Corpora for Bilingual Terminology Extraction in Cybersecurity Domain

Dvikalbis automatinis terminų atpažinimas (DVITAS)

Andrius Utka, Sigita Rackevičienė, Liudmila Mockienė, Aivaras Rokas, Marius Laurinaitis and Agnė Bielinskienė
Aims

• The main aims of the project:
  1. to prepare language resources for bilingual terminology extraction (BiTE),
  2. to apply current bilingual terminology extraction methods,
  3. create an English-Lithuanian termbase of cybersecurity (CS) terminology.
Cybersecurity domain

• Cybersecurity (CS) domain is chosen, as we think:
  - that for today’s digital world this domain is particularly relevant;
  - it is also very dynamic,
  - and Lithuanian CS terminology needs standardization.
BiTE

• **Bilingual Terminology extraction (BiTE)** is understood as a method when terms and their translations are extracted from parallel or comparable texts.

• Term extraction from parallel corpora has been already applied for several decades.

• But the importance of comparable data is increasing...
Parallel vs comparable data

Parallel data

• Translated language is influenced by the source language;
• Shortage of parallel data;
• Lack of diversity: mostly EU documents;
• Expensive to build (needs aligning);
• **BUT**: it’s easier to do BiTE.

Comparable data

• The original language is more natural;
• More data is available;
• The data is more diverse;
• Cheaper to build;
• **BUT**: it’s more difficult to do BiTE.
Corpora system for BiTE

CS Parallel Corpus (1)
- English (0.7m)
- Lithuanian (0.7m)

CS Comparable Corpus (2)
- English (2m)
- Lithuanian (2m)

CS Training Corpora
- Training Corpus (3) (Gold standard) 0.1m
  - Neural network model-1
  - English-Lithuanian terms

- Training Corpus (4) (Gold standard) 0.1m
  - Neural network model-2
  - Neural network model-3
  - Training Corpus (5) (Gold standard) 0.1m
  - English terms
  - Lithuanian terms

Bilingual Termbase
Texts types in corpora

EN-LT Parallel corpus

- Non-binding EU documents; 40%
- Legally binding EU documents; 60%

1.4m words

Period: 2010-2020

EN-LT Comparable corpus

- Legislative; 10%
- Non-binding; 15%
- Academic; 25%
- Media; 50%

4m words
Thank you!

Ask questions in Q&A session and in the afternoon breakout session
15:30-16:00 Room 2 (Resources)