BILINGUAL TERMINOLOGY EXTRACTION

Vít Baisa, Michal Cukr, Barbora Ulipová

INTRODUCTION

- automatic bilingual terminology extraction (ABTE)
- terminology extraction + parallel corpora
- implemented in Sketch Engine
- a few parameters to tune and play with
- evaluation
- gold standard English-Czech

RELATED WORK

- ABTE available in several commercial tools: SDL MultiTerm, Araya, MemoQ
- either alignment of pre-extracted terms
- or an alternative approach
  - extracting parallel phrase rules (N+ADJ = ADJ+N)
  - phrase-based machine translation approach
- majority of terms: noun phrases
- majority (70%) MWE

BITERMS IN SKETCH ENGINE, DEMO

L1 term	L2 term	Logdice	Co-freq	L1 freq	L2 freq
prevalence	prévalence	-0.0257005103	306	316	307
soap	savon	-0.0580571016	207	220	211
survival	survie	-0.0683060134	165	170	176
education	éducation	-0.0705785710	1815	1968	1844
adolescence	adolescence	-0.0711610289	89	91	96
condom	préservatif	-0.0840642648	125	139	126
primary prevention	prévention primaire	-0.0840642648	25	27	26
chronological age	âge chronologique	-0.0848888976	33	36	34
basic information	informations de base	-0.0874628413	16	17	17
acid	acide	-0.0874628413	16	17	17
rotavirus	rotavirus	-0.0931094044	15	16	16
universal access	accès universel	-0.0981803939	142	151	153
international guidance	directives internationales	-0.0995356736	14	15	15
stigma	stigmatisation	-0.1040724541	127	133	140
fish	poisson	-0.1043366598	20	21	22
pregnancy	grossesse	-0.1059334447	210	230	222
alcohol	alcool	-0.1110313124	25	28	26
vol	vol	-0.1168136650	83	87	93
syphilis	syphilis	-0.1233824155	28	32	29
public health	santé publique	-0.1235746851	123	133	135

ALGORITHM I

- first step: terminology extraction
- currently supported languages: Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish
- alignment 1:1 or m:n
- TMX import
- matching grammar rules (noun phrases)
- lexical form is made and stored

ALGORITHM II

- bilingual alignment
- co-occurrence statistics
- sorted by logDice or co-frequency

$$logdice = 14 + \log_2 \frac{2f_{ab}}{f_a + f_b}$$

GOLD STANDARD

- authors of two papers contacted
- failed to obtain their data
- DGT vs. IATE could not be used
- English-Czech from DGT
- manually cleaned 1000 top pairs
- lexically correct translations were selected
- “dry linen” ≠ “suchého prádla”
- 328 term pairs

EXPERIMENTS

- minimum co-frequency (4)
- preferring shorter / longer terms in L1,2
- prefer a ratio of character-length between terms

EVALUATION

Ratio	Precision	Recall	F-score
N/A	0.3282	0.3232	0.3257
2.00	0.3994	0.3933	0.3963
0.98–1.5	0.4025	0.3964	0.3994
0.92–0.97	0.4056	0.3994	**0.4025**
0.91	0.4025	0.3964	0.3994
0.90	0.3994	0.3933	0.3963
0.80	0.3993	0.3933	0.3963
0.75	0.3963	0.3902	0.3932

CONCLUSION

- evaluation is hard
- recall is preferred
- ABTE tests all processing steps
- a language-dependent improvement was proposed
- logDice vs. co-freq
- big data vs. alignment granularity
- gold standard CC BY-SA
- TMX export supported, integration with CAT tools
- ABTE not recognized yet
- rather a toy for terminologists

FUTURE WORK

- plugin integration in CAT tools
- multilingual evaluation
- richer TMX export
  - contexts
  - frequencies
  - metadata
- other improvements
  - splitting long sentences
  - cognates
  - hierarchical