Collecting Aligned Textual Corpora from the Hidden Web

Boštjan Pajntar
bostjan.pajntar@ijs.si
Aligned Parallel Corpus

Definition (wikipedia):

- "A parallel text is a text placed alongside its translation or translations"

Usage:

- Translation Memory
- Machine Translation
- Natural Language Processing

Standards:

- TMX – Translation Memory eXchange
- TBX – TermBase eXchange
- UTX – Universal Terminology eXchange
- (SRX, GMX-GILT, OLIF, XLIFF, TransWS, ...)

([Image])
But Where to Get the Data?

- Non-English professional websites
- Huge amounts of translated text
- Generally quality translations

- We call this the Hidden Web
Problems

- Translation Memory is hard / expensive to obtain
  - Idea: Automatic harnessing of existing data

- Data should have very high precision
  - What precision is needed?

- No standard fully supports automatic:
  - Harnessing of the data
  - Cleaning of the data
Proposed Solution

Parallel Corpora Extraction

Crawling → Relational Database → Candidates Extraction → List of HTML candidates → Parsing → List of text candidates → Filtering → Parallel Corpora

Available at: http://kameleon.ijs.si/t4me
Discussion on Standards

- We build on TMX:
  - Is this the right choice?
  - Source language must be defined!
  - An optional parameter to define the source of each segment

- Proposals for automatic harnessing of TM:
  - Provide a new standard
  - Build on an existing one
  - Ideas?
Future Work

- Optimizing Crawling:
  - Two phase crawling
  - Character Encodings
  - Enhanced candidates extraction

- Optimizing Extraction:
  - Segmentation
  - Language identification
  - Enhanced filtering

- Web service / Web application
  - Translation Memory distribution
  - Filtering (Web 2.0 style)