi bir cavabda siz problem və ya tibbi vəziyyət haqqında

danışa bilərsiniz. Siz bu dərman və ya prosedur və onun nəticəsi haqqında danışmaq

istəyə bilərsiniz. Bu tip sual tiplərini bilmək bizə sualın cavabında nə olacağını bilməkdə

kömək edir. Bundan sonra biz kiçik detektor düzəldə bilərik hansı ki, qenus cümlələri sənəddən

tapıb çıxarda bilir. Məsələn, mənə qenus cümləni (ən yüksək ranklı cümlə), bu şəxsin doğulduğu

və ya öldüyü tarixi, və ya təhsil tarixini tap. Onları harda məktəbə getdiklərini tap. Əgər

biz üzərində araşdırma aparılmış dərmandan danışırıqsa onun tərkibi nədən ibarətdir,

proseduru nədir və s. Başqa sözlə desək biz bu tip informasiyanın çıxardılması üçün

klassifikatorlar düzəldirik. Tərif suallarına kompleks cavablar verən Blair-Goldensohn'un

sisteminin nümunəsinə baxaq. Bizim sualımız var: "What is the Hajj?" Biz qeyd edirik ki, bizim

qısa məzmun çıxartmaq üçün 20 sənədimiz var və 8 cümlə uzunluğunda qısa məzmun çıxartmaq

istəyirik. Biz bütün sənədlərdən informasiya çıxarışı edəcəyik və biz bilirik ki, bu təriflə

bağlı sualdır, deməli biz qenus növlü cümlələri axtaracağıq. Klassifikatorumuzu işə salırıq,

bütün sənədlərdən qenus növlü cümlələri və hər bir informasiya hissəsi üçün digər

cümlələri də tapırıq, cümlələr klasteri düzəldirik. Daha sonra LLR və MMR-ı və ya digər cümlə

seçmə yanaşmalarını tətbiq edirik və nəticəni tərifimizə əlavə edirik. Müxtəlif sənədlərə

baxmağı tələb edərək kompleks suallara cavab verən iki metoddan danışdıq. Biz müxtəlif

sənədlərdən loqrifma ehtimal nisbəti və MMR-dan istifadə edərək artıqlığı olmayan

düzgün informativ cümlələri seçə bilən aşağıdan yuxarıya mətn parçası əsaslı metoddan

danışdıq. Sonra informasiya çıxarılması metodundan danışdıq, hansında ki, hər bir sual tipi

ayrı qəlib düzəldirdik və həmin sual üçün önəmli şeylər olan xüsusi atributları seçirdik.

Əlbəttə ki, müxtəlif sənədlərə baxaraq bu tipli ağır suallar üzərindəki araşdırmalar

hələ başlanğıcdadır. Araşdırmanın hara gedəcəyini görmək üçün bu önəmli və həyəcan verici

bir sahə olacaqdır.