COMP4650 / COMP6490 Document Analysis 2018

Information Extraction

Gabriela Ferraro

DATA 61

CSIRO

Australian National University

Overview of IE lectures

- Introduction to Information Extraction (IE)
- Sequence labeling methods 1
- Sequence labeling methods 2
- Automatic Summarization

* Acknowledgement: Some of the content originates from the Stanford NLP course at Coursera.org

What is a summary?

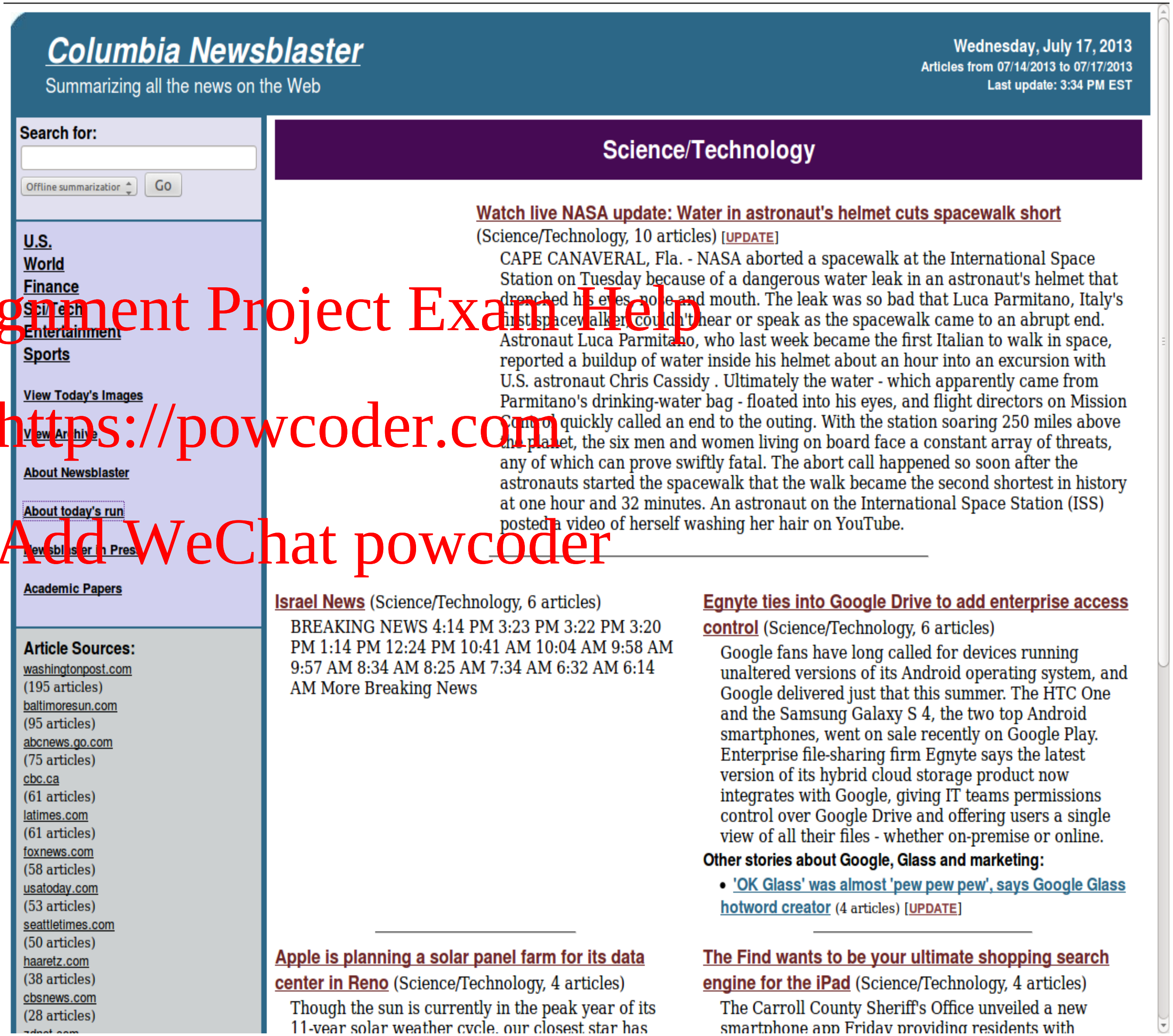
Is a brief statement of the main points of something, usually a text (Oxford Dictionary).

Automatic summarization?

Is an brief statement about the main points of something generated by an algorithm.

Automatic summarization is a classical Natural Language Processing problem with more than 60 years of history and still a HOT topic!

News summaries



Columbia Newsblaster

Summarizing all the news on the Web

Wednesday, July 17, 2013
Articles from 07/14/2013 to 07/17/2013
Last update: 3:34 PM EST

Article Sources:
washingtonpost.com (195 articles)
baltimoresun.com (95 articles)
abcnews.go.com (75 articles)
cbc.ca (61 articles)
latimes.com (61 articles)
foxnews.com (58 articles)
usatoday.com (53 articles)
seattletimes.com (50 articles)
haaretz.com (38 articles)
cbsnews.com (28 articles)
zdnet.com

Science/Technology

Watch live NASA update: Water in astronaut's helmet cuts spacewalk short (Science/Technology, 10 articles) [UPDATE]

CAPE CANAVERAL, Fla. - NASA aborted a spacewalk at the International Space Station on Tuesday because of a dangerous water leak in an astronaut's helmet that drenched his eyes, nose and mouth. The leak was so bad that Luca Parmitano, Italy's first spacewalker, couldn't hear or speak as the spacewalk came to an abrupt end. Astronaut Luca Parmitano, who last week became the first Italian to walk in space, reported a buildup of water inside his helmet about an hour into an excursion with U.S. astronaut Chris Cassidy . Ultimately the water - which apparently came from Parmitano's drinking-water bag - floated into his eyes, and flight directors on Mission Control quickly called an end to the outing. With the station soaring 250 miles above the planet, the six men and women living on board face a constant array of threats, any of which can prove swiftly fatal. The abort call happened so soon after the astronauts started the spacewalk that the walk became the second shortest in history at one hour and 32 minutes. An astronaut on the International Space Station (ISS) posted a video of herself washing her hair on YouTube.

Israel News (Science/Technology, 6 articles)

BREAKING NEWS 4:14 PM 3:23 PM 3:22 PM 3:20 PM 1:14 PM 12:24 PM 10:41 AM 10:04 AM 9:58 AM 9:57 AM 8:34 AM 8:25 AM 7:34 AM 6:32 AM 6:14 AM More Breaking News

Egnyte ties into Google Drive to add enterprise access control (Science/Technology, 6 articles)

Google fans have long called for devices running unaltered versions of its Android operating system, and Google delivered just that this summer. The HTC One and the Samsung Galaxy S 4, the two top Android smartphones, went on sale recently on Google Play. Enterprise file-sharing firm Egnyte says the latest version of its hybrid cloud storage product now integrates with Google, giving IT teams permissions control over Google Drive and offering users a single view of all their files - whether on-premise or online.

Other stories about Google, Glass and marketing:

- 'OK Glass' was almost 'pew pew pew', says Google Glass hotword creator (4 articles) [UPDATE]

Apple is planning a solar panel farm for its data center in Reno (Science/Technology, 4 articles)

Though the sun is currently in the peak year of its 11-year solar weather cycle, our closest star has

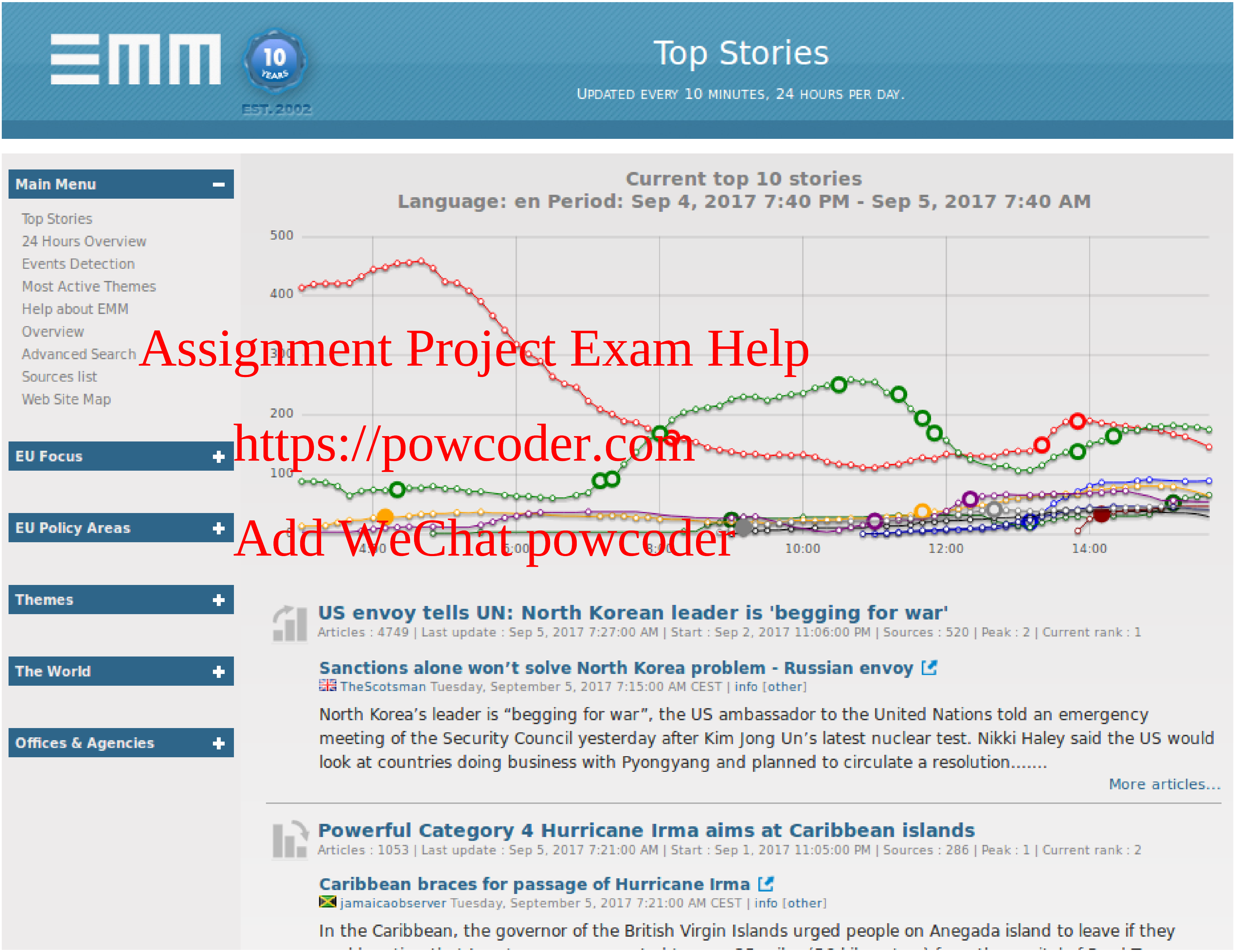
The Find wants to be your ultimate shopping search engine for the iPad (Science/Technology, 4 articles)

The Carroll County Sheriff's Office unveiled a new smartphone app Friday providing residents with

Multiple sources

Multi-modal (text, tables, map graphics, etc.)

Selection and placement of stories are determined automatically



Top Stories

Updated every 10 minutes, 24 hours per day.

Current top 10 stories
Language: en Period: Sep 4, 2017 7:40 PM - Sep 5, 2017 7:40 AM

US envoy tells UN: North Korean leader is 'begging for war'
Articles : 4749 | Last update : Sep 5, 2017 7:27:00 AM | Start : Sep 2, 2017 11:06:00 PM | Sources : 520 | Peak : 2 | Current rank : 1

Sanctions alone won't solve North Korea problem - Russian envoy
TheScotsman Tuesday, September 5, 2017 7:15:00 AM CEST | info [other]

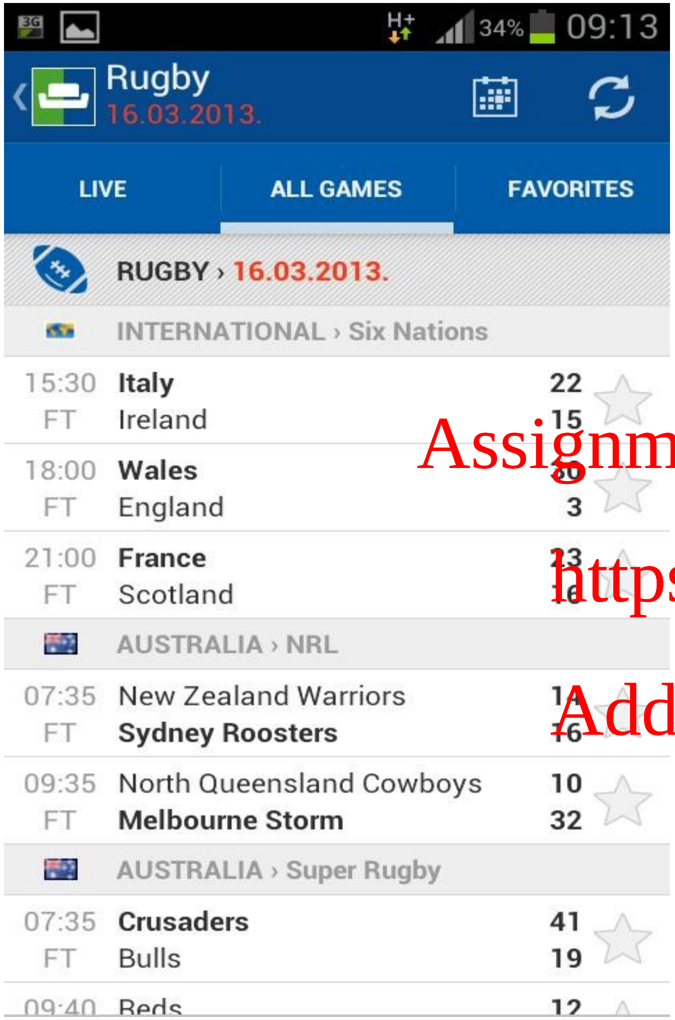
North Korea's leader is "begging for war", the US ambassador to the United Nations told an emergency meeting of the Security Council yesterday after Kim Jong Un's latest nuclear test. Nikki Haley said the US would look at countries doing business with Pyongyang and planned to circulate a resolution.......

Powerful Category 4 Hurricane Irma aims at Caribbean islands
Articles : 1053 | Last update : Sep 5, 2017 7:21:00 AM | Start : Sep 1, 2017 11:05:00 PM | Sources : 286 | Peak : 1 | Current rank : 2

Caribbean braces for passage of Hurricane Irma
jamaicaobserver Tuesday, September 5, 2017 7:21:00 AM CEST | info [other]

In the Caribbean, the governor of the British Virgin Islands urged people on Anegada island to leave if they

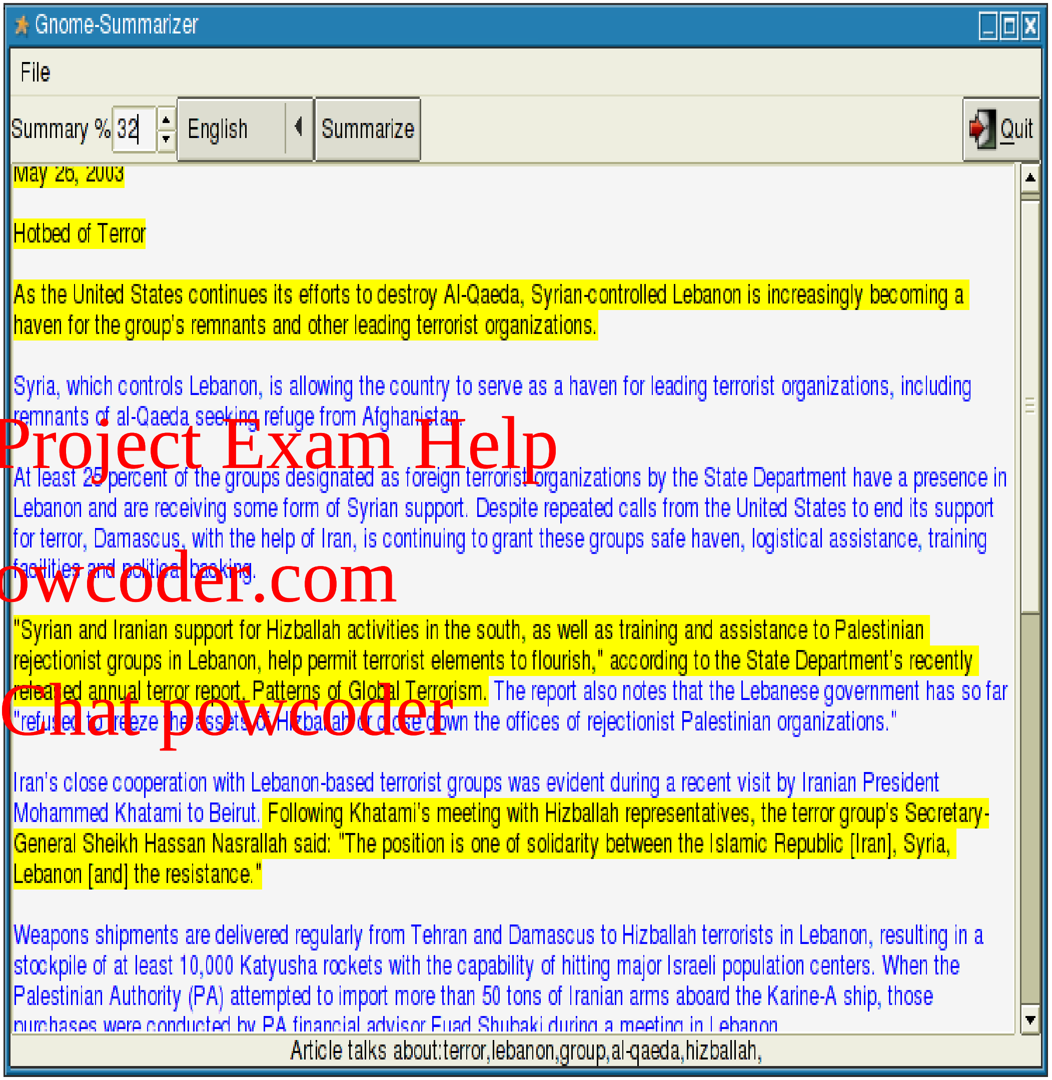
-



Rugby
16.03.2013.

RUGBY › 16.03.2013.

INTERNATIONAL › Six Nations		
15:30 FT	Italy Ireland	22 15
18:00 FT	Wales England	30 3
21:00 FT	France Scotland	23 16
AUSTRALIA › NRL		
07:35 FT	New Zealand Warriors Sydney Roosters	14 16
09:35 FT	North Queensland Cowboys Melbourne Storm	10 32
AUSTRALIA › Super Rugby		
07:35 FT	Crusaders Bulls	41 19
09:40	Reds	12



Summary % 32 English

May 26, 2003

Hotbed of Terror

As the United States continues its efforts to destroy Al-Qaeda, Syrian-controlled Lebanon is increasingly becoming a haven for the group's remnants and other leading terrorist organizations.

Syria, which controls Lebanon, is allowing the country to serve as a haven for leading terrorist organizations, including remnants of al-Qaeda seeking refuge from Afghanistan.

At least 25 percent of the groups designated as foreign terrorist organizations by the State Department have a presence in Lebanon and are receiving some form of Syrian support. Despite repeated calls from the United States to end its support for terror, Damascus, with the help of Iran, is continuing to grant these groups safe haven, logistical assistance, training facilities and political backing.

"Syrian and Iranian support for Hizballah activities in the south, as well as training and assistance to Palestinian rejectionist groups in Lebanon, help permit terrorist elements to flourish," according to the State Department's recently released annual terror report, Patterns of Global Terrorism. The report also notes that the Lebanese government has so far "refused to freeze the assets of Hizballah or close down the offices of rejectionist Palestinian organizations."

Iran's close cooperation with Lebanon-based terrorist groups was evident during a recent visit by Iranian President Mohammed Khatami to Beirut. Following Khatami's meeting with Hizballah representatives, the terror group's Secretary-General Sheikh Hassan Nasrallah said: "The position is one of solidarity between the Islamic Republic [Iran], Syria, Lebanon [and] the resistance."

Weapons shipments are delivered regularly from Tehran and Damascus to Hizballah terrorists in Lebanon, resulting in a stockpile of at least 10,000 Katyusha rockets with the capability of hitting major Israeli population centers. When the Palestinian Authority (PA) attempted to import more than 50 tons of Iranian arms aboard the Karine-A ship, those purchases were conducted by PA financial advisor Fuad Shubaki during a meeting in Lebanon

Article talks about:terror,lebanon,group,al-qaeda,hizballah,

Coverpage
Discover new and trending summaries
Saved Summaries
U.S. News
apple
Apple Appeals Tokyo Court's Ruling...
Business
ECB's Knot says getting euro debt ba...
Manage Topics
yahoo
Design, Cool Hunting, Design Week,...
Business
Election 2012
AP Washington, Fox News Politics, LA Times Politics, NBC News Election 20...
Entertainment and Arts

Summary Typology

Single document summary

Multi-document summary

Generic summary

- ➔ contains information about the main topics

Query-focused summary

- ➔ e.g. make a summary about today news that talk about climate change and global warming

Indicative summary

- ✔ e.g. this document is about climate change and global warming

Informative summary

- e.g. global warming has a very serious impact on vulnerable ecosystems

Multi modal summary

- Include tables, maps, graphs, etc.

Multi-lingual summary

- systems capable to summarize in several languages
- **cross-language**: were source and target languages are different

Comparative summarization

- provide short summaries from multiple comparative aspects

Update summarization

- Assumes the user already read some earlier documents on the same topic

Summarizing spoken data or transcripts

Opinion summarization

- Combines summarization and opinion mining

Summarizing **emails, community question answering, movie scripts, entity descriptors in knowledge graphs, source code descriptors,...**

Examples

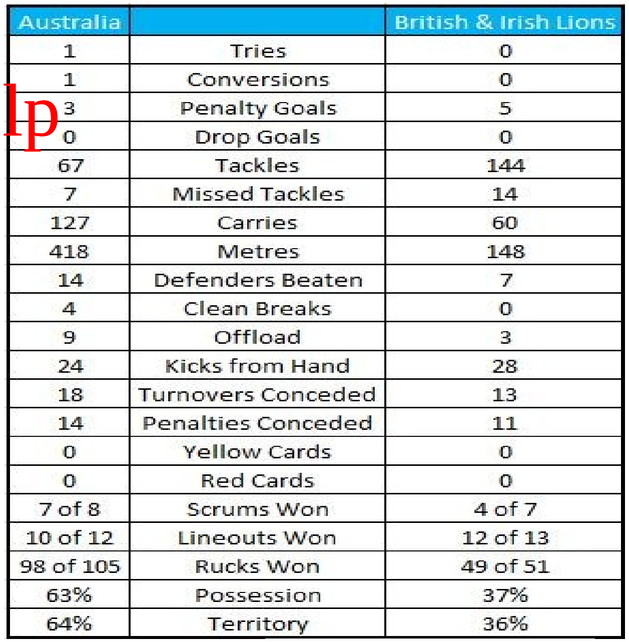
- headlines (from around the world)
- outlines (notes for students)
- minutes (of a meeting)
- previews (of movies)
- synopses (soap opera listings)
- reviews (of a book, CD, movie, etc.)
- digests (TV guide)
- biography (resumes, obituaries)
- abridgments (Shakespeare for children)
- bulletins (weather forecasts/stock market reports)
- sound bites (politicians on a current issue)
- histories (chronologies of salient events)

Summarization Techniques

- **Extractive summarization**
  - Copy the most important information to the summary (e.g.: key phrases, clauses, sentences, paragraphs, etc.)

- **Abstractive summarization**

  Abstractive text summarization involves generating entirely new phrases and sentences to capture the meaning of the source document
  - Involves paraphrasing, aggregation, text simplification and/or text generation
  - Harder to develop



Australia		British & Irish Lions
1	Tries	0
1	Conversions	0
3	Penalty Goals	5
0	Drop Goals	0
67	Tackles	144
7	Missed Tackles	14
127	Carries	60
418	Metres	148
14	Defenders Beaten	7
4	Clean Breaks	0
9	Offload	3
24	Kicks from Hand	28
18	Turnovers Conceded	13
14	Penalties Conceded	11
0	Yellow Cards	0
0	Red Cards	0
7 of 8	Scrums Won	4 of 7
10 of 12	Lineouts Won	12 of 13
98 of 105	Rucks Won	49 of 51
63%	Possession	37%
64%	Territory	36%

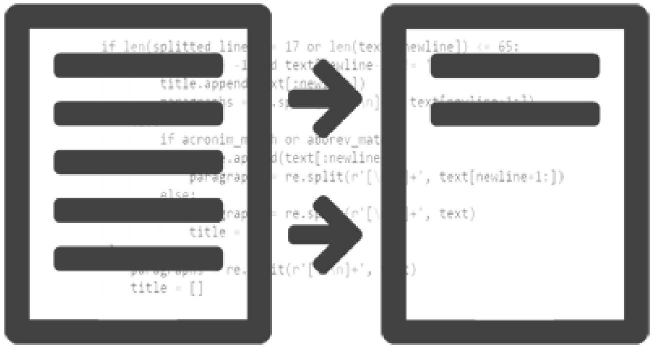
Extractive Summarization

Sentence Extraction Summarization

- Subset of the sentences from the original document

- Sentences that contained the most relevant information

- The extracted sentences are usually ordered as in the original document



Extractive Summarization

- Sentence ranking

- Sentence selection

- Sentence reformulation (in novel methods)

- Sentence ordering

Sentence Extraction Summarization

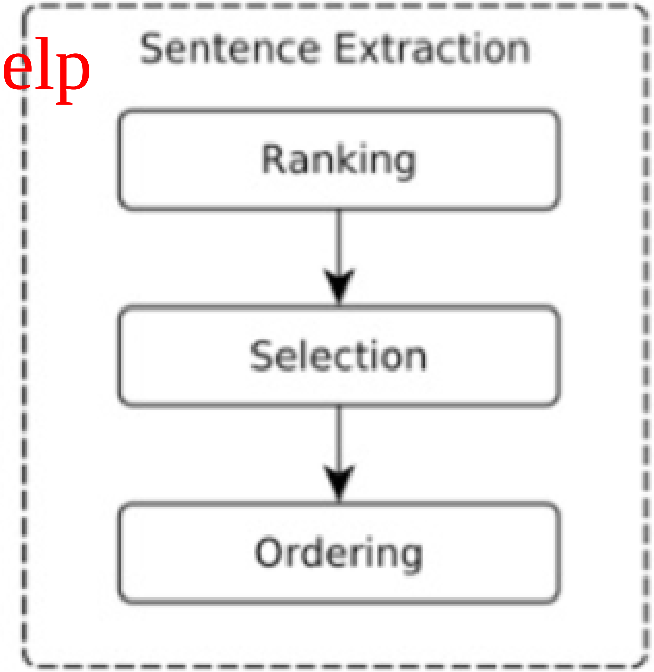
Generic algorithm

- Compression parameter
  - Number of words of the summary, e.g.: 200 words.
  - Desired percentage, e.g. 10% of the original text.

- Create a list of sentences *L*

- Assign to each sentence a score (relevance method)

- Order the sentences according to the score
- While desire compression is false
  - Save the next sentence in *L*
- Show the sentences in L order according their position in the original document



Sentence Extraction

Ranking

Selection

Ordering

What is Relevant?

We need relevance methods to assess which sentences are the most important

Common relevance methods

- Keywords
- Position
- Titles
- Indicative phrases
- Hybrid
- Syntax based
- Discourse based
- As a learning problem (supervised, unsupervised)

Relevance

Early unsupervised approaches rely on two ideas:

- **Frequency**: more important information is more frequently

- **Centrality**: sentences more similar to other sentences are assumed to carry central ideas

Relevance Function

$$\mathbf{R}(\boldsymbol{C}, \boldsymbol{Q}, \phi)$$

C is a document collection (or document stream)

Q is a query or user profile or topic

ϕ ranking threshold (below which the system will not retrieved docs or sentences, e.g.: degree of match)

Relevance Method: Keywords

Hypothesis:

- The repetition of a concept is indicative of its relevance
  - But counting concepts is not easy because the same concepts can be expressed by different words (dog-car, woman-she, etc.)

General steps:

- Apply a stemmer algorithm to normalize the words (`orange = oranges`)
- Remove stop words (`a, an, the, at, from, on, etc.`)
- Calculate the distribution of each word
  - in the document, *term frequency* ***tf(t)***
  - in a corpus, *inverted document frequency* ***tf(t) * idf(t)***

- But frequency is not enough to produce a good summary...

Relevance Method: Position

The most important sentences usually appeared in fixed positions

- ✓ Brandow (1995) show that on **news articles** the **first sentences** of the text are the most relevant

- ✓ Others show that for **scientific articles** the **last sentences of the abstract** are usually the most relevant
- ✓ **Position at the paragraph level**: usually the first and last sentence are the important ones
- ✓ Note that the **position feature is domain/genre dependent**

Relevance Method: Title

Hypothesis:

- The title of a document is indicative of its topic

How:

- Use the words in the title to find relevant sentences
  - Create a list with the title words and remove stop words,

$$title(S) = |TIT \cap S|$$

Relevance Method: Indicative Phrases

Hypothesis:

➔ Important sentences contain indicative phrases

Examples:

✓ *The aim of this research is to describe...*

✓ *The purpose of this paper is to demonstrate..*

✓ *In this report, we outline...*

It is possible to use a list with words to assess the sentence relevance

+ *comparatives*, *superlatives*, *conclusions*, *etc*.

- *negation*, *pronouns*, *etc*.

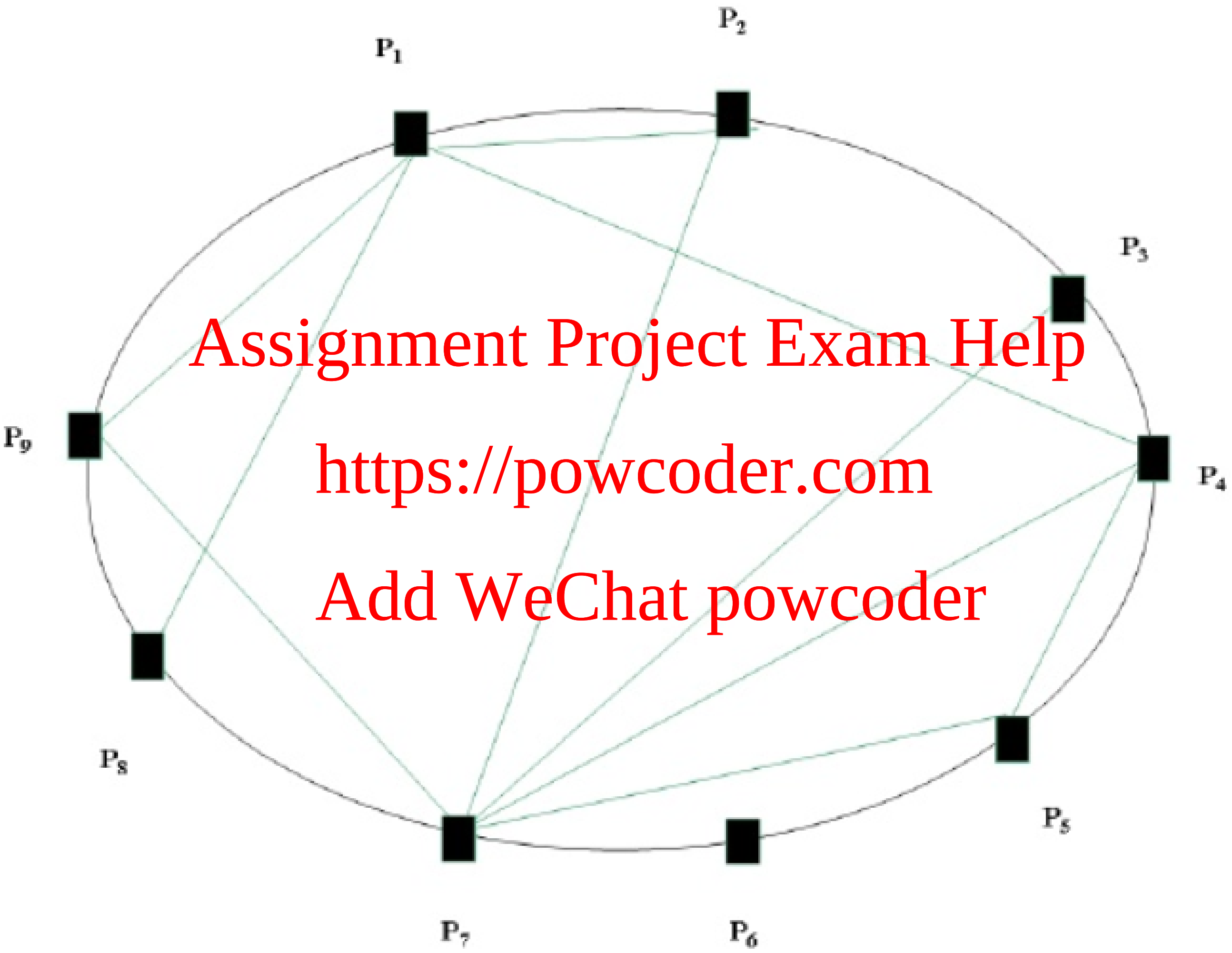
Relevance method: hybrid

- Combination of 4 methods (Edmundson, 1996)
  - keywords, title, indicative phrase and position
  - linear equation with weights
  - selects a part/portion of the text to adjust the equation parameters

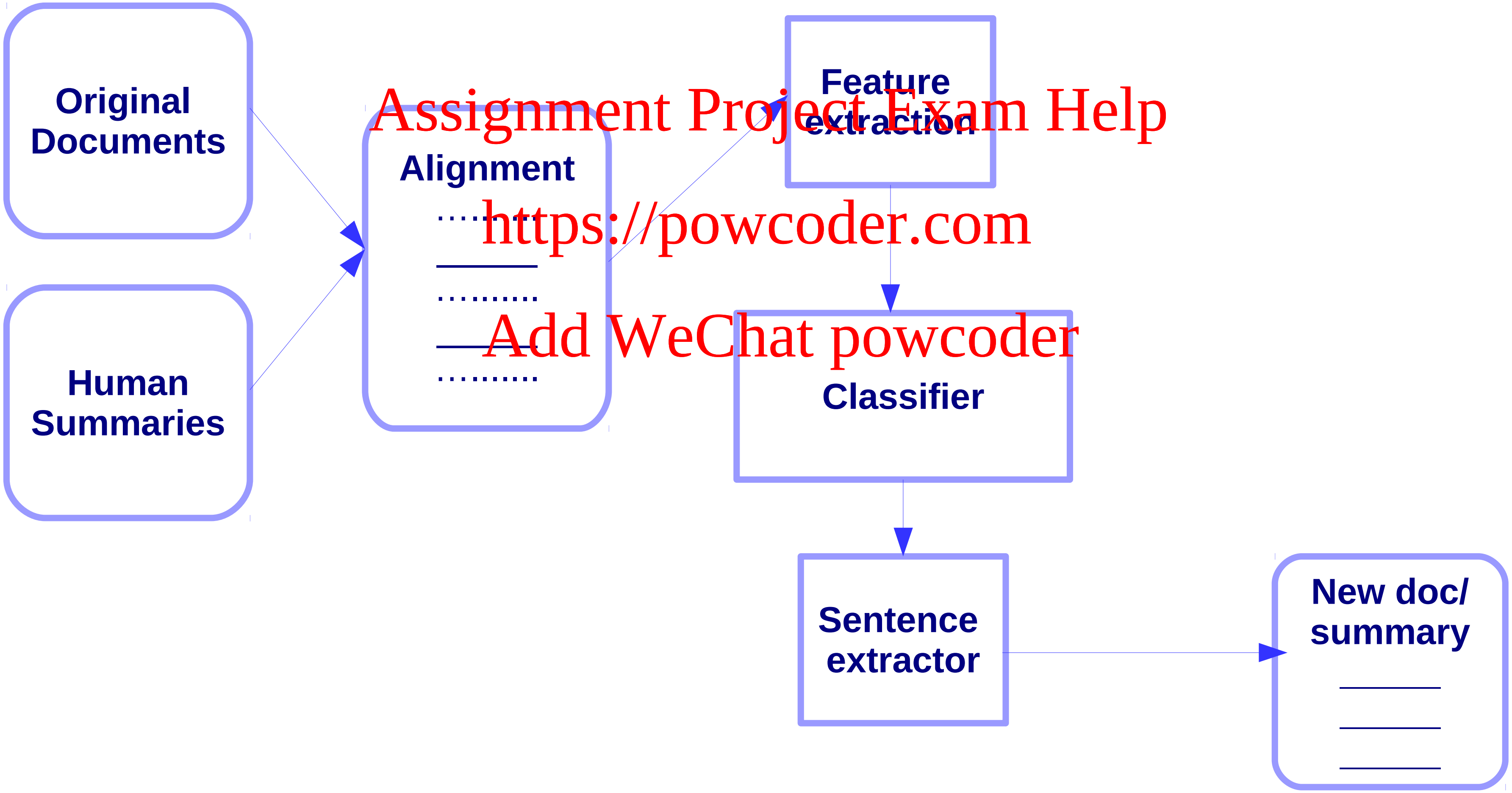
$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

Methods inspired from IR (Salton et al. 1997)

- Graph-based summarization frameworks, inspired from link analysis algorithms in network analysis.
- Computes the similarity between sentences/paragraphs and represent their link strength.
- Similar paragraphs are considered those who have a similarity above a threshold
- Paragraphs can be extracted according to different strategies (e.g. the number of links they have, select connected paragraphs, etc.)


P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9

Sentence selection as learning



Sentence selection as learning

Each sentence in the set to be learn is described by a set of features:

- The **features** are different properties of the sentences (e.g. position, keywords distribution, length of the sentence, named entities distribution, indicative phrase, etc.)
- Two classes: **extract | do-not-extract**

- **Regression models** for importance prediction
- **Learning to rank models** that assign high ranks to important sentences
- **Sequence labeling models**: model inter-sentence dependency

Sentence selection with HMM

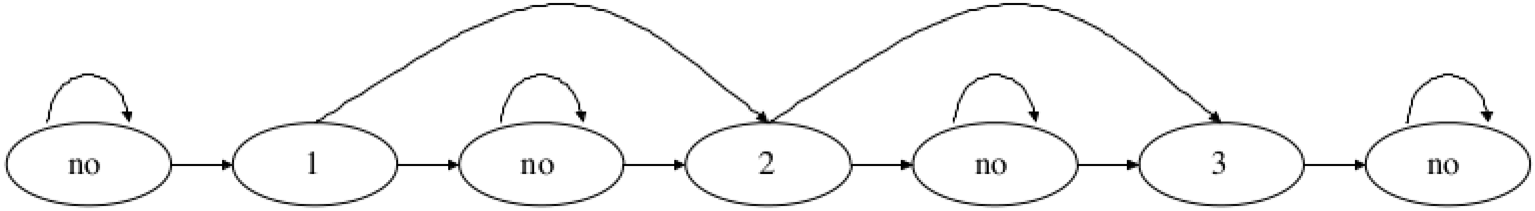
Conroy and O'Leary (2001)

This model takes into account local dependency between sentences

- 2 states: summary state | non summary state

Features :

- Position of the sentence in the doc
- Number of terms in the sentence
- Likelihood of the sentence terms given the document terms



Sentence selection: relevant + diverse

Maximal Marginal Relevance (MMR) Carbonell & Goldstein, 1998

λ[0, 1] trades of relevance and similarity

S is a subset of documents in *R* already selected

R/S is the set difference (e.g. the set of as yet unselected documents in R)

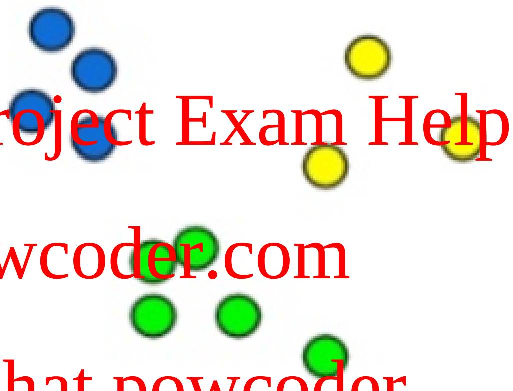
Sim1 measures the relevance between an item (e.g. sentence) and a query

Sim2 measures the similarity between two items (e.g. relevant sentences)

* Note: good performance typically relies in careful tunning of the parameter λ

$$MMR \overset{\mathrm{def}}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda (Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

Sentence selection using K-means clustering



Approach:

- Sentences as points
- Divide into clusters
- Select sentences from each cluster
- Diverse summaries

Sentence reformulation

Modify sentences in order to produce more clear, coherent and concise summaries

- rule-based sentence compression
- sentence fusion or aggregation
- sentence simplification
- paraphrasing

COMPLICATED!!!

Sentence ordering

Single document summarization

- Original order

Multi-document summarization

- More difficult: inferring order from a weighted sentence graph, use timestamps and position

Multi Document Summarization

- Multi document summarization is the extension of single-doc summarization to collections of related documents
- Very rarely, methods from single-doc summarization can be directly used
- It is possible to produce single-doc summaries from every single document in collection and then to concatenate them
- Normally, they are user-focused summaries

Multi Document Summarization

- The size of the collection might require different methods
- A much higher compression rate is needed
- Redundancy
- Similarities between different texts need to be considered
- Contradiction between information
- Fragmentary information

Summarization Evaluation

Intrinsic evaluation

- ✓ Humans read the documents and decide which are the most relevant sentences
- ✓ **ROUGE measure**: calculate the recall between human and automatic summaries in terms of n-grams (n-gram overlap)

Extrinsic evaluation

- ✓ Verify that the summaries are useful for an specific task, e.g.: text classification

Issues regarding the evaluation

- Humans usually do not agree in which are the most important sentences of a document

- Usually, there is more then one summary for the same document

- Humans generated summaries are costly

- The comparison between human and automatic summaries based on n-grams has been strongly criticized (ROUGE, Lin 2004)

- New evaluation measures without human models, which are based on probability distributions (FRESA, Saggion et al., 2010)

Limitations of Extractive Summ.

- **Redundancy**
  - The content of a summary must be diverse: apply methods that incorporate diversity (Grasshopper algorithm, MMR)

- **Coherence**
  - Part of the summaries extracted can be out of the content (anaphora gaps, missing references, lack of discourse analysis, etc.)

Take away

- Think about the best summarization approach according to the summary type and the available data (training sets?)

Extractive summ. main tasks:

- Sentence ranking
- Sentence selection
- Sentence reformulation (in novel methods)
- Sentence ordering

Abstractive summarization

Involves re-writting sentences

- paraphrasing
- simplification
- compression

or/and

generating novel content

- Natural Language Generation (NLG)

Abstractive summarization

Natural Language Generation steps:

- Content determination (what information?)
- Text/Doc structuring (ordering)
- Sentence aggregation (merging sents. = readability, naturaless)
- Lexicalization (from concepts to words)
- Referring expressions generation (pronouns, anaphora)
- Realization (acoording to syntax and morphology)

Deep Learning For Text Summarization

- Advanced **abstractive summ.** approaches

- Inspired by the application of deep learning methods for **automatic machine translation**

- Summarization as a **sequence-to-sequence** learning problem

- **End-to-end**, entirely **data-driven**

- Results are not yet state-of-the-art compared to extractive methods

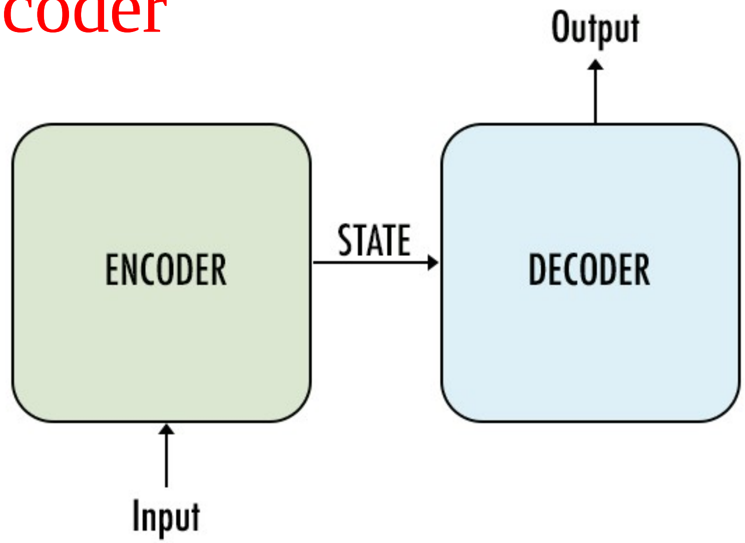
Neuronal Abstractive Summarization

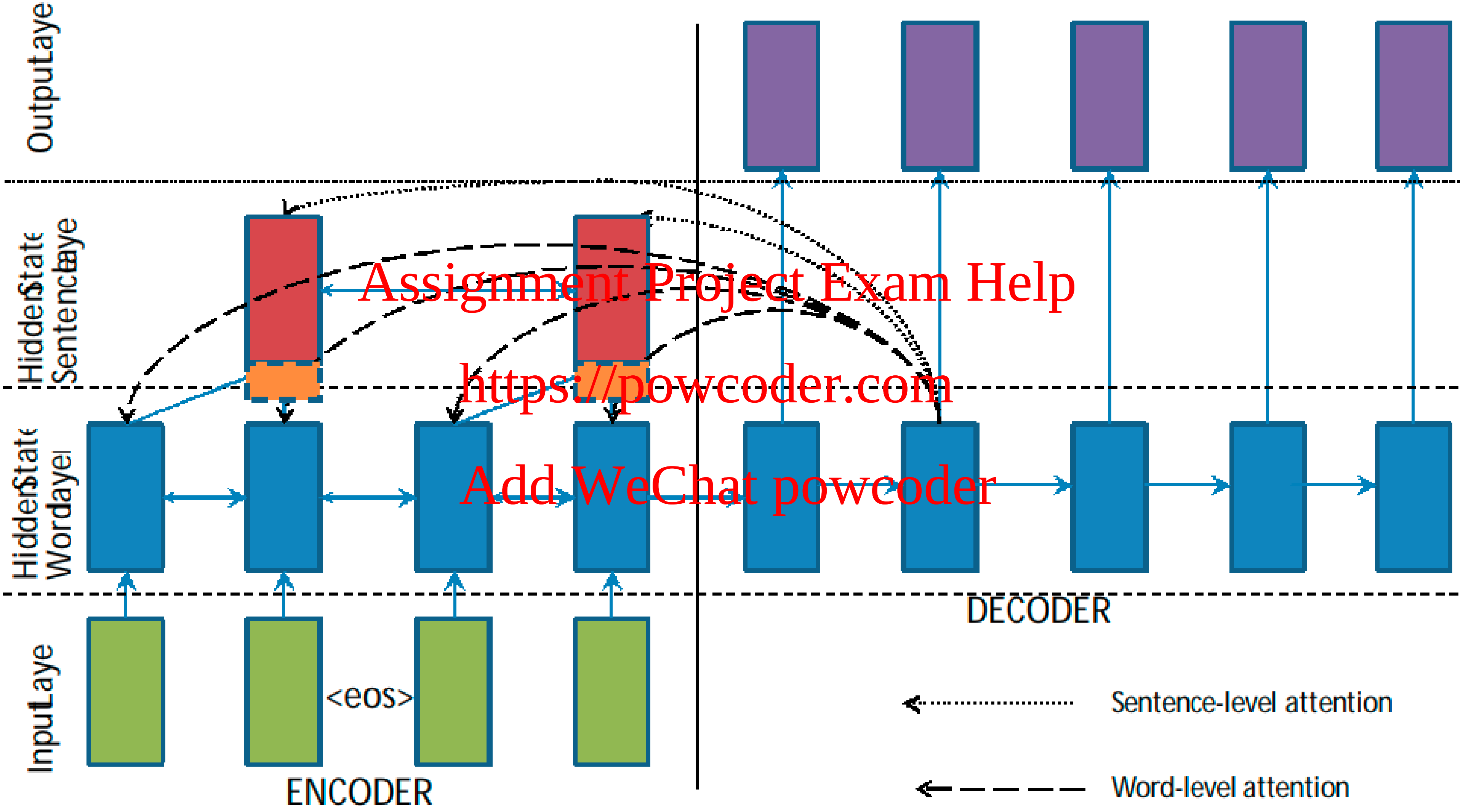
Encoder: how to represent the whole document by the encoder

- Bag-of-words-encoder: summ word embs
- …

Decoder: how to generate the word sequence

- Language model for estimating the prob. distribution that generates the word at each time step *t*
- ...





Ramesh Nallapati, et al. from IBM Watson in their 2016 paper “
Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond“.

Neuronal Abstractive Summarization

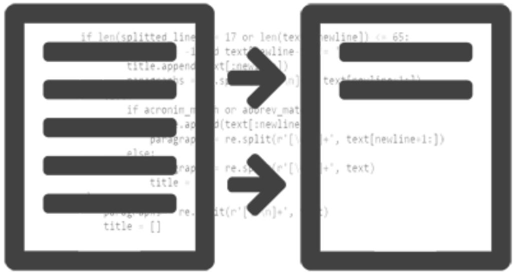
Limitations

- Unable to deal with deal with sequences longer than a few thousand word → due to the memory requirment of these model

- Unable to work well on small datasets → due to the large amount of parameters these models have

- Slow training → due to the complexity of the models

Conclusion



- Research in summarization is **still very active!!**

- Evaluation is still a problem

- The current state of the art is still sentence extraction

- More language understanding should be add to the summarization systems

Demo

News Article Summarization Ryan Endacott and Krit Pattamadit

- http://nlpsummarize.herokuapp.com/
- https://github.com/ryan-endacott/nlp

Resources

- **Online examples**
  - News explorer
    - http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html
  - News blaster
    - http://newsblaster.cs.columbia.edu/index.html

- **Other tools**
    - Summly http://summly.com/index.html
  - Open Source software
    - Meeds http://www.summarization.com/mead/
    - Open Text Summarizer http://libots.sourceforge.net/

References

- Dipanjan Das and Andre F. T. Martins. (2007). **Survey on Automatic Text Summarization**.

- Yao et al. (2017) **Recent Advances in Document Summarization**. In proceedings of Expert Systems with Applications, 2017.

- **Abstractive Text Summarization** using Sequence-to-sequence RNNs and Beyond by IBM Watson, published Aug 10, 2016. Only paper, no source code.

- A Neural Attention Model for **Abstractive Sentence Summarization** by Facebook AI Research, published Sep 3, 2015. Paper. Source code.

- Sequence-to-Sequence with Attention Model for Text Summarization (textsum) by Google Brain, published Aug 4, 2016. Only source code, no paper.