SCRC Test Collection for Evaluation of Chinese OCR Text Retrieval

Yuen-Hsien Tseng
Fu Jen Catholic University,
Taipei, Taiwan, R.O.C.
tseng@blue.lins.fju.edu.tw

Retrospective Information Access Approaches

- Full-text search based on manually re-keying the text
  - Prohibitively expensive at large scale
- Search based on bibliographic metadata
  - May be difficult to adequately describe the materials.
- Full text based on Optical Character Recognition (OCR)
  - Inexpensive and relatively rapid
  - Sensitive to OCR accuracy
  - Effectiveness evaluation is needed to justify the cost advantage, **which requires a test collection**.

Socio-Cultural Research Center (SCRC) Collection

- 800,000 newspaper clippings from 1950-1976
  - Scanned over 300,000 at 300 dpi
- 30 China, Hong Kong, and Taiwan news agencies
  - Mostly simplified Chinese, some traditional Chinese
- Selected images focus on diplomatic and military activities

64 10 9 人 (1) P13 10

我政府关于中印边界問題的声明

64.10.9.人 (1) P13

新华社九日訊 中华人民共和国政府关于中印边界问题的声明

一九六四年十月九日

连日来，印度总理和外交部长在开罗相继发表谈话，就中印边界问题对中国进行攻击。印度领导人竟然利用不结盟国家在开罗举行会议的机会叫嚣反华，中国政府对此不能不感到遗憾。中国政府深信，印度的这种作法是同绝大多数参加不结盟会议的国家的愿望背道而驰的。

驳斥印度总理利用出席不结盟国家会议的机会，就中印边界问題对中国进行歪曲和诬蔑。指出策动科伦坡会议国家背着中国进行磋商，是不公正的，会增加中印直接谈判的障碍，使科伦坡会议六国将来更难进行调解工作。任何没有得到中国同意和没有中国在场的调解及其所作出的建议，都是中国政府不能接受的。印度政府所以不断提出所谓全盘接受科伦坡建议等等，完全是为了反华，以便转移国内人民的视线和争取美苏军援。重申如果印度政府真正愿意谈判，中国政府准备在任何时候任何地点以科伦坡建议为基础同印度政府开始谈判。

关于中印边界问题，中国政府已经发表过大量文件，不准备再在这里重复。现在，为了驳斥印度领导人的歪曲和诬蔑，仅作如下声明：

一、印度总理说中国对科伦坡建议没有作出积极的或友好的响应，这是对事实的嘲弄。为了促进中印边界问题的和平解决，中国方面采取了停火、后撤等一系列主动措施，这些措施远远超过了科伦坡建议的要求。科伦坡会议国家知道，印度政府也清楚，如果不是中国采取了这些主动措施，中国边防部队就根本不会沿整个中印边界实际控制线向中国境内后撤二十公里，中印边境局势也不会象现在这样和缓。印度政府口口声声说它全盘接受科伦坡建议，事实上它不仅没有做过一件和缓边境局势的事，而且不断向中国境内进行骚扰和挑衅，企图制造新的紧张。

二、中国政府从一开始就表示接受科伦坡建议作为中印进行直接谈判的基础。谈判至今未能开始，其责任完全在印度方面。印度外长说，中国采取这个立场，就意味着想从侵略中得到好处。这是颠倒是非。恰恰相反，事实是印度至今还非法侵占着所谓麦克马洪线以南的中国九万多平方公里的领土，而中国方面却从来没有占领过印度一寸领土。究竟是谁在侵略，难道还不清楚吗？至于印度方面提出中国撤出七个民政点的要求，作为谈判的先决条件，这是完全没有道理的。这些民政点的所在地一直是在中国政府有效管辖之下的中国领土，印度军队连到都没有到过，印度有什么权利要中国撤出，这七个民政点，中国一个也不能撤。相反，中国倒完全有权利要求印度撤出非法的麦克马洪线以南的九万多平方公里的中国领土。但是，为了争取通过谈判和平解决中印边界问题，中国到现在还没有提出印度撤出这块九万多平方公里的中国领土的要求，作为谈判的先决条件。

三、中国政府一向欢迎科伦坡会议国家推动中印双方直接谈判而不介入纠纷的公正的调解努力。中国政府也不反对科伦坡会议国家为此而进行新的磋商。但是，谁都知道，任何有效的调解活动，都必须取得当事双方的同意。而调解国家提出的任何建议，只能是一种供双方考虑的推荐，而决不可能是一种强加于任何一方的裁决。目前，印度总理正在利用他在开罗出席不结盟国家会议的机会，就中印边界问題对中国进行歪曲和诬蔑，并且力图利用科伦坡会议国家对中国施加压力。同时，印度总理在开罗，而中国总理不在开罗。中国政府认为，在这种情况下，策动科伦坡会议国家背着中国进行磋商，是不公正的，因而也是中国政府不能同意的。这种磋商，不仅不可能得到任何足以促进中印直接谈判的结果，反而会增加中印直接谈判的障碍，使科伦坡会议六国将来更难进行调解工作。任何关于中印边界问題的实质讨论，都必须有中国在场，任何没有得到中国同意和没有中国在场的调解及其所作出的建议，都是中国政府不能接受的。

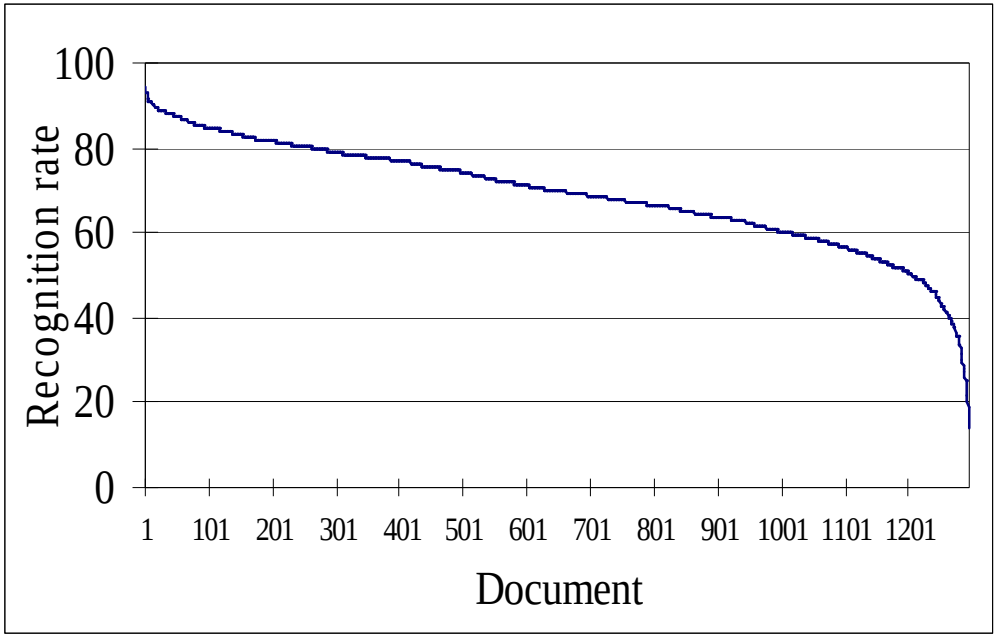
四、其实，印度并不真正想通过谈判解决中印边界问題。如果真正想谈判，它可以在谈判中提出它认为应该提出的问題，正如中国可以在谈判中提出中国认为应该提出的问題一样。印度政府明明知道，中国不会同意它提出中国撤出七个民政点作为谈判先决条件的无理要求。它也明明知道，科伦坡会议国家不会同意把科伦坡建议当作裁决要求中国全盘接受。它同样知道，中国不会在任何国际压力下屈服。印度政府之所以不断提出所谓全盘接受科伦坡建议，中国撤出七个民政点等等，完全是为了反华，以便转移国内人民的视线和争取美国军事援助和苏联军事援助，来推行它在不结盟幌子下实行双重结盟的政策。印度领导人的这一套，已经被越来越多的国家看穿了。中印边境局势基本上是和缓的。得到大量外援的印度威胁不了中国，真正受到威胁的是印度的其他邻邦。中国政府重申，如果印度政府真正愿意谈判，中国政府准备在任何时候、任何地点以科伦坡建议为基础同印度政府开始谈判。否则，空谈和解，那是无济于事的。

SCRC Test Collection for Chinese OCR Text Retrieval Evaluation

- Document Set
  - 8438 images and OCR texts
- Query Topics
  - 30 topics in TREC-like format
- Relevance Judgment
  - Exhaustive judgment by 3 persons

Document Preparation

- Selected 11,108 scanned document images
- OCR yielded 8,438 valid documents
  - OCR software package: Presto! OCR Pro, Big-5
  - Average valid document had a 69% system-reported “recognition rate”
    - Computed on a sample of 1,300 documents
- Second version prepared using Big-5 to GB conversion



Topic Preparation

- Based on contemporaneous Chinese journal articles
  - From 100 paper titles, 30 were selected and rewritten as Chinese topics
- Made English translations for cross-language experiments
  - Translated by native speakers of Chinese

<top>
<num> 12
<title> Anti-Chinese Movements
<description>
Activities related to the anti-Chinese movements in Indonesia
<narrative>
Articles must deal with activities related to the anti-Chinese movement in Indonesia; case reports or articles dealing with PRC's criticism of the Anti-Chinese movement will be considered partly relevant.
</top>

Relevance Judgments

- Exhaustive tri-state relevance judgments
  - Irrelevant (=0), partially relevant (=1), fully relevant (=2)
- Every topic-document pair judged by 3 assessors
  - 2 majored in history, 1 majored in library science
  - Averaged 4 minutes per document image (for all 30 topics)
- Sum of the judgments provides a final estimate
  - 0=not relevant, 1…5=partially relevant, 6=fully relevant
  - Threshold as desired to reflect the intended application
    - In our experiments, any score > 0 is treated as “relevant”

Query ID	Doc. ID	1st Assessor	2nd Assessor	3rd Assessor	Total Score
01	0053487	1	1	0	2
01	0053489	1	2	1	4
…					
02	0054425	2	2	2	6
02	0054452	1	1	1	3
…					

How to Get the Test Collection

Please visit

http://www.lins.fju.edu.tw/~tseng/Collections/Chinese_OCR_IR.html

for details