Lecture 10: Generation

Overview of Natural Language Generation

An extended example: cricket reports

Learning from existing text

Text summarisation

Generation

Generation from what?! (Yorick Wilks)

Generation

Starting points:

- Some semantic representation, e.g. logical form
- Formally-defined data: databases, knowledge bases
- Semi-structured data: tables, graphs etc.
- Numerical data: e.g., weather reports.
- User input in assistive communication.

Generating from data often requires domain experts.

Regeneration: transforming text

- Text from partially ordered bag of words: statistical MT.
- Paraphrase
- Summarization (single- or multi- document)
- Text simplification

Components of a generation system

Content determination deciding what information to convey (selecting important or relevant content)

Discourse structuring overall ordering, sub-headings etc

Aggregation deciding how to split information into sentence-sized chunks

Referring expression generation deciding when to use pronouns, which modifiers to use etc

Lexical choice which lexical items convey a given concept (or predicate choice)

Realization mapping from a meaning representation (or syntax tree) to a string (or speech)

Fluency ranking

Approaches to generation

- Classical (limited domain): hand-written rules for first five steps, grammar for realization
- Templates: most practical systems. Fixed text with slots, fixed rules for content determination.
- Statistical (limited domain): components as above, but use machine learning (supervised or unsupervised).
- Regeneration: symbolic or statistical or mixed.

An example: generating cricket reports

Kelly et al 2009. *Investigating Content Selection for Language Generation using Machine Learning*

Input: cricket scorecard

Result India won by 63 runs						
India innings (50 overs maximum)	R	M	B	4s	6s	SR
SC Ganguly run out (Silva/Sangakarra)	9	37	19	2	0	47.36
V Sehwag run out (Fernando)	39	61	40	6	0	97.50
D Mongia b Samaraweera	48	91	63	6	0	76.19
SR Tendulkar c Chandana b Vaas	113	141	102	12	1	110.78
. . .						
Extras (lb 6, w 12, nb 7) 25						
Total (all out; 50 overs; 223 mins) 304						

An example: generating cricket reports

Output: match report

India beat Sri Lanka by 63 runs. Tendulkar made 113 off 102 balls with 12 fours and a six. . . .

Actual report:

The highlight of a meaningless match was a sublime innings from Tendulkar, . . . he drove with elan to make 113 off just 102 balls with 12 fours and a six.

An example: generating cricket reports

Output: match report

> *India beat Sri Lanka by 63 runs. Tendulkar made 113 off 102 balls with 12 fours and a six. . . .*

Actual report:

> *The highlight of a meaningless match was a sublime innings from Tendulkar, . . . he drove with elan to make 113 off just 102 balls with 12 fours and a six.*

Representing the data

Predicates to express the factoids:

name(team1/player4, Tendulkar),
balls-faced(team1/player4, 102),
result(win, team1, 63)

- Granularity: we need to be able to consider individual information chunks
- Abstraction: generalize over instances?
- Inferences over data (e.g., amalgamation of scores)?

Content selection

There are thousands of factoids in each scorecard: we need to select the most important.

name(team1, India),
total(team1, 304),
name(team2, Sri Lanka),
result(win, team1, 63),
name(team1/player4, Tendulkar),
runs(team1/player4, 113),
balls-faced(team1/player4, 102),
fours(team1/player4, 12),
sixes(team1/player4, 1)

Statistical content selection

Treat content selection as a classification problem:

- derive all possible factoids from the data source and decide whether each is in or out, based on training data.
- categorise factoids into classes, group factoids
- learning from aligned scorecards and reports

Learning from aligned scorecards and reports

Result India won by 63 runs						
India innings (50 overs maximum)	R	M	B	4s	6s	SR
SC Ganguly run out (Silva/Sangakarra)	9	37	19	2	0	47.36
V Sehwag run out (Fernando)	39	61	40	6	0	97.50
D Mongia b Samaraweera	48	91	63	6	0	76.19
SR Tendulkar c Chandana b Vaas	113	141	102	12	1	110.78
. . .						
Extras (lb 6, w 12, nb 7) 25						
Total (all out; 50 overs; 223 mins) 304						

The highlight of a meaningless match was a sublime innings from Tendulkar, . . . he drove with elan to make 113 off just 102 balls with 12 fours and a six.

Learning from aligned scorecards and reports

Annotate reports with corresponding data structures:

> *The highlight of a meaningless match was a sublime innings from Tendulkar (team1 player4), . . . and this time he drove with elan to make 113 (team1 player4 R) off just 102 (team1 player4 B) balls with 12 (team1 player4 4s) fours and a (team1 player4 6s) six.*

- *Either* by hand
- *or* write rules to create training set automatically, using numbers and proper names as links.

Train a classifier to determine whether factoids in the score cards should be included

Discourse structure and ordering

Distribute data into sections and decide on overall ordering:

Title: name(team1, India), name(team2, Sri Lanka), result(win,team1,63)

First sentence: name(team1/player4, Tendulkar), runs(team1/player4, 113), fours(team1/player4, 12), sixes(team1/player4, 1), balls-faced(team1/player4, 102)

Reports often state the highlights and then describe events in chronological order.

Statistical discourse structuring: generalising over reports to see where particular information types are presented.

Lexical choice and realisation

Mapping rules from the initial scorecard predicates:

result(win,t1,n) $\mapsto$ *_beat_v(e,t1,t2), _by_p(e,r), _run_n(r), card(r,n)*

Realisation:

India beat Sri Lanka by 63 runs.

Text summarisation[1]

Task: generate a short version of a text that contains the most important information

Single-document summarisation:

- given a single document
- produce its short summary

Multi-document summarisation:

- given a set of documents
- produce a brief summary of their content

[1] This part of the lecture is based on Dan Jurafsky's summarisation lecture, and is (quite appropriately) a summary thereof. The full lecture can viewed online at `https://class.coursera.org/nlp/lecture/preview`

Generic vs. Query-focused summarisation

Generic summarisation:

- identifying important information in the document(s) and presenting it in a short summary

Query-focused summarisation:

- summarising the document in order to answer a specific query from a user

A simple example of query-focused summarisation

Google

what is natural language processing?

Web Videos News Images Shopping More Search tools

About 30,200,000 results (0.38 seconds)

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages. As such, **NLP** is related to the area of human–computer interaction.

Natural language processing - Wikipedia, the free ...
https://en.wikipedia.org/wiki/**Natural_language_processing**

Feedback

Natural language processing - Wikipedia, the free ...
https://en.wikipedia.org/wiki/**Natural_language_processing**
Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages. As such, **NLP** is related to the area of human–computer interaction.
Outline of natural language ... - Natural language understanding

Approaches

Extractive summarisation:

- extract important / relevant sentences from the document(s)
- combine them into a summary

Abstractive summarisation:

- interpret the content of the document (semantics, discourse etc.) and generate the summary
- formulate the summary using other words than in the document
- very hard to do!

Extractive summarisation

Three main components:

- Content selection: identify important sentences to extract from the document
- Information ordering: order the sentences within the summary
- Sentence realisation: sentence simplification

Content selection – unsupervised approach

- Choose sentences that contain informative words
- Informativeness measured by:
  - tf-idf: assign a weight to each word i in the doc j as

$$weight(w_i) = tf_{ij} * idf_i$$

tf_{ij} – frequency of word i in doc j
idf_i – inverse document frequency

$$idf_i = \log \frac{N}{n_i}$$

N – total docs; n_i docs containing w_i

  - mutual information
  - log-likelihood ratio (LLR)

Content selection – supervised approach

- start with a training set of documents and their summaries
- align sentences in summaries and documents
- extract features:
  - position of the sentence (e.g. first sentence)
  - sentence length
  - informative words
  - cue phrases
  - etc.
- train a binary classifier: should the sentence be included in the summary?

Content selection – supervised vs. unsupervised

Problems with the supervised approach:

- difficult to obtain data
- difficult to align human-produced summaries with sentences in the doc
- doesn't perform better than unsupervised in practice

An example summary

from Nenkova and McKeown (2011):

As his lawyers in London tried to quash a Spanish arrest warrant for Gen. Augusto Pinochet, the former Chilean Dictator, efforts began in Geneva and Paris to have him extradited. Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean Dictator has diplomatic immunity is ridiculous. Margaret Thatcher entertained former Chilean Dictator Gen. Augusto Pinochet at her home two weeks before he was arrested in his bed in a London hospital, the ex-prime minister's office said Tuesday, amid growing diplomatic and domestic controversy over the move.

Query-focused multi-document summarisation

Example query: "*Describe the coal mine accidents in China and actions taken*"

Steps in summarization:

1. find a set of relevant documents
2. simplify sentences
3. identify informative sentences in the documents
4. order the sentences into a summary
5. modify the sentences as needed

Sentence simplification

- parse sentences
- hand-code rules to decide which modifiers to prune
  - appositives: e.g. *Also on display was a painting by Sandor Landeau, ~~an artist who was living in Paris at the time.~~*
  - attribution clauses: e.g. *Eating too much bacon can lead to cancer, ~~the WHO reported on Monday.~~*
  - PPs without proper names: e.g. *Electoral support for Plaid Cymru increased ~~to a new level.~~*
  - initial adverbials: e.g. *~~For example,~~ ~~On the other hand,~~*
- also possible to develop a classifier (e.g. satelite identification and removal)

Content selection from multiple documents

Select informative and non-redundunt sentences:

- Estimate informativeness of each sentence (based on informative words)
- Start with the most informative sentence:
  - identify informative words based on e.g. tf-idf
  - words in the query also considered informative
- Add sentences to the summary based on maximal marginal relevance (MMR)

Content selection from multiple documents

Maximal marginal relevance (MMR): iterative method to choose the best sentence to add to the summary so far

- Relevance to the query: high cosine similarity between the sentence and the query
- Novelty wrt the summary so far: low cosine similarity with the summary sentences

$$\hat{s} = \underset{s_i \in D}{\text{argmax}} \left[\lambda sim(s_i, Q) - (1 - \lambda) \max_{s_j \in S} sim(s_i, s_j) \right]$$

Stop when the summary has reached the desired length

Sentence ordering in the summary

- Chronologically: e.g. by date of the document
- Coherence:
  - order based on sentence similarity (sentences next to each other should be similar, e.g. by cosine)
  - order so that the sentences next to each other discuss the same entity / referent
- Topical ordering: learn a set of topics present in the documents, e.g. using LDA, and then order sentences by topic.

Example summary

Query: "*Describe the coal mine accidents in China and actions taken*"

Example summary (from Li and Li 2013):
(1) In the first eight months, the death toll of coal mine accidents across China rose 8.5 percent from the same period last year. (2)China will close down a number of ill-operated coal mines at the end of this month, said a work safety official here Monday. (3) Li Yizhong, director of the National Bureau of Production Safety Supervision and Administration, has said the collusion between mine owners and officials is to be condemned. (4)from January to September this year, 4,228 people were killed in 2,337 coal mine accidents. (5) Chen said officials who refused to register their stakes in coal mines within the required time