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" $\rightarrow$ 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
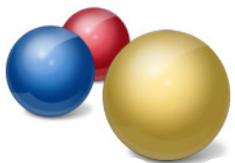 - Natural language tasks - machine translation, speech recognition
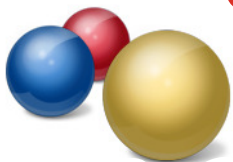
Google's Mission

"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

Overview

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

- Task in focus: summarization
  - News summarization im 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
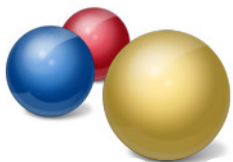
Machine translation @ Google

Google

Google translate

Translation

Translated Search

Translator Toolkit

Tools and Resources

Search websites in other languages

| Snowshoeing in switzerland | [Translate and Search] |

Search pages written in: [German ▾]   Translated to: **Schneeschuhwandern in der Schweiz** - Not quite right? Edit

My language: [English ▾]

Translated results from German web pages   Results **1 - 10** of about **118,000** for **Schneeschuhwandern in der Schweiz**.

English translation   **Original German** - Hide German results

Snowshoeing - a special kind of winter fun...   **Schneeschuhwandern** - ein Wintervergnügen der besonderen Art ...
**Snowshoeing**, the re-discovered method to kräftesparend in...Search for all of   **Schneeschuhwandern**, die wieder entdeckte Methode, sich kräftesparend im ... Suche
Switzerland. Aletsch: Aletsch Forest Amden: **Snowshoe** tour with huskies...   ganze **Schweiz** für. Aletsch: Aletschwald Amden: Schneeschuhtour mit Huskys ...
www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee... - 36k - Cached   www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/schnee... - 36k - Im
                                                                             Cache

The Jura Vaudois by snowshoe Explore - Switzerland Tourism   **Den Waadtländer Jura per Schneeschuh entdecken - Schweiz Tourismus**
23 km of marked circuit between Sainte-Croix, Les and Le Rasses Chasseron. At the   23 km markierter Rundkurs zwischen Sainte-Croix, Les Rasses und Le Chasseron. Zu
beginning of a trail of 10 km from the Col des Etroits between the Mont-des-Cerfs...   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 - Cached   www.myswitzerland.com/de/home/winterferien/schneeschuhlaufen/den-wa... - 37k - Im
                                                                             Cache
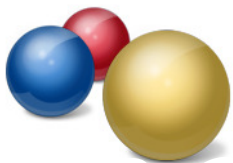
Machine translation @ Google

Machine translation @ Google
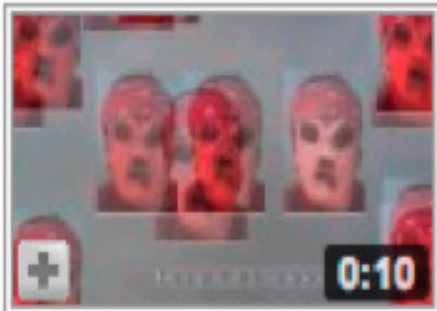
View original **Hello everyone**

Hello everyone Sweasy26, www.sweasy26.com

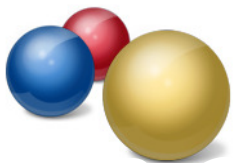no rating   2 months ago   16 views   Sweasy26
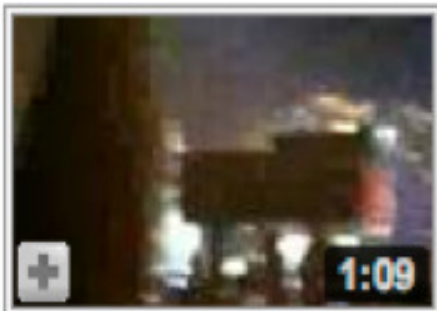
Translate **bonjour tout le monde**

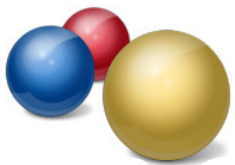devenu très vite une tradition :D

no rating   2 years ago   107 views   marine93340

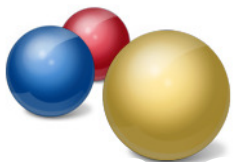# Machine translation @ Google

Machine translation @ Google

Google

Machine translation @ Google

Machine translation @ Google

Google

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

Google translator toolkit

All translations
- 📄 **Active**
- 📄 Hidden
- 🗑 Trash
⊞ Shared with…
All translation tools
- 🛢 Translation memories
- 📘 Glossaries

| Upload | Download | Hide | Delete | Share |

Select: All, None, Completed, Incomplete

Name	Words	Date↓	Sharing

Machine translation tools

Google

☆ 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

Google

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.
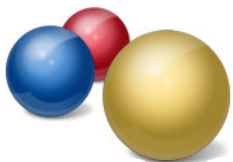
Table of Contents
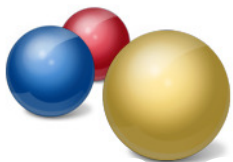
Speech @ Google

- 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 Eropean languages (each trained on 100B tokens)
  - Chinese (5-gram, 883B tokens)

Information extraction

Information extraction

Information extraction

Information extraction

Information extraction

Information extraction

Information extraction

Information extraction

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

Google

Google squared labs

currencies [Square it] [Add to this Square]

Google Squared

Unsaved &Share [Export ▼] [Save]
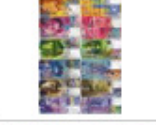
currencies   20 item

	Item Name ▼	Image ✕	Description ✕	Symbol ▼✕	Central Bank ▼✕	Inflation ▼
✕	**Euro**		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	EC	European Central Bank	0,9%, February 2010
✕	**Japanese yen**		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**, ...	JPY	bank of japan	−1.3%
✕	**Australian dollar**		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 **...**	AUD	reserve bank of australia	2.1%
✕	**Canadian dollar**		**Canadian Dollar** Currency Exchange Forecast. **Canadian Dollar** to US Dollar Currency Exchange Forecast. **Canadian Dollars** to One U.S. Dollar. Average of	$ or C$	Bank of Canada	1%
✕	**Swiss franc**		The **franc** (German: Franken, French and Romansh: **franc**, Italian: franco; code: CHF) is the currency and legal tender of Switzerland and Liechtenstein;	CHF	swiss national bank	-0.5%
✕	**Indonesian rupiah**		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**	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)
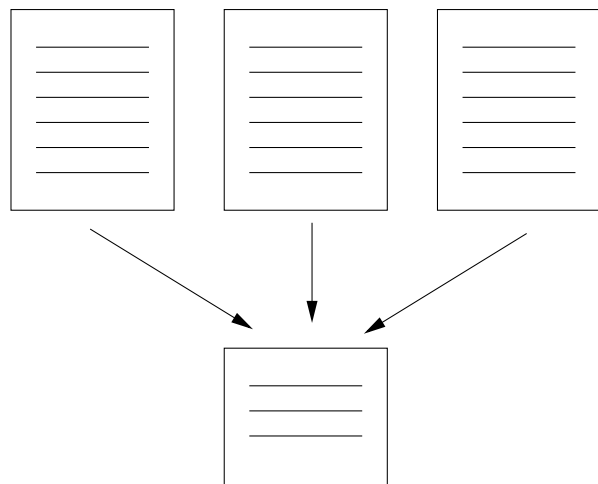
- information retrieval

- 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"

Text summarization

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.

Text summarization
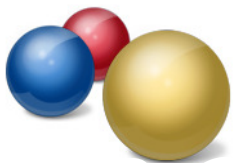
INFORMATIVE

- 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
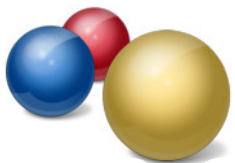
Text summarization

- 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
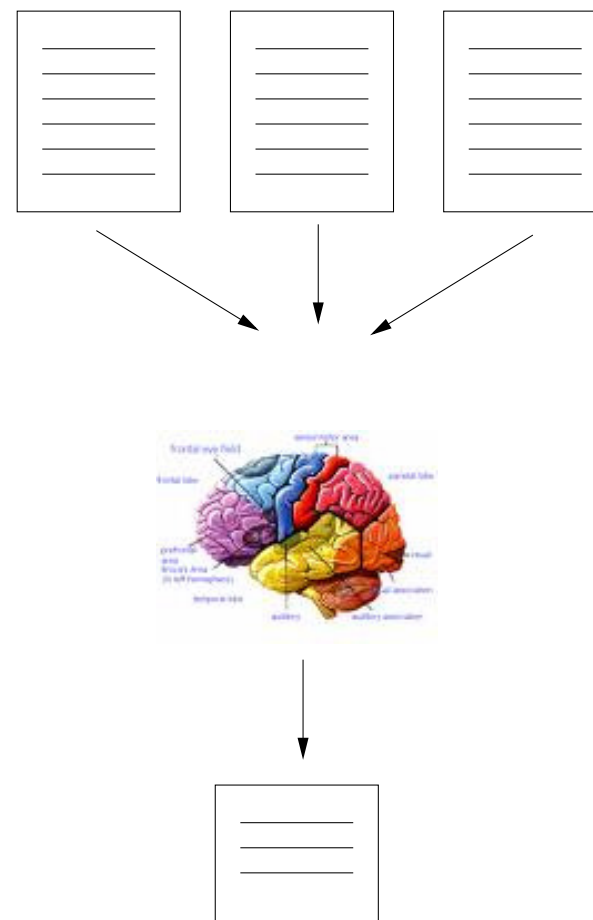
Extraction vs. abstraction

- 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:
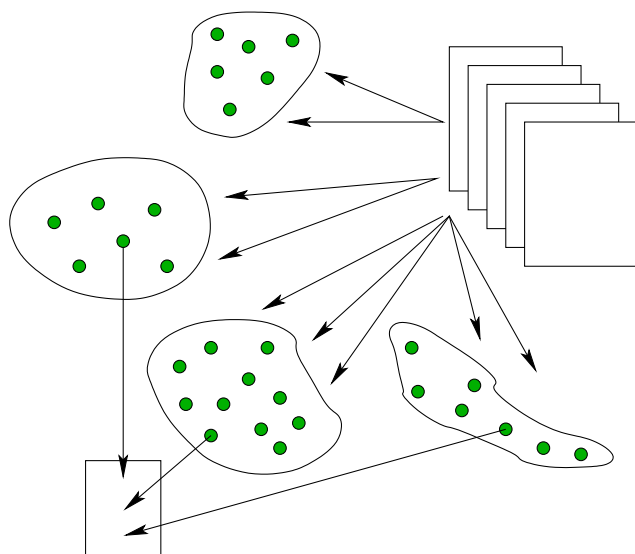
  - similar sentences are grouped into clusters

  - 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.

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.
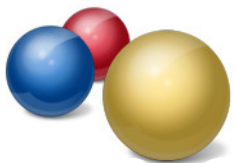
Sentence compression

- 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.

- 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.

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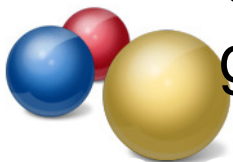
The US's highest court ruled by 5-4 that a ban on handgun ownership in Chicago was unconstitutional.

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?
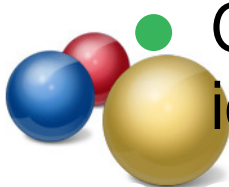
Multi-sentence compression

- 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)
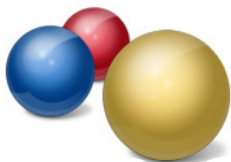
Word graph
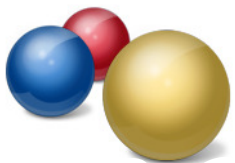
1. Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week.

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

pos: N
{1:1}

S → Hillary → Clinton → wanted → to → visit → China

China ↓

E ← week ← last ← Monday ← till ← (1) ← month ← last
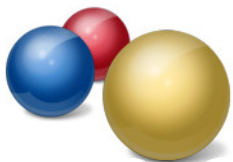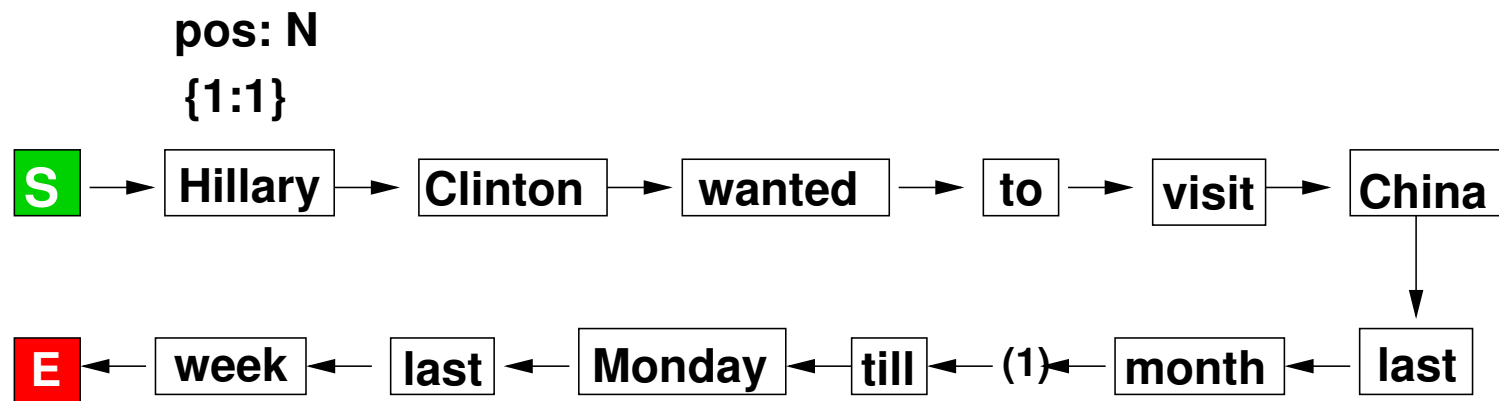
[1: *but postponed her plans*]

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

pos: N
{1:1,2:1}

S → Hillary → Clinton → wanted → to → visit → China

Clinton → paid → a → visit → to → the → People's → Republic → of → China

China → last → month → (1) → till → Monday → last → week → E

China → on → Monday
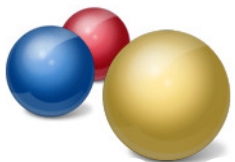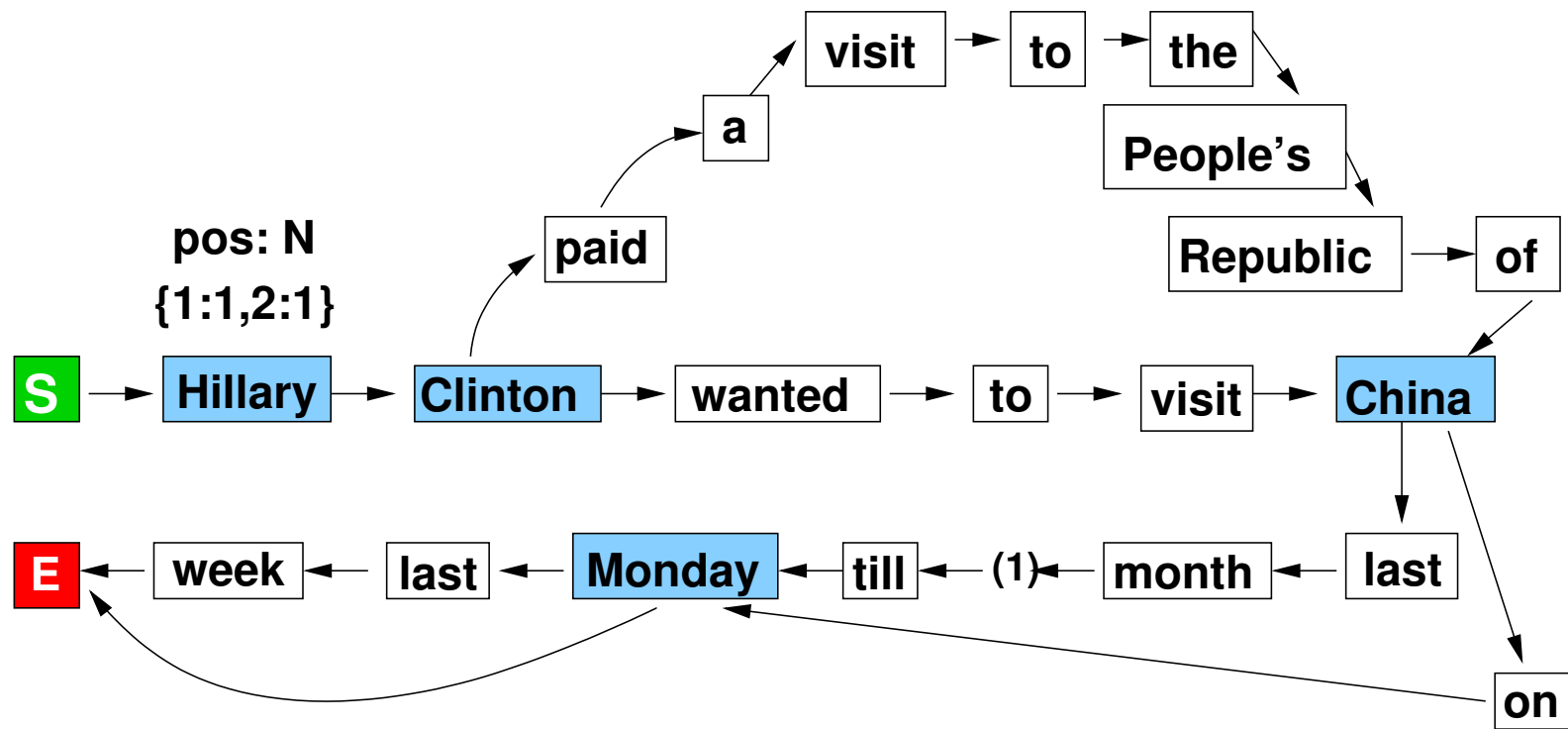
[1: *but postponed her plans*]

Word graph
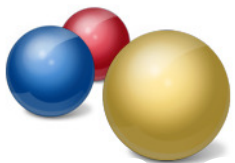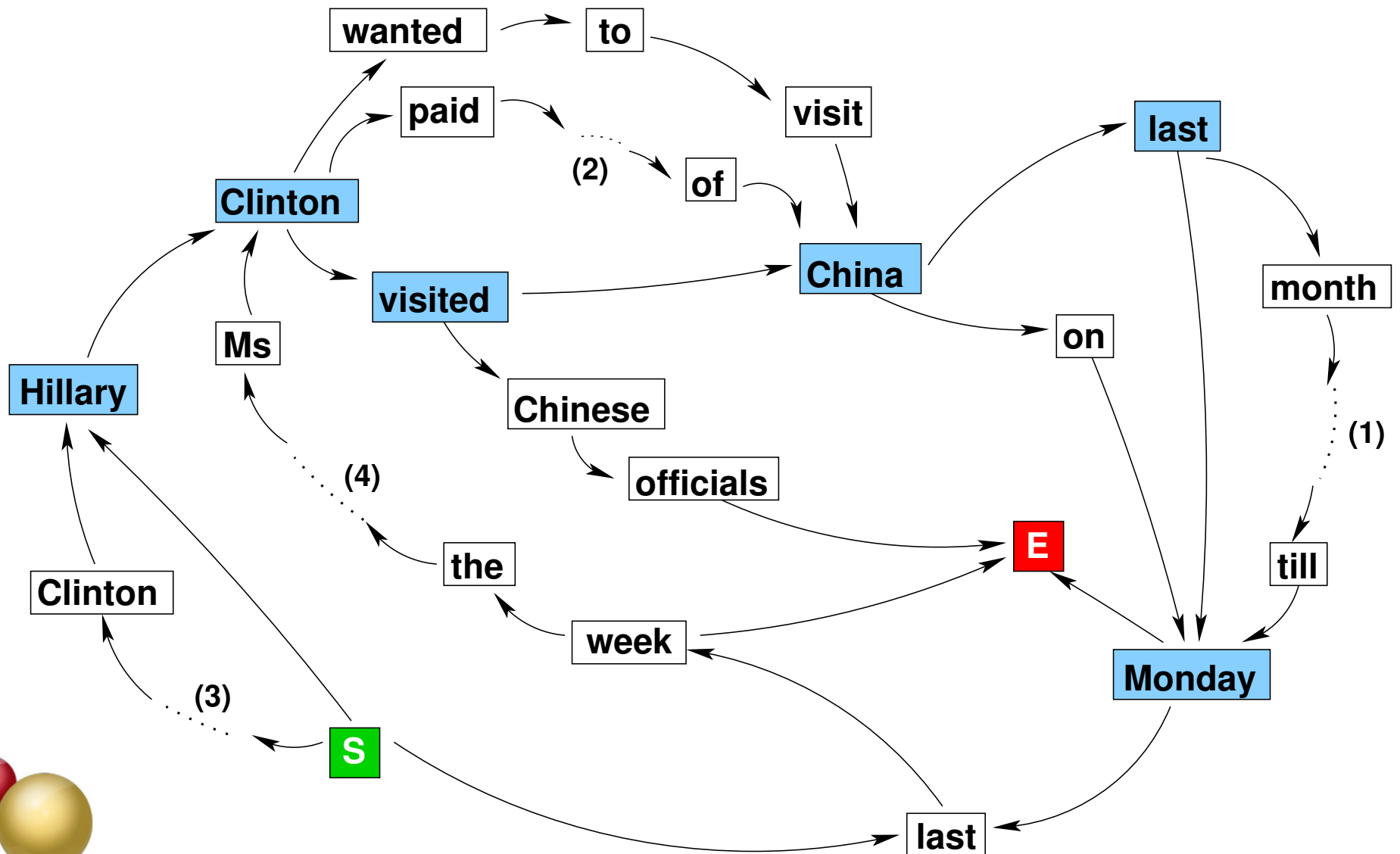
K shortest paths

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

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

K shortest paths

Google

freq(e)

u   **v**
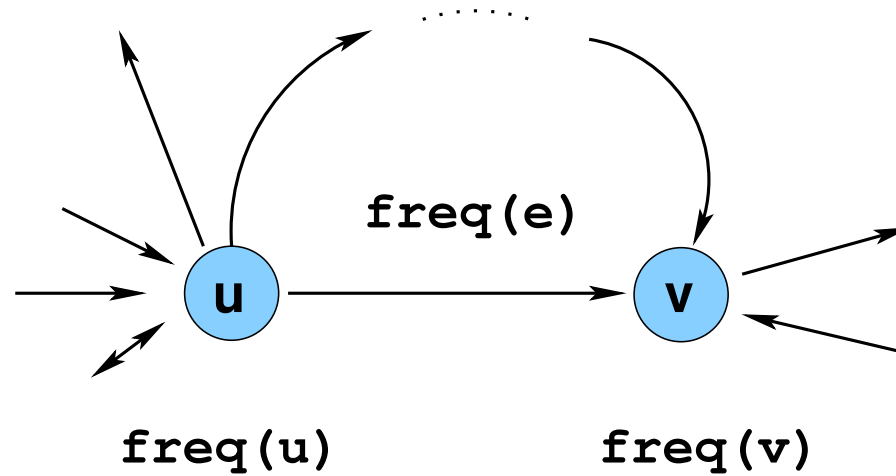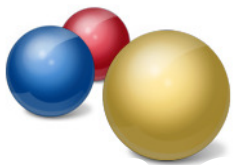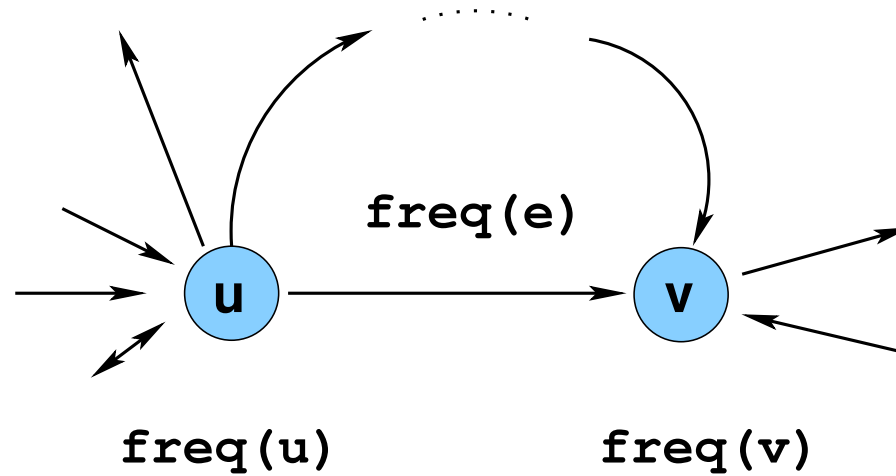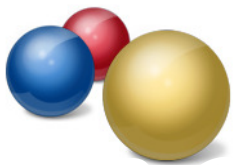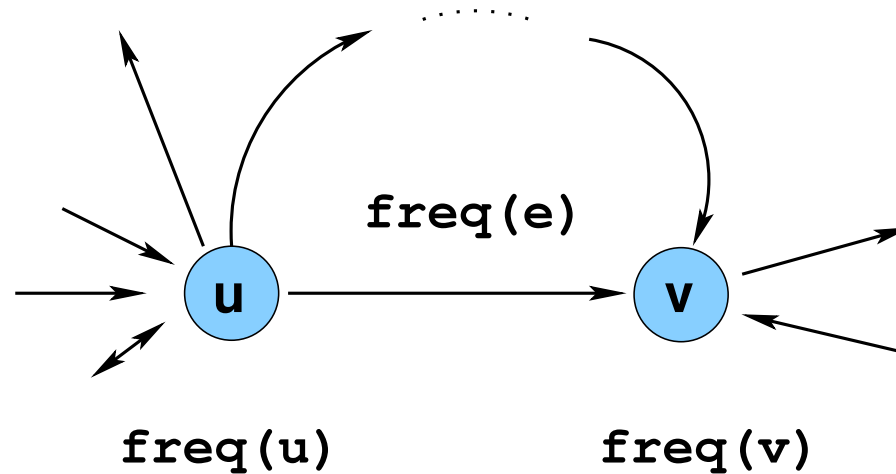
freq(u)   **freq(v)**

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

K shortest paths

freq(e)

u → **v**
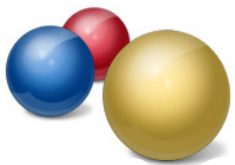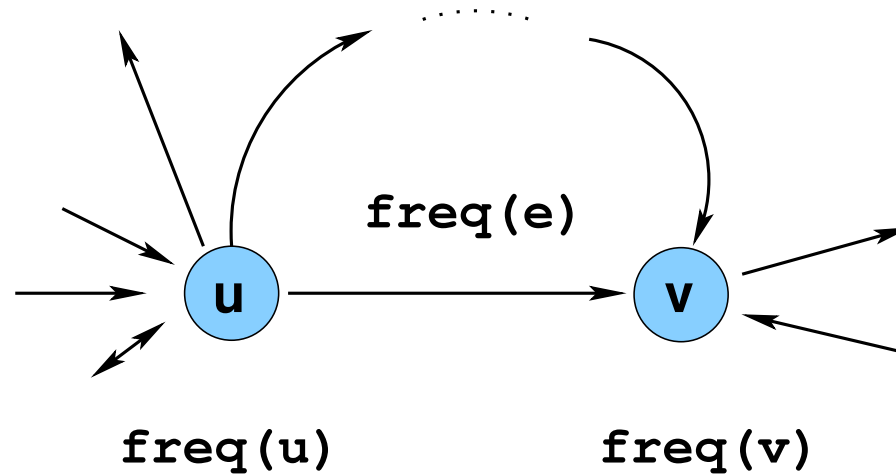
freq(u)   **freq(v)**

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

K shortest paths

$freq(e)$

u   v

$freq(u)$ $freq(v)$
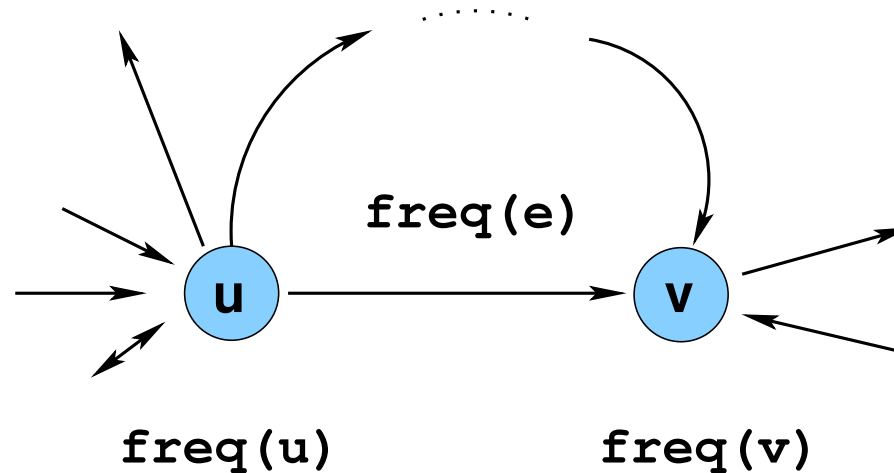
- Paths shorter than eight edges are discarded.
- Paths not passing a verb are filtered out.
- The total path length is normalized by the number of edges.

Data: Google News

Data: Google News

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
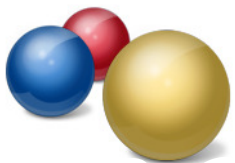
KABUL, Afghanistan A Taliban suicide squad stormed the compound of a US-based development group in northern Afghanistan on Friday, killing at least three expatriate workers, a security guard and an Afghan police officer, officials said. **...**

" "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 **...**

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
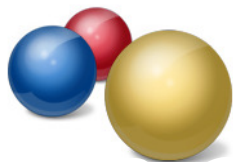 - considerably longer than other sentences

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. 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

System	Gram-2	Gram-1	Gram-0	Avg. Len.
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
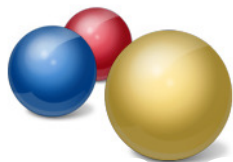
Results

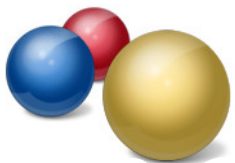| System	Info-2	Info-1	Info-0	Avg. Len.
Baseline (EN)	18%	10%	73%	8 / 28
Shortest path (EN)	36%	33%	31%	10 / 28
Shortest path++ (EN)	52%	32%	16%	12 / 28
Baseline (ES)	9%	19%	72%	8 / 35
Shortest path (ES)	23%	26%	51%	10 / 35
Shortest path++ (ES)	40%	40%	20%	12 / 35

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 surprizingly 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 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

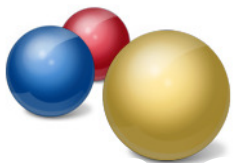- Why bother about user comments?
  - The title and description can be uninformative (e.g., *IMG_2947219.avi*).
  - Many videos do not have tags.
  - Description tells us about the video from the uploader's perspective.
  - What do the viewers think about the video?

- Comments are **very** different from news:
  - spelling errors
  - poor grammar
  - lots of meaningless noise
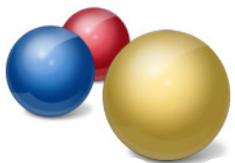
YouTube comment cloud

- 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.
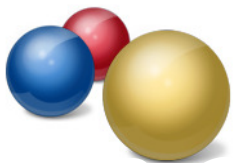
YouTube comment cloud

- 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
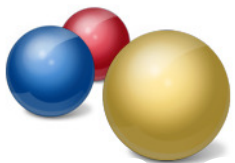
YouTube comment cloud

- 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

YouTube comment cloud

YouTube comment cloud

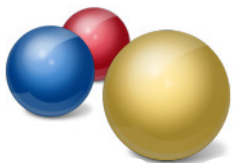**Nokia N97 Demo**

NokiaConversations | 479 videos | Subscribe

Nokia N97

0:06 / 1:11 | 360p

4,822,762 views

NokiaConversations | December 02, 2008

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
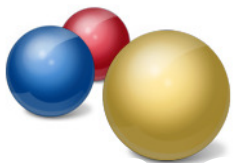
YouTube comment cloud

Aug 26, 2010   Sep 8, 2010

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?