



1999 GWU SEAP

FALCon: Evaluation of OCR and Machine Translation Paradigms

Presented by:

Kathleen Swam

Raquel Burgos

US Army Research Laboratory (ARL)

Information Science and Technology Directorate (ISTD)

Aberdeen Proving Ground, MD 21005-5067

13 August 1999



Factors Affecting OCR Accuracy

- Scanning Process - partially controlled
- Document Quality - **low, medium, high**
- Font Style - **serif**, sans serif
- Font Size - **10, 12, 14**
- FALCon System - **AHSfalcon1**, AHSfalcon2
- Imaging Parameter - color, fax, **grayscale**
- Document Type - **letter, newspaper, magazine**, fax, tech article
- Zone - header, caption, footer, **main body**, signature block
- Page Length - **paragraph, half page, full page**



Evaluation Process Steps

1. Identify factors affecting OCR quality
2. Find documents that fit these characteristics
3. Perform character and word counts
4. Type groundtruth version
5. Scan all documents and groundtruths
6. OCR the original and groundtruth documents
7. Pass data through automated scoring software and generate OCR accuracy reports
8. Pass documents through machine translation software
9. Complete human translation and evaluation performance reports



Common Machine Translation Errors

- Word Order - translating words in an arrangement that makes no sense
- Context - words with multiple meanings that aren't used in the proper context
- Pronoun - misplacement of he, she, it, them, etc.
- Dictionary - using the wrong word
- OCR - occur because characters were unrecognizable
- Missing Word - not translating certain words
- Extra Word - adding unnecessary text that results in the wrong translation
- Translation - leaving the word in the original language
- Conjugation - a verb conjugated in the wrong tense
- Proper Name - translating a proper name into another word that makes no sense



Summary

Scanning Process

- quality of scan does affect OCR accuracy
- recommendations for improvement

OCR

- accuracy appeared to be independent of the factors evaluated
- character accuracy ranged from 90.79% - 99.76% across the sample documents

MT

- accuracy appeared to increase with the decrease in the length of the document
- word accuracy ranged from 43.09% - 84.74% across the sample documents