HexTac: the Creation of a Manual Extractive Run

Pierre-Etienne Genest
Guy Lapalme
Mehdi Yousfi-Monod

RALI, Université de Montréal

November 16 2009

Outline

Introduction

What is a manual extractive run ?

- Extractive summaries written by humans
- Pure sentence extraction from source documents
- No editing, no re-writing, no compression

Why create a manual extractive run ?

- Determine how well/poorly extracts perform compared to abstracts
- Verify if human extracts beat the best current systems
- Provide a model similar to current systems
- Applications in domains which require quotable summaries

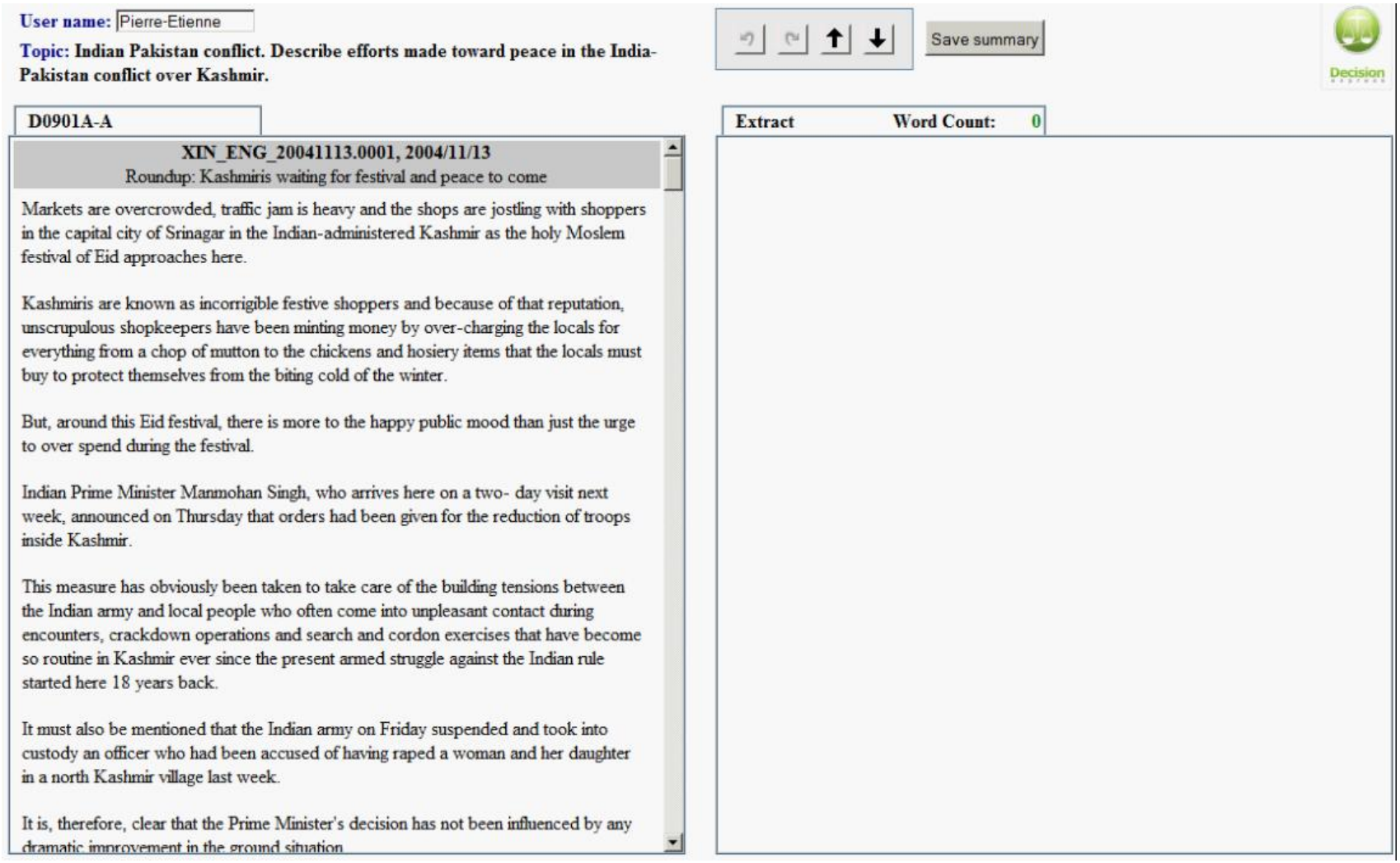
Experimental Context of HexTac

- 5 human extractors
- 88 extractive summaries (18 each on average)
- 3.88 extracted sentences per summary on average
- 20 minutes spent to complete a summary on average
- 30 man-hours of work in total

Interactive Human Extraction Interface

- Make manipulating sentences during extraction easy and fast
- Guarantee pure sentence extraction and respect of 100-word limit
- Automatize data handling
  - Loading clusters, saving completed summaries
- Keep records of metadata
  - User name, date, time spent, ID of selected sentences

Interactive Human Extraction Interface



Guidelines for Human Extractors

1. Begin with part A
2. Always read the topic and all 10 articles
3. Extract sentences that answer the topic and summarize the cluster
4. Favor sentences that can be understood on their own
5. Respect the limit of 100 words
6. Maximize the information content
7. Re-order the sentences of the extract to improve readability
8. Complete part B the same way immediately after part A
9. Avoid repetition of information that appears in cluster A

Feedback from HexTac participants

- Thankful for the interface
  - It saves time and helps with the task
- Frustration at the unability to make modifications
  - Solving referential clarity problems (time, person)
  - Removing a few words
- Difficulty to choose how to answer list-like topic requests
- Difficulty with the tradeoff between content and linguistic quality
- Boredom, repetiveness of the task

Results in TAC 2009

Part A	Pyramid Score	Linguistic Quality	Overall Responsiveness
Abstracts	0.683	8.915	8.830
HexTac	0.352	7.477	6.341
Best System	0.383	5.932	5.159

Part B			
Abstracts	0.606	8.807	8.506
HexTac	0.324	7.250	6.114
Best System	0.307	5.886	5.023

- Manual extraction performs better than any automatic system
- Pure extraction performs very significantly worse than abstraction
- Greatly superior linguistic quality, even with pure extraction
- Shows room for improvement in automatic sentence extraction models

Inter-Extractor Agreement

- Based on 12 additional, redundant extractive summaries
- Very low inter-extractor agreement
- Roughly 15% sentence-agreement between human extractors
- Widely varying scores between extractors, though using a small sample

	Pyramid Score	Linguistic Quality	Overall Responsiveness
HE1	0.278	8.222	7.556
HE2	0.297	7.611	5.333
HE3	0.340	7.000	5.917
HE4	0.378	7.583	7.125
HE5	0.392	6.063	4.125

HexTac as a ROUGE model

- HexTac-ROUGE
  - ROUGE score of a system using HexTac as the model

- Similarity with other metrics

Correlation coefficients	Part A	Part B
HexTac-ROUGE–ROUGE	0.80	0.85
HexTac-ROUGE–Overall Responsiveness	0.78	0.91
ROUGE–Overall Responsiveness	0.97	0.94

- A less costly alternative to other evaluation metrics ?
  - Made from only one manual "run"
  - Extracts easier to make than abstracts

Conclusion

HexTac

- Successful, reusable methodology to manual extraction
- Requires 30 man-hours for 88 extracts using an interactive interface
- Approximation of an upper-bound on purely extractive summarization
  - Better extracts/extractors definitely exist

Food for thought

- A tool for supervised training of sentence selection ?
- Manual sentence ranking / sentence evaluation, the next step ?
- Necessity of a linguistically richer approach than sentence selection to achieve significant improvements ?

Questions ? Comments ?