Cross-Language IR

CISC489/689-010, Lecture #23

Monday, May 11th

Ben Carterette

Cross-Language IR

- User submits a query in one language, gets results in a different language
- Documents are semi-structured and heterogeneous (as almost all data in IR), and also in multiple languages
- Information may only be available in documents written in one of the languages
- Highly useful to intelligence community

Approaches to CLIR

- Translate the documents into the users' language, and let the users submit queries in their own language
- Translate the users' queries into target language(s) and use the translated query for retrieval
- Translate both queries and documents to an "intermediate" language

Automatic Translation

- What are some approaches to automatic translation?
  - Language-to-language dictionaries
- Languages do not translate precisely
  - One word with several meanings in one language might translate to several different words in the other
  - Many words with the same meaning might all translate to a single word
  - A word in one language might only be expressible as a phrase in another (or vice-versa)
  - etc…

Example

- English queries to retrieve Spanish documents
- System works by translating query to Spanish
- Query: "bank fraud"
- Translations of "bank":
  - *Orilla* (river bank)
  - *Terraplen* (bank of earth)
  - *Banco* (bank of clouds)
  - *Bateria* (bank of lights)
  - *Banco* (financial institution)
  - *Banca* (casino bank)
- Translations of "fraud":
  - *Impostor* (fraudulent person)
  - *Fraude* (deception)
- How would a dictionary-based system know which pair of translations to use?
- Possibly correct translation:
  - *Fraude bancario*

Statistical Approach

- Instead of trying to translate directly, apply statistical methods
- Learn "translation probabilities" P(f | e) – probability of translating string e in language E to string f in language F
- E.g.:
  - P(orilla fraude | bank fraud), P(orilla impostor | bank fraud), P(banco fraude | bank fraud), …