Terminology-finding in the Sketch Engine

Miloš Jakubíček, Adam Kilgarriff,

Vojtěch Kovář, Pavel Rychlý, Vit Suchomel

Lexical Computing Ltd., Brighton, UK & Masaryk University, Brno, Czech Republic

Terminology

- Problem #1
  - Finding it

Terminology

- Problem #1
  - Finding it
- Existing lists
- Ask experts
- Corpora

To find terms in a corpus

- Unithood
  - For multi-word terms
  - Do the words form a unit?
- Termhood
  - Does it belong to the domain?

Unithood

- Grammar
- Terms are noun phrases
  - (in canonical form, without the article)
- Requirements
  - Noun phrase grammar
  - Prerequisites: tokeniser, lemmatiser, POS-tagger
  - Parsing machinery

Termhood

- Frequency
  - in domain corpus *vs* reference corpus
- Same as keywords
- Requirements
  - Formula for keyness
  - Domain corpus
  - Reference corpus

In the Sketch Engine

Unithood

- Grammar
- Terms are noun phrases
  - (in canonical form, without the article)
- Requirements
  - Noun phrase grammar
  - First: Chinese English French Japanese Korean Spanish
  - Last additions: German Portuguese Russian
  - Prerequisites: tokeniser, lemmatiser, POS-tagger
  - Available/installed for languages above and several others
  - Parsing machinery
  - In place: variant on word sketches infrastructure

Termhood

- Frequency
  - in domain corpus *vs* reference corpus
- Same as keywords
- Requirements
  - Formula for keyness
  - Kilgarriff 2009: Simple maths for keywords
  - Ratio of normalised frequencies with simple maths parameter
  - Domain corpus
  - Existing machinery for
    - Instant corpora from the web: WebBootCaT
    - Uploading/installing your own corpus
  - Reference corpus
  - Large web corpora: sixty languages

Current status

- Lead customer
  - WIPO (World Intellectual Property Organisation)
  - terminology group of their translation dept
  - Five languages: delivered
  - Added functionality, blacklists etc
- All customers
  - First version in beta

Current challenges

- Identical processing chain for
  - Reference corpus (batch mode)
  - Domain corpus (runtime)
- Lemmas and word forms
  - When to use singular, when plural
  - Adjective-noun agreement
- How to evaluate

Thank you

http://www.sketchengine.co.uk