Acquiring Comparable Corpora

BRUNO REZENDE LARANJEIRA
BRUNO.REZENDELARANJEIRA@INF.UFRGS.BR

THIRD CAMELEON WORKSHOP

20/12/2012

Motivation

- Machine Translators
- Multilingual Ontology Enrichment
- Cross-Language Information Retrieval

Parallel and Comparable Corpora

- Parallel corpus
  - Texts with exact **translations**
- Comparable corpus
  - Texts which are not exact translations, but are about the **same subject**
  - May contain **parallel subcorpora**

Machine Translation Systems

- Rule Based
  - Rules defined **manually**, for translating through an intermediate representation of the text.
- Statistical
  - **Probability** based.
  - Highly dependent on the **corpus** and its **quality**.

Talvensaari et al., (2008)

- Compare many translation algorithms.
  - Among them, two statistical machine translations were tested:
    - One trained on a **large generic corpus**.
    - The other trained on a **smaller** corpus with texts belonging to the **same domain** as the input texts.

Talvensaari et al., (2008)

- Compare many translation algorithms.
  - Among them, two statistical machine translations were tested:
    - One trained on a **large generic corpus**.
    - The other trained on a **smaller** corpus with texts belonging to the **same domain** as the input texts.

Universal X Focused Crawling

- Universal
  - **Any** document can be collected.
- Focused
  - The main goal is to only collect documents that belong to a **specific set of topics**.

Focused Crawling

- Main issue
  - Determine **which pages** to fetch.
  - In **which order** it will be done.

Focused Crawling

- Main algorithms
  - **Best-First**.
  - Focused Crawling via **Classifiers** and **Distillers**.
  - **Adaptive** Focused Crawlers.

Granada et al. (2012)

- Use the labels of a multilingual ontology to build a comparable corpus.
  - Combine them and send to a **search engine**, like Google or Yahoo!.
  - Highly dependent on search engine **idiosyncrasies**.

Work in Progress

- Replicate the work of Granada et al. , with some changes.
  - URL **normalization**
  - Near **Duplicate** Detection

- In the end, we aim to analyze:
  - **Intersection** between documents in different languages;
  - Corpus reduction after
    - URL normalization
    - Near duplicate elimination;
  - Documents per query; sentences per document; words per sentence;
  - Token type repetition;

Future Work

- Get rid of search engine dependency
  - Build a focused crawler.
    - **Exploit hub pages** during the crawl and eliminate them in the end.
    - Detect hub pages by their **tag structure**.
  - Sentence alignment
    - Finer granularity than paragraphs (Talvensaari et al (2008))

Case Study

- Languages:
  - English, French, and Portuguese.
- Topics:
  - Conferences and Tourism.

Thank You!

- Questions?
- Suggestions?