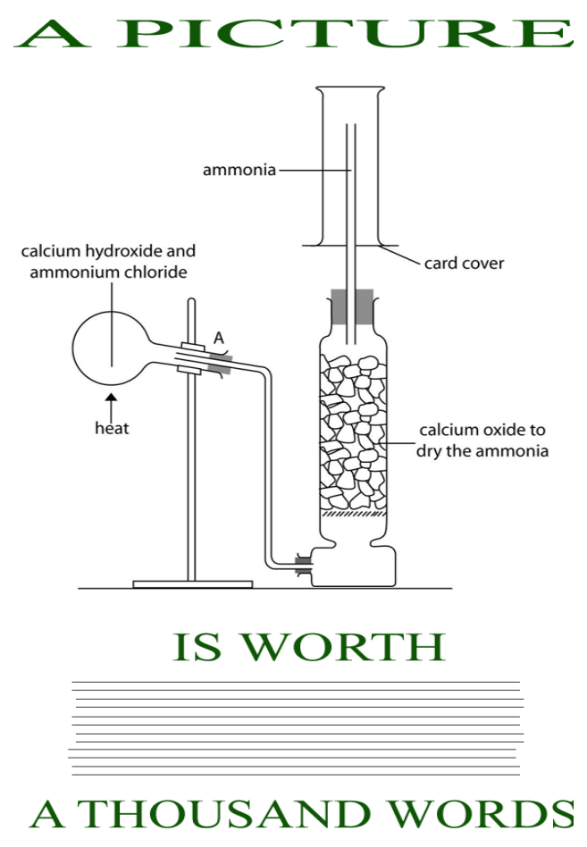

A PICTURE
ammonia
calcium hydroxide and ammonium chloride
card cover
heat
calcium oxide to dry the ammonia
IS WORTH
A THOUSAND WORDS

Extraction of Relevant Figures and Tables For Multi-Document Summarization

Ashish Sadh, Amit Sahu, Devesh Srivastava, Ratna Sanyal, Sudip Sanyal

Indian Institute of Information Technology - Allahabad, India

A picture is worth a thousand words

Figures and tables enhance the understanding and easily gain focus of the readers. Furthermore, these figures/tables convey a large chunk of information in relatively condensed form. Hence, these units characterize an excellent choice as components in a document summary. Given such significance, one must find a way to extract important figures and tables for effective summarization of digital documents.

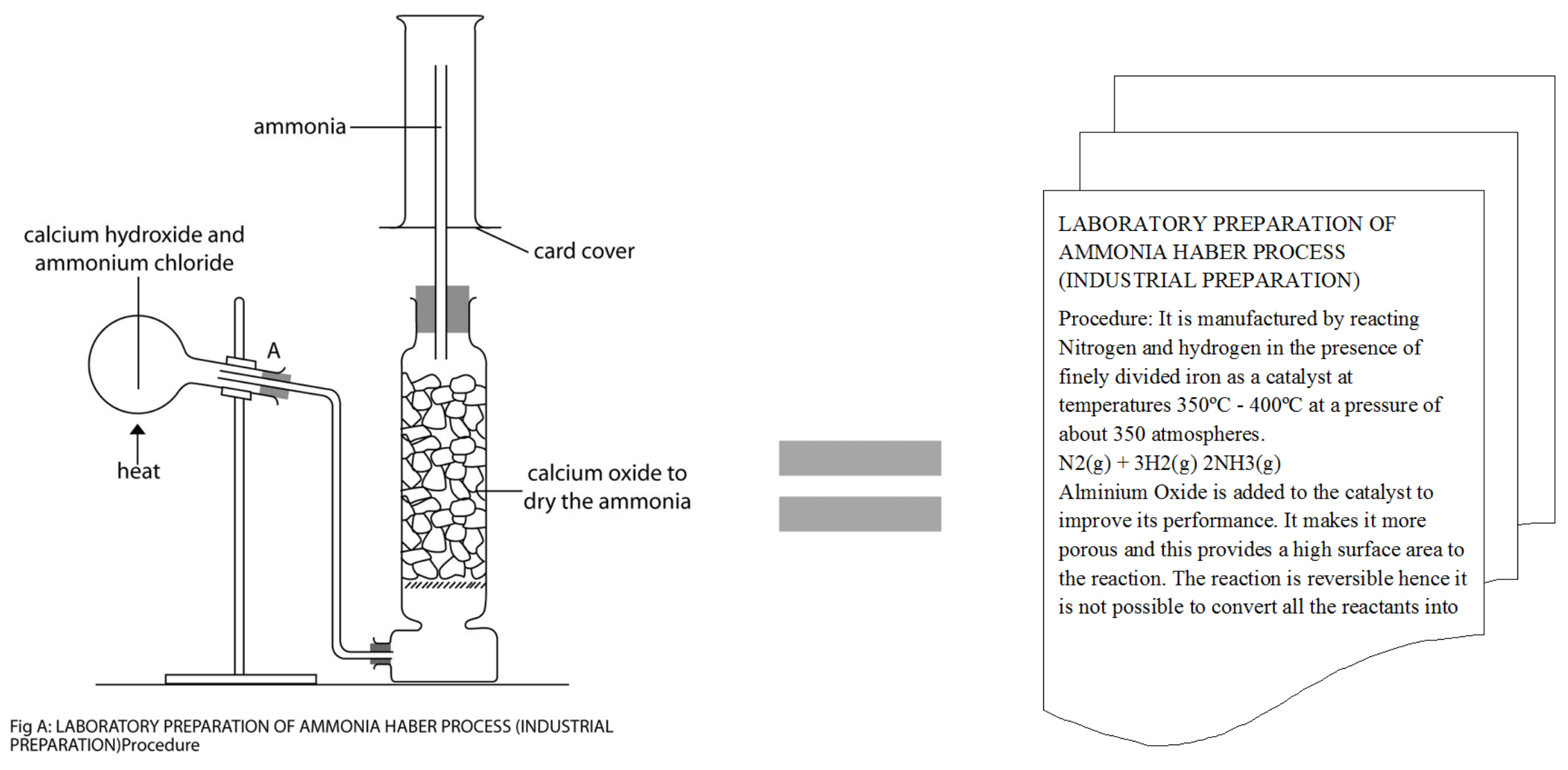

ammonia
calcium hydroxide and ammonium chloride
card cover
A
heat
calcium oxide to dry the ammonia

Fig A: LABORATORY PREPARATION OF AMMONIA HABER PROCESS (INDUSTRIAL PREPARATION)Procedure

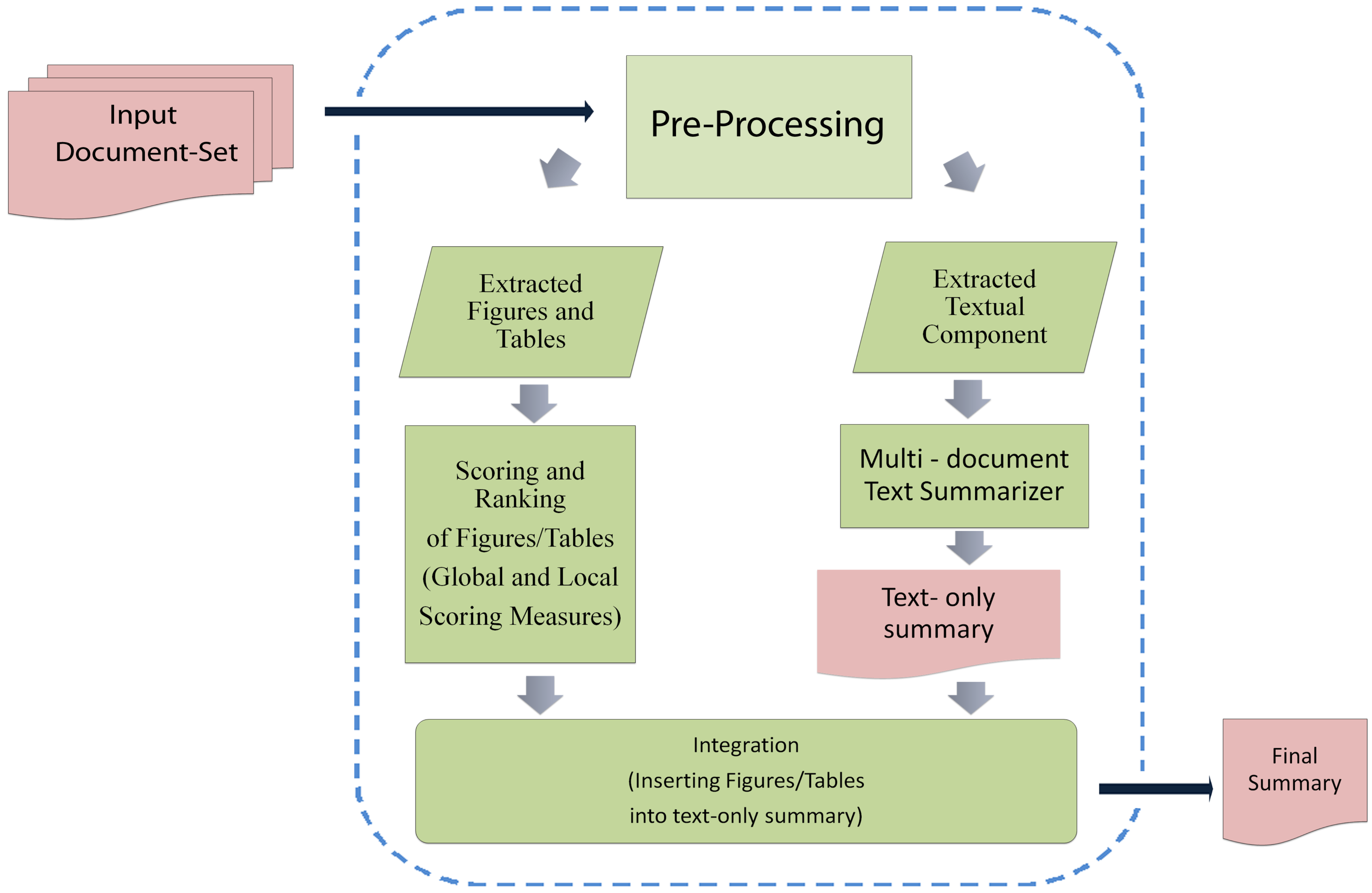
LABORATORY PREPARATION OF AMMONIA HABER PROCESS (INDUSTRIAL PREPARATION)

Procedure: It is manufactured by reacting Nitrogen and hydrogen in the presence of finely divided iron as a catalyst at temperatures 350°C - 400°C at a pressure of about 350 atmospheres.

N2(g) + 3H2(g) 2NH3(g)

Alminium Oxide is added to the catalyst to improve its performance. It makes it more porous and this provides a high surface area to the reaction. The reaction is reversible hence it is not possible to convert all the reactants into

Figure 1: Overview Diagram



Input Document-Set

Pre-Processing

Extracted Figures and Tables

Extracted Textual Component

Scoring and Ranking of Figures/Tables (Global and Local Scoring Measures)

Multi - document Text Summarizer

Text- only summary

Integration (Inserting Figures/Tables into text-only summary)

Final Summary

Results

Document Collection	Domain/Topic	No. of Documents	No. of elements	
			Figures	Tables
Doc-Set 1	Scientific (Artificial Neural Network)	5	7	0
Doc-Set 2	Medical (Effect of the Sun on Skin)	3	2	8
Doc-Set 3	Geography (Nile River)	4	12	0

Table 1. Description of the Document Collections .

We devise an experimental evaluation, in which human evaluators rank the elements(figures/tables) of these document collections (Table 1). Ranking is done based on their relevance to the text-summary of the collection. We calculate Spearman's rank correlation coefficient and Kendall's τ coefficient for these ranked-lists with the system generated ranks. An aggregated score using these correlation values is calculated using methods WCA and RBA as described in [2]. Values are considerable close to the perfect correlation measure.

The major findings of the experiments are summarized in the table below:

Document Collection	WCA-τ Score	WCA-Sp Score	RBA-τ Score	RBA-Sp Score
Doc-Set 1	0.767	0.848	0.952	0.982
Doc-Set 2	0.759	0.839	0.878	0.951
Doc-Set 3	0.871	0.935	0.964	0.988

Table 2. Scores for the system rankings .

Methodology

System extracts important figures and tables from topically related documents by exploiting their association with the textual component. Typically, any figure or table can be associated with its corresponding text using a direct reference or an indirect reference.

A ranked list of figures/tables is generated using relevancy of associated text. Integration into the summary is assisted by this ranked list and done in a way to improve cohesion and coherence of the extractive text summary.

Relevancy Measure

Relevance score of any figure/table can be measured using following equations:

$$\mathrm{scs} = \sum_{k=0}^{n} \mathrm{w_j}. \qquad (1)$$

where, scs is the score of a sentence containing n words and j^{th} word occurs w_j times within the document.

The contribution of a direct references can be calculated as:

$$DR = m * \sum_{i=1}^{m} scs_i \; . \qquad (2)$$

where m is number of direct references and scs_i is the score of i^{th} direct reference (equation 1).

The contribution of a indirect references can be calculated as:

$$IR = CS * \sum_{j=1}^{n} scs_j. \qquad (3)$$

where CS is the sum of cosine similarity scores of indirect references with the caption, n is number of indirect reference and scs_j is the score of j^{th} indirect reference (equation 1).

Final Relevance Score, S = IR + DR .

Source Code

The source code and compiled executables with an interactive interface are available at : `http://goo.gl/YgoIy`

Acknowledgements

The authors gratefully acknowledge the support from the Indian Institute of Information Technology, Allahabad and Ministry of Communication and Information Technology, New Delhi for carrying out this research.

References

[1] Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility based evaluation, and user studies. In: ANLP/NAACL Workshop on Summarization, vol. 40, pp. 21-29. ACL, Seattle, (2000)

[2] Kim, H.D., Zhai, C., Han, J.: Aggregation of Multiple Judgments for Evaluating Ordered Lists. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Ruger, S.M., Rijsbergen, K.V. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 166-178. Springer, Milton Keynes (2010)