Multilingual Search

Shibamouli Lahiri

(shibamouli@cse.psu.edu)
Existing Work in CLIR

• Query Translation

• Document Translation

• No translation – cognate matching, transliteration

• Interlingual Techniques (LSI[5], decipherment)
Existing Work in CLIR (Contd.)

- **Query Stemming** – before/after translation

- **Translation Knowledge** – Bilingual Dictionaries (online, offline)[1], Parallel Corpora, the Web[6]

- **Translation Disambiguation** – word co-occurrence statistics, translation probability, POS tags, user and pseudo relevance feedback and user-assisted translation, structured queries[9], bidirectional translation, using web resources[11, 12]

- **Evaluation** – TREC CLIR track, NTCIR, TDT, CLEF
Existing Work in CLIR (Contd.)

- **Ranking** – vector space model, language model

- **Pivot Language (Triangulation) Approach**[^8]

- **Named Entity Extraction and Translation**[^7]
CyDAR – Basic Premises and Requirements

- Corpus is in Ancient Greek
- Query might be in English, Modern Greek or Ancient Greek
Issues

- Unit of Retrieval (Pages? Part of a page? Paragraphs?)

- An intelligible summary, preferably in all three languages

- Ranking

- Evaluation

- Query and translation disambiguation
User Interface Example
Lexicon

- Lexilogos

- Kypros (also offers Ancient Greek)
Parallel Corpora

- Perseus Project has English ↔ Ancient Greek and English ↔ Latin

- Europarl

- ELRA (European Language Resources Association)


Thank you!