Information Retrieval and Dialogue Systems

IR for building interactive dialogue systems
IR for analysis of interactions
interactive IR

Anton Leuski
What’s Information Retrieval?

- Google?

- Topics in IR
  - tasks: search, filtering & routing, topic detection & tracking, question answering,
  - content: cross-lingual IR, multimedia IR (images, speech, video, music), genes
  - applications: web, enterprise, desktop, legal

- IR is about matching text with text
  - or text with images, or text with video, or ...
What’s Dialogue Systems? An Example:
Dialogue System

• Core: get the correct response

• Given a question, I want the system to reply appropriately
  – we talk about Q&A,
  – it’s not required, but easier to memorize
    ▶ user: question, machine: answer

• How? By example?
  – give machine sample questions and link them to required responses
  – the machine finds the most similar example question to the user’s question
  – the machine returns the corresponding answer
Response Selection

• What does it mean “the most similar question?”
  – Users asks: “What do you do here girls?”
  – Questions to compare:
    ▶ “What do you do here?”,
    ▶ “What are your names?”,
    ▶ “What can I do here?”,
    ▶ “Do you like working here?”,
    ▶ “What’s your job?”
    ▶ ...

• The problem of text content matching

• Assumption:
  – The meaning is in the words. The more words in common two texts have, the more similar is the content!
What is this about?

- 16 × said
- 14 × McDonalds
- 12 × fat
- 11 × fries
- 8 × new
- 6 × company french nutrition
- 5 × food oil percent reduce taste Tuesday
- 4 × amount change health Henstenburg make obesity
- 3 × acids consumer fatty polyunsaturated US
- 2 × amounts artery Beemer cholesterol clogging director down eat estimates expert fast formula impact initiative moderate plans restaurant saturated trans win
- 1 × ... added addition adults advocate affect afternoon age Americans Asia battling beef bet brand Britt Brook Browns calorie center chain chemically ... crispy customers cut ... vegetable weapon weeks Wendys Wootan worldwide years York
McDonald's slims down spuds
Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.


Copyright © James Allan
Text Processing

- **Tokenization**: break utterance into words
  - “what”, “do”, “you”, “here”, “girls”

- **Stemming**: find roots
  - “girls” = “girl”

- **Stopping**: remove frequent words
  - “what”

- **Ignore word order**

- **Result**: “bag of words” that we convert to “term vectors”

<table>
<thead>
<tr>
<th>can</th>
<th>do</th>
<th>girl</th>
<th>here</th>
<th>i</th>
<th>is</th>
<th>job</th>
<th>you</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>db</td>
<td>girl</td>
<td>here</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>ydu</td>
</tr>
<tr>
<td>0</td>
<td>db</td>
<td>0</td>
<td>here</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>ydu</td>
</tr>
<tr>
<td>cân</td>
<td>db</td>
<td>0</td>
<td>here</td>
<td>î</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>i$</td>
<td>job</td>
<td>ydu</td>
</tr>
</tbody>
</table>
More text processing secrets

• Weighting
  – the more frequent the word is, the more important it is (high term frequency)
  – some words are frequent “by nature”, e.g., “a”, “the” (high document frequency)

  \[
  w_{i,j} = \begin{cases} 
  1 & \text{word } i \text{ is present in string } j \\
  0 & \text{otherwise}
  \end{cases}
  \]

  \[
  w_{i,j} = tf_{i,j}
  \]

  \[
  w_{i,j} = tf_{i,j}/df_i
  \]

  \[
  w_{i,j} = tf_{i,j}/\log df_i
  \]

  \[
  w_{i,j} = \frac{tf_{i,j}}{tf_{i,j} + 0.5 + 1.5 \frac{doclen}{avgdoclen}} \cdot \frac{\log(\frac{colsiz}{docf_i})}{\log(colsiz + 1)}
  \]

• N-grams
  – “do you work here?” vs. “you do work here”
  – capture ordering
  – bi-grams: “do you”, “you do”, “do here”, “here girl”
  – tri-grams: “do you do”, “you do here”, “do here girl”
Approach 1: Text Classification

- **Answer = class**
  - as a simple label, e.g., “a1”, “a2”, etc.

- **Each sample Question = instance of the class**
  - as a term vector (tf-idf). terms will be called “features” here
  - it deals nicely with multiple questions linked to an answer
    ▶ What do you do here?”
    ▶ “What’s your job?”
    ▶ are linked to “We are Virtual Museum guides”

- **New question: which answer? = which class does it belong to?**

- **A classification algorithm will assign a class to the new question**
  - For example, SVM is a state-of-the-art classification technique and works very well for NL applications
  - Alternatives: naïve Bayes, decision trees, etc.

**Limitations:**
- ignores answer text
- ad-hoc features
- binary classification - cannot easily handle many-to-many relations

Can we do better?
Dealing with Limitation 3: Regression

- Binary classification (SVM, etc.) requires us to select 1 class out of 2.
  - That's a character with 2 answers only!

- Alternative: regression
  - We introduce a similarity function
  - Use the similarity to compare all vectors to the new question
  - Select the best match

- Compare vectors
  - Compare the angles between vectors
  - Cosine \[ \cos(\theta) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \]

- BTW, how do we handle multiple questions linked to the same answer?
Dealing with Limitation 2: Language Models

Question

What do you do here girls?

We are Virtual Museum guides

Our looks are based on a model named Bianca.

Answer 1

“content” A1

Answer 2

“content” A2
**Language Model**

What are your jobs here? 0.2000

What are you here for? 0.1000

What do you know 0.0100

What time is it? 0.0005

...

Probability Distribution

\[ P(W|M_Q) \]
Language Model Comparison

What do you do here girls?

We are Virtual Museum guides.

Our looks are based on a model named Bianca.

\[ P(W|M_{A1}) \]

\[ P(W|M_{A2}) \]
Language Model Comparison in IR

- **Answer (document in IR)**
  - estimate $M_A$: $P(w|M_A)$

- **Question (query)**
  - estimate $M_Q$: $P(w|M_Q)$

- **Compare questions against answers: similarity score**
  - cross-entropy: number of bits to “encode” $M_Q$ with $M_A$

\[
H(M_Q || M_A) = - \sum_w P(w|M_Q) \log P(w|M_A)
\]

- **Rank all answers by the similarity score**
- **Cut the ranking at some threshold**
- **Return the set (or you can return the top ranked answer)**
Estimations

- **Unigram language model**

\[ P(W) = P(w_1 \ldots w_n) = \prod_{i=1}^{n} P(w_i) \]

- **Jelinek-Mercer estimation**
  - Interpolated Maximum-likelihood

\[ P(w|M_A) = \pi_A(w) \]

\[ \pi_s(w) = \lambda_\pi \cdot \frac{\#(w,s)}{|s|} + (1 - \lambda_\pi) \cdot \frac{\sum_s \#(w,s)}{\sum_s |s|} \]

- **Other approaches exist**

- **You can improve the model by creating better estimations!**
Language Models in Question Answering

- IR assumes that language of queries is the same as language of documents
  - a query is like a document - will have common words

- This is incorrect for question answering
  - questions and answers may have no common words:
    - “What do you do here girls?” → “We are Virtual Museum guides.”
  - questions have specific grammar
  - questions are not answers!

- Questions and answers are two different “languages”
Approach 2: Single Language Retrieval

What do you do here girls?

We are Virtual Museum guides.

Our looks are based on a model named Bianca.

$P(W|M_{A1})$

$P(W|M_{A2})$
Approach 2: Single Language Retrieval

What do you do here girls?

We are Virtual Museum guides.

Our looks are based on a model named Bianca.

What do you do here?

Are you a girl?

What do you do here girls?

Training questions
What do you do here?

$P(W|M_Q)$

$P(W|M_{Q_{A1}})$

$P(W|M_{Q_{A2}})$
Approach 2: Single Language Retrieval

- Retrieve a training question, and select the matching answer

- How do we deal with multiple linked questions?
  - document = a training question
  - document = text of all questions linked to a single answer

- Limitation: ignores the answer text (Limitation 1, right?)
Dealing with Limitation 1: Cross-Language IR

• find Chinese documents with English query
  – using a dictionary
  – or, using a “parallel” corpora: a set of the same documents in both languages

• we do treat questions and answers as two different languages
Approach 3: Cross-Lingual Retrieval

What do you do here girls?

$P(W|M_Q)$

LM of the best possible answer

Answer 2 can be returned because of its similarity to Answer 3

$P(A(Q)|M_Q)$

$P(W|M_A1)$

I’m Ada. And, I’m Grace.
We are Virtual Museum guides.

$P(W|M_A2)$

Our looks are based on a model named Bianca.

$P(W|M_A3)$

We are Virtual Museum guides.
Approach 3: Cross-Lingual Retrieval

• **Expected answer model:**
  - If I see some known words in the question, how likely I’d see a particular word in an answer?

\[
P(w|M_Q) = E_{(Q,A)_T} \pi_a(w) = \frac{\sum_{(q,a) \in (Q,A)_T} \pi_a(w) \prod_{i=1}^{|q|} \pi_q(q_i)}{\sum_{(q,a) \in (Q,A)_T} \prod_{i=1}^{|q|} \pi_q(q_i)}
\]

• **Given a question, estimate the LM of the best possible answer, compare to known answers**

\[
H(M_Q||M_A) = - \sum_w P(w|M_Q) \log P(w|M_A) \quad P(w|M_A) = \pi_A(w)
\]

\[
\pi_s(w) = \lambda \pi \cdot \frac{(w, s)}{|s|} + (1 - \lambda) \cdot \frac{\sum_{s} (w, s)}{\sum_{s} |s|}
\]
Approach 3: Cross-Lingual Retrieval
Comparing three Approaches

- **Systems**
  - SVM: support vector machine, tf•idf weighting
  - LM: comparing questions to training questions
  - CLM: “translated” question to answers

- **1261 questions**
- **60 answers**
- **10-fold cross-validation**
- **t-test (p<0.05)**
- **differences are statistically significant**

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>LM</th>
<th>CLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>45%</td>
<td>50%</td>
<td>60%</td>
</tr>
<tr>
<td>Improvement</td>
<td>+8.8%</td>
<td>+16.7%</td>
<td></td>
</tr>
</tbody>
</table>
• **User study**
  – 20 participants
  – 20 questions each
  – appropriateness scale 1-6
  – 400 questions transcribed by ASR and by human transcribers (TRS)

• **Expected appropriateness score**
  – what would be the average appropriateness of the response if WER is below a certain level
Issues

• What to do when the list is empty?
• What to do when the list has more than one answer?
• How do you deal with context?
• Does the system really understand the question?
Semantic Parsing

- “What do you do here girls?” →
  - mapping from text to semantic frames

Two approaches:
- Frame retrieval
  ▶ Pros: well-formed frames
  ▶ Cons: cannot create a new one
- Frame building
  ▶ Retrieving slot-value pairs
  ▶ Cut the ranking
  ▶ Pros: Can construct a new frame
  ▶ Cons: The frame may have problems

\[
P(\sigma|M_S) = \frac{\sum_{(s,f) \in (S,F)_T} \pi_f(\sigma) \prod_{i=1}^{|s|} \pi_s(s_i)}{\sum_{(s,f) \in (S,F)_T} \prod_{i=1}^{|s|} \pi_s(s_i)}
\]
Summary

• Using IR techniques for conducting a dialogue
  – e.g, language understanding
  – but, we also have shown how to learn dialogue strategies using IR-like approaches

• Cross-lingual Language Modeling is
  – effective
  – robust
IR for Interaction Analysis

- MMORPG conversations and activities

- Chat log
  - 01/12/03, 22:22:02, 64, 56, rradeon says, magic helps us out
  - 01/12/03, 22:22:06, 64, 56, rradeon says, plz
  - 01/12/03, 22:22:07, 78, 30, Burto tells everyone, WE need to kill it before he reaches the sacred stones !!!!! or there will be Hell to pay!!
  - 01/12/03, 22:22:08, 64, 56, Agnes says, hey magic fair... help us get to spirit?

- Game log
  - Green Dragon killed! 1/12, 23:22; attending: cool_jaws,
  - Gavron killed! 1/12, 23:22; attending: ThunderBoobs, Bloodwarlord, lord_samot, Warrior Arcana, Infused, Nova,

- Questions:
  - Activity detection: e.g., is there a raid?
  - Player forensics: who is involved? What are their roles?
  - What the relationships among players? Their expertise? Status? Personality?
Interaction for IR
Questions?