

# **NLP @Google Overview**

## **News Summarization with Word Graphs**

### **Word Clouds for YouTube**

**Katja Filippova**

`katjaf@google.com`

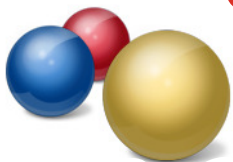
Google Inc.



# 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

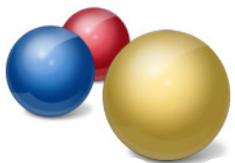


“To organize the world’s information and make it universally accessible and useful” → understanding the web

- 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



- 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



# Machine translation @ Google



## Google translate

Translation

[Translated Search](#)

[Translator Toolkit](#)

[Tools and Resources](#)

### Translate text, webpages and documents

Enter text or a webpage URL, or [upload a document](#).

Another beer, please.

Translate from: English

Translate into: German

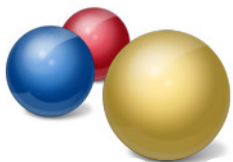


Translate

#### English to German translation

Noch ein Bier, bitte.

[+ Contribute a better translation](#)



# Machine translation @ Google



Google translate

[Translation](#)

[Translated Search](#)

[Translator Toolkit](#)

[Tools and Resources](#)

## Search websites in other languages

Snowshoeing in switzerland

Search pages written in:  Translated to: **Schneeschuhwandern in der Schweiz** - [Not quite right? Edit](#)

My language:

Translated results from German web pages

Results 1 - 10 of about 118,000 for Schneeschuhwandern in der Schweiz.

### 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 Amden: **Snowshoe** tour with huskies...  
[www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee...](http://www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee...) - 36k - [Cached](#)

### Original German - [Hide German results](#)

#### [Schneeschuhwandern - ein Wintervergnügen der besonderen Art...](#)

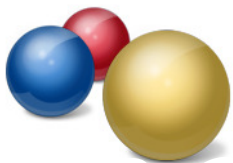
**Schneeschuhwandern**, die wieder entdeckte Methode, sich kräftesparend im ... Suche ganze **Schweiz** für. Aletsch: Aletschwald Amden: Schneeschuhtour mit Huskys ...  
[www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee...](http://www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee...) - 36k - [lm](#)  
[Cache](#)

#### [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...](http://www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/den-wa...) - 37k - [Cached](#)

#### [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...](http://www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/den-wa...) - 37k - [lm](#)  
[Cache](#)



# Machine translation @ Google



**You Tube** Broadcast Yourself™  
Worldwide | English Si

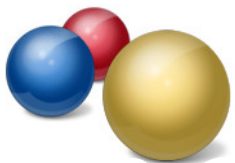
[Home](#) [Videos](#) [Channels](#) [Community](#)

**2/28/09: Your Weekly Address**

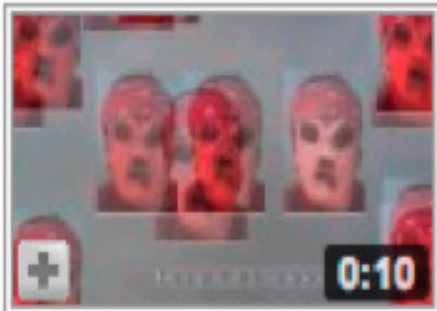


في القضاء على برامج اننا لسنا في حاجة ل

2:36 / 4:53



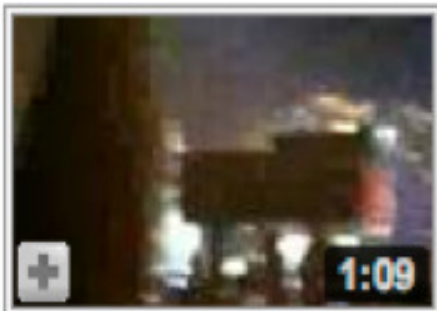
# Machine translation @ Google



[View original](#) [Hello everyone](#)

Hello everyone Sweasy26, [www.sweasy26.com](http://www.sweasy26.com)

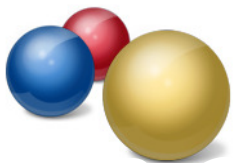
no rating 2 months ago 16 views [Sweasy26](#)



[Translate](#) [bonjour tout le monde](#)

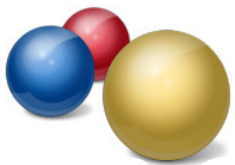
devenu très vite une tradition :D

no rating 2 years ago 107 views [marine93340](#)





# Machine translation @ Google



# Machine translation @ Google



fin de semaine Inbox | X

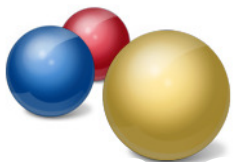
☆ ● **Hiking Buddy** to me [show details](#) 4:30 PM (3 minutes ago) ↩ Reply | ▼

[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 !

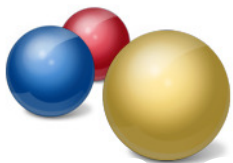
À demain j'espère,  
Hiking Buddy



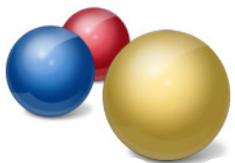
# Machine translation @ Google



The screenshot shows a Google search result page for the question "What are tourist attractions in Southeast Asia?". The page is displayed in English, as indicated by the Google Toolbar which says "Toolbar has automatically translated this page to English". The search bar contains the Chinese text "东南亚有什么旅游景点?". The search result is from the website "wenda.tianya.cn" and is titled "What are tourist attractions in Southeast Asia?". The question is "What are tourist attractions in Southeast Asia?". The answer is from user "liushuwen0706" and reads: "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...".



# Machine translation @ Google





# Machine translation tools





## Google translator toolkit

All translations


 **Active**


 Hidden

 Trash

 Shared with...

All translation tools

 Translation memories

 Glossaries

Select: [All](#), [None](#), [Completed](#), [Incomplete](#)

Name	Words	Date ↓	Sharing
------	-------	--------	---------



# Machine translation tools



## ★ Google AJAX Language API

### 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.

Добро пожаловать!



# Machine translation tools



## 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 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](#) **New Languages!**

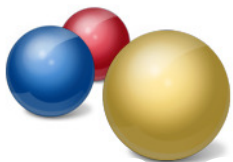
[Supported Translation Pairs](#) **New Languages!**

#### Flash and other Non-Javascript Environments

[Code Snippets](#)

#### Troubleshooting

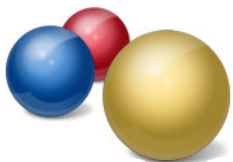
#### Class Reference



# 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



new york

Search

About 3,120,000,000 results (0.21 seconds)

Advanced search

Everything

Images

Videos

Maps

News

Discussions

More

Any time

Latest

Past 3 weeks

All results

Related searches

Timeline

Page previews

More search tools

Something different

[new jersey](#)

[washington](#)

[pennsylvania](#)

[oklahoma](#)

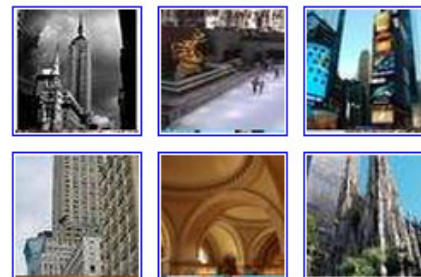
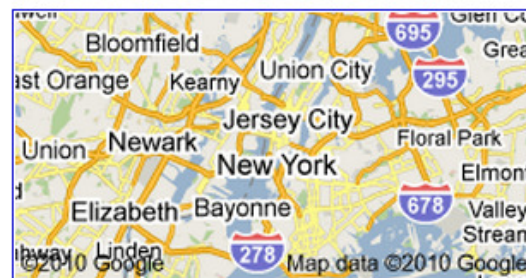
[texas](#)

## Städtereise nach New York

[travel.ch/staedtereisen\\_newyork](http://travel.ch/staedtereisen_newyork) Wählen Sie aus über 600 Hotels und 22 Flügen pro Tag mit Travel.ch!  
Dublin ab CHF 298.- - Berlin ab CHF 213.- - Amsterdam ab CHF 225.-

Sponsored link

## New York, NY [maps.google.com](http://maps.google.com)



[Hotels](#) - [Restaurants](#) - [Empire State Building](#) - [Rockefeller center](#) - [Statue of Liberty](#) - [Ellis Island](#) - [Grand Central Terminal](#) - [Metropolitan Museum of Art](#)

## ILOVENY | The Official New York State Tourism Site-Home ☆

Visit ILOVENY.com for all things **New York** state tourism. ILOVENY.com is the official **New York** state tourism site. Book events, find places to eat and stay, ...

[See & Do](#) - [Travel Guides](#) - [What To Do](#) - [Where To Stay](#)

[www.iloveny.com/](http://www.iloveny.com/) - [Cached](#) - [Similar](#)

## The Official New York City Site ☆

Provides the public with quick and easy access to information about **New York** City agencies, programs and services. Also provides, through links to external ...

[www.nyc.gov/](http://www.nyc.gov/) - [Cached](#) - [Similar](#)

## WWW.NY.GOV ☆

Locate **New York** State Government information and services available on the Web.

[www.state.ny.us/](http://www.state.ny.us/) - [Cached](#) - [Similar](#)

## New York - Wikipedia, the free encyclopedia ☆

# Information extraction



new york times

X Search

About 87,900,000 results (0.13 seconds)

[Advanced search](#)

**Everything**

**Images**

**Videos**

**News**

**More**

**Any time**

[Latest](#)

[Past 2 weeks](#)

[More search tools](#)

Something different

[wall street journal](#)

[washington post](#)

[los angeles times](#)

[la times](#)

[boston globe](#)

## [The New York Times - Breaking News, World News & Multimedia](#) ☆

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](#)

## [The New York Times - Breaking News, World News & Multimedia](#) ☆

Find global breaking news, multimedia, reviews & opinion, business, sports, movies, travel, books, jobs, education, real estate, cars & more.

[global.nytimes.com/](#) - [Similar](#)





# Information extraction



new york times square

Search

About 14,100,000 results (0.22 seconds)

Advanced search

Everything

Images

Videos

News

Maps

More

Any time

Latest

Past 2 days

All results

Sites with images

More search tools

## Images for new york times square - Report images



## Videos for new york times square



[Times Square Cam - EarthCam](#)  
Jul 6, 2009  
earthcam.com



[Times Square Cam - EarthCam](#)  
Aug 26, 2009  
earthcam.com

## Home | Times Square New York City ☆

Information on **Times Square** in **New York** City including new year's, restaurant, hotels, theater, museums, travel and visitor guides, tours, billboards, ...

[www.timessquare.com/](#) - Cached

## Times Square - Wikipedia, the free encyclopedia ☆

As the growth in **New York** City continued, **Times Square** quickly became a .... "**Times Square New York** City, **New York** City **Times Square** by NYCTourist.com". ...

[en.wikipedia.org/wiki/Times\\_Square](#) - Cached - Similar

## Times Square Cam - EarthCam ☆

View a large choice of webcams from **New York** City's **Times Square**. Broadway's huge number of animated neon and television-style signage have long made it one ...

[EarthCam Network - Panoramas - My EarthCam](#)

[www.earthcam.com/usa/newyork/timessquare/](#) - Cached - Similar

## News for new york times square

[Frank Chanke: the new year's Eve in Times Square tonight](#) ☆ - 23 hours ago



# Information extraction



paris hilton

Search

About 62,500,000 results (0.33 seconds)

[Advanced search](#)

**Everything**

[News](#)

[Images](#)

[Videos](#)

[Blogs](#)

[More](#)

**Any time**

[Latest](#)

[Past 2 days](#)

**All results**

[Sites with images](#)

[More search tools](#)

Something different

[pamela anderson](#)

[angelina jolie](#)

[lindsay lohan](#)

[jessica simpson](#)

[jessica alba](#)



## [News for paris hilton](#)



[Paris Hilton Eats, Plays, Loves in Maui](#) ☆ - 18 hours ago

Xo Paris :) " At some point after her vacation in paradise **Hilton** will have to face the music. She was charged with felony cocaine possession following her ...

Kansas  
City Star

[People Magazine - 752 related articles »](#)

[Paris Hilton burglary suspect pleads not guilty](#) ☆ -

[Daily News & Analysis - 998 related articles »](#)

[Charges dismissed in John Travolta extortion case: Paris Hilton ...](#) ☆ -

[Washington Post - 93 related articles »](#)

## [Images for paris hilton](#) - [Report images](#)



[Paris Hilton - Wikipedia, the free encyclopedia](#) ☆

**Paris Whitney Hilton** (born February 17, 1981) is an American socialite, heiress,



# Information extraction



hilton paris

Suche

Ungefähr 40'000'000 Ergebnisse (0.38 Sekunden) [Erweiterte Suche](#)

- Alles**
- Maps**
- News**
- Bilder**
- Videos**
- Mehr**

## Das Web

[Seiten auf Deutsch](#)  
[Seiten aus der Schweiz](#)

## Alle

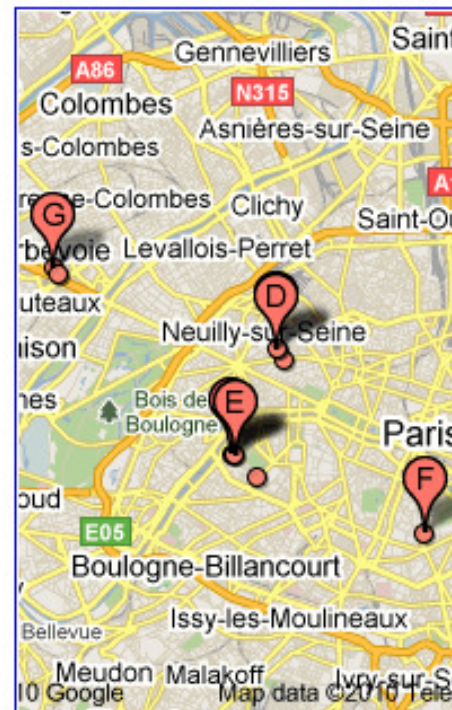
[Neueste](#)  
[Letzte 2 Tage](#)

## Alle Ergebnisse

[Websites mit Bildern](#)

[Mehr Optionen](#)

## Lokale Branchenergebnisse für **hilton** im Umkreis von **Paris, Frankreich**

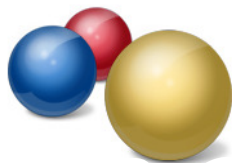


- Hilton Arc de Triomphe Paris Hotel** ☆  
[hiltonmap.com](#) - 01 58 36 67 00 - 497 Bewertungen
- Hilton Paris** ☆  
[maps.google.ch](#) - Mehr
- Hilton** ☆  
[maps.google.ch](#) - 01 42 73 92 00 - 2 Bewertungen
- Hilton Arc de Triomphe Paris Hotel** ☆  
[hiltonmap.com](#) - 01 58 36 67 00 - Mehr
- Hilton International Orly** ☆  
[maps.google.ch](#) - 08 00 04 08 03 - Mehr
- Hilton Holly** ☆  
[maps.google.ch](#) - 09 54 33 66 49 - Mehr
- Hilton Paris La Defense Hotel** ☆  
[hiltonmap.com](#) - 01 46 92 10 10 - 228 Bewertungen

[Weitere Ergebnisse im Umkreis von Paris, Frankreich](#) »

[Hilton Arc de Triomphe Paris hotel](#) ☆ - [[Diese Seite übersetzen](#)]

The **Hilton** Arc de Triomphe **Paris** hotel sits among iconic **Paris** landmarks, 10 minutes' walk from the Champs-Élysées. Indulge in the mosaic-decorated spa or ...



# Information extraction



what is the height of everest

Search

About 1,230,000 results (0.41 seconds)

[Advanced search](#)

 **Everything**

More

Show search tools

[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](http://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](http://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](http://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) ...

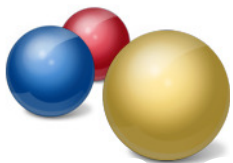
[www.mnteverest.net/history.html](http://www.mnteverest.net/history.html) - [Cached](#) - [Similar](#)

[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](http://pweb.jps.net/~prichins/everest.htm) - [Cached](#) - [Similar](#)

[Mount Everest \(mountain, Asia\) -- Britannica Online Encyclopedia](#) ☆



# Information extraction



~fishing

Search

About 177,000,000 results (0.28 seconds)

[Advanced search](#)

 **Everything**

 Images

 Videos

 News

 Maps

 Discussions

 Blogs

 Books

 More

**Any time**

Latest

Past 3 weeks

**All results**

[Related searches](#)

[Wonder wheel](#)

[Page previews](#)

 More search tools

[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](#)

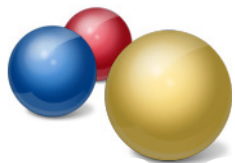
[en.wikipedia.org/wiki/Fishing](#) - [Cached](#) - [Similar](#)

[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



~fishing email

Search

About 401,000,000 results (0.25 seconds)

[Advanced search](#)

Go to Google Home



Everything



News



More

Any time

Latest

[More search tools](#)

[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](http://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](http://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](mailto:phishing-report@us-cert.gov) ...

[www.us-cert.gov/nav/report\\_phishing.html](http://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](http://direct.gov.uk). If you receive a **phishing email** purporting to come ...

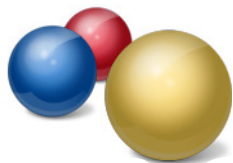
[Which?](#) - [814 related articles](#) »

[Phishing Emails, Scams and Hoaxes | Credit Card Safety & Security ...](#)



Another popular **phishing** scam is to **email** you to claim that your account will be closed unless you confirm confidential personal information or update ...

[www.capitalone.co.uk/.../phishing-emails-and-scams.jsf](http://www.capitalone.co.uk/.../phishing-emails-and-scams.jsf) - [Cached](#)



# 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.



# Google Squared





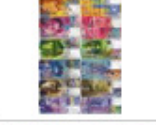










**currencies** 20 items

<input type="checkbox"/>	Item Name	<input type="checkbox"/> Image	<input type="checkbox"/> Description	<input type="checkbox"/> Symbol	<input type="checkbox"/> Central Bank	<input type="checkbox"/> Inflation
<input checked="" type="checkbox"/>	<b>Euro</b>		The <b>euro</b> (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	EC	European Central Bank	0,9%, February 2010
<input checked="" type="checkbox"/>	<b>Japanese yen</b>		The <b>yen</b> is the <b>Japanese</b> currency unit. There are four kinds of bills (10000 <b>yen</b> , 5000 <b>yen</b> , 2000 <b>yen</b> 1000 <b>yen</b> ) and six kinds of coins (500 <b>yen</b> , 100 <b>yen</b> , ...	JPY	bank of japan	-1.3%
<input checked="" type="checkbox"/>	<b>Australian dollar</b>		The <b>Australian dollar</b> is currently the fifth-most-traded currency in the world foreign exchange markets behind the US dollar, the euro, the yen and the ...	AUD	reserve bank of australia	2.1%
<input checked="" type="checkbox"/>	<b>Canadian dollar</b>		<b>Canadian Dollar</b> Currency Exchange Forecast. <b>Canadian Dollar</b> to US Dollar Currency Exchange Forecast. <b>Canadian Dollars</b> to One U.S. Dollar. Average of	\$ or C\$	Bank of Canada	1%
<input checked="" type="checkbox"/>	<b>Swiss franc</b>		The <b>franc</b> (German: Franken, French and Romansh: <b>franc</b> , Italian: franco; code: CHF) is the currency and legal tender of Switzerland and Liechtenstein;	CHF	swiss national bank	-0.5%
<input checked="" type="checkbox"/>	<b>Indonesian rupiah</b>		The rupiah (Rp) is the official currency of Indonesia. Issued and controlled by the Bank of Indonesia, the ISO 4217 currency code for the <b>Indonesian</b>	rp	bank indonesia	3,65%



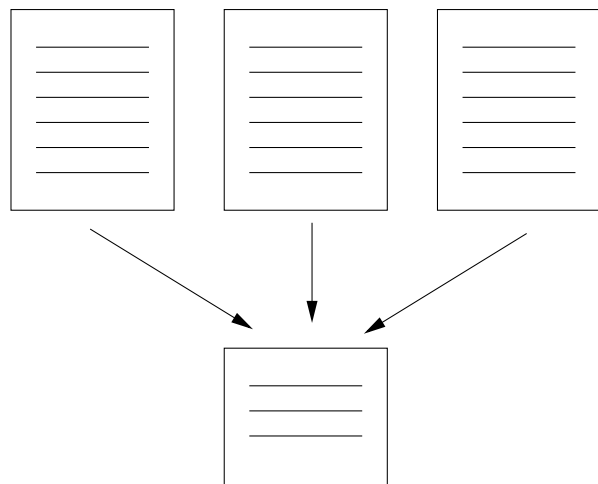
# Summarization



# 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)



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



# Text summarization

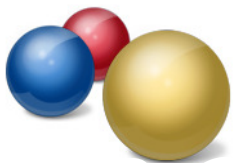


- A **summary** is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s)
- Indicative
  - indicates types of information
  - “alerts”
- Informative
  - includes quantitative/qualitative information
  - “informs”



## INDICATIVE

- 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.



## 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 committees, shared premises, joint publicity referral and the sharing of professional expertise.

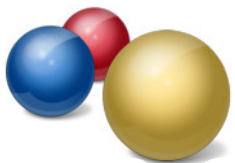




# Text summarization



- 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

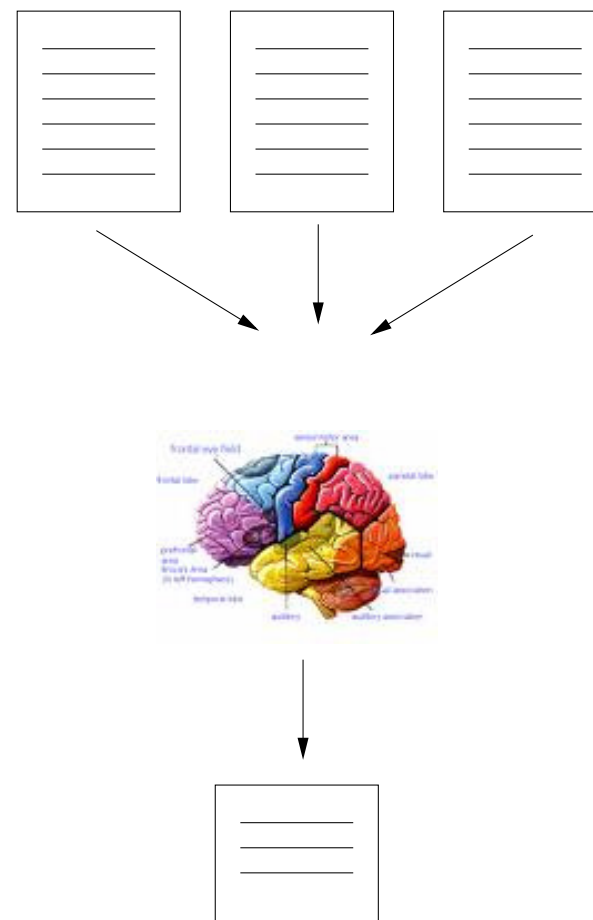


# Extraction vs. abstraction



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



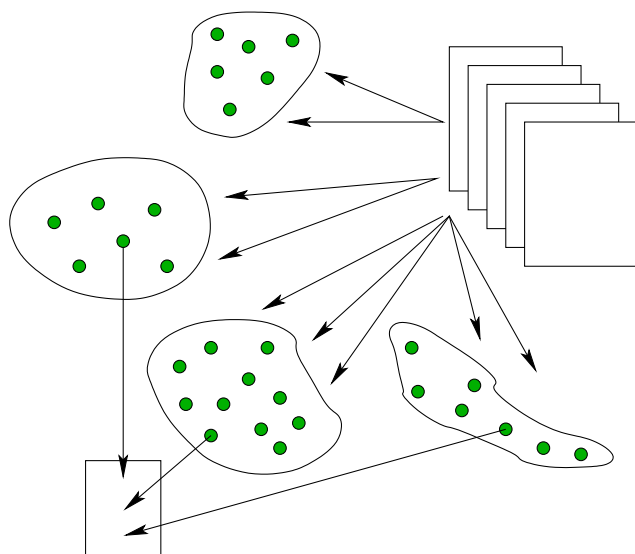
- Unfortunately, a rich semantic representation is not possible yet.
- To date, most summarization systems are extractive.
- Usually, extraction units are sentences.
- Low cost solution: could work without ontologies, complex representations, etc.
- Extractive summaries are usually incoherent.
- Trade-off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



# 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



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.

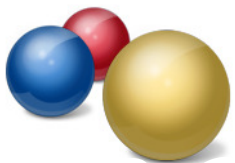


# Sentence compression



- 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.

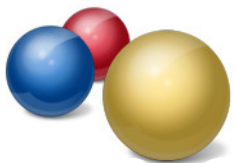


# Sentence compression



- 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.



# 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.





# 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.

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 ban on gun ownership.



# Sentence fusion



- Fusing information from different sentences in a single one:

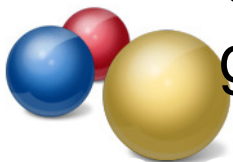
The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

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.



# Sentence fusion



- Fusing information from different sentences in a single one:

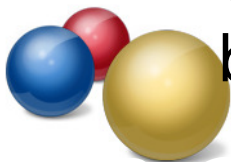
The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

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. Chicago ban on handgun 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.

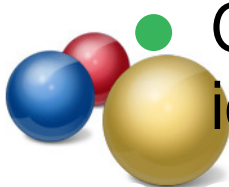
On Monday, the Supreme Court reversed a 28 y. o. Chicago ban on handgun ownership in 5-to-4 decision.



# 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?



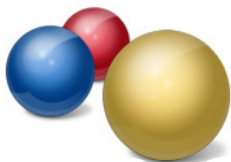


# Multi-sentence compression



- A word graph built from related sentences:
  - vertices = tokens  $\cup \{Start, End\}$
  - 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

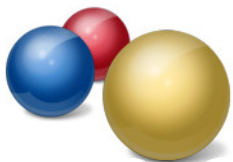
(for more details see Filippova, Coling'10)



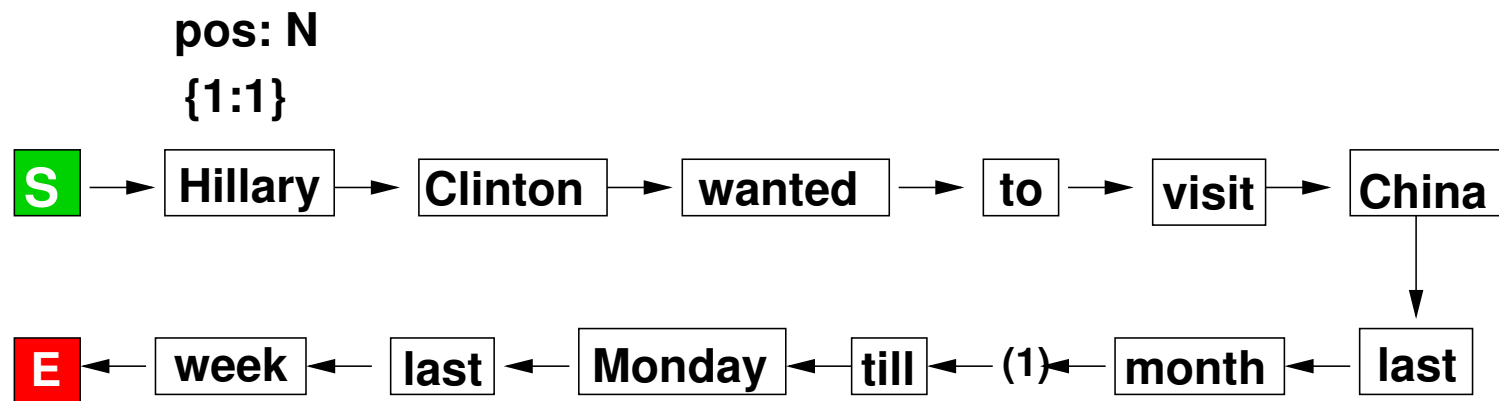
# Word graph



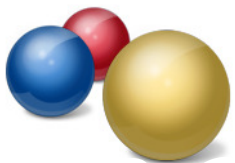
1. Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week.
2. Hillary Clinton paid a visit to the People's Republic of China on Monday.
3. The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday.
4. Last week the Secretary of State Ms. Clinton visited Chinese officials.



# Word graph



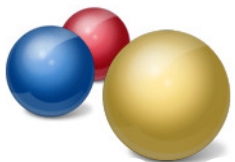
[1: *but postponed her plans*]



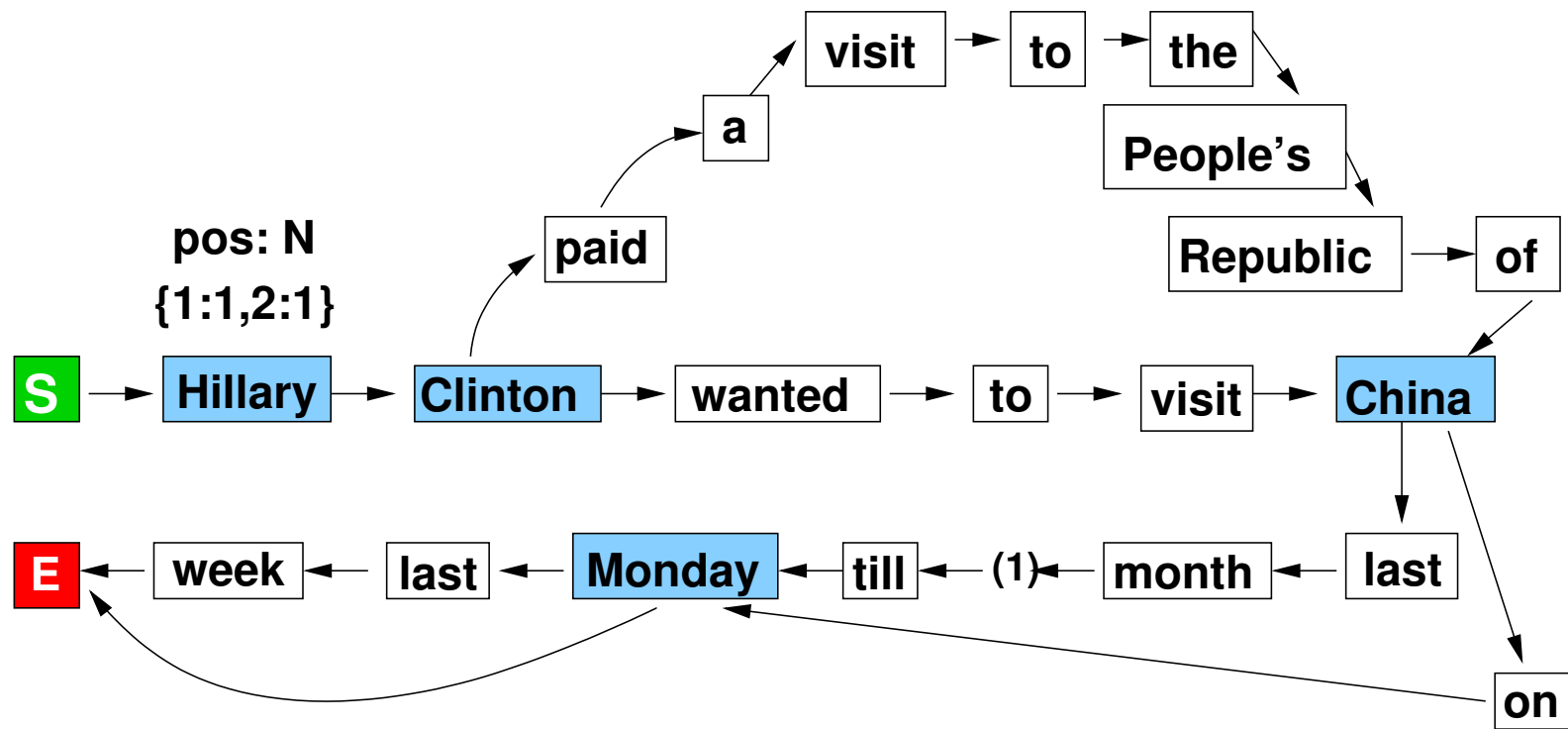
# Word graph



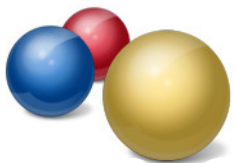
- Words from a new sentence are added in three steps:
  - unambiguous non-stopwords – either merged with an existing node or a new node is created;
  - ambiguous non-stopwords – a node with some overlap in neighbors is preferred;
  - stopwords – only added if the following word matches the out-neighbor of the node.
- The graph permits loops.
- Words from the same sentence are never merged in one node.



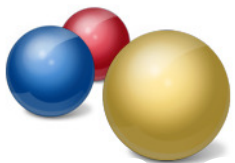
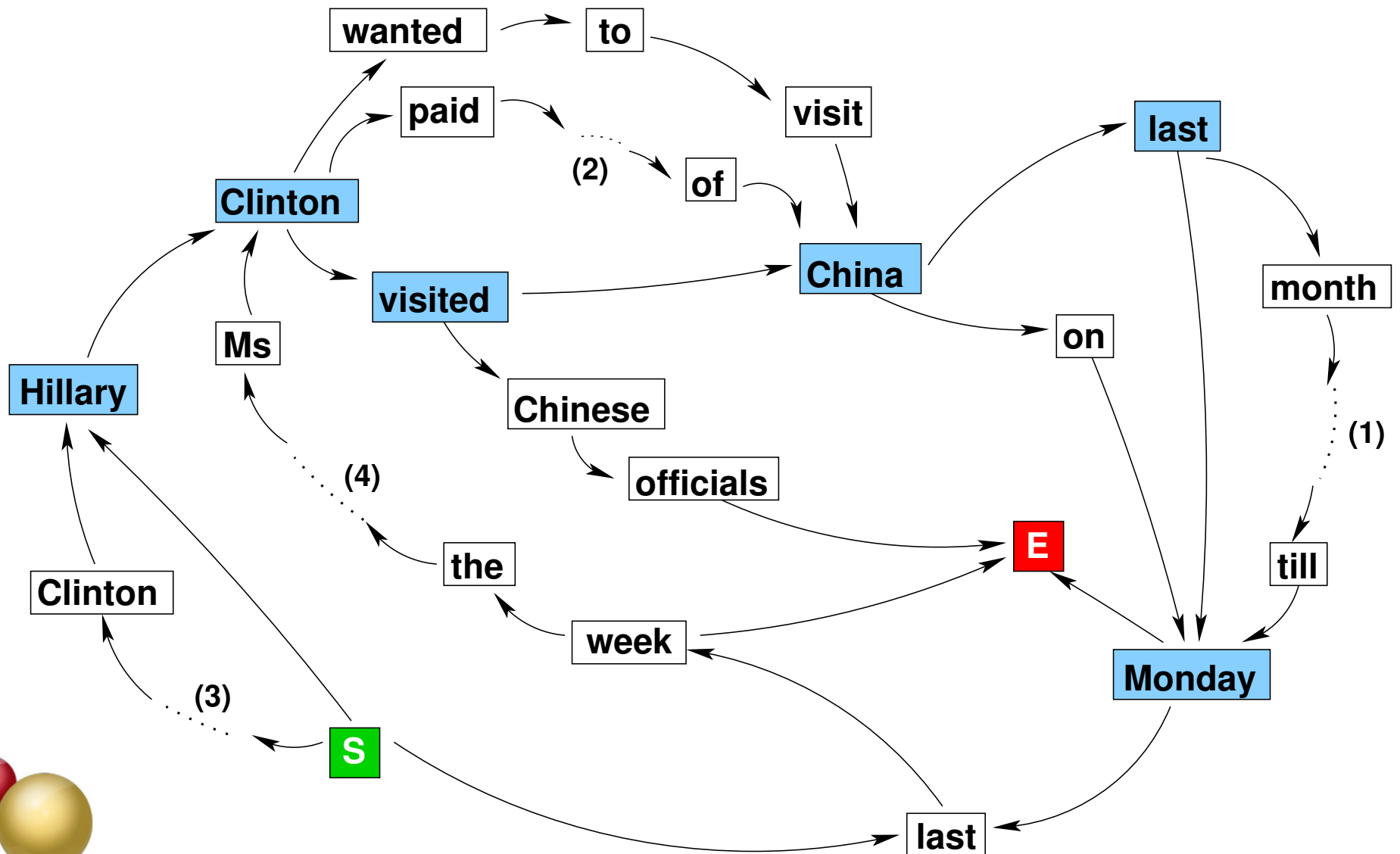
# Word graph



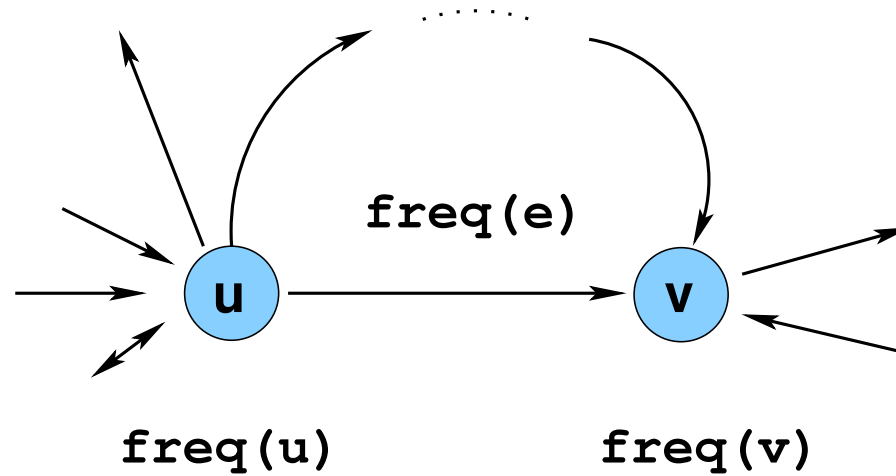
[1: *but postponed her plans*]



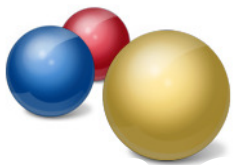
# Word graph



# K shortest paths

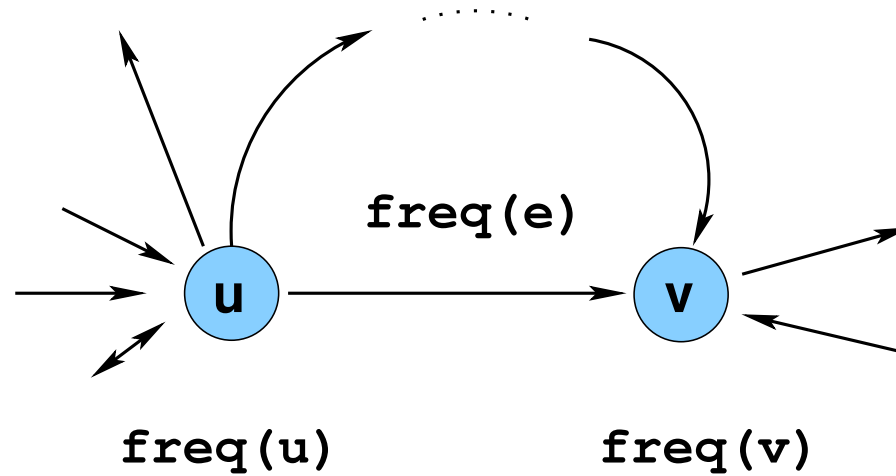


- $w(e) = \frac{1}{\text{freq}(e)}$

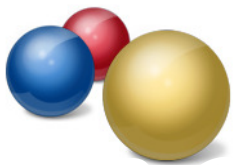




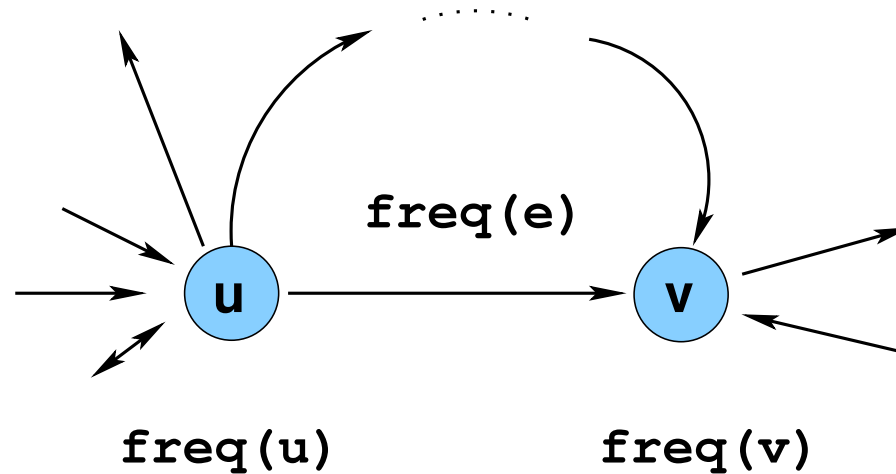
# K shortest paths



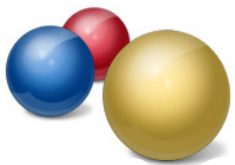
- $w(e) = \frac{1}{\sum_{s \in S} \text{distance}(s, u, v)^{-1}}$



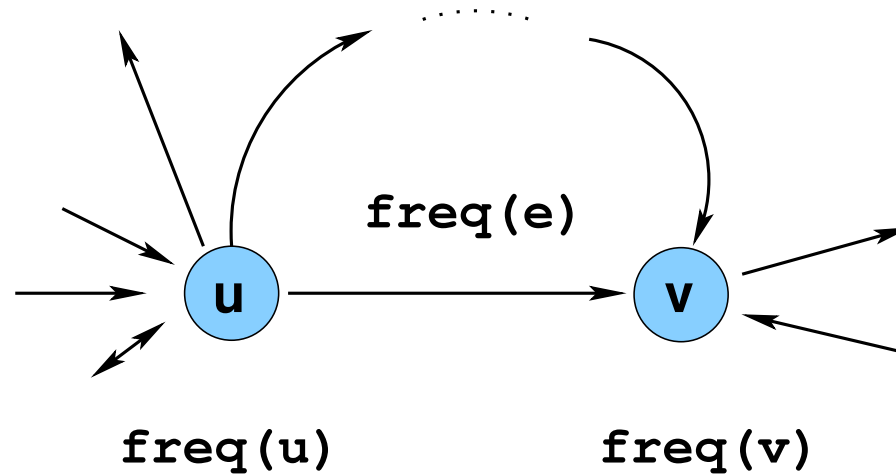
# K shortest paths



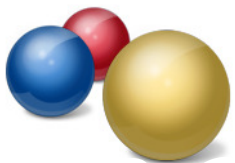
- $w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\sum_{s \in S} \text{distance}(s, u, v)^{-1}}$



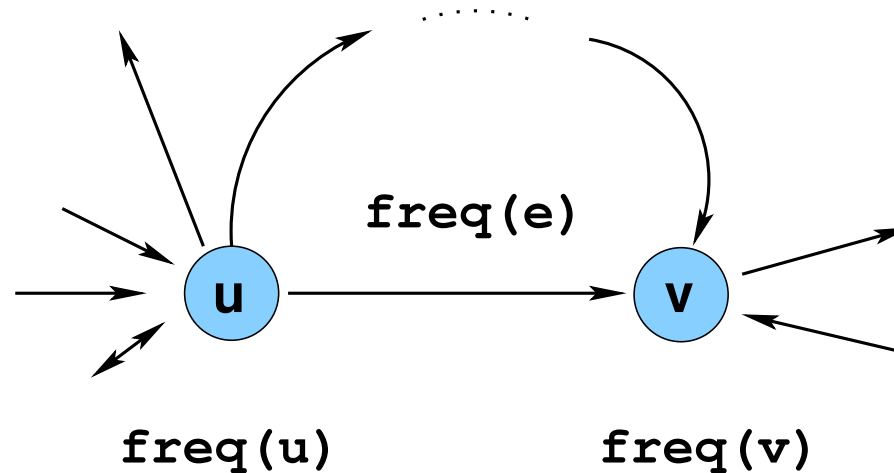
# K shortest paths



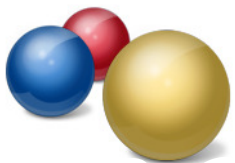
- $w(e) = \frac{\text{freq}(u) + \text{freq}(v)}{\text{freq}(u) \times \text{freq}(v) \times \sum_{s \in S} \text{distance}(s, u, v)^{-1}}$



# 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

Search News

[Advanced news search](#)

## Top Stories

- Gun politics
- Justine Henin
- Vince Neil
- Chris Brown
- Robert Byrd
- Apple
- Elena Kagan
- Tropical storm
- Alex
- iPhone
- Diabetes mellitus

## Starred

- Zurich, Switzerland
- World
- U.S.
- Business
- Sci/Tech
- Entertainment
- Sports
- Health
- Spotlight
- Most Popular

All news  
[Headlines](#)  
[Images](#)

## 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](#) »



Telegraph.c...

[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

[The Atlantic - Christian Science Monitor - Voice of America -](#)

[Wheeling News Register - Wikipedia: Robert Byrd](#)

[all 2,998 news articles](#) »



Globe and M...

[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 former dean of Harvard Law School, would be the fourth woman



New York Da...



## [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. Ooppel](#), [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 ...



# Data: Google News



- 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

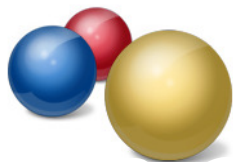




# Evaluation



- 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. Ooppel](#), [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



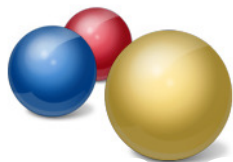
- Is there a main event in the cluster? (yes/no)
- Is the compression grammatical?
  - perfect (2)
  - minor mistake (1)
  - otherwise (0)
- Does it summarize the main event, if present?
  - summarizes the main event (2)
  - related to the main event but misses smth important (1)
  - otherwise (0)



# Results



<b>System</b>	<b>Gram-2</b>	<b>Gram-1</b>	<b>Gram-0</b>	<b>Avg. Len.</b>
Baseline (EN)	21%	15%	65%	8 / 28
Shortest path (EN)	52%	16%	32%	10 / 28
Shortest path++ (EN)	64%	13%	23%	12 / 28
Baseline (ES)	12%	15%	74%	8 / 35
Shortest path (ES)	58%	21%	21%	10 / 35
Shortest path++ (ES)	50%	21%	29%	12 / 35



# Results



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



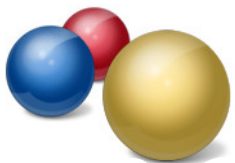
- Sentence compression in the context of MDS – multi-sentence compression.
- Experiments with English, French, Italian, Spanish, German and Russian.
- Evaluation on English and Spanish.
- A simple, syntax-lean method with surprisingly good results.



# 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, ccccoooollll!!!*
  - *i luv tihs vid coz its sooo coooolll!*





# YouTube comments



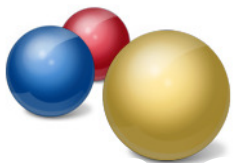
- Why bother about user comments?
  - The title and description can be uninformative (e.g., *IMG\_2947219.avi*).
  - Many videos do not have tags.
  - Description tells us about the video from the uploader's perspective.
  - What do the viewers think about the video?



# YouTube comments



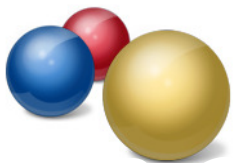
- Why bother about user comments?
  - The title and description can be uninformative (e.g., *IMG\_2947219.avi*).
  - Many videos do not have tags.
  - Description tells us about the video from the uploader's perspective.
  - What do the viewers think about the video?
- Comments are **very** different from news:
  - spelling errors
  - poor grammar
  - lots of meaningless noise



# YouTube comment cloud



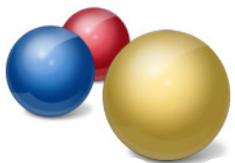
- Task: select most salient, representative words from the comments on a video.
- Simple approach: tokenize comments, count word frequencies.
- Most frequent words are not representative of the video: *lol*, *cool*, *the*
- Filter (YouTube-specific) stopwords.



# YouTube comment cloud



- 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

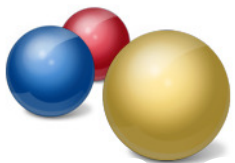


- Most frequent words:

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



# YouTube comment cloud



## Steve Jobs unveils the Apple iPad

**networkworld** 567 videos



**networkworld** | January 27, 2010

Steve Jobs unveils Apple's iPad tablet computer at an event in San Francisco.

**Like** **Comment**

**2,935**   
views



# YouTube comment cloud

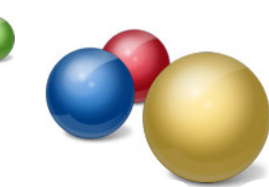


## Words most used in comments on this video

sexy touch wasnt yeaaaaa imagine multi-tasking whistling

screen sposed mid products macbook nor removable-

replaceable inbetween ly ipod sentence metre picture





# YouTube comment cloud



## Nokia N97 Demo

**NokiaConversations**

479 videos

Subscribe

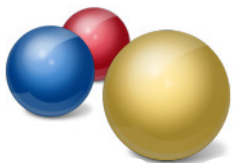
Nokia N97

0:06 / 1:11 360p

**NokiaConversations** | December 02, 2008

4,822,762 views

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 [the NokiaConversations Nokia N97 \(part 1\)](#)

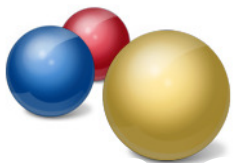


# YouTube comment cloud



## 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?

