Challenges to Document Recognition and Retrieval

Workshop - SIGIR 2002

Tampere Finland

Paul B. Kantor, Rutgers.

Differing Goals

- Document Recognition -- to identify what a document ***is*** which may mean
  - – layout
  - – content
  - – image/music/play/report/phone_book ….
- Difficult -- usually done by humans.
  - – Weibel at OCLC -- reverse engineer Book Title pages into T$_E$X.

Differing Goals

- Document Recognition OCR
  - to see what the specific characters (notes; chemical diagrams, floor plans, engineering diagrams ….) of a document are
- Difficult for text
- Enormously difficult for non-text.

Differing Goals

- Document Retrieval (**not** Information Retrieval, Question answering, summarization, ….)
  - put the “most likely to be useful” documents into the hands of the human who wants to use them.
  - Quality is conditioned on both aspects of the documents, and the purpose for which it is retrieved

Differing Goals

- Recognition
  - entity typing correct
  - characters correct
  - layout features correct
  - entity identified (tiger burning bright, or burning tire)

- Recall/ Precision/
  - (a) much of the useful documents found
  - much of the found documents useful
  - (a) is usually just estimated.

The challenge (*my version)*

- Which aspects of the Recognition criteria are important in achieving the Retrieval criteria.
- Reconstruction 1: “On Aug**v**st 15, President Ge**a**rge W. Bush visited the SIGIR Conference in Tampere Finlan**g**”.
- Reconstruction 2: “On Aug**v**st 15, **R**resident Ge**a**rge W. Bu**5**h visited the SIGI**P** Conference in Tamp**o**re Finlan**g**”.

It depends on how the questions are asked:

- Reconstruction 1: “On Aug**v**st 15, President Ge**a**rge W. Bush visited the SIGIR Conference in Tampere Finlan**g**”.
- Find documents that relate **President Bush** to the **SIGIR Conference** in **Tampere** this year.
- Excellent match. But not so good for
- Reconstruction 2: “On Aug**v**st 15, **R**resident Ge**a**rge W. Bu**5**h visited the SIGI**P** Conference in Tamp**o**re Finlan**g**”.

Classification of Errors

- Errors that have no effect on retrieval logic.
  - Errors of white space
  - errors in stop words
- Errors that can probably be corrected
  - misspelled common words (not proper names)
  - misspellings when correct form exists elsewhere in document

Two kinds of questions

- Questions of method
  - are there methods in OCR that might have correlates or translations to IR -- is there a “de-skewing”
  - could OCR type techniques for layout -- eg. spanning large parts of text in defining structure, be applied to texts
  - are there IR methods that can help OCR - most methods are linear classifiers in features such as word.

Method (2)

- – Relevance feedback (even pseudo relevance feedback) is very valuable -- is there a corresponding concept applicable to OCR
- – statistical methods used for guessing broken characters --- edit distance, etc. -- can they be helpful in IR

relations

- questions of measure
  - character level in OCR
  - document level in IR
  - can moving to an IR type overall measures lead to different assessment of OCR effectiveness -- trees versus forest

Why do we care?

- Huge cultural legacy (what makes us civilized) in print form
  - ought to be available on the desktop (dining room table)
  - if it is not, our children will not look for it
  - the web must become the world's libraries together -- IR access to image files

Care more

- Legacy print materials that are “long lived ephemera”
  - – manuals for the VCR we just bought
  - – the Patriot missile system
  - – airliners that are 20-30 years old
  - – the manuals weight more than the plane
  - – the POH is unique to the plane

Why we are here

- SPIE (society for photo optic instrumentation and electronics “the international society for optical engineering)
  - more proceedings that Springer LNCS
  - DRR -- Baird, Spitz, LoPresti
  - [google: spie document recognition; 3rd rranked is this year’s call]

Why ...

- Many untapped areas for corporate activity
- profit margins too small
- need breakthrough methods
- problem ahs importance but is difficult
- not good economic potential
- like the “orphan diseases”
- what can we do to improve conversation between the two communities

SPIE has invited

- Via LoPresti
  - Kantor, for DRR IX
  - David Lewis -- gave overview
- Tapas Kanungo has added
  - Jamie CAllan
  - David Grossman
  - Alex Hauptmann

This meeting

- First talk on the IR side
  - Kanungo unable to attend
- We need a memo that can begin the dialog
  - and any here are welcome to present late submitted papers to the DRR IX, in Santa Clara CA Jan 2003.
- Callan will lead discussion on memo

Thank you

- Collaborators -- E. Voorhees, T. Kanungo, many conversations with folks at TREC and SIGIR and at SPIE DRR
- References: in other talks
- SPIE web site

* Warning

Microsoft Windows has now been operating continuously for 88 minutes and 23 seconds. This is the limit allowed under your lease agreement. Windows Operating System will now crash in flames, destroying all of the content in your presentation, and humiliating you in front of your peers. For a better upgrade version contact Microsoft Sales Promotion at any computer store.