Data Mining and Machine Learning

Topic Analysis

Peter Jančovič

Objectives

- Statistical modelling of topics
- Identifying topics in a document
  - Latent Dirichlet Allocation (LDA)
- Topic Spotting
  - Salience and Usefulness
  - Example: The AT&T "How May I Help You?" system

Motivation

- **Example 1**:You are responsible for competitor analysis in a large company. You need to monitor all media for press-releases, news items and other articles relating to your company's product range.

- **Example 2**: You work for the police. You are given the task of monitoring 500 telephone lines for 12 months. You have to identify calls on these lines which are about illegal drug trafficking.
- **Example 3**: You manage a call centre. You are concerned that some staff are being rude to the people that they are calling. You need to monitor all calls for a period of 6 months and detect all instances of 'rudeness'.

Topics

- “your company’s product range”, “illegal drug trafficking” and “rudeness” are all examples of topics
- A typical document typically covers multiple topics
- Topic Analysis is about decomposing a document into its component topics
- Topic Spotting is about identifying documents that are relevant to a particular topic
- The previous slide is a list of Topic Spotting problems

Topics as “bundles of words”

- For any term *w*, P(*w*) is the probability of *w*
  - Choose a document at random, and then choose a term at random from the document, P(*w*) is the probability that the term is *w*
  - We know about P(*w*) from Zipf’s Law
- If *T* is a topic, P(*w*|*T*) is the conditional probability of *w* given the topic *T*
  - Choose a document about topic *T* at random, then choose a term at random from the document, P(*w*|*T*) is the probability that the term is *w*

Statistical modelling of topics

- The conditional distribution P(*w*|*T*) is a “bundle of words” model of the topic *T*
- A typical document is made up of multiple topics
  - Example: Wikipedia entry on the London Marathon (next slide)

- Latent Dirichlet Allocation (LDA) expresses a document as a combination of topics

- The simplest way to understand LDA is to see how the LDA model generates a document

Documents have multiple topics

Topics include: London, marathons, fund-raising

The race was founded by the former Olympic champion and journalist Chris Brasher and athlete John Disley. It is organised by Hugh Brasher (son of Chris) as Race Director and Nick Bitel as Chief Executive. Set over a largely flat course around the River Thames, the race begins at three separate points around Blackheath and finishes in The Mall alongside St. James's Park. Since the first marathon, the course has undergone very few route changes. In 1982, the finishing post was moved from Constitution Hill to Westminster Bridge due to construction works. It remained there for twelve years before moving to its present location at The Mall.

In addition to being one of the top six international marathons run over the distance of 26 miles and 385 yards, the IAAF standard for the marathon established in 1921 and originally used for the 1908 London Olympics, the London Marathon is also a large, celebratory sporting festival, third only to the Great North Run in South Shields and Great Manchester Run in Manchester in terms of the number of participants. The event has raised over £450 million for charity since 1981,[2][3] and holds the Guinness world record as the largest annual fund raising event in the world, with the 2009 participants raising over £47.2 million for charity. In 2007, 78% of all runners raised money. In 2011 the official charity of the London Marathon was Oxfam. In 2014, the official charity was Anthony Nolan, and in 2015, it will be Cancer Research UK.

Overview of the London Marathon, Wikipedia, January 2017

Latent Semantic Analysis

- Latent Semantic Analysis can be seen as a method for automatically discovering topics in a corpus
- $W = USV^T$
- In LSA the topic vectors are the columns of V
- So, a topic is described as a document vector
- If d is a document and v_i is the i^{th} “topic” (column of V), then

$$vec(d) \cdot v_i$$

is a measure of the contribution of the i^{th} topic to d

Latent Dirichlet Allocation

- Consider the document *d*:

  "I eat sandwiches in a deck-chair on the sand by the sea" → "eat sandwiches deck-chair sand sea"

- Intuitively *d* is made up of two topics, A and B:

  - A: food, corresponding to "eat" and "sandwiches"

  - B: seaside, corresponding to "deck-chair", "sand" and "sea"
- It looks like *d* is made up approximately of 40% topic A (food) and 60% topic B (seaside)

Latent Dirichlet Allocation

- According to LDA, *d* might be generated as follows:
  - Decide number of topics: *N*=2 “food” (A) and “seaside” (B)
  - Decide the document length: *M*=5
  - Decide the prior probabilities of the topics:

$$P_T(A) = 0.4, \quad P_T(B) = 0.6$$

  - For *i*=1 to *M*
    - Choose the topic T_i randomly according to P_T
    - Choose word w_i randomly according to $P(w|T)$

Latent Dirichlet Allocation

- So, according to this model the document *d* was generated as follows:
  - i=1, T_1 = A (“food”), w_1 = “eat”
  - i=2, T_2 = A (“food”), w_2 = “sandwiches”
  - i=3, T_3 = B (“seaside”), w_3 = “deck-chair”
  - i=4, T_4 = B (“seaside”), w_4 = “sand”
  - i=5, T_5 = B (“seaside”), w_5 = “sea”

Latent Dirichlet Allocation

- This is simple because we know the two topics and their associated word probability distributions
- Given a corpus *C* and a number of topics *N*, a much bigger problem is to derive a set of *N* topics that cover *C* in some optimal sense
- This is the clever part of LDA
- LDA uses an "E-M" type algorithm to do this

Latent Dirichlet Allocation

- Basically:
  1. Make an initial estimate of *N* topics (remember, a topic is just a probability distribution over words)
  2. Decompose each document in *C* into its component topics
  3. Use this decomposition to re-estimate the topic word probability distributions
  4. Go back to 2.

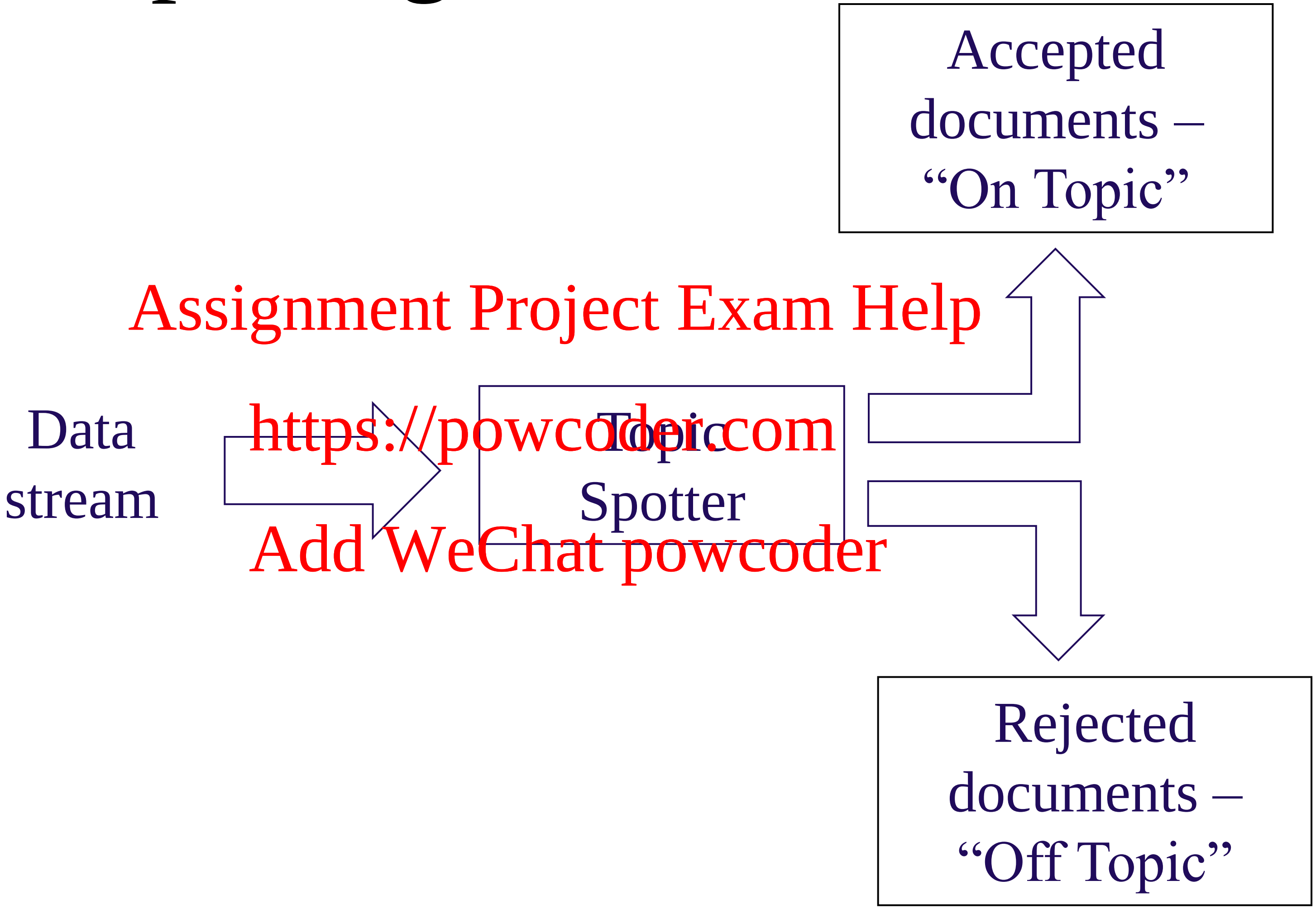
Latent Dirichlet Allocation

- See Edwin Chen's blog "Introduction to Latent Dirichlet Analysis" for an explanation
- The method is called "Latent Dirichlet Allocation" because the prior probabilities of the different topics, $P_T(A)$, is assumed to be a Dirichlet distribution

Topic Spotting

- Topic Spotting is a type of 'dedicated' IR
  - The task is to find documents that are about a particular topic
  - Corpus from which data is retrieved is dynamic
- Other examples
  - Detect all weather forecasts in BBC radio 4 broadcasts
  - Find all documents written by Charlotte Bronte
  - Find all requirements in new EU railway legislation
- Topic Spotting vs IR
  - Because a topic is richer than a query we can calculate probabilities $P(t|T)$ and not just $IDF(t)$

Topic spotting



Data stream

Topic Spotter

Accepted documents – “On Topic”

Rejected documents – “Off Topic”

TF-IDF weights

- Recall the definition of the TF-IDF weight for a term *t* relative to a document *d:*

$$w_{t,d} = f_{t,d} \times IDF(t), \text{ where, } IDF(t) = \log\left(\frac{ND}{ND_t}\right)$$

- *IDF*(*t*) indicates how useful *t* is for discriminating between documents
- $f_{t,d}$ ensures that *t* occurs sufficiently often to be useful
- For Topic Spotting we can define a more sophisticated criterion to identify words that are indicative of a given topic

Usefulness

- Given a term ***t*** and a topic ***T***, define the usefulness of ***t*** (relative to ***T***) by:

$$U(t) = P(t|T)\log\frac{P(t|T)}{P(t)}$$

- If $\log\frac{P(t|T)}{P(t)}$ is large t is characteristic of the topic
- If $P(t|T)$ is large, then t occurs sufficiently often "on topic" to be useful for topic spotting

Usefulness and IDF

- Recall $IDF(t) = \log\left(\frac{ND}{ND_t}\right)$

- Given a set of documents *S*, write $S = S_t \cup S_{t'}$, where S_t is the set of documents that contain *t* and $S_{t'}$ is the set of documents that don't contain *t*

- Then $P(t) = P(t \mid S_t)P(S_t) + P(t \mid S_{t'})P(S_{t'})$

$$= P(t \mid S_t)P(S_t) = P(t \mid S_t)\frac{ND_t}{ND}$$

Usefulness and IDF

- Hence

$$\frac{P(t \mid S_t)}{P(t)} = \frac{ND}{ND_t}, \text{ and } IDF(t) = \log\left(\frac{P(t \mid S_t)}{P(t)}\right)$$

Usefulness and IDF

- $IDF(t) = \log\left(\frac{P(t \mid S_t)}{P(t)}\right)$ is a measure of how useful the term *t* is for general information retrieval (or for retrieving documents that contain it?)

- So, $\log\left(\frac{P(t \mid T)}{P(t)}\right)$ is a measure of the usefulness of *t* for retrieving documents about topic *T*

‘Salience’

- Similarly, given a term ***t*** and a topic ***T***, define the salience of ***t*** (relative to ***T***) by:

$$S(t) = P(T \mid t) \log \frac{P(T \mid t)}{P(T)}$$

- Using Bayes’ Theorem (below) it is easy to establish a relationship between salience and usefulness

$$P(T \mid t) = \frac{p(t \mid T)P(T)}{p(t)}$$

Salience and Usefulness

$$S(t) = P(T \mid t)\log\left(\frac{P(T \mid t)}{P(T)}\right)$$

$$= \frac{p(t \mid T)P(T)}{p(t)}\log\left(\frac{p(t \mid T)P(T)}{P(T)p(t)}\right)$$

$$= \frac{P(T)}{p(t)}\,p(t \mid T)\log\left(\frac{p(t \mid T)}{p(t)}\right) = \frac{P(T)}{p(t)}U(t)$$

Salience and Usefulness

$$S(t) = \frac{P(T)}{p(t)} U(t)$$

- Now, *T* is the topic, so *P*(*T*) is fixed. Therefore

$$S(t) \propto \frac{1}{p(t)} U(t)$$

Salience and Usefulness

- So, main difference between Salience and Usefulness is that to have high usefulness, a term must occur frequently
- Sometimes this means that the most useful words for a topic are not the ones you would immediately suspect:
  - E.G. For Weather Forecast spotting, "north", "south", "east" and "west" turned out to be more 'useful' than "rain" and "sun" – why?

Example

- A term *w* occurs:
  - t_1 times in documents about topic *T*
  - t_2 times in documents which are not about topic *T*
- Total number of terms:
  - in documents about topic *T* is N_1
  - in documents not about topic *T* is N_2
- The corpus contains C_1 documents about *T* and C_2 documents not about *T*
- Then
  - $P(w|T) = t_1/N_1$, $P(w|not\text{-}T) = t_2/N_2$
  - $P(w) = P(w|T)P(T) + P(w|not\text{-}T)P(not\text{-}T)$

$$= \frac{t_1C_1}{N_1(C_1+C_2)} + \frac{t_2C_2}{N_2(C_1+C_2)} = \frac{t_1N_2C_1 + t_2N_1C_2}{N_1N_2(C_1+C_2)}$$

Example

- A term *w* occurs:
  - $t_1 = 150$ times in documents about topic *T*
  - $t_2 = 230$ times in documents which are not about topic *T*
- Total number of terms:
  - in documents about topic *T* is $N_1 = 12{,}500$
  - in documents not about topic *T* is $N_2 = 23{,}100$
- Suppose that only 10% of documents are "on topic"
- So:
  - $P(w|T)=0.012$, $P(w)=0.0102$, $\log(P(w|T)/P(w)) = 0.0706$
  - $U(w) = 0.000847$
  - $S(w) = (P(T)/P(w))\times U(w) = (0.1/0.0102)\times 0.000847 = 0.0083$

Application to Topic Spotting

1. Start with a training corpus of documents $d_1, \ldots, d_N$. Each d_n could be a separate document, or a section (e.g. paragraph) from the same document.
2. For each n decide whether d_n is on-topic (T) or off-topic (not-T)
3. Apply stemming and stop-word removal if required
4. Identify the set of terms (the vocabulary) in the corpus: $w_1, \ldots, w_V$.
5. For each v, calculate $U(w_v)$ – the usefulness of w_v for the topic T.

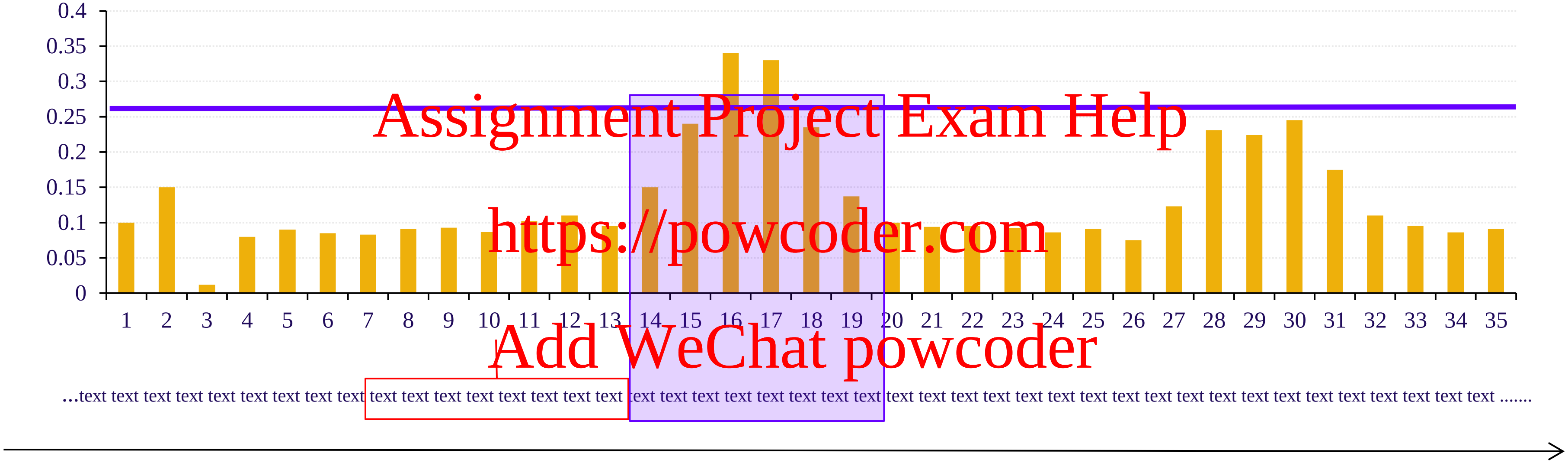
Application (continued)

6. If required, choose a threshold *X* and discard any terms with usefulness less than *X*
7. For each document d_n in the training set:
   - Let $v_1,\ldots, v_{I(n)}$ be the terms in d_n
   - Calculate $AU(d_n) = \frac{1}{I(n)}\sum_{i=1}^{I(n)} U(v_i)$
   - $AU(d_n)$ is the average usefulness of terms in d_n

Application (continued)

8. For a threshold *W* define a classification rule by:
   - If $AU(d_n) > W$, then d_n is classified as “topic”
   - If $AU(d_n) \leq W$, then d_n is classified as “not-topic”
9. Choose a suitable value of *W* using the training documents. For example *W* could correspond to the Equal Error Rate
10. Classification: Given a new document *d*:
    - Calculate $AU(d)$
    - Classify *d* as “topic” if $AU(d) > W$, otherwise d is “not-topic”

Topic Spotter



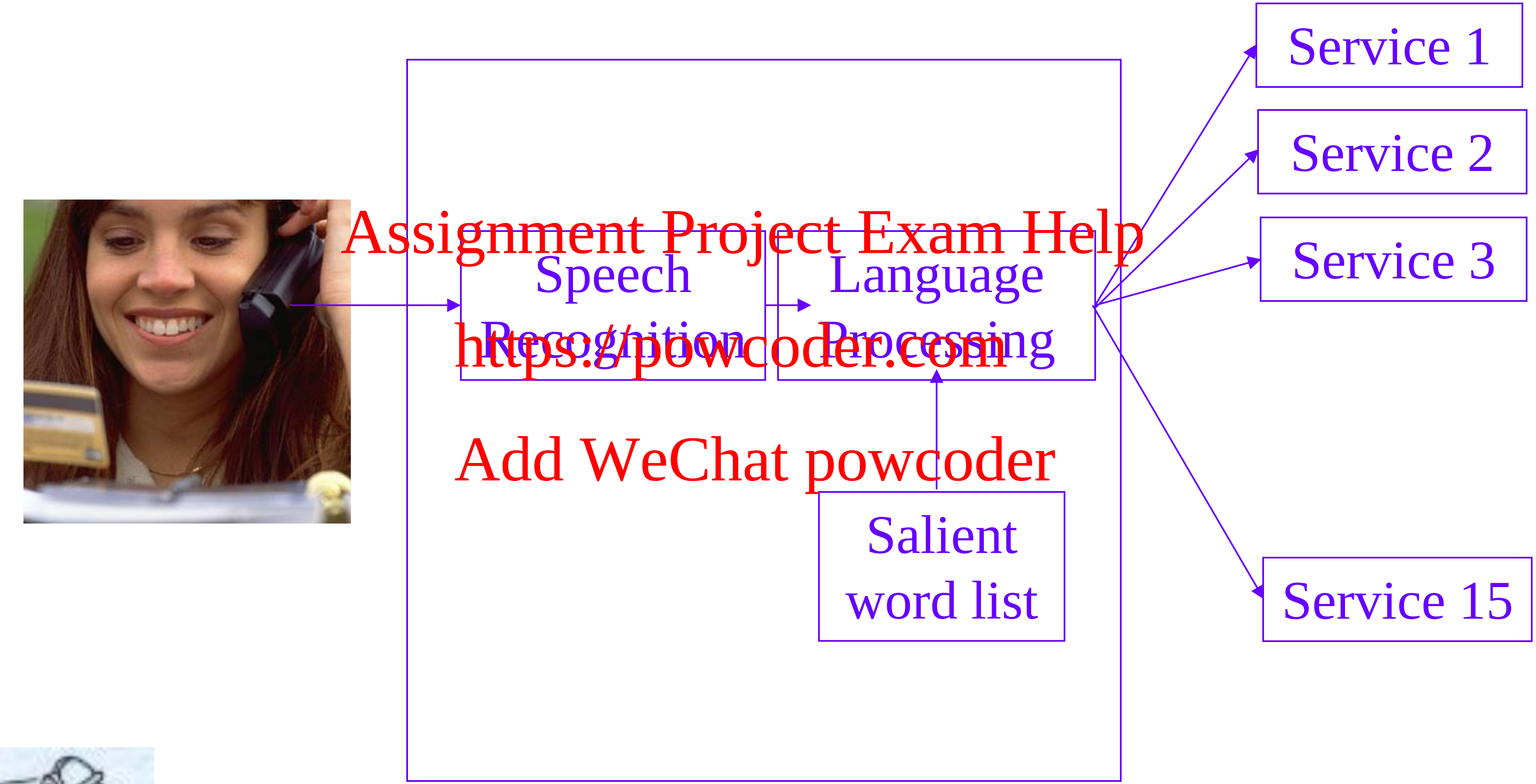
Spotting topics in speech

- First convert audio stream into a text stream using automatic speech recognition
- Consider overlapping sections of text corresponding to, say, 30 seconds of speech (depends on the application)
- Calculate the Average (or Total) Usefulness or Average (or Total) Salience of words in the section of text for the topic
- Signal whenever this value exceeds a threshold

Example

- The AT&T "How May I Help You?" system
- Task: to understand what AT&T customers' messages are about sufficiently well to connect them to the correct service
- Services can be human operators (who deal with a specific problem or speak a specific language) or automated services.
- Look HMIHY? Up on the web

AT&T How May I Help You?



Speech Recognition

Language Processing

Salient word list

Service 1

Service 2

Service 3

Service 15

AT&T How May I Help You?

- HMIHY? Treats telephone network services as <u>topics</u> or <u>documents</u>, to be detected or retrieved
- Example salient words:

Word	Salience		Word	Salience
Difference	4.04		Dialed	1.29
Cost	3.39		Area	1.28
Rate	3.37		Time	1.23
Much	3.24		Person	1.23
Emergency	2.23		Charge	1.22
Misdialed	1.43		Home	1.13
Wrong	1.37		Information	1.11
code	1.36		credit	1.11

Allen Gorin, "Processing of semantic information in fluent spoken language, Proc. ICSLP 1996

HMIHY Demonstrations

- See http://www.research.att.com/~algor/hmihy/samples.html

Summary

- Topics
- Modelling a document as a mixture of topics
- Latent Dirichlet Allocation
- Topic spotting
- Salience and usefulness
- How May I Help You?