Ongoing research on sentence retrieval and novelty detection

Dr. David E. Losada

Grupo de Sistemas Inteligentes
Departamento de Electrónica y Computación
Universidade de Santiago de Compostela (Spain)
http://www-gsi.dec.usc.es/~dlosada
dlosada@dec.usc.es

Outline

The importance of sentence retrieval & novelty

Our current research lines

Language modeling for sentence retrieval: Multiple-Bernoulli distribution.
Hierarchical query-biased summaries
Proximity between relevant sentences and query difficulty
Language modeling for sentence retrieval: Study of smoothing.
The importance of SR & novelty

- To facilitate effective web information access
- To focus Question Answering processes on a set of well selected sentences
- To assist (query-biased) summarization
- To aid Topic Detection and Tracking methods
- Just to name a few ...


- Web searchers typically fail to view results beyond the 1st page
- Doc surrogates can be uninformative and difficult to interpret. Hard to assess the relevance of the returned docs.
The importance of SR & novelty

Focusing on web retrieval...

- Searchers forced to make 2 steps: 1) assess the surrogate.

  *Is this title relevant? Are these terms in the correct context? What comes after the ellipses? Shall I click this title?*

  and 2) analyze exhaustively the doc to locate the relevant material, if any

  [Krish, 2000]: associated cost (time, effort and stress)

---

The importance of SR & novelty

Focusing on web retrieval...

- Why not to present directly the document contents? The required information may be even found directly at the results interface...

- White et al. shift away from surrogates to actual doc content (query-relevant top ranking sentences)

- Encourages a deeper examination of the contents of the doc retrieved set.

- Increased contextual coherence (surrogates are rarely composed of full sentences)

- Highly relevant content from lower ranking docs has now more chance to be viewed.
The importance of SR & novelty

Focusing on web retrieval...

- In [White et al. 2005] user studies for factual searches (e.g. find a named person’s current email address), decision search (e.g. choose the best impressionist art museum) and background searches (e.g. finding information on dust allegies).

- Users do not need the top-ranking sentences for the factual queries.

- But useful for decision and background searches. A general overview of the topic usually needed to make reasonable search decisions. The presentation of top sentences coming from different docs helps to supply the user with a general view on the query subject.

Searchers are fully aware of what they are looking for \( \Rightarrow \) top ranked sentences not needed

Searchers are not fully aware of what they are looking for \( \Rightarrow \) top ranked sentences useful

The ranked sentences also encouraged more page views outside the top 10 docs and a reduced number of query iterations.
The importance of SR & novelty

Focusing on web retrieval...

- In [White et al. 2005], they didn't apply any method to filter out redundant sentences

Our research lines

- Language modeling for sentence retrieval: Multiple-Bernoulli distribution.
- Hierarchical query-biased summaries
- Proximity between relevant sentences and query difficulty
- Language modeling for sentence retrieval: Study of smoothing.
Motivation

- Multinomial (MN) vs multi-variate Bernoulli (MB) for sr.
- The pioneering LM proposal was based on a query generation process modelled by a MB distribution ([Ponte & Croft, 1998]).
- But, following that, MN models became popular
- In general, there is no good reason for choosing MB (see e.g. [Metzler et.al. 2004]).
- However, the granularity of sr and its particular characteristics could be suitable for a MB approach

Language Modeling

MN and MB can be formally analyzed in the context of Bayesian Statistics:

\[ P(\theta_D|D) = \frac{P(D|\theta_D)P(\theta_D)}{P(D)} \]

- \( P(\theta_D) \): prior belief about the adequacy of the distribution \( \theta_D \)
- \( P(D|\theta_D) \): likelihood of the data \( D \) under distribution \( \theta_D \)
- \( P(\theta_D|D) \) is the posterior distribution.
- \( P(D) \): prob. of generating the doc. It is independent on \( \theta_D \).
Language Modeling

Applying MAP ...

- $P(D|\theta_D)$ (MN) and $P(\theta_D)$ (Dirichlet) leads to a posterior distribution which is also Dirichlet and...

$$\hat{\theta}_i = P(w_i|\widehat{\theta}_D) = \frac{t_{f,i,D} + \alpha_i - 1}{|D| + \sum_{i=1}^{|V|} \alpha_i - |V|}$$

- $P(D|\theta_D)$ (MB) and $P(\theta_D)$ (Multiple Beta) leads to a posterior distribution which is also Multiple Beta and...

$$\hat{\theta}_i = P(w_i|\widehat{\theta}_D) = \frac{\delta_{i,D} + \alpha_i - 1}{\alpha_i + \beta_i - 1}$$

Query likelihoods

- Standard unigram LM

$$P(Q|\widehat{\theta}_D) = \prod_{w_i \in Q} P(w_i|\widehat{\theta}_D)^{qtf(w_i)}$$

- MB likelihood

$$P(Q|\widehat{\theta}_D) = \prod_{w_i \in Q} P(w_i|\widehat{\theta}_D) \prod_{w_i \not\in Q} (1 - P(w_i|\widehat{\theta}_D))$$

Different space of events (binary vectors, such as in BIM)
Product across non query terms. Kind of off-topic correction.
Sentence retrieval

- Lack of a non-binary tf component in MB seems less important
- MB takes into account the non-query terms:
  - the terms in the sentence (especially those ones having \(P(w_i|\theta_S)\) high) which are missing in the query text ⇒ penalty in the retrieval score.
  - Intuition: The sentence will probably deviate from the query topic.

Sentence retrieval

- But MB is not efficient for doc retrieval, why?
  - Docs are usually multi-topic whereas sentences deal with a single topic.
  - MB selects sentences very focused on query topics.
  - In doc retrieval, most relevant docs will mention many non-query terms
  - The lack of non-binary tf is undoubtedly an issue for doc retrieval
Experiments

(Main findings...)

- MB was always better than (or at least as good as) MN
- MB is more stable w.r.t the smoothing levels
- In most of the cases the MB performance was significantly better than the MN performance (> 10%)

Our research lines

- Language modeling for sentence retrieval: Multiple-Bernoulli distribution.
- Hierarchical query-biased summaries
- Proximity between relevant sentences and query difficulty
- Language modeling for sentence retrieval: Study of smoothing.
Hierarchical query-biased summaries

Joint collaboration

University of Santiago de Compostela
University of Strathclyde

Grupo de Sistemas Inteligentes
Information Access lab

Fabio Crestani & Simon Sweeney

Summarization with novelty detection

Two basic aims:
- Incremental length summaries vs fixed length summaries
  (interesting e.g. in WAP mobile phones)
- Incorporating novelty detection does really help?
Hierarchical query-biased summaries

- Whilst summarisation paired with novelty detection is not a new concept, we are concerned with the mechanism of delivery.

- Is there an optimal strategy for showing summaries in response to the request to 'show me more'?

- In previous work we took 'more' to mean an increase in summary length.

- An intuitive approach 'more' as a function of the summary length and information content.

Hierarchical query-biased summaries

- Compare user groups performance with both systems (increasing vs constant length), and baseline systems that do not use novelty.
Hierarchical query-biased summaries

Research questions...

1. Do query-biased summaries that take account of novelty (SumN) perform better or worse than those without novelty (SumB)?
2. Do query-biased summaries that have a fixed, or constant length (Sumc) perform better or worse than those with an increasing length (Suml)?
3. Which of the summary configurations (SumNL, SumNc, SumBL, SumBc) achieves the highest level of performance?

1. Start from a rank of sentences in decreasing similarity to the query (e.g. [Tombros & Sanderson 98] )
2. Top X sentences produce the 1st level summary (re-ordered as they appear in the doc)
3. If the user wants to see more...
User studies ongoing to test the 4 alternatives.

Given the summaries test the ability to identify correctly relevant documents.

User groups are shown summaries 2 out of the 4 configurations.

The first summary shown (at level 1) is generic, and the same for all users.
Hierarchical query-biased summaries

- To generate the novel summaries...
  - Sentences have a relevance score (e.g. [Tombros & Sanderson 98]) and a novelty score
  - Novelty score measures how novel they are with respect to the previously seen summaries (e.g. wordsSeen)

Our research lines

- Language modeling for sentence retrieval: Multiple-Bernoulli distribution.
- Hierarchical query-biased summaries
- Proximity between relevant sentences and query difficulty
- Language modeling for sentence retrieval: Study of smoothing.
Proximity and query difficulty

Many questions but very few answers...

- How to retrieve relevant sentences from a set of top retrieved docs?
- Many different methods tried out in the context of the TREC novelty tracks (2002, 2003, 2004). VSP, LMs (2-stage, KLD, ...)
- A regular tf/idf technique works consistently better than any other approach [Allan et al 03].

Some recent (tiny) improvements for sentence retrieval using named entities, phrases and combinations of query words [Li & Croft 05]

Hence, it is a challenging problem and an effective solution is still to come.
Proximity and query difficulty

- There is some evidence that users tend to locate relevant sentences in close proximity one to each other [collins-thompson et al 02] (CMU)
- CMU applied a window of nearby sentences (2-3 sentences before and 2-3 sentences after) to adjust the tf/idf score of a sentence.
- It didn’t improve sentence retrieval

Anyway, there is a lack of exhaustive reports addressing this issue...

- Aim: test different operators to check whether or not a combination with nearby sentences is good.
- For each sentence, we only consider the sentence before and the sentence after.
- Regular tf/idf as a baseline and proximity method for re-ranking the sentences.
Proximity and query difficulty

We first conducted some preliminary analytical study to check the working hypothesis (relevant sentences are in close proximity one to each other).

Very different datasets: TREC-2002 (old TREC topics, very few relevant sentences), TREC-2003/4 (new topics, AQUAINT collection). But...

| Year | \( P(R) \) | \( P(R|\text{prev is rel}) \) | \( P(R|\text{next is rel}) \) | \( P(R|\text{prev & next are rel}) \) |
|------|----------|----------------|----------------|----------------|
| 2002 | 0.024    | 0.285          | 0.281          | 0.660          |
| 2003 | 0.391    | 0.802          | 0.790          | 0.928          |
| 2004 | 0.159    | 0.572          | 0.561          | 0.766          |

Relevant sentences tend to occur nearby.

But how to come out with an effective sentence retrieval method able to handle proximity?
Proximity and query difficulty

So far, we tried out...

1. \( rsv(s_i) = \lambda \text{sim}(s_i) + (1 - \lambda) \frac{\text{sim}(s_{i-1}) + \text{sim}(s_{i+1})}{2} \)

2. \( rsv(s_i) = \lambda \text{sim}(s_i) + (1 - \lambda) \frac{\text{sim}(s_{i-1}) + \text{sim}(s_{i+1})}{2} \) (only \( \text{sim} \)s\( <> \)0 are considered)

3. \( rsv(s_i) = \lambda \text{sim}(s_i) + (1 - \lambda) \max(\text{sim}(s_{i-1}), \text{sim}(s_{i+1})) \)

4. \( rsv(s_i) = \min(\text{sim}(s_i), \lambda \text{sim}(s_i) + (1 - \lambda) \frac{\text{sim}(s_{i-1}) + \text{sim}(s_{i+1})}{2}) \)

5. \( rsv(s_i) = \min(\text{sim}(s_i), \lambda \text{sim}(s_i) + (1 - \lambda) \max(\text{sim}(s_{i-1}), \text{sim}(s_{i+1}))) \)

- Last two methods to avoid that a sentence with high initial score gets significantly penalized when there are low score surrounding sentences.

Proximity and query difficulty

- Evaluated for both long and short queries
- Main evaluation ratios: F measure (std metric in the novelty track), P@10 and P@5
- No major difference among proximity methods
- Small average improvements in performance (but most of them are not stat. significant).
- Anyway, in most of the cases, the number of queries whose performance is improved w.r.t the baseline is larger than the number of queries whose performance is decreased
Proximity and query difficulty

- Is there any query feature that helps to adjust the proximity-based methods?
- Correlation between sentence retrieval performance and query difficulty measures? [He & Ounis, 04]
  - Average inverse collection term frequency (ICTF)
  - Query scope
- Suitable for predicting trends in sentence retrieval?
- Adequate for adjusting the proximity-based approach?

---

- No correlations were found between F and ICTF/Query scope (still need to check P@5 and P@10)
- Proximity methods tend to work better when the avg rsv of the retrieved set of sentences is high
- Conclusion: Rel sens tend to be close to each other but an effective proximity-based SR method is still to come
Our research lines

- Language modeling for sentence retrieval: Multiple-Bernoulli distribution.
- Hierarchical query-biased summaries
- Proximity between relevant sentences and query difficulty
- Language modeling for sentence retrieval: Study of smoothing.

Smoothing for SR

- Joint collaboration
  - University of Santiago de Compostela
  - University of A Coruña
  - Grupo de Sistemas Inteligentes
  - IRLab

Smoothing for SR

- Re-examine smoothing strategies in the context of a sentence retrieval problem.

Conclusions

- Multiple-Bernoulli LMs look promising for sentence retrieval
- Effective novelty techniques at the sentence level are promising for improving current doc summarization methods
- Relevant sentences tend to be close one to each other but still don’t know how to effectively model this fact