NLP @Google Overview
News Summarization with Word Graphs
Word Clouds for YouTube

Katja Filippova
katjaf@google.com

Google Inc.
Natural Language and Google

- Natural Language – the language used by humans to communicate, the human languages.

- Google’s mission: “To organize the world’s information and make it universally accessible and useful” → understanding the web

- Why is Google interested in natural language processing?
  - Trillions of web pages (?) billions of these containing natural language)
  - Natural language technologies - “understanding” the meaning of web content for better Information Retrieval
  - Natural language tasks - machine translation, speech recognition
Google’s Mission

“To organize the world’s information and make it universally accessible and useful” $\rightarrow$ understanding the web

- Applied techniques for scalable NLP
  - Vector-space similarity
  - Bag-of-words models
  - TF.IDF
  - Regular expressions

- Natural language understanding
  - Part of speech tagging
  - Syntactic parsing
  - Semantic analysis
  - Coreference resolution
  - Discourse processing
Overview

- NLP @ Google
  - Machine translation
  - Speech
  - Large-scale language modeling
  - Information extraction

- Task in focus: summarization
  - News summarization in many languages
  - Video summary from user comments
Another beer, please.

Translate from: English
Translate into: German

Noch ein Bier, bitte.

Contribute a better translation
Machine translation @ Google

Search websites in other languages

Translated results from German web pages

English translation

Snowshoeing - a special kind of winter fun...
Snowshoeing, the re-discovered method to kräftesparend in...Search for all of Switzerland. Aletsch: Aletsch Forest Amten: Snowshoe tour with huskies...
www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee... - 38k - Cached

The Jura Vaudois by snowshoe Explore - Switzerland Tourism
23 km of marked circuit between Sainte-Croix, Les and Le Rasses Chasseron. At the beginning of a trail of 10 km from the Col des Etroits between the Mont-des-Cerfs...
www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/den-wa... - 37k - Cached

Original German - Hide German results

Schneeschuhwandern - ein Wintervergnügen der besonderen Art...
Schneeschuhwandern, die wieder entdeckte Methode, sich kräftesparend im... Suche ganze Schweiz für: Aletsch: Aletschwald Amten: Schneeschuhtour mit Huskys ...
www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee... - 38k - Cache

Den Waadtländer Jura per Schneeschuh entdecken - Schweiz Tourismus
23 km markierter Rundkurs zwischen Sainte-Croix, Les Rasses und Le Chasseron. Zu Beginn eine Spur von 10 km vom Col des Etroits zwischen dem Mont-des-Cerfs ...
www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/den-wa... - 37k - Cache
Machine translation @ Google

- **View original**: Hello everyone
  Hello everyone Sweasy26, www.sweasy26.com
  no rating  2 months ago  16 views  Sweasy26

- **Translate**: bonjour tout le monde
  devenu très vite une tradition :D
  no rating  2 years ago  107 views  marine93340
Machine translation @ Google

Contacts

Group Chat

- Gagan Gupta
  Available
- Qi Ke
  Available
- zh2en
  Invite as Friend

me: happy holidays!
en2zh: 节日快乐！
Qi: 谢谢。
zh2en: Thank you.
me: do you want to go skiing?
en2zh: 你想不想去滑雪吗？
Gagan: sure!
en2zh: 肯定！
fin de semaine

**Hiking Buddy** to me

*show details* 4:30 PM (3 minutes ago)

**Translate message to: English**  Change language  Learn more

Salut !

On est en train de préparer une randonnée du côté de Davos demain, ça te dirait de participer ? N'oublie pas d'apporter un pique-nique !

À demain j'espère,

Hiking Buddy
Machine translation @ Google

What are tourist attractions in Southeast Asia?

Answer:
Thailand: 1 "Temple of the Reclining Buddha The Grand Palace is located next to the Temple of the Reclining Buddha Temple, also known as Buddha, the Buddhist temple of a series of 'most' the history of Bangkok's oldest and largest temples, largest Reclining Buddha in Thailand and the largest st
Translation: ‘hello world’ in English means ‘Bonjour tout le monde’ in French.

Tip: get airline contact nos; send HELP AIRLINE to learn more.
### Machine translation tools

#### Google translator toolkit

<table>
<thead>
<tr>
<th>All translations</th>
<th>Active</th>
<th>Upload</th>
<th>Download</th>
<th>Hide</th>
<th>Delete</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trash</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shared with...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All translation tools:
- Translation memories
- Glossaries

Select: All, None, Completed, Incomplete

<table>
<thead>
<tr>
<th>Name</th>
<th>Words</th>
<th>Date</th>
<th>Sharing</th>
</tr>
</thead>
</table>

---

[Image of Google logo and machine translation tools interface]
What is the Google AJAX Language API?

With the AJAX Language API, you can translate and detect the language of blocks of text within a webpage using only Javascript. In addition, you can enable transliteration on any textfield or textarea in your web page. For example, if you were transliterating to Hindi, this API will allow users to phonetically spell out Hindi words using English and have them appear in the Hindi script.

The language API is designed to be simple and easy to use to translate and detect languages on the fly when offline translations are not available. We plan on adding more exciting capabilities to the AJAX Language API in the future, so stay tuned.

So what is the difference between Translation and Transliteration? Transliteration is the process of phonetically converting a word written in one script into another. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. With transliteration, it is the sound of the words that are converted from one alphabet to the other.
AJAX Language API for Translation and Detection

With the AJAX Language API, you can translate and detect the language of blocks of text within a web page using JavaScript.

The API is new, so there might be some issues and slightly less than perfect documentation. Bear with us as we fill in the holes, and join the [AJAX API forum](https://ajax.googleapis.com/ajax/services/language/forum) to give feedback and discuss the API.

Table of Contents

**Audience**

**Introduction**
- The "Hello, World" of the Google AJAX Language API
- Including the AJAX Language API on Your Page
- API Updates

**Examples**
- Language Translation
- Language Detection
- Detection of Unicode Font Rendering Support
- Source Detection during Translation
- Branding
- Additional Examples

**API Details**
- [Supported Languages](https://ajax.googleapis.com/ajax/services/language/supportedLanguages)
- [Supported Translation Pairs](https://ajax.googleapis.com/ajax/services/language/supportedTranslationPairs)
- [Flash and other Non-Javascript Environments](https://ajax.googleapis.com/ajax/services/language/flash)
- Code Snippets

**Troubleshooting**

**Class Reference**
Speech @ Google

- VoiceSearch - Google search from your spoken query (Android, iPhone, Blackberry)
- Voice spoken input for Maps
- Voicemail transcripts for Google Voice
- YouTube video captioning
- Text-to-speech Google Translate (into English)
- API for Android developers
Large-scale language models

- 7-gram LMs trained on more than 2 trillion tokens
  - MapReduce training
  - Simplified smoothing (Brants et al., EMNLP’07)
  - Randomized data structures (for compression and fast lookup)
- Google n-grams distributed through LDC
  - English trained on 1T tokens
  - Japanese (from 255B tokens)
  - 10 European languages (each trained on 100B tokens)
  - Chinese (5-gram, 883B tokens)
Information extraction

Find breaking news, multimedia, reviews & opinion on Washington, business, sports, movies, travel, books, jobs, education, real estate, cars & more.
www.nytimes.com/ - Similar - Add to iGoogle

Today's Paper   World
Opinion         Jobs
Sports          Obituaries
Business        Real Estate

Today's Paper - New York Times
Use the Today's Paper page to see all the headlines from the Final City Edition of The New York Times organized in the same sections as they appeared in...
www.nytimes.com/pages/todayspaper/index.html - Similar - Add to iGoogle

The New York Times - Wikipedia, the free encyclopedia
The New York Times is an American daily newspaper founded and continuously published in New York City since 1851. Although it remains both the largest local...
en.wikipedia.org/wiki/The_New_York_Times - Cached - Similar

New York Times Learning Network
Sep 10, 2010... An online current and special events news site for grades 3-12; updated weekdays during the school year.
learning.blogs.nytimes.com/ - Cached - Similar

Find global breaking news, multimedia, reviews & opinion, business, sports, movies, travel, books, jobs, education, real estate, cars & more.
global.nytimes.com/ - Similar
Information extraction

Google search results for "new york times square"

Images for new york times square

Videos for new york times square

Home | Times Square New York City

Times Square - Wikipedia, the free encyclopedia

Times Square Cam - EarthCam

Times Square Cam - EarthCam

News for new york times square
Information extraction

Google search results for "hilton paris" showing local branch results for "hilton im Umkreis von Paris, Frankreich". The Hilton Arc de Triomphe Paris hotel is highlighted, described as sitting among iconic Paris landmarks, 10 minutes' walk from the Champs-Élysées. It features a mosaic-decorated spa or ...
Information extraction

Mount Everest - Wikipedia, the free encyclopedia
In 1856, the Great Trigonometric Survey of India established the first published height of Everest, then known as Peak XV, at 29002 ft (8840 m). ... Identifying the highest mountain - Naming - Measurement en.wikipedia.org/wiki/Mount_Everest - Cached - Similar

WikiAnswers - What is the height of Mount Everest
What is height of Mount Everest? 8848m. What is the real height of Mount Everest? 12000 ft!!! Everest is, to begin with, 18000 ft above sea level. ... wiki.answers.com/Q/What_is_the_height_of_Mount_Everest - Cached - Similar

BBC News - Nepal and China agree on Mount Everest's height
8 Apr 2010 ... China and Nepal agree on an official height for Mount Everest - ending a long-running disagreement over the issue. news.bbc.co.uk/2/hi/8608913.stm - Cached

Mt Everest History and facts
Sir George Everest was the first person to record the height and location of Mt. Everest, this is where Mt."Everest" got its name from(In american language) ... www.mnteverest.net/history.html - Cached - Similar

Mount Everest Height has been determined to be seven feet higher
Recent measurements place the elevation of Mount Everest 7 feet higher than previous measurements. pweb.jps.net/~prichins/everest.htm - Cached - Similar

Mount Everest (mountain,Asia) -- Britannica Online Encyclopedia
Information extraction

Bass Pro Shops Outdoors Online: Offering the best in Fishing ... 
The world's most exciting sporting goods available online. Truly everything an outdoor sportsman could dream of online. World's largest fishing and hunting ... www.basspro.com/ - Cached - Similar

Fishing - Wikipedia, the free encyclopedia
Fishing is the activity of catching fish. Fish are normally caught in the wild. Techniques for catching fish include hand gathering, spearing, netting, ...
Rod - Netting - Industry - Vessels en.wikipedia.org/wiki/Fishing - Cached - Similar

BASS Fishing Membership and Tournament News, Information ... 
Lures, storage and terminal tackle you need to carry. ... I finally got the fish within arm's reach and was about to grab her when she made a surge ...
Tournaments - Live Coverage - Games & Contests - Opens sports.espn.go.com/outdoors/bassmaster/index - Cached - Similar
Information extraction

Phishing - Wikipedia, the free encyclopedia
An example of a phishing e-mail, disguised as an official e-mail from a (fictional) bank. The sender is attempting to trick the recipient into revealing...
en.wikipedia.org/wiki/Phishing - Cached - Similar

E-mail Hoaxes and Fraudulent Links | Microsoft Security
Tips to help recognize fraudulent or hoax emails and links.
www.microsoft.com/protect/fraud/phishing/symptoms.aspx - Cached - Similar

Report Phishing Sites
US-CERT is collecting phishing email messages and web site locations so... You can report phishing to us by sending email to phishing-report@us-cert.gov...
www.us-cert.gov/nav/report_phishing.html - Cached - Similar

News for ~fishing email

Tax demands fraud risk - 2 hours ago
The Which? website has more information on phishing and ID Fraud, or you can also visit direct.gov.uk. If you receive a phishing email purporting to come...
Which? - 814 related articles »

Phishing Emails, Scams and Hoaxes | Credit Card Safety & Security...
Google Squared

www.google.com/squared

- Project aims:
  - Web scale: extract from tens of billions of pages.
  - Open domain: answer questions on any topic.
  - Automatic extraction, no manual intervention.
  - Solve real problems, learn from user feedback.
<table>
<thead>
<tr>
<th>Item Name</th>
<th>Description</th>
<th>Symbol</th>
<th>Central Bank</th>
<th>Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro</td>
<td>The euro (sign: €; code: EUR) is the official currency of the eurozone: 16 of the 27 Member States of the European Union (EU). It is also the currency used</td>
<td>EUR</td>
<td>European Central Bank</td>
<td>0.9%, February 2010</td>
</tr>
<tr>
<td>Japanese yen</td>
<td>The yen is the Japanese currency unit. There are four kinds of bills (10000 yen, 5000 yen, 2000 yen, 1000 yen) and six kinds of coins (500 yen, 100 yen, ...</td>
<td>JPY</td>
<td>bank of japan</td>
<td>-1.3%</td>
</tr>
<tr>
<td>Australian dollar</td>
<td>The Australian dollar is currently the fifth-most-traded currency in the world foreign exchange markets behind the US dollar, the euro, the yen and the ...</td>
<td>AUD</td>
<td>Reserve Bank of Australia</td>
<td>2.1%</td>
</tr>
<tr>
<td>Canadian dollar</td>
<td><strong>Canadian Dollar</strong> Currency Exchange Forecast. <strong>Canadian Dollar</strong> to US Dollar Currency Exchange Forecast. <strong>Canadian Dollar</strong> to One U.S. Dollar. Average of ...</td>
<td>$ or C$</td>
<td>Bank of Canada</td>
<td>1%</td>
</tr>
<tr>
<td>Swiss franc</td>
<td>The franc (German: Franken, French and Romansh: franc, Italian: franco; code: CHF) is the currency and legal tender of Switzerland and Liechtenstein;</td>
<td>CHF</td>
<td>Swiss National Bank</td>
<td>-0.5%</td>
</tr>
<tr>
<td>Indonesian rupiah</td>
<td>The rupiah (Rp) is the official currency of Indonesia. Issued and controlled by the Bank of Indonesia, the ISO 4217 currency code for the Indonesian</td>
<td>rp</td>
<td>Bank Indonesia</td>
<td>3.65%</td>
</tr>
</tbody>
</table>
Summarization
A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s).

- information retrieval
- stock market prediction
- generation of abstracts
- online news summarization
- ...

Text summarization
Text summarization

- A **summary** is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s).

- **Indicative**
  - indicates types of information
  - “alerts”

- **Informative**
  - includes quantitative/qualitative information
  - “informs”
The work of Consumer Advice Centres is examined. The information sources used to support this work are reviewed. The recent closure of many CACs has seriously affected the availability of consumer information and advice. The contribution that public libraries can make in enhancing the availability of consumer information and advice both to the public and other agencies involved in consumer information and advice, is discussed.
In informative

An examination of the work of Consumer Advice Centres and of the information sources and support activities that public libraries can offer. CACs have dealt with pre-shopping advice, education on consumers’ rights and complaints about goods and services, advising the client and often obtaining expert assessment. They have drawn on a wide range of information sources including case records, trade literature, contact files and external links. The recent closure of many CACs has seriously affected the availability of consumer information and advice. Libraries can cooperate closely with advice agencies through local coordinating committed, shared premises, join publicity referral and the sharing of professional expertise.
Text summarization

Form:
- headlines
- snippets
- abstracts
- answers
- outlines
Text summarization

- Source: single-document vs. multi-document
  - research paper
  - proceedings of a conference

- Content: generic vs. query-based vs. user-focused
  - equal coverage of all major topics
  - based on a question “what are the causes of the war?”
  - users interested in chemistry

- Approach: extract vs. abstract
  - fragments from the document
  - newly re-written text
Extraction vs. abstraction

How should a text summarization system proceed?

• read the documents

• understand them – build a semantic representation

• generate a summary from this representation
Extraction vs. abstraction

- Unfortunately, a rich semantic representation is not possible yet.
- To date, most summarization systems are extractive.
- Usually, extraction units are sentences.
- Low cost solution: could work without ontologies, complex representations, etc.
- Extractive summaries are usually incoherent.
- Trade-off between non-redundancy and completeness.
Extraction vs. abstraction

- A common extractive approach to multi-document summarization:
  - similar sentences are grouped into clusters
  - the clusters are ranked
  - a sentence is selected from each of the top clusters

- Sentences often contain irrelevant information.
- Better wording might exist in different sentences.
Three sentences from related documents (Oct. 27 2009):

- The Syrian foreign minister today condemned the killing of eight civilians in a US raid as an act of "criminal and terrorist aggression". (The Guardian)

- Syria accused the United States on Monday of carrying out a "terrorist aggression" after a deadly raid near its border with Iraq which it said killed eight civilians. (Reuters)

- Lebanese President Michel Suleiman on Monday contacted his Syrian counterpart Bashar Assad to denounce "Sunday's American aggression" against the Syrian village of Abu Kamal near the border with Iraq, local ElNashra website reported. (Aljazeera)
Three sentences from related documents (Oct. 27 2009):

- The Syrian foreign minister today condemned the killing of eight civilians in a US raid as an act of "criminal and terrorist aggression". (The Guardian)

- Syria accused the United States on Monday of carrying out a "terrorist aggression" after a deadly raid near its border with Iraq which it said killed eight civilians. (Reuters)

- Lebanese President Michel Suleiman on Monday contacted his Syrian counterpart Bashar Assad to denounce "Sunday's American aggression" against the Syrian village of Abu Kamal near the border with Iraq, local Elnashra website reported. (Aljazeera)
Extraction vs. abstraction

Three sentences from related documents (Oct. 27 2009):

- The Syrian foreign minister today condemned the killing of eight civilians in a US raid as an act of "criminal and terrorist aggression". (The Guardian)

- Syria accused the United States on Monday of carrying out a "terrorist aggression" after a deadly raid near its border with Iraq which it said killed eight civilians. (Reuters)

- Lebanese President Michel Suleiman on Monday contacted his Syrian counterpart Bashar Assad to denounce "Sunday's American aggression" against the Syrian village of Abu Kamal near the border with Iraq, local Elnashra website reported. (Aljazeera)
Extraction vs. abstraction

Three sentences from related documents (Oct. 27 2009):

- The Syrian foreign minister today condemned the killing of eight civilians in a US raid as an act of "criminal and terrorist aggression". (The Guardian)

- Syria accused the United States on Monday of carrying out a "terrorist aggression" after a deadly raid near its border with Iraq which it said killed eight civilians. (Reuters)

- Lebanese President Michel Suleiman on Monday contacted his Syrian counterpart Bashar Assad to denounce "Sunday's American aggression" against the Syrian village of Abu Kamal near the border with Iraq, local Elnashra website reported. (Aljazeera)
Extraction vs. abstraction

- Extractive summaries are not coherent – sentences pulled out from different documents make sense each but sound awkward when put together.
  - unresolved pronouns may distort the meaning;
  - beginning with a sentence which starts with *However, ...* is not a good idea.
Extraction vs. abstraction

- Extractive summaries are not coherent – sentences pulled out from different documents make sense each but sound awkward when put together.
  - unresolved pronouns may distort the meaning;
  - beginning with a sentence which starts with *However, ...* is not a good idea.

- Can this problem be solved without doing abstraction?
  - sentence compression;
  - sentence fusion.
Summarization on the sentence level:

As The Labour leadership congratulates itself on a virtually unprecedented exhibition of unity and moderation, they should be aware that knives are being sharpened at Conservative Central Office.
Summarization on the sentence level:

As The Labour leadership congratulates itself on a virtually unprecedented exhibition of unity and moderation, they should be aware that knives are being sharpened at Conservative Central Office.
Fusing information from different sentences in a single one:

The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois’, 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago's ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment's guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.
Fusing information from different sentences in a single one:

The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois', 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago's ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment’s guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.

The Supreme Court reversed a ban on gun ownership.
The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois’ 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago's ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment’s guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.

On Monday, the Supreme Court reversed a ban on gun ownership.
Sentence fusion

Fusing information from different sentences in a single one:

The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois’, 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago’s ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment’s guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.

On Monday, the Supreme Court reversed a 28 y. o. ban on gun ownership.
Fusing information from different sentences in a single one:

The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois’, 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago's ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment’s guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.

On Monday, the Supreme Court reversed a 28 y. o. Chicago ban on handgun ownership.
The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

In another dramatic victory for firearm owners, the Supreme Court has ruled unconstitutional Chicago, Illinois’, 28-year-old strict ban on handgun ownership, a potentially far-reaching case over the ability of state and local governments to enforce limits on weapons.

The Supreme Court reversed a ruling upholding Chicago’s ban on handguns Monday and extended the reach of the 2nd Amendment as a nationwide protection against laws that infringe on the "right to keep and bear arms."

The Second Amendment’s guarantee of an individual right to bear arms applies to state and local gun control laws, the Supreme Court ruled on Monday in 5-to-4 decision.

On Monday, the Supreme Court reversed a 28 y. o. Chicago ban on handgun ownership in 5-to-4 decision.
Challenges

- How can important content be identified?
  - Word scoring: words recurring in this but not many other documents
  - Syntactic clues: sentence subject is likely to be more important than a prepositional phrase

- How can grammatical sentences be generated?
  - Language models (high-scoring strings should be preferred)
  - Syntactic rules (e.g., there must be a subject in a sentence)

- Can redundancy be used not only for important content identification but also for generating grammatical sentences?
Multi-sentence compression

- A word graph built from related sentences:
  - vertices = tokens $\cup \{\text{Start, End}\}$
  - edges represent token adjacency
- A compression is a path in the graph from Start to End.
- Identical lowercased tokens are merged if:
  - they have the same part of speech;
  - they have some overlap in neighbors

(for more details see Filippova, Coling’10)
1. Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week.

2. Hillary Clinton paid a visit to the People’s Republic of China on Monday.

3. The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday.

4. Last week the Secretary of State Ms. Clinton visited Chinese officials.
Hillary Clinton wanted to visit China last month till last week, but postponed her plans.
Word graph

- Words from a new sentence are added in three steps:
  - unambiguous non-stopwords – either merged with an existing node or a new node is created;
  - ambiguous non-stopwords – a node with some overlap in neighbors is preferred;
  - stopwords – only added if the following word matches the out-neighbor of the node.

- The graph permits loops.

- Words from the same sentence are never merged in one node.
Hillary Clinton wanted to visit China but postponed her plans last week.
Hillary Clinton wanted to visit China last month of the week till the last E official. Clinton visited China on Monday.
K shortest paths

\[ w(e) = \frac{1}{\text{freq}(e)} \]
K shortest paths

\[ w(e) = \frac{1}{\sum_{s \in S} \text{distance}(s, u, v)^{-1}} \]
K shortest paths

\[ w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\sum_{s \in S} \text{distance}(s, u, v) - 1} \]
$K$ shortest paths

$w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\text{freq}(u) \times \text{freq}(v) \times \sum_{s \in S} \text{distance}(s, u, v)^{-1}}$
K shortest paths

- Paths shorter than eight edges are discarded.
- Paths not passing a verb are filtered out.
- The total path length is normalized by the number of edges.
Data: Google News

Google News

Top Stories
Gun politics
Justine Henin
Vince Neil
Chris Brown
Robert Byrd
Apple
Elena Kagan
Tropical storm
Alex
Iphone
Diabetes
mellitus

Top Stories

Gun politics

Justices Say Gun Rights Apply Locally
New York Times - Adam Liptak - 26 minutes ago
The Supreme Court ruled on Monday that the Second Amendment, which forbids Congress from infringing the right to keep and bear arms, applies to state and local governments as well.
Supreme Court extends 2nd Amendment protection to state, local levels Los Angeles Times
Supreme Court extends gun rights BBC News
CNN - The Associated Press - Wall Street Journal - Reuters
all 686 news articles »

Robert Byrd

Sen. Robert Byrd dies at 92
Los Angeles Times - 1 hour ago
The West Virginia lawmaker was the longest-serving member of Congress in US history and cast more votes than any other member since taking office in 1959.
In Byrd's home county, recipients of pork largesse mourn their benefactor Washington Post
Testimonials to Sen. Robert C. Byrd The Associated Press
all 2,998 news articles »

Elena Kagan

Live coverage of the Elena Kagan Senate confirmation hearings
Los Angeles Times - 19 minutes ago
Kagan, President Obama's nominee for the US Supreme Court, appears before the Senate Judiciary Committee on Monday. If confirmed, Kagan, solicitor general of the United States and a professor of Harvard Law School, would be the fourth woman
**Taliban Suicide Attack Kills At Least 5 in Afghanistan**
Voice of America - 54 minutes ago
Photo: AP Taliban suicide bombers have killed at least five people, including three foreigners, in an attack on the compound of a USAID contractor in northern Afghanistan. Afghan officials in the city of Kunduz say one bomber blew up his vehicle at the ...

**Bombers Storm US Aid Compound in Afghanistan**
New York Times - Richard A. Oppel, Sharifullah Sahak - 4 hours ago
KABUL, Afghanistan — A half-dozen suicide bombers stormed the compound of a American contractor working for the United States Agency for International Development in the northern city of Kunduz on Friday, ...

**Taliban claims Afghan attack that killed five**
Los Angeles Times - Laura King - 4 hours ago
KABUL, Afghanistan A Taliban suicide squad stormed the compound of a US-based development group in northern Afghanistan on Friday, killing at least three expatriate workers, a security guard and an Afghan police officer, officials said. ...

"The building has been destroyed," Mr. Omar said.
more by Mullah Mohammad Omar - 4 hours ago - New York Times (2 occurrences)

**Suicide Bombers Attack USAID Compound in Northern Afghanistan**
Voice of America - 4 hours ago
Photo: AP Taliban suicide bombers have killed at least two people and wounded several others in an attack on a USAID compound in northern Afghanistan. The governor of northern Kunduz province says one bomber blew up his vehicle at the entrance to allow ...
A news cluster consists of related articles from different sources:
- published at about the same time
- about the same event
- contains duplicates
- can be noisy

In news, first sentences are known to summarize the content of the article:
- competitive baseline (DUC, TAC)
- expected to be similar
- considerably longer than other sentences
Evaluation

- Baseline: sequence which has the maximum product of bigram and unigram probabilities.
- Two configurations of Shortest path:
  - inverted edge frequency;
  - the final formula.
- 80 news clusters for English, 40 for Spanish.
- Four native speakers per cluster-compression pair.
Evaluation

**Taliban Suicide Attack Kills At Least 5 in Afghanistan**
Voice of America - 54 minutes ago
Photo: AP Taliban suicide bombers have killed at least five people, including three foreigners, in an attack on the compound of a USAID contractor in northern Afghanistan. Afghan officials in the city of Kunduz say one bomber blew up his vehicle at the ...

**Bombers Storm US Aid Compound in Afghanistan**
New York Times - Richard A. Oppel, Sharifullah Sahak - 4 hours ago
KABUL, Afghanistan — A half-dozen suicide bombers stormed the compound of a American contractor working for the United States Agency for International Development in the northern city of Kunduz on Friday, ...

**Taliban claims Afghan attack that killed five**
Los Angeles Times - Laura King - 4 hours ago
KABUL, Afghanistan A Taliban suicide squad stormed the compound of a US-based development group in northern Afghanistan on Friday, killing at least three expatriate workers, a security guard and an Afghan police officer, officials said. ...

**Suicide Bombers Attack USAID Compound in Northern Afghanistan**
Voice of America - 4 hours ago
Photo: AP Taliban suicide bombers have killed at least two people and wounded several others in an attack on a USAID compound in northern Afghanistan. The governor of northern Kunduz province says one bomber blew up his vehicle at the entrance to allow ...
Evaluation

- Is there a main event in the cluster? (yes/no)
- Is the compression grammatical?
  - perfect (2)
  - minor mistake (1)
  - otherwise (0)
- Does it summarize the main event, if present?
  - summarizes the main event (2)
  - related to the main event but misses smth important (1)
  - otherwise (0)
## Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (EN)</td>
<td>21%</td>
<td>15%</td>
<td>65%</td>
<td>8 / 28</td>
</tr>
<tr>
<td>Shortest path (EN)</td>
<td>52%</td>
<td>16%</td>
<td>32%</td>
<td>10 / 28</td>
</tr>
<tr>
<td>Shortest path++ (EN)</td>
<td>64%</td>
<td>13%</td>
<td>23%</td>
<td>12 / 28</td>
</tr>
<tr>
<td>Baseline (ES)</td>
<td>12%</td>
<td>15%</td>
<td>74%</td>
<td>8 / 35</td>
</tr>
<tr>
<td>Shortest path (ES)</td>
<td>58%</td>
<td>21%</td>
<td>21%</td>
<td>10 / 35</td>
</tr>
<tr>
<td>Shortest path++ (ES)</td>
<td>50%</td>
<td>21%</td>
<td>29%</td>
<td>12 / 35</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>Info-2</th>
<th>Info-1</th>
<th>Info-0</th>
<th>Avg. Len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (EN)</td>
<td>18%</td>
<td>10%</td>
<td>73%</td>
<td>8 / 28</td>
</tr>
<tr>
<td>Shortest path (EN)</td>
<td>36%</td>
<td>33%</td>
<td>31%</td>
<td>10 / 28</td>
</tr>
<tr>
<td>Shortest path++ (EN)</td>
<td>52%</td>
<td>32%</td>
<td>16%</td>
<td>12 / 28</td>
</tr>
<tr>
<td>Baseline (ES)</td>
<td>9%</td>
<td>19%</td>
<td>72%</td>
<td>8 / 35</td>
</tr>
<tr>
<td>Shortest path (ES)</td>
<td>23%</td>
<td>26%</td>
<td>51%</td>
<td>10 / 35</td>
</tr>
<tr>
<td>Shortest path++ (ES)</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
<td>12 / 35</td>
</tr>
</tbody>
</table>
Results

- Sentence compression in the context of MDS – multi-sentence compression.
- Experiments with English, French, Italian, Spanish, German and Russian.
- Evaluation on English and Spanish.
- A simple, syntax-lean method with surprisingly good results.
YouTube comments

- YouTube - video-sharing website: upload, share, view

- For every video, the uploader can provide:
  - title
  - description
  - tags
  - category

- Viewers can provide comments:
  - lolololo
  - omg, ccccooooollll!!!
  - i luv tihs vid coz its sooo cooooolll!
Why bother about user comments?
- The title and description can be uninformative (e.g., `IMG_2947219.avi`).
- Many videos do not have tags.
- Description tells us about the video from the uploader’s perspective.
- What do the viewers think about the video?
YouTube comments

- Why bother about user comments?
  - The title and description can be uninformative (e.g., IMG_2947219.avi).
  - Many videos do not have tags.
  - Description tells us about the video from the uploader’s perspective.
  - What do the viewers think about the video?

- Comments are very different from news:
  - spelling errors
  - poor grammar
  - lots of meaningless noise
Task: select most salient, representative words from the comments on a video.

Simple approach: tokenize comments, count word frequencies.

Most frequent words are not representative of the video: *lol, cool, the*

Filter (YouTube-specific) stopwords.
Extract the list of YouTube stopwords:
- 10K videos from each of the 15 YouTube categories
- videos with at least 10 comments are considered
- only first 500 comments are taken (balanced dataset)
- video count for every word
YouTube comment cloud

- Most frequent words:
  a, i, the, is, to, and, it, you, in, this, that, of, so, for, me, on, like, but, was, my, have, video, are, with, what, do, lol, just, not, be, good, all, your, one, at, no, can, if, love, get, how, u

- From top-200:
  love, nice, really, wow, awesome, thanks, first, haha, song, shit, please, ur, omg, dude, funny, god, amazing, guys, fuck*, ya, yeah
Words most used in comments on this video

sexy touch wasn't yeaaaaaaa imagine multi-tasking whistling

screen sposed mid products macbook nor removable-

replaceable inbetween ly ipod sentence metre picture
YouTube comment cloud

Nokia N97 Demo

Nokia Conversations 479 videos Subscribe

Nokia N97

The N97 isn't a device that will trigger knee-jerk hysteria, but instead it should breed cool-headed excitement at the prospect of a new era in mobile experience. Story to continue in "Beware: Nokia N97 Unleashed!"
YouTube comment cloud

Words most used in comments on this video

nokia  mobile  finland  keyboard  os  touch  months
n95  symbian  n97  screen  week  software
phone  xperia  apple  apps  slow  iphone
n900
Wrap-up

- NLP is crucial to organize the information available on the web.
- Possible applications: machine translation, speech processing, information extraction, summarization.
- Large-scale distributed processing, language-independent methods.
- Real-world tasks with lots of challenges.
Thank you! Questions?