COMP4650 / COMP6490
Document Analysis 2018

Info ction

Gabriela Ferraro

DATA 61 CSIRO

Australian National University

Overview of IE lectures

- Introduction to Information Extraction (IE)
- Sequence labeling methods 1
- Sequence la
- Automatic Summarizati

* 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 sum

Is an brief state oints 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

Multiple sources ⇐

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

Selection and
placement of
stories are
determined
automatically

-

Summary Typology

Single document summary

Multi-document summary

Generic summary

- contains information about th

Query-focused summary

- e.g. make a summary about today news that talk ab ge 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 documen[illegible]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 movi
- synopses (soap o
- reviews (of a book, CD, movie, et
- 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.)

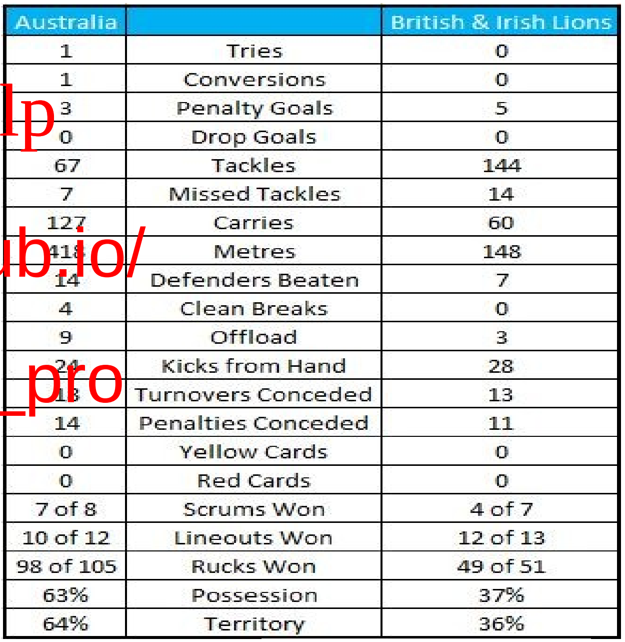
- **Abstractive summariz**

  Abstractive text summa

  generating entirely new phrases and sent

  to capture the meaning of the source doc

  - 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
13	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%

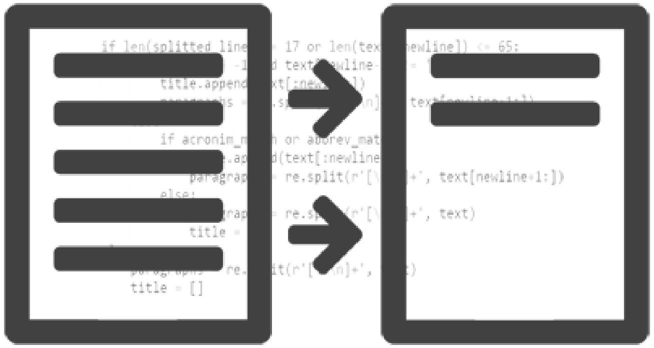
Extractive Summarization

Sentence Extraction Summarization

- Subset of the sentences from the original document

- Sentences that contained the m information

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



Extractive Summarization

- Sentence ranking
- Sentence selection
- Sentence reformulation 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 sente

- ✓ Assign to each sentence a score (releva

- ✓ 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

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**: mation is more frequently

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

Relevance Function

$$\mathrm{R}(C, Q, \phi)$$

C is a document [illegible] tream)

Q is a query or user profile or to [illegible]

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-can, woman-she, etc.)

General steps:

- Apply a stemmer algorithm to normalize all [illegible]`ge = 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 a** **last sentences of the abstract** are usually the most rel

- **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 ntences
  - 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 rese*
- ✓ *The purpose of this paper is to dem*
- ✓ *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 wi
  - selects a part/port quation parameters

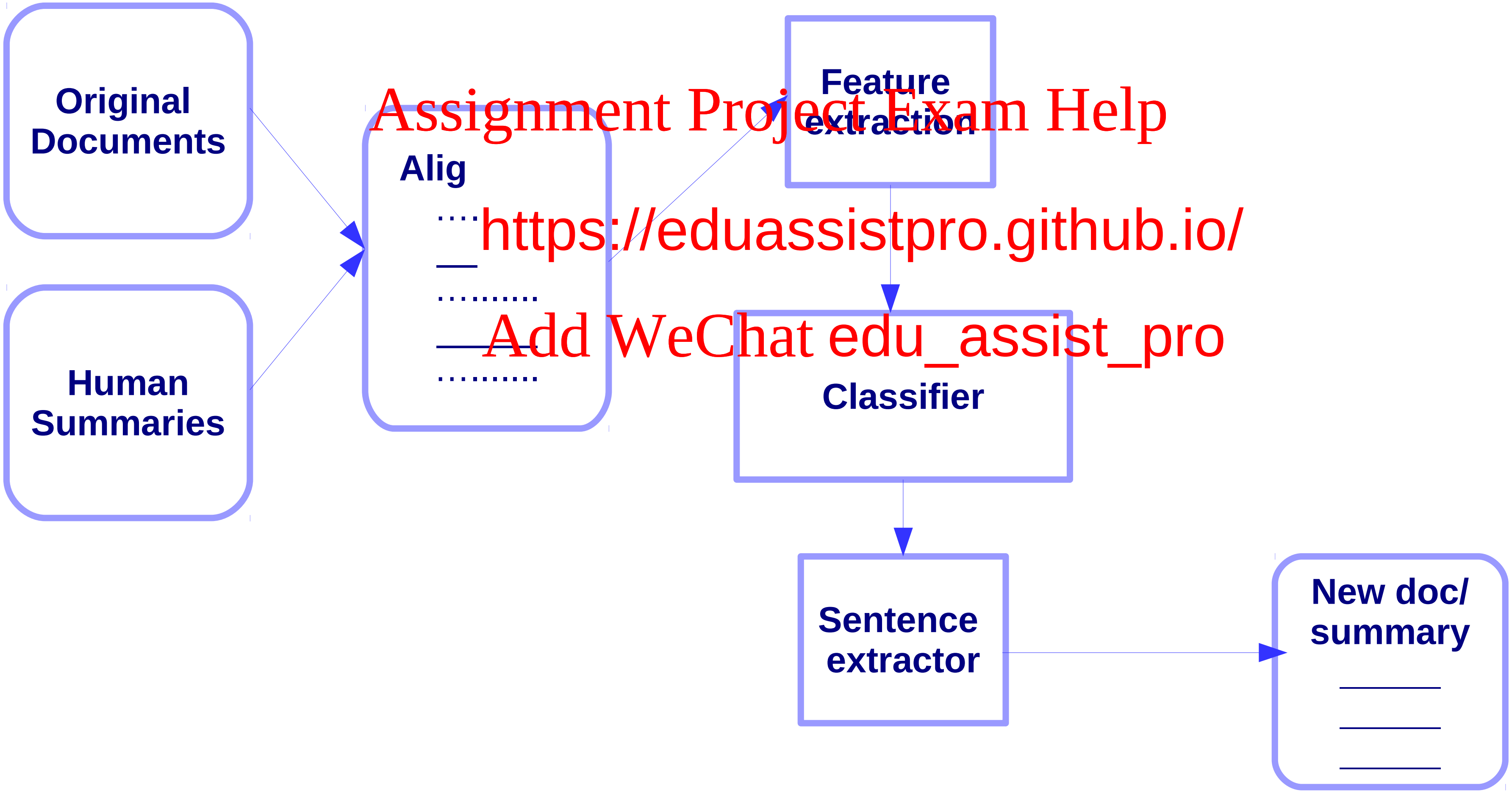
$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

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 th

- 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.)

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, d 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)

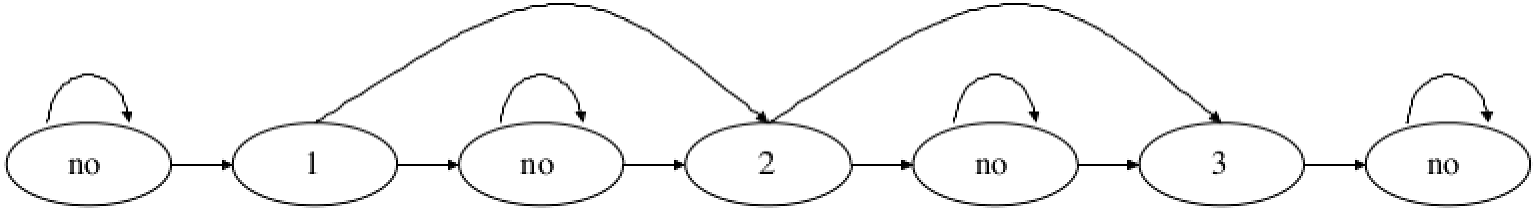
This model takes into account local dependency between sentences

- 2 states: summary state | non summary state

Features :

- Position of the sentence in the d

- 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 documen [illegible]

R/S is the set difference (e.g. the set of as ye[illegible]cuments 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

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

Sentence selection using K-means clustering

Sentence reformulation

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

- rule-based s n
- sentence fusion or aggre
- sentence simplification
- paraphrasing

COMPLICATED!!!

Sentence ordering

- Single document summarization
  - Original order

- Multi-docume [illegible]
  - More difficult: inferring or [illegible] 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-d es 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 o

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 m same document

- Humans generated summaries are c

- 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 sum

- 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 struct
- Sentence aggre s. = readability, naturales
- 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 machi**
- Summarization as a **sequence-** **ce** 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 [illegible]quence

- 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

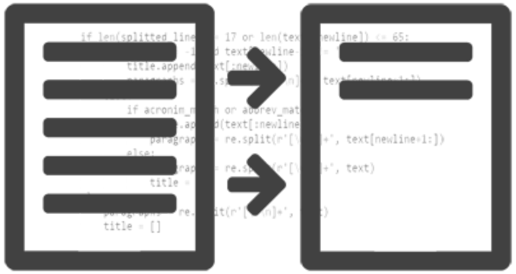
Limitations

- 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 d due to the large amount of parame e 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 i ence 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 nlp

Resources

- **Online examples**
  - News explorer
    - http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html
  - News blaster
    - http://newsbla [illegible] tml

- **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 S 7.
- **Abstractive Text Summarization** using equence RNNs and Beyond by IBM Watson, published Aug 1 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.