MODEL SOLUTIONS

SETTER: Mark Stevenson

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2011-2012

TEXT PROCESSING 2 hours

Answer THREE questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

1. a) Describe *manual* and *automatic* approaches to indexing for Information Retrieval. What are the advantages and disadvantages of the different approaches? [15%]

ANSWER:

In manual approaches documents are indexed using terms that are identified manually, often from some pre-defined controlled vocabulary or taxonomy. In automatic approaches the the index terms are identified directly from the documents.

Students could mention some of the following points when discussing the advantages and disadvantages of each approach.

The advantage of manual indexing is that the index terms that are added is more closely controlled than is possible with automatic indexing. They are more likely to be appropriate for the document and unambiguous. The (human) indexer may also be able to use judgement and experience to identify index terms which are appropriate but could not be easily identified by simple analysis of the document. On the other hand manual indexing is expensive in terms of effort that is required from human indexers which can make this approach impractical. Manual indexing also allows the use of hierarchical indexing terms which can be used to create complex queries and this is difficult to achieve automatically. However, the index terms may be drawn from some controlled vocabulary which a user may not be familiar with and training may be required to search a manually indexed collection effectively.

The advantages and disadvantages of automatic approaches are largely the opposite of manual indexing. Automatic indexing does not require manual effort and can be efficiently applied to large document collections but, on the other hand, the quality of the index terms is unlikely to be as good as those that are added manually.

b) Consider the following documents and query:

Document 1: They sailed to the port for a good dinner and sailed home after.

Document 2: Ruby port is very good after a meal, any meal.

Query: good ports after dinner

(i) Show how Documents 1 and 2 would be stored in an inverted index, using stemming and the following list of stopwords: {`a, and, any, for, is, the, they, to` } [20%]

ANSWER:

Using the given stoplist and stemming (which reduces e.g. *sailed* to *sail*), and assuming that there are no other documents (i.e. only terms in these document need be considered) we get the following inverted index:

term-id	word	doc
1	after	d1:1, d2:1
2	dinner	d1:1
3	good	d1:1, d2:1
4	home	d1:1
5	meal	d2:2
6	port	d1:1, d2:1
7	ruby	d2:1
8	sail	d1:2
9	very	d2:1

Here, entries record frequency counts, e.g. the entry for *after* tells us that the term appears once in Document 1 (d1:1) and once in Document 2 (d2:1).

(ii) Compute the similarity between the query and each document using the cosine metric and and using term frequency values for the term weights in the document vectors. [30%]

ANSWER:

The cosine between two vectors $\vec{x}$ and $\vec{y}$ is computed as:

$$cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Using the term order taken from the inverted index above, we can represent the two documents and the query as vectors in the following way:

d1: $\langle 1,1,1,1,0,1,0,2,0 \rangle$
d2: $\langle 1,0,1,0,2,1,1,0,1 \rangle$
q: $\langle 1,1,1,0,0,1,0,0,0 \rangle$

The vector magnitudes and cosine values are then:

$$|\mathrm{d1}| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 2^2 + 0^2} = \sqrt{9} = 3$$

$$|\mathrm{d2}| = \sqrt{1^2 + 0^2 + 1^2 + 0^2 + 2^2 + 1^2 + 1^2 + 0^2 + 1^2} = \sqrt{9} = 3$$

$$|\mathrm{q}| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2} = \sqrt{4} = 2$$

$$cos(\mathrm{d1}, \mathrm{q}) = \frac{1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 0 + 0 \cdot 0}{3 \cdot 2} = \frac{4}{3 \cdot 2} = \frac{2}{3}$$

$$cos(\mathrm{d2}, \mathrm{q}) = \frac{1 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 0}{3 \cdot 2} = \frac{3}{6} = \frac{1}{2}$$

(iii) Explain why the vector space model for Information Retrieval would not identify Document 2 as the top ranked document for the query. [15%]

ANSWER:

The model would not identify Document 2 since the query does not contain terms which would lead to it being identified. The terms in the query which matter for retrieval are "good", "port" and "dinner". "good" appears in both documents and does not distinguish between them. "Port" also appears in both documents but is ambiguous between two meanings (alcoholic drink and location). However, the vector space model is unable to distinguish between these so this does not effect the outcome. Although the term "meal" appears in Document 2, the vector space model is unable to recognise the connection between this and the closely associated term in the query ("dinner") but since "dinner" appears in Document 1 this is given a higher score by the vector space model.

c) Define the precision and recall measures used in Information Retrieval. Explain why it is easier to compute precision than recall for web-based Information Retrieval systems such as Google or Bing. [20%]

ANSWER:

Assuming that: RET is the set of all documents the system has retrieved for a specific query; REL is the set of relevant documents for a specific query; $RETREL$ is the set of the retrieved relevant documents, i.e., $RETREL = RET \cap REL$. Precision is defined as $|RETREL|/|RET|$ and recall is defined as $|RETREL|/|REL|$.

Precision is easier to compute in the web-based scenario since it only requires the documents returned by the system to be analysed while recall requires knowledge of all relevant documents. Since precision is computed as the proportion of returned documents that are relevant to the query all that is required is to determine which of those are relevant (i.e. $RELRET$). However, computing recall also involves determining the number of relevant documents in the collection (ie. the internet) which is unlikely to be practical or even possible.

2. a) Describe the three main approaches to Machine Translation and explain the differences between them. [20%]

ANSWER:

The three main approaches are *direct*, *transfer* and *interlingua*.

The key difference between the three approaches is the level of analysis which is applied to the source text.

Direct approaches apply very little analysis to the the source text and rely on simple translation of each word in the source text. Statistical MT could be considered to be a direct approach.

Transfer approaches attempt to analyse the structure of the source text to produce an intermediate representation. The intermediate representation of a sentence from the input text is then used in generating the translated sentence in the target language. Transfer approach can employ syntactic and/or semantic representations.

Interlingual approaches rely on a representation of the meaning which is independent of both the source and target language. The source sentence goes through syntactic and semantic analysis to be translated into the interlingua. This representation is then used to generate a sentence in the target language. The difference between transfer approaches which use semantic representations and interlingua approaches rests on the independence of the system used to represent meaning; interlinguas are completely independent of the source and target languages while the representation used in semantic transfer simply aims to capture enough information to allow translation.

b) Describe the main features of the statistical approach to Machine Translation. [30%]

ANSWER:

Here are some of the general points that the students should discuss in answering this question:

Machine translation can be cast in terms of the noisy-channel model. The approach was pioneered by IBM in the early 90's. If we want to translate French into English, we shall assume that someone has E in his head but by the time he writes it down it's *corrupted* and becomes F. To recover the original E, need to reason about:

(a) the kinds of things people say in English and (b) how English gets turned into French. Given a French sentence F the approach searches for English E that maximises $P(E|F)$.

$$\begin{aligned} \text{argmax}_E\, P(E|F) &= \text{argmax}_E\, \frac{P(E)\cdot P(F|E)}{P(F)} \\ &= \text{argmax}_E\, P(E)\cdot P(F|E) \end{aligned}$$

We choose not to model $P(E|F)$ directly. Instead, we break things apart to get good translations even if the probability numbers are not that accurate. $P(F|E)$ will ensure that a good E will have words that generally translate to words in F. $P(E)$ will help save us from bad English and $P(F|E)$ only needs to say whether a bag of English words corresponds to a bag of French words. $P(E)$ and $P(F|E)$ can be trained independently.

In order to model $P(F|E)$ we need a sentence-aligned parallel (bilingual corpus) and a model of translation $\sim P(F|E)$. The latter assumes word alignments which we do not have. We can approximate those by using EM, an unsupervised learning algorithm that can learn alignment probabilities.

c) Give an example of a language pair for which Statistical Machine Translation is likely to work well and another example for which is it likely to work badly. Explain your choices with reference to the approach used by Statistical Machine Translation systems. [20%]

ANSWER:

SMT is best suited to languages which are structurally similar (e.g. English-French, Spanish-Portuguese) and least suited to those which are structurally different (e.g. English-Chinese). The reason is that it is difficult to align parallel text when the languages include pairs which are structurally different and accurate alignment is important for the translation model used by SMT. Reasons structural differences can make alignment difficult include (1) different notions of what a words is (so the alignment between words is not one to one), (2) different word orders (leading to a huge number of possible alignments which would be impossible to compute), (3) different morphologies (making it difficult to learn translation probabilities).

d) Describe how Round-trip translation can be used to evaluate Machine Translation systems. What are the advantages and disadvantages of this approach? [15%]

ANSWER:

Round-trip translation translates the original text, written in L1, into another language L2 and then back into L1. The quality of the translation is evaluated by checking how close the text produced is to the original text.

The advantages are that this approach is that it does not require knowledge of another language (L2) or significant resources (such as reference translations). The disadvantages are that two Machine Translation systems are involved (L1 to L2) and (L2 to

L1). The approach cannot distinguish between them and cannot tell which Machine Translation system introduced any errors. In addition, one Machine Translation system may fix errors made by the other but this will not be detected. In addition the basic premise behind this approach, that the text produced by round trip translation should be the same as the original text is flawed since human translators would not necessarily produce the same text.

e) Describe the approach used by the BLEU system for evaluation of Machine Translation systems. [15%]

ANSWER:

The BLEU system relies on multiple reference translations, each of which represents a possible way in which a text could be translated into a target language. The translation being evaluated (the candidate) is compared against the reference translations by counting the number of possible ngrams (strings of words) which occur in them. BLEU assumes that there are several possible ways to translate a text, each of which are equally valid, and uses multiple reference reference translations to provide these alternatives.

3. a) What is summarisation? Does summarisation always involve text? [10%]

ANSWER:

In its most general sense, summarisation is the art of abstracting key content from one or more information sources. This has most commonly been done for textual sources (e.g. newpaper articles, journal papers), but this is not necessarily so. In fact, either, or both, of the input to/output from summarisation might be non-textual. For example, the story told by a movie might be summarised in a critical review (video>text), A news report on troop movements might be summarised by a short animation (text>graphic), or a full-length video of a football match might be summarised by a short video of edited highlights (video>video).

b) Outline three of the different types of ways in which summaries have been classified and explain how this effects the properties of the summary that should be produced by an automatic summariser. [15%]

ANSWER:

Some of the ways in which summaries have been distinguished are as follows and any three would acceptable. Answers should explain the terms used.

1. abstract vs. extract. Abstract summary is a new document while an extractive summary is a set of sentences from the original document. An abstractive summary system will be required to generate text while the goal of an extractive system is to identify the important sentences in a document.
2. indicative vs. informative vs. evaluative (critical) summaries. Indicative summaries provide an indication of the content of the document and help users to decide whether to read the full document or not. Informative summaries cover all key information in the document and act as a convenient replacement for the full document that conveys the important information quickly. Critical summaries evaluate the content of the document.
3. single vs. multi-document summaries. A single document summarisation system requires only one document as input while a multi-document summariser will make use of multiple documents. Single document summarisation decides which information to include by examining only one document while multi-document may consider whether a piece of information is repeated several times in a set of documents.
4. generic vs. query-based vs. user-profile-based summaries. A since generic summary would be generated for any document while an automatic summariser would generate different query-based summaries and user-profile-based summaries. Query-based summaries are tailored to an individual query (such as the snippets returned by Google) while user-profile-based summaries make use of some pre-defined model of the user to tailor the summary to their needs.

5. summary for expert vs. novice users. Summaries for novice users should avoid the use of complex terminology and generate summaries that do not require background knowledge to be understood, although this can be assumed to be known by expert users.

c) What is meant by *deep* and *shallow* approaches to automatic text summarisation. Discuss the differences between them. [15%]

ANSWER:

Deep approaches:

- are knowledge-rich
- involve a significant extent of 'semantic analysis'
- employ a sizeable knowledge-base of rules that are costly to create/maintain
- are domain dependent
- are difficult/costly to adapt to new domains
- produce abstracts

By contrast, shallow approaches:

- use an analysis based primarily on words and other shallow document features
- no significant 'semantic analysis'
- are knowledge-poor (or '-light')
- have little domain dependence and so are easily ported to new domains
- produce extraction-based summaries

d) Discuss the various features which have been used by *shallow* approaches to automatic summarisation and explain why they are useful for identifying segments of text containing important information in a document. [20%]

ANSWER:

Methods used include

- **keywords**:
  - assume frequent content words are often indicative of the topic of a document
  - identify set of keywords by (e.g.) TF.IDF criterion
  - score sentences w.r.t. occurrence of keywords within them
- **title method**:
  - titles and subheadings are intended to indicate the content of a document
  - use words from them to identify useful sentences

- **position**:
  - sentences at certain places in a document are likely to provide an overview of its content
  - use position in document as indicator of importance of text segment
  - notable positions include: (i) being close to start of text, (ii) in certain sections (e.g. intro/conclusion), (iii) first sentence of paragraph, etc.
- **cue method**:
  - certain phrases may signal important content (e.g. "this paper presents . . . ")
  - may distinguish both positive and negative indicators, as 'bonus' and 'stigma' words/phrases

e) Consider the following short document and summaries. Discuss the quality of the two summaries with particular reference to coherence and cohesion.

> **Original document**: John went to the park last week. He took his daughter, Sally. She likes going to the park. John also took his dog Fido. Fido also likes going to the park. He likes to chase the ducks.
>
> **Summary 1**: John went to the park last week. She likes going to the park.
>
> **Summary 2**: John went to the park last week. He likes to chase the ducks.

[15%]

ANSWER:

Summary 1 is not well formed or easily interpretable. In particular it is not cohesive since there is no antecedent for "she" in the second sentence.

Summary 2 is cohesive, since the text is connected in a linguistically plausible way. It is also coherent, since the text does make sense, however, it is misleading since it implies that John like to chase the ducks while the original text does not say this and is therefore not a good summary since it is not faithful to the original document.

f) Explain the difference between *intrinsic* and *extrinsic* approaches to the evaluation of summaries, providing details about the two approaches. [25%]

ANSWER:

At the highest level intrinsic evaluation evaluates the quality of the summary itself while extrinsic evaluation assesses the impact of subjects using the summary on their performance at some task.

Tasks used in extrinsic evaluations include the subject's ability to answer questions about the topic of the original text, or to assign a topic category to the original text, when given the automatic summary, as opposed to using the original text or a human summary.

Intrinsic evaluation may be done by human judgement (e.g. of whether key topics of original text are present, whether summary is coherent) or be done automatically. Human judgements are subjective, and often vary between judges, and the approach is time-consuming (and therefore costly). Automatic evaluation requires the creation of gold standard materials, i.e. collections of documents with associated reference summaries (which are costly to create). It primarily addresses whether the content of reference summaries is represented in automated summaries, not whether the latter are coherent. For sentence extraction approaches, evaluation against extraction-based reference summaries may be done using measures such as precision and recall. This is not possible for abstraction approaches. Alternative approaches measure textual overlap between automatic and reference summaries, as an indicator of conceptual overlap. A key approach is the ROUGE system (Recall Oriented Understudy for Gisting Evaluation), which measures (a version of) word n-gram overlap between automatic and reference summaries.

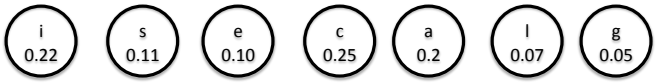
4. a) The fictitious language *Hartish* uses a script that consists of 7 characters: `a, c, e, g, i, l, s`. Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
a	0.20
c	0.25
e	0.10
g	0.05
i	0.22
l	0.07
s	0.11

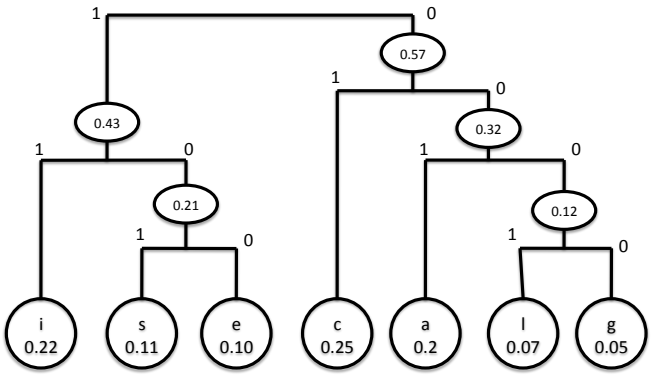
(i) Show how to construct a Huffman code tree for Hartish, given the above probabilities for its characters. [25%]

ANSWER:

Start off by creating a leaf node for each character, with associated probability:



Then join two nodes with smallest probabilities under a single parent node, whose probability is their sum, and repeat untill only one node left. Finally, 0's and 1's are assigned to each binary split, giving the result shown below.



(ii) Use your code tree to encode the message "`gailsails`" and show the resulting binary encoding. How many bits are required to encode the message? [10%]

ANSWER:

Encoding for "`gailsails`" will be

`g a i l s a i l s`

`0000 001 11 0001 101 001 11 0001 101`

The encoding uses 28 bits.

(iii) How many bits are required to encode the message "`gailsails`" using a minimal fixed length binary character encoding for a 7 character alphabet? Explain why fewer bits are required to compress this string using a minimal fixed length binary character encoding than using Huffman encoding. [15%]

ANSWER:

For a seven character alphabet a minimal fixed length binary encoding is 3 bits per character. There are 9 characters in the string, so a fixed length encoding would require 27 bits.
In Huffman coding the code length for a character is based on its probability with shorter strings being used for more frequent characters and longer ones for less frequent characters. This string includes a large proportion of low frequency characters and therefore a high number of longer than average codes.

b) Explain the difference between **static**, **semi-static** and **adaptive** techniques for text compression, noting their key advantages and disadvantages. [15%]

ANSWER:

Compression techniques can also be distinguished by whether they are

- **Static** – use a fixed model or fixed dictionary derived in advance of any text to be compressed
  - adv: model does not need to be transmitted
  - disadv: model may not be well suited to text currently being compressed
- **Semi-static** – use current text to build a model or dictionary during one pass, then apply it in second pass
  - adv: model should be well suited to current text
  - disadv: model must also be transmitted, reducing effectiveness of compression
- **Adaptive** – build model or dictionary adaptively during one pass
  - adv: model does not need to be transmitted
  - disadv: decoder determines model used at each stage from data decoded so far, so cannot do random access into data

c) (i) Explain how the LZ77 compression method works. [25%]

ANSWER:

Students could discuss the following points in relation to LZ77.

The **key idea** underlying the LZ77 adaptive dictionary compression method is to replace substrings with a pointer to previous occurrences of the same substrings in same text. The encoder output is a series of triples where

- the first component indicates how far back in decoded output to look for next phrase
- the second indicates the length of that phrase
- the third is next character from input (only necessary when not found in previous text, but included for simplicity)

Issues to be addressed in implementing an adaptive dictionary method such as LZ77 include

- how far back in the text to allow pointers to refer
  - references further back increase chance of longer matching strings, but also increase bits required to store pointer
  - typical value is a few thousand characters
- how large the strings referred to can be
  - the larger the string, the larger the width parameter specifying it
  - typical value $\sim$ 16 characters
- during encoding, how to search window of prior text for longest match with the upcoming phrase
  - linear search very inefficient
  - best to index prior text with a suitable data structure, such as a trie, hash, or binary search tree

A popular high performance implementation of LZ77 is **gzip**

- uses a hash table to locate previous occurrences of strings
  - hash accessed by next 3 characters
  - holds pointers to prior locations of the 3 characters
- pointers and phrase lengths are stored using variable length Huffman codes, computed semi-statically by processing 64K blocks of data at a time
- pointer triples are reduced to pairs, by eliminating 3rd element
  - first transmit phrase length – if 1 treat pointer as raw character; else treat pointer as genuine pointer

(ii) How would the following LZ77 encoder output

$$\langle 0,0,d\rangle\langle 0,0,o\rangle\langle 0,0,m\rangle\langle 2,1,d\rangle\langle 3,2,o\rangle\langle 3,3,d\rangle\langle 12,2,o\rangle\langle 0,0,m\rangle$$

be decoded, assuming the encoding representation described in the lectures for this module (Text Processing)? Show how your answer is derived. [10%]

ANSWER:

1. $\langle 0, 0, d \rangle$ Go back 0 copy for length 0 and end with d: $\underline{d}$
2. $\langle 0, 0, o \rangle$ Go back 0 copy for length 0 and end with o: $d\underline{o}$
3. $\langle 0, 0, m \rangle$ Go back 0 copy for length 0 and end with m: $do\underline{m}$
4. $\langle 2, 1, d \rangle$ Go back 2 copy for length 1 and end with d: $dom\underline{od}$
5. $\langle 3, 2, o \rangle$ Go back 3 copy for length 2 and end with o: $domod\underline{moo}$
6. $\langle 3, 3, d \rangle$ Go back 3 copy for length 3 and end with d: $domodmoo\underline{mood}$
7. $\langle 12, 2, o \rangle$ Go back 12 copy for length 2 and end with o: $domodmoomood\underline{doo}$
8. $\langle 1, 5, o \rangle$ Go back 1 copy for length 5 and end with m: $domodmoomooddoo\underline{ooooom}$

Thus, the decoded string is:

`domodmoomooddooooooom`

END OF QUESTION PAPER