Centre for
Data Analytics

Insight

Domain Adaptation for Ontology Localisation

Journal of Web Semantics, Vol 36, Jan 2015

John P. McCrae[a,b], Mihael Arcan[a], Kartik Asooja[b,c], Jorge Gracia[c], Paul Buitelaar[a] and Philipp Cimiano[b]

[a]Insight Centre for Data Analytics, National University of Ireland, Galway

[b]CIT-EC, Bielefeld University

[c]Ontology Engineering Group, Universidad Politécnica de Madrid

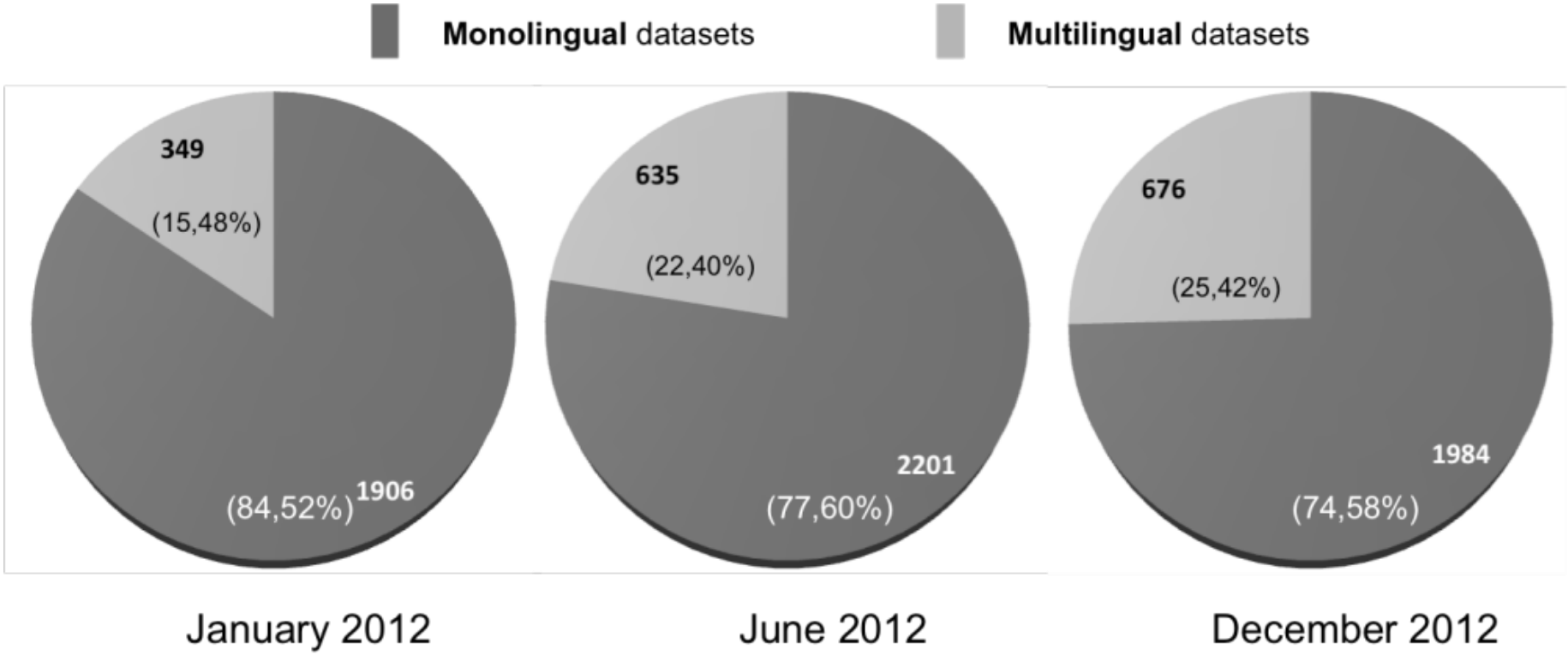
Universität Bielefeld

NUI Galway
OÉ Gaillimh

monnet

sfi

Is the Semantic Web multilingual?



"Guidelines for multilingual linked data" Gómez-Pérez et al. 2013

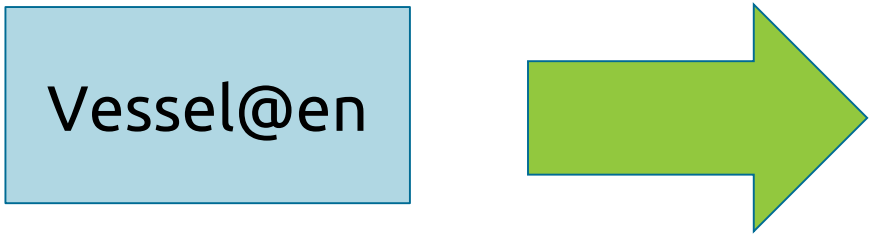
Ontology Localisation

- Different languages (cultures) have different concepts
  - Cross-lingual ontology alignment
  - Not the focus of this paper
- Translation of labels
  - Manually is time-intensive and costly
  - Standard MT has poor performance for ontologies
- Domain ontologies
  - Small, focused ontologies
  - Large, general purpose ontologies (e.g., DBpedia) should be divided into more sections

Why is ontology translation hard?

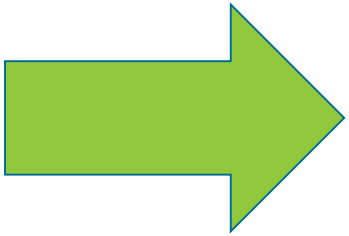
Vessel@en

Why is ontology translation hard?



Why is ontology translation hard?

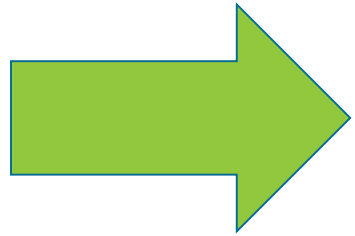
Vessel@en



船@ja

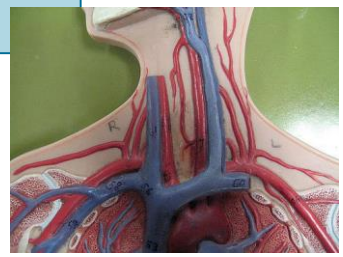
Why is ontology translation hard?

Vessel@en

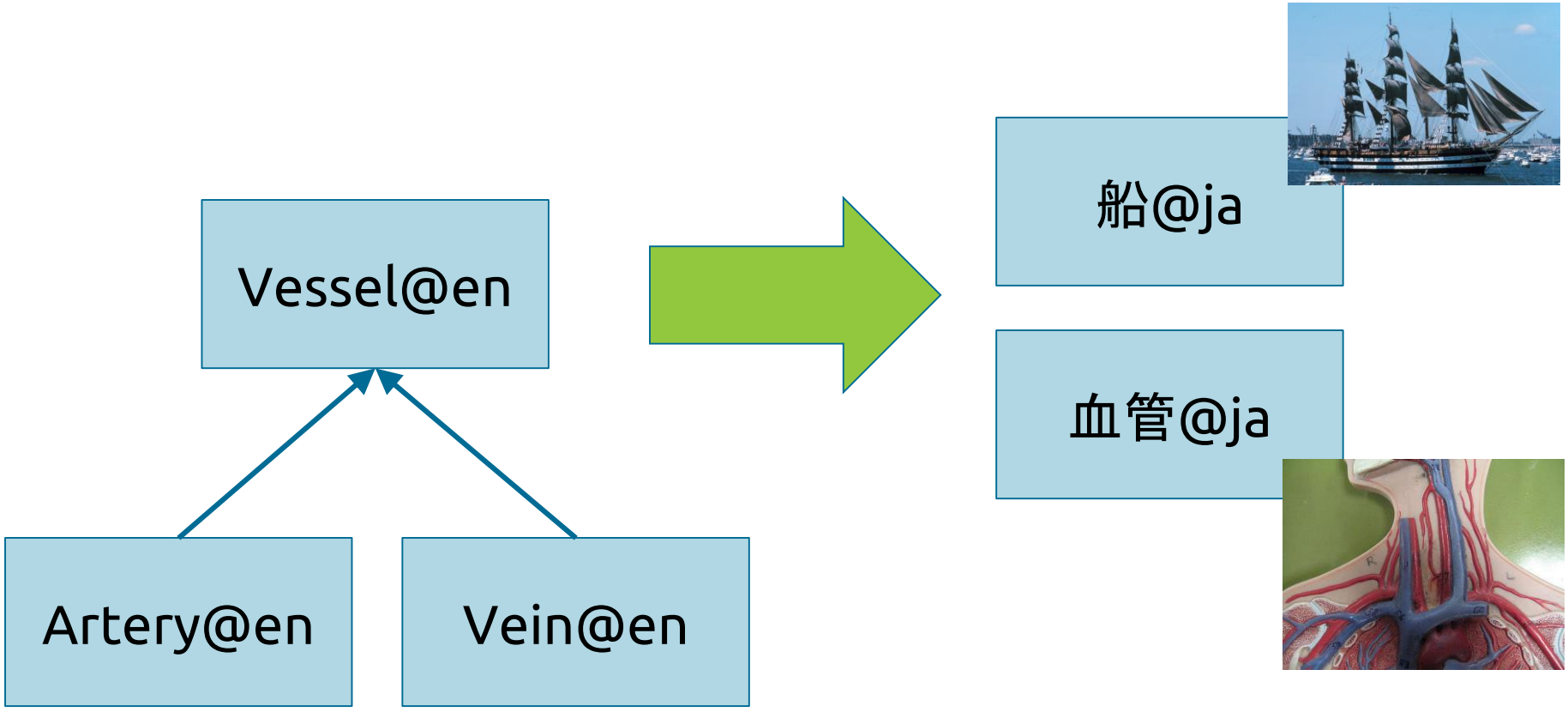


船@ja

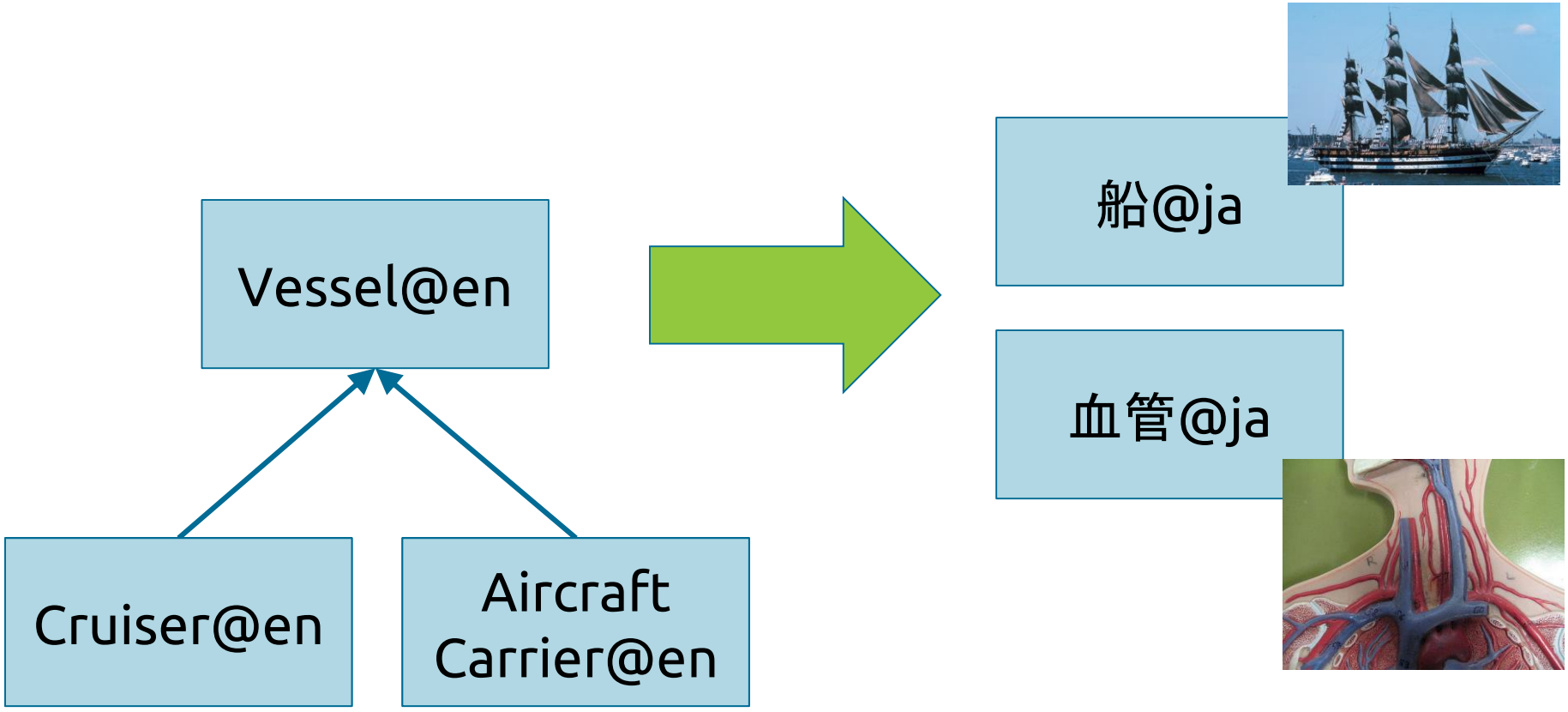
血管@ja



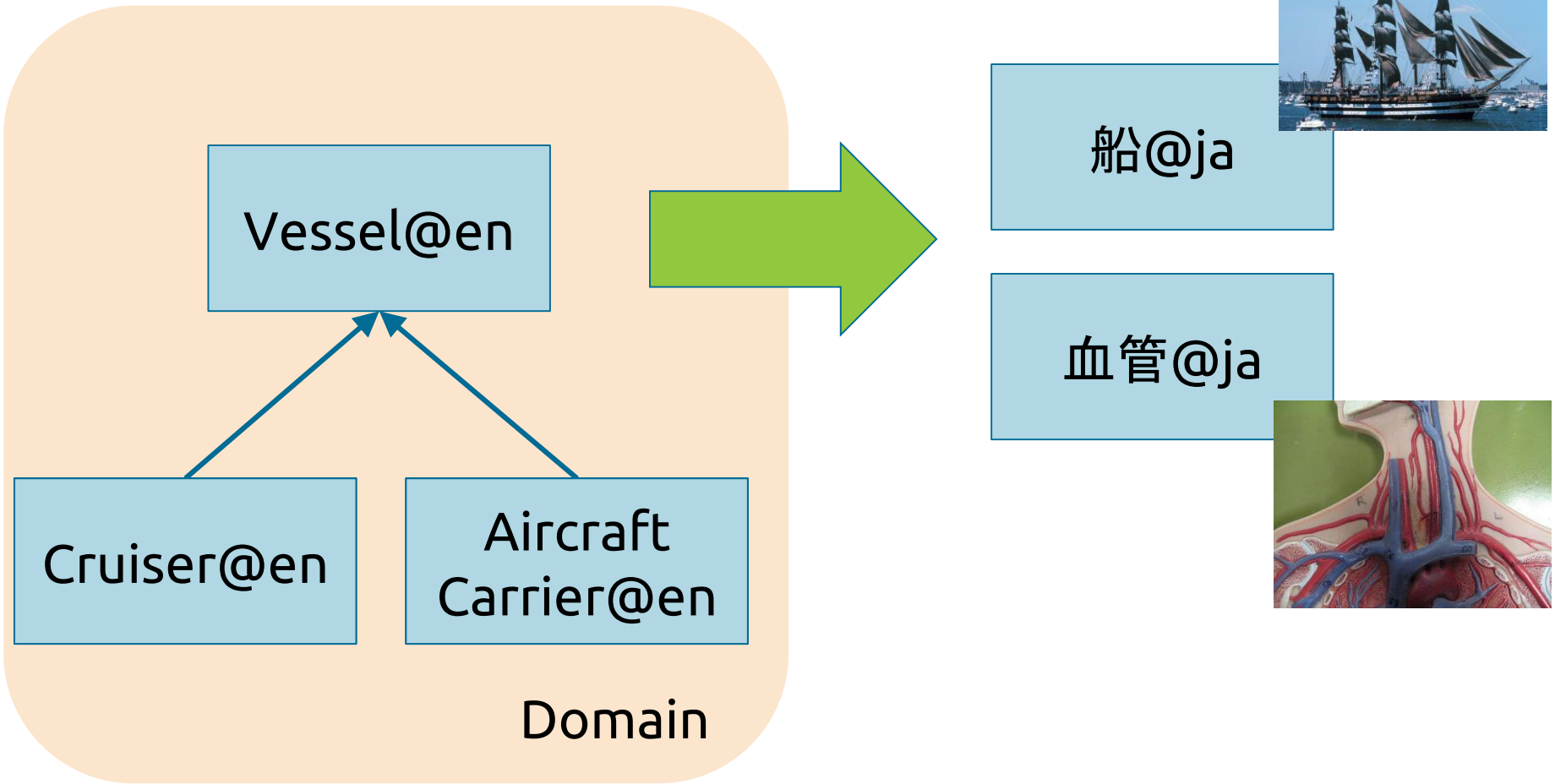
Why is ontology translation hard?



Why is ontology translation hard?



Why is ontology translation hard?



Three-pronged attack

Find new translations
(Domain Lexicon)

Prefer in-domain translation
(Domain Features)

Choose in-domain translations
(Domain Language Model)

Phrase-based Machine Translation

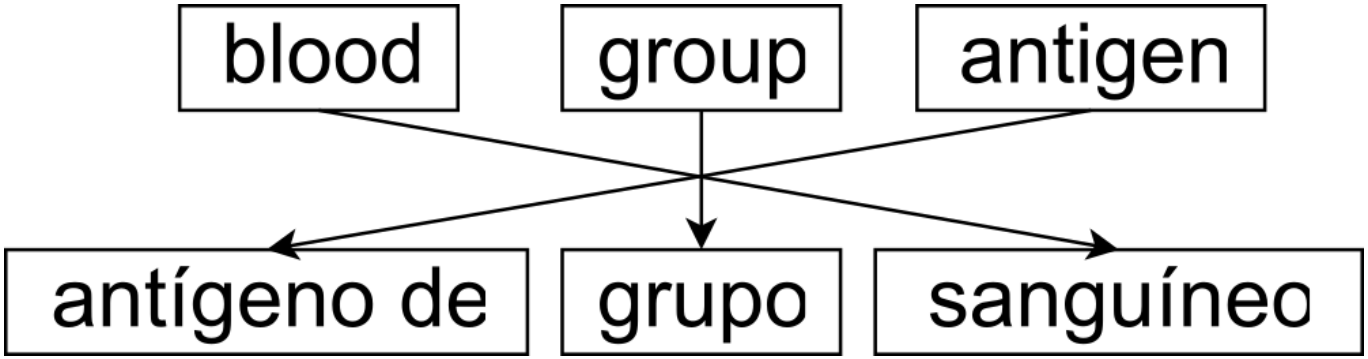
- The ‘workhorse’ of machine translation is the phrase-based model (Koehn et al., 2003, 2007, 2010)
- We want to translate a *foreign* sentence, **f**, into a *translation*, **t**.
- We divide **f** into a sequence of phrases (one or more consecutive words)
  - $\mathbf{f} = \{f_1, \ldots, f_n\}$
- Generate a permuted sequence of translations from a *phrase table*
  - $\mathbf{t} = \{t_{d(1)}, \ldots, t_{d(n)}\}$

Phrase-based Machine Translation II

- We search for a sequence of phrases that maximizes
  - $[\Sigma_i \Sigma_j \alpha_j \varphi_j(f_i, t_i)] + \alpha_l l(\mathbf{t}) + \alpha_d d(\mathbf{f}, \mathbf{t})$
- Where $\varphi_j(f_i, t_i)$ are the *feature scores*
  - Log probability of f_i given t_i
  - Log probability of t_i given f_i
  - Lexical weighting of f_i given t_i
  - Lexical weighting of t_i given f_i
  - Out-of-vocabulary?
  - Constant 1 (to bias towards using fewer phrases)
- $l(\mathbf{t})$ is a language model score
  - Prefer fluent translations

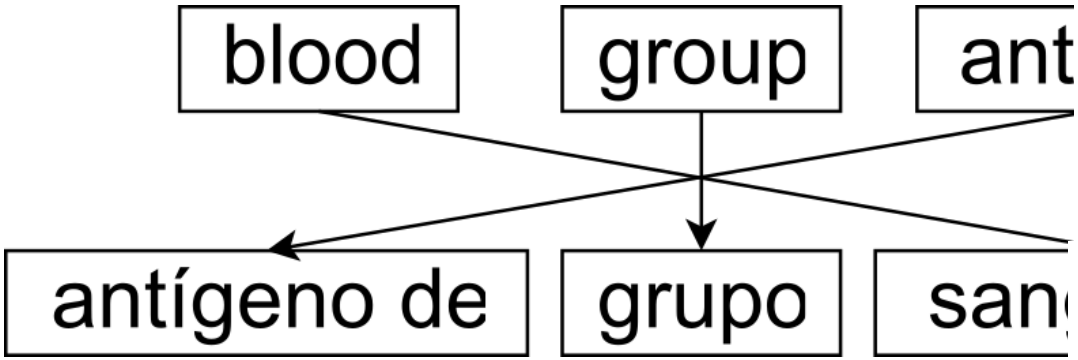
Phrase-based Machine Translation III

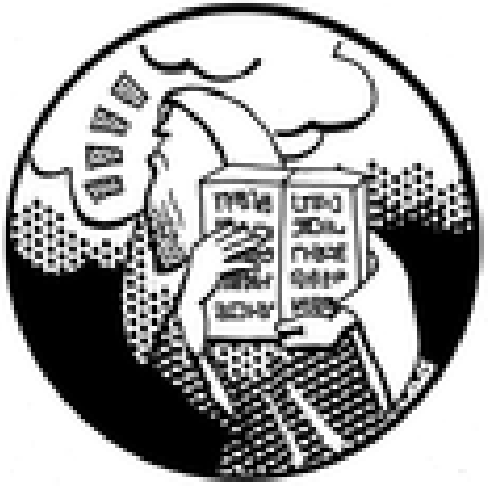
- d(f, t) is the *distortion* score
  - Measures how much the translation has been rearranged
- A *decoder* heuristically finds the optimal division into phrases and the permutation of phrases



Phrase-based Machine Translation III

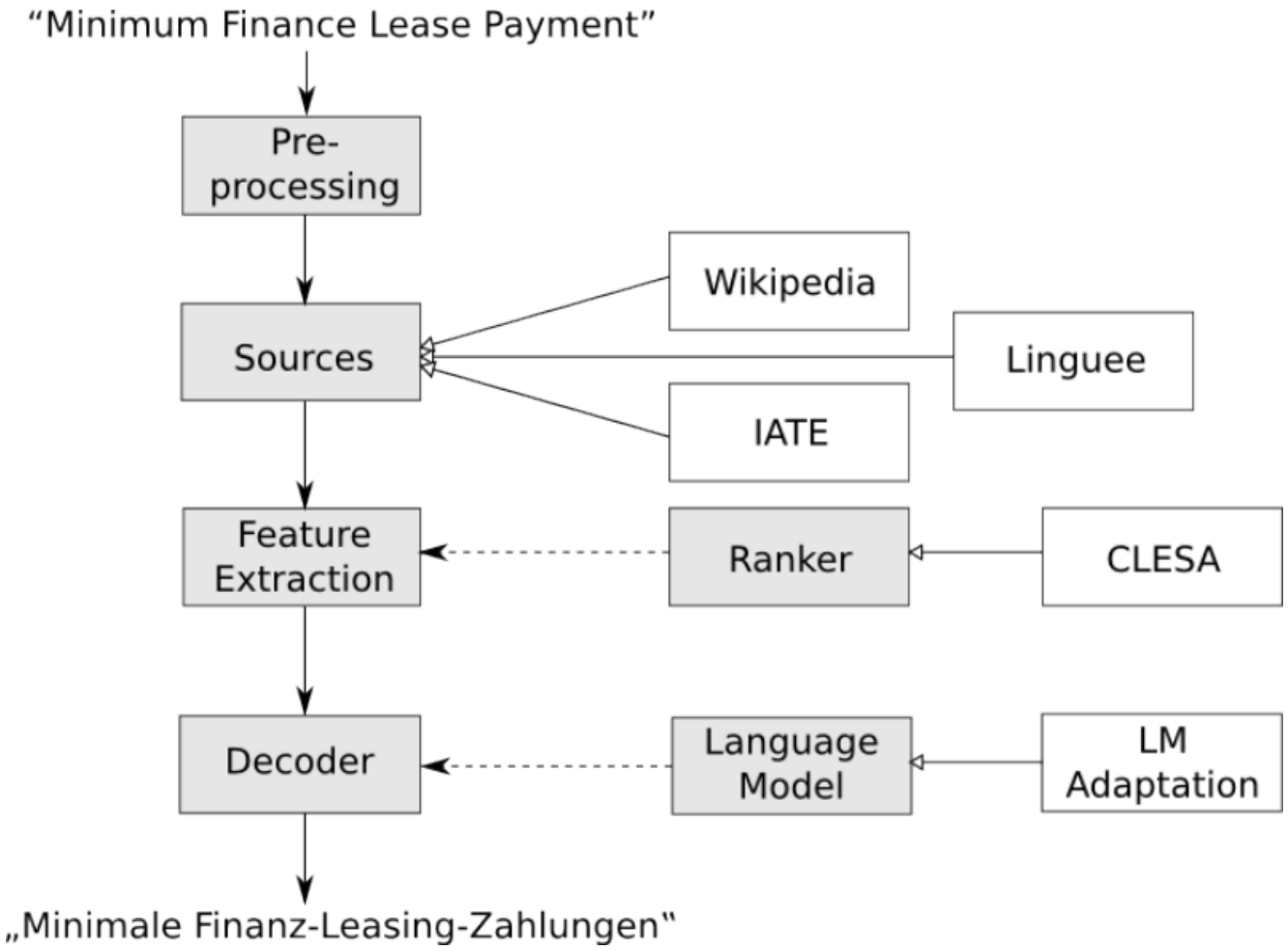
- d(f, t) is the *distortion* score
  - Measures how much the translation has been rearranged
- A *decoder* heuristically finds the optima
  phrases and the permutation of phrase





Moses SMT

Monnet Architecture for MT



Domain lexicon

- Parallel text may not have domain translations
- We add additional data
- Wikipedia
  - We find all articles with names matching an ontology label
  - Find all categories for these articles
  - Filter categories by threshold
  - Add all translations from DBpedia for these categories

Domain lexicon

- Parallel text may not have domain translations
- We add additional data
- Wikipedia
  - We find all articl
  - Find all categori
  - Filter categories
  - Add all translati

Frequency	Wikipedia Category
95	Economics Terminology
62	Generally Accepted Accounting Principles
61	Macroeconomics
55	Accounting Terminology
47	Finance
44	Economic Theories
42	International Trade

Domain lexicon II

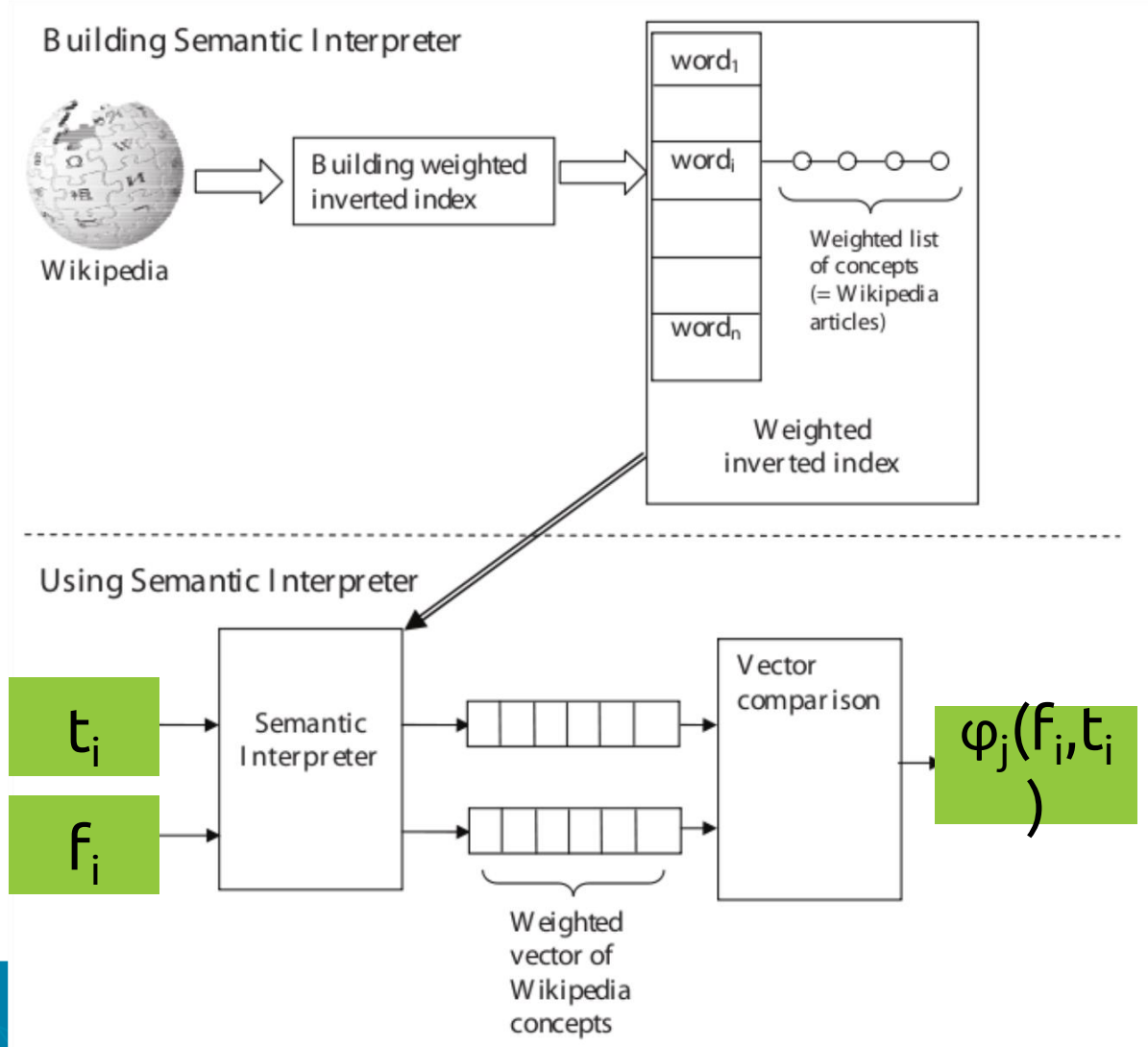
- Linguee
  - Large source of parallel data from the Web
  - We queried ontology labels
  - Generates domain parallel corpus (Financial Domain: ~24,000 sentences)
  - Train phrase table (Moses)
- IATE (Interactive Terminology for Europe)
  - Translations in EU languages
  - Weighted by Cross-lingual Explicit Semantic Analysis (Sorg et al., 2008)

Domain features

We used Cross-lingual extension (Sorg & Cimiano, 2008)

of

Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007)

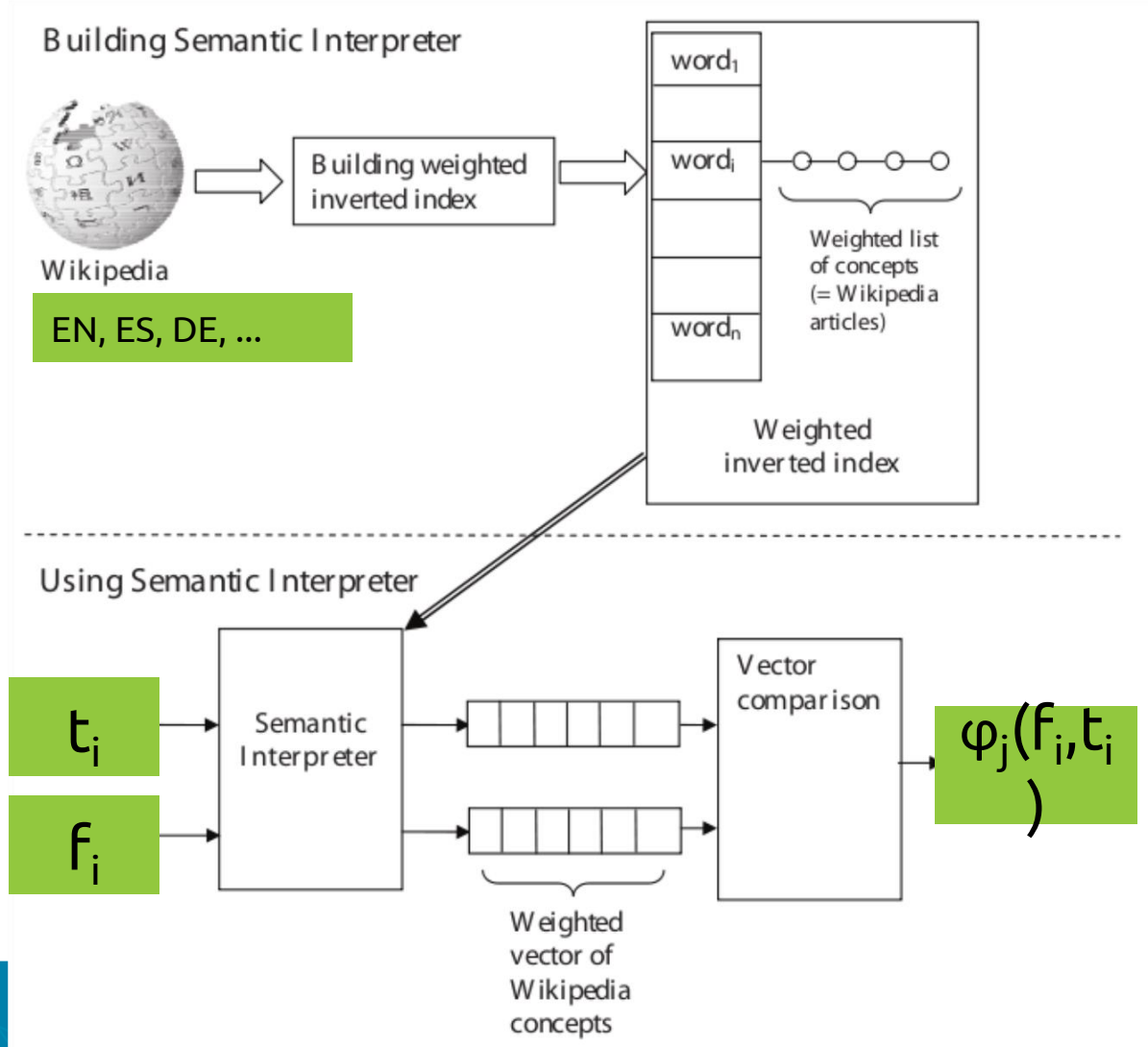


Domain features

We used Cross-lingual extension (Sorg & Cimiano, 2008)

of

Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007)



Domain Language Model

- The language model estimates the likelihood of the target sentence
  - $p(w_1 \dots w_n) = \Pi\, p(w_i|w_{i-n} \dots w_{i-1})$
  - $p(w_i|w_{i-n} \dots w_{i-1}) = c(w_{i-n} \dots w_{i-1}w_i) / c(w_{i-n} \dots w_{i-1}*)$
- For each document in our corpus we estimate its domain relevance (using ONETA , McCrae et al., 2014)
  - $s_O(d)$
- The count is weighted by document relevance
  - $c(w_{i-n} \dots w_i) = \Sigma\, s_O(d)\, c_d(w_{i-n} \dots w_i)$

Domain Language Model

- The language model estimates the likelihood of the target
  - p(
  - p(
- For e
  dom
  - s_O(
- The c
  - $c(w_{i-n} \ldots w_i) = \Sigma\, s_O(d)\, c_d(w_{i-n} \ldots w_i)$

$$p(\text{Welcome to Kobe}) = p(\text{Welcome}) \times p(\text{to}|\text{Welcome}) \times p(\text{Kobe}|\text{Welcome to})$$

$$p(\text{Kobe}|\text{Welcome to}) = c(\text{Welcome to Kobe}) \div c(\text{Welcome to } *)$$

14)

Domain Language Model

- The language model estimates the likelihood of the target sentence
  - $p(w_1 \dots w_n) = \Pi\, p(w_i|w_{i-n} \dots w_{i-1})$
  - $p(w_i|w_{i-n} \dots w_{i-1}) = c(w_{i-n} \dots w_{i-1}w_i) / c(w_{i-n} \dots w_{i-1}*)$
- For each document in our corpus we estimate its domain relevance (using ONETA , McCrae et al., 2014)
  - $s_O(d)$
- The count is weighted by document relevance
  - $c(w_{i-n} \dots w_i) = \Sigma\, s_O(d)\, c_d(w_{i-n} \dots w_i)$

Evaluation

Metrics

- BLEU
- BLEU-2
- METEOR
- NIST
- PER
- WER
- TER

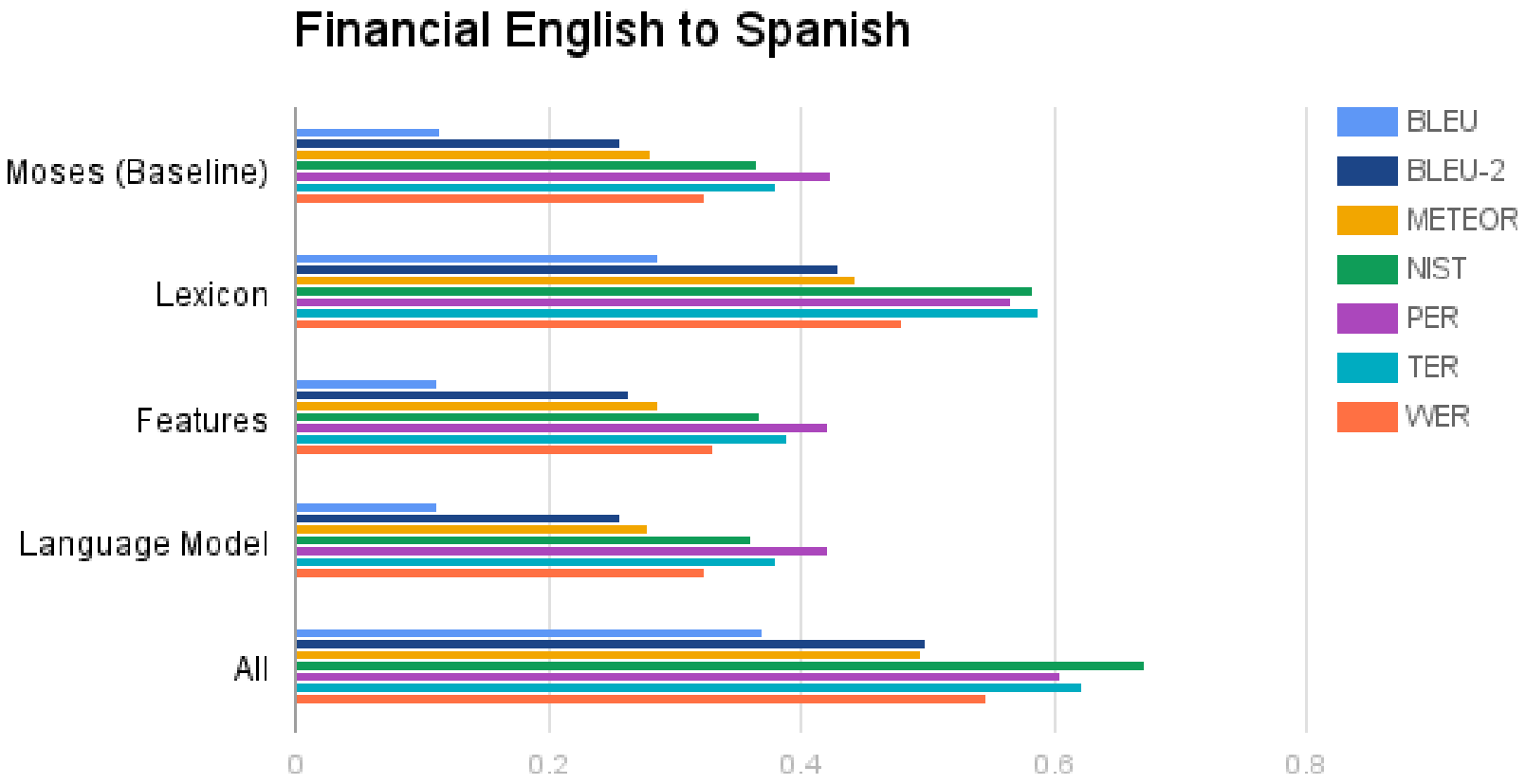
Ontology	Size	Language
IFRS 2009	2,757	
DE-GAAP	2,782	
LAG	196	
RB	1,449	
HB	857	

Finance Ontologies

Public Service Ontologies

BLEU is gold standard for machine translation but has been shown to perform very poorly on short texts (McCrae et al., 2011)

Results


Financial English to Spanish
Moses (Baseline)
Lexicon
Features
Language Model
All
0
0.2
0.4
0.6
0.8
BLEU
BLEU-2
METEOR
NIST
PER
TER
WER

Discussion

- Strongest improvement is provided by domain lexicon
  - Generalized for any corpus/domain (Arcan et al., 2015)
- Other methods alone did not significantly improve over the baseline
- The combination of all approaches better than domain lexicon in most settings (33/42 settings)
  - Main exception was English to Dutch for public services
- Domain adaptation was more effective for financial domain than public services domain

Conclusion

- Domain lexicon improves translation
  - Capable of suggesting new translations
  - Requires parallel text
- Domain selection weaker but still useful
- Code is open source:
  - https://github.com/monnetproject/translation
- Online demo (OTTO Ontology Translator)
  - http://server1.nlp.insight-centre.org/otto/

References

A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, G. Aguado de Cea, Guidelines for multilingual linked data, in: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, New York, NY, USA, 2013, pp. 14–25. doi:10.1145/2479787.2479867.

P. Koehn, F. J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003, pp. 48–54.

P. Koehn, Statistical machine translation, Cambridge University Press, 2010.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., Moses: Open source toolkit for statistical machine translation, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, 2007, pp. 177–180.

E. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in: Proceedings of the 20th international joint conference on artificial intelligence, Vol. 6, 2007, p. 12.

References

P. Sorg, P. Cimiano, Cross-lingual information retrieval with explicit semantic analysis, in: Working Notes for the CLEF 2008 Workshop, 2008.

J. McCrae, P. Cimiano, R. Klinger, Orthonormal explicit topic analysis for cross-lingual document matching, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1732–1740.

J. McCrae, E. Montiel-Ponsoda, G. Aguado de Cea, M. J. Espinoza Mejía,P. Cimiano, Combining statistical and semantic approaches to the translation of ontologies and taxonomies, in: Proceedings of 7th Workshop on Syntax, Structure and Semantics in Statistical Translation, 2011, pp. 116–125.

M. Arcan, M. Turchi, P. Buitelaar, Knowledge portability with semantic expansion of ontology labels, in: The 53rd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 708–718.