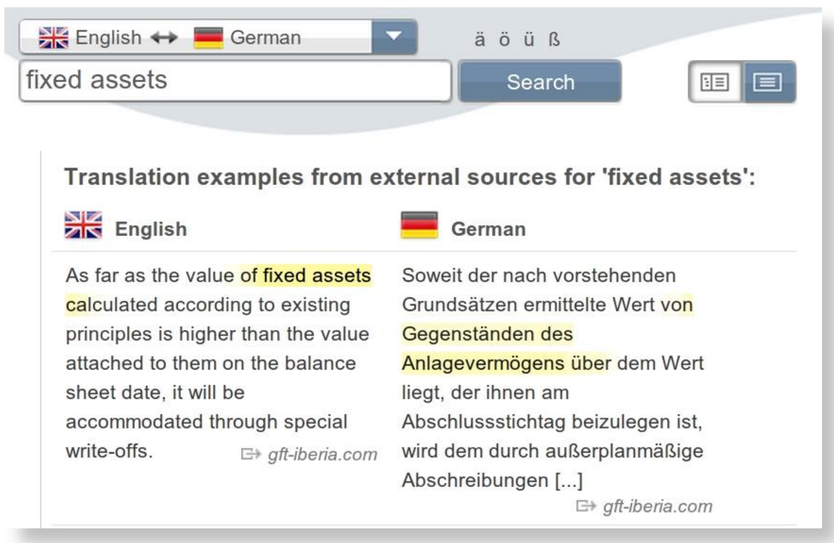
Using Domain-specific and Collaborative Resources for Term Translation

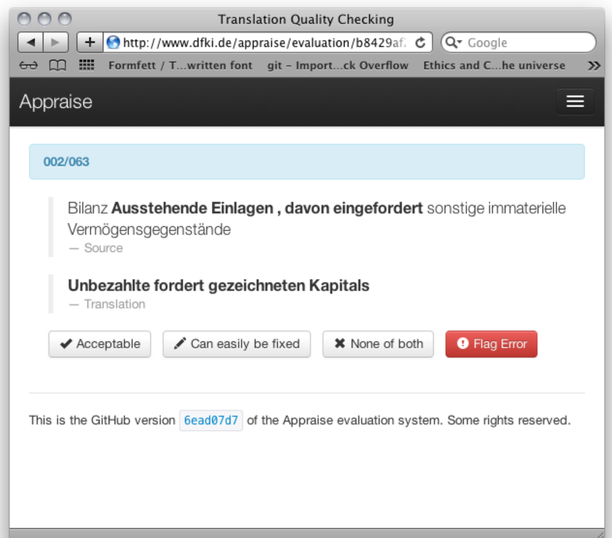
Mihael Arcan, Paul Buitelaar
Unit for Natural Language Processing & eBusiness Unit
Digital Enterprise Research Institute, National University of Ireland, Galway
Christian Federmann
Language Technology Lab, German Research Center for AI, Saarbrücken, Germany

Motivation & Objective:

- translation of terms in the isolation of an ontology vocabulary
- sharing and transparency of financial and business information across Europe

Data Set: concept taxonomy xEBR (XBRL Europe Business Registers) to describe business entities





English: 24K sentences (1M tokens)
German: 24K sentences (0.85M)

7388 English-German translation pairs

2520 classification items (six annotators)

Source	# correct	Scoring Metric BLEU	NIST	Meteor
Google Translate	18	0.264	4.382	0.369
JRC-Acquis	12	0.167	3.598	0.323
EuroParl	4	0.113	2.630	0.326
Linguee	25	0.347	4.567	0.408
Lexical substitution	4	0.006	0.223	0.233
Linguee+Wiki	25	0.324	4.744	0.432

System	Classes A	C	N
Linguee+Wiki	58%	27%	15%
Google Translate	55%	31%	14%
Linguee	51%	37%	12%
JRC-Acquis	32%	28%	40%
EuroParl	5%	25%	70%

System	Agreement Metric S	π	κ	α
Linguee+Wiki	0.599	0.528	0.533	0.530
Google Translate	0.698	0.655	0.657	0.657
Linguee	0.484	0.416	0.437	0.419
JRC-Acquis	0.412	0.406	0.413	0.408
EuroParl	0.515	0.270	0.269	0.273

Source	# correct	Scoring Metric BLEU	NIST	Meteor
Google Translate	21	0.452	4.830	0.641
JRC-Acquis	9	0.127	2.458	0.480
EuroParl	5	0.021	1.307	0.412
Linguee	15	0.364	3.938	0.631
Lexical substitution	4	0.006	0.243	0.260
Linguee+Wiki	22	0.348	3.993	0.644

System	Classes A	C	N
Linguee+Wiki	56%	32%	12%
Linguee	56%	31%	13%
Google Translate	39%	40%	21%
JRC-Acquis	39%	31%	30%
EuroParl	15%	30%	55%

System	Agreement Metric S	π	κ	α
Linguee+Wiki	0.532	0.452	0.457	0.454
Linguee	0.599	0.537	0.540	0.539
Google Translate	0.480	0.460	0.465	0.463
JRC-Acquis	0.363	0.359	0.366	0.360
EuroParl	0.552	0.493	0.499	0.495

Discussion:

- Domain-specific resource gives better results than a bigger, but more general one
- Our approach outperforms Google Translate for translating German terms into English

Future Work / Next Steps:

- Evaluation on larger data set (IFRS, finance domain) and vocabularies (BioCaster, medical domain)
- Extracting terms (translation pairs) from non-parallel text

Acknowledgements: This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.:249119) and in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No.SFI/08/CE/I1380 (Lion-2). The authors would like to thank Susan-Marie Thomas, Tobias Wunner, Nitish Aggarwal and Derek De Brandt for their help with the manual evaluation.

Enabling **Networked** Knowledge