Compiling terminological data using comparable corpora: from term extraction to dictionaries

Marion Weller, Anita Gojun, Ulrich Heid
IMS - Universität Stuttgart
(weller@ims.uni-stuttgart.de)

Beatrice Daille, Emmanuel Morin
LINA - Université de Nantes
(Beatrice.Daille@emmanuel.Morin@univ-nantes.fr)

Overview

- Current situation: "Terminology Bottleneck" in translation
  - few resources for automatic bilingual terminology acquisition
  - lack of stable head term logics for new/updated domains

- TTC solutions: Term extraction from comparable corpora
  - semi-automatic tool chain
  - languages: DE, EN, ES, FR, LV, RU, ZH

- Project Goals:
  - Development of tool for corpus crawling, monolingual term extraction and bilingual term alignment
  - Philosophy of tool development:
    - Assessment of slim solutions: as little linguistic knowledge as possible

Terminology Processing Chain

(1) Collecting domain-specific texts
- Thematic Web crawlers for collecting domain documents from the Web
  - [de Groc, 2011]

(2) Monolingual term extraction
- Input: monolingual crawled domain-specific texts
  - Freq. scores
    - token ranking + bagging = term ranking
  - Monolingual extraction of single-word and multi-word term candidates
  - Identification of domain-relevant terms using frequency-based and statistical approaches

(3) Identifying term variants
- Grouping related terms using pre-defined language-specific variation patterns
- Output: Groups of synonymous and related monolingual term variants

(4) Bilingual term alignment
- Input: for a given source language term, find a translation in target term lists
  - Output: bilingual and monolingual term lists

Project Partners:

- OpenSource UIMA-based application: TTC Term Suite
  - Code: googlecode.com/p/ttc-project/
  - Term extraction Web-service
  - http://greenhouse.syllabs.com/ttc/

Method

- Individually translate the parts of a multi-word term
- Combine all translation possibilities
- Compare generated translation candidates with target language terms

Example

- Input term: elektrisch, ADJ
  - Lexicon look-up:
    - Netz → electric
    - Netz → wire
    - Netz → cage
    - sektor → sector

- Re-combine & compare:
  - elektrotechniek → not in target term list
  - rets électrique → not in target term list
  - sektor → sector

- Output:
  - elektrisches, ADJ → electric
  - Netz → electric
  - sektor → sector

Term alignment: Lexical strategy

Method

- Identifying term variants
- Grouping related terms using pre-defined language-specific variation patterns
- Output: Groups of synonymous and related monolingual term variants

Term alignment: Context vector strategy

Method

- Lexical context analysis:
  - terms and their translations tend to appear in the same lexical contexts
  - For each term, co-occurance frequency of terms in a window is used.

- Bilingual context vector strategy
  - For each source term, co-occurance frequency of source language terms is used.
  - Output: bilingual context vector pairs
  - Bilingual context vector analysis
  - Terms with the most similar context vectors are a likely to be translations

Problems

- Limited coverage of the general language dictionary
- This method is only suited for single-word terms

Output and evaluation

- For each input term, the top n results are shown
- For each target language term with the most similar context vectors

Example: output of the alignment component in TTC Term Suite

References

[de Groc, 2011] Clément de Groc: Babouk: "Focused Web crawling for corpus compilation and automatic terminology extraction"

The project has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under Grant Agreement Number 248005.