

# Rookie NLP spinoff notes/draft

Abram Handler

March 21, 2016

## 1 Potential introduction

Current algorithms for query-focused multi-document summarization are too slow for interactive systems. At worst, exact linear programming approaches require hundreds of seconds to find a summary [mcdonald2007study]. At best, approximate MMR methods make a linear pass over the data, but still require seconds of real time [mcdonald2007study]. Yet in interactive systems, a user will perform many queries and create many filtered subsets of query-responsive documents. SOMETHING FROM HCI ON ACCEPTABLE RESPONSE TIME THRESHOLDS [doherty1982economic] This calls for new, anytime approaches, which attempt to provide the best-possible summary within some real time threshold of acceptable delay.

## 2 Events definition

Filatova et. al use “events” to create summaries. Copying this approach requires a definition of an “event”. Seems reasonable to say “event” is defined by the particular task at hand – any other claim requires serious linguistics or cognitive psychology. One of their papers assumes 1 sentence = one event.

**Rookie-like idea:** noun phrase clusters? I suspect that finding high tf-idf noun phrases in sentences that contain Q or F or both will sketch out an event. Q = “Mitch Landrieu”, F = “Marlin Gusman”. High tf-idf noun phrases: “Department of Justice,” “Orleans Parish Prison”

**Temporal and spatial granularity.** Event based summarization for news has to deal with temporal and spatial granularity. Some of the Hovy papers address this. (1) The US has been fighting in Iraq since 2003. (2) In October, 2010 there was an ambush of US forces outside Fallujah. News reader project addressed this.

**Head words** Read Nate Chambers. more linguistics.

## 3 Form of summary

DUC and TREC summarization competitions have long focused on a particular form of single or multi-document summary. However, one could imagine that different choices about white space or text layout might make for more or less useful summaries. These choices about white space would in turn impact the choice of algorithm.

1. **Search engine results** A form of summary.

2. **Traditional summary (DUC/TREC)** I think: no line breaks. I think the goal is a human readable summary, like what an analyst would sit down to write. DUC website requires login.
3. **Bullet point summary** One short point per line. With a temporal corpus, these could be events. But for scientific summarization, each bullet point might represent a particular approach or theory. I think this is most suitable for Rookie.
4. **KWIK viewer summary**
5. **Sentence selection:** one sentence per line.
6. **Sub-sentential clause or phrase selection:** one phrase per line, both with or without line breaks

## 4 Current Rookie

Current Rookie, 3/15/16, takes approach #5 defined in section 3. After a search engine returns a list of documents in response to a query, Rookie places each sentence from each of the query-responsive documents,  $R$  (results), into exactly 1 of the following categories:

1. Sentence contains  $Q \wedge F$ <sup>1</sup>
2. Sentence contains  $Q \vee F$
3. Sentence  $\neg$  contains  $Q \vee F$

Each document will contribute  $N$  eligible sentences for potential inclusion in a multi document summary. Rookie chooses greedily from categories 1, 2, and 3 (in that order) until it finds the  $N$  sentences from a document. Rookie uses sentence position to break ties within a category, within a document. The set of all eligible sentences from  $R$  is called  $E$ .

Not all eligible sentences from  $E$  are displayed in the summary. Rookie’s UI has a “summarization area” that is  $P$  pixels tall. The  $P$  pixels are divided into horizontal **stripes**. Each stripe,  $s_i$  is filled with one sentence. The sentence is allowed to wrap within its stripe, so the height of one stripe  $s_i$  may or may not equal the height of another,  $s_j$ .<sup>2</sup> Rookie picks sentences from  $E$  to fill the stripes until the height of the chosen stripes + the height of the next proposed stripe is greater than  $P$ . Rookie chooses sentences from  $E$  by picking greedily from categories 1, 2 and 3 (in that order). If a category contains more than one eligible sentence, Rookie chooses randomly.

This is not an anytime algorithm, but it is imperceptibly fast, even w/ React overhead. It could be turned into an anytime algorithm.

HCI + NLP note: this algorithm was designed to support “interactive summarization” along the axis of time. That is: a user might be interested in a particular ambush in Fallujah 2010, which is part of a broader US invasion from 2003-2015. “Events-based summarization with context?” Matching an event to its context seems like an unexplored problem. Pretty hard.

---

<sup>1</sup>If no  $F$  is selected, no sentences will be placed into category 1.

<sup>2</sup>Informally, the sentences in the summarization area should reflect important information about  $Q$  and  $F$  in  $T$ , although this notion is poorly defined.

## 5 Paper ideas & related work

### 5.1 Clause dropping and human readability

Pittler 2010 defines sentence compression as producing a grammatical, shorter version of the original. Addition: compression with a set of objectives [Q, F, F' ]

## 6 Evaluation

Brendan 3/15: I think we'll need to compare alternative summaries side-by-side and think about what information they convey, relative to the user's expressed information need (which I think is the (Q,F,T) tuple). then eventually operationalize our criteria into an mturk-able (or otherwise annotatable) task.

## 7 Results

Brendan 3/15: The failure of "naive multidoc snippets", if I understand it correctly, is a good finding – it's a great argument for joint inference being important, since joint inference helps ensure the selected text spans give the right context for each other.

## 8 A "judgments" evaluation

One use case for summarization: should I read this? A good summary helps you make the correct choice about reading (as opposed to best reflecting the contents of the document). If you have a particular reason for being interested in a document (a Q or F) then maybe these are different goals. There are lots of IR corpora about "judgments" (i.e. and expert says which docs are relevant or not relevant), which might be a way to approach this.

## 9 Related work

Filatova

Toward Abstractive Summarization Using Semantic Representations

Event-Centric Summary Generation

Talk to Emma and Luke who wrote a paper on this.