FONDAZIONE
BRUNO KESSLER

Translating Ontologies in Real-World Settings with ESSOT

Mihael Arcan[1], Mauro Dragoni[2], and Paul Buitelaar[1]

[1]Insight Centre for Data Analytics, NUI Galway
[2]Fondazione Bruno Kessler (FBK), Shape and Evolving Living Knowledge Unit (SHELL)

International Semantic Web Conference 2016
Kobe, Japan, October 19th 2016

Presentation's main points

- Our Goal:
  to demonstrate why a domain-aware machine translation system can help domain experts in translating ontologies

Presentation's main points

- Our Goal:
  to demonstrate why a domain-aware machine translation system can help domain experts in translating ontologies

ESSOT: a collaborative knowledge management architecture for ontology translation

Presentation's main points

- Our Goal:
  to demonstrate why a domain-aware machine translation system can help domain experts in translating ontologies

ESSOT: a collaborative knowledge management architecture for ontology translation

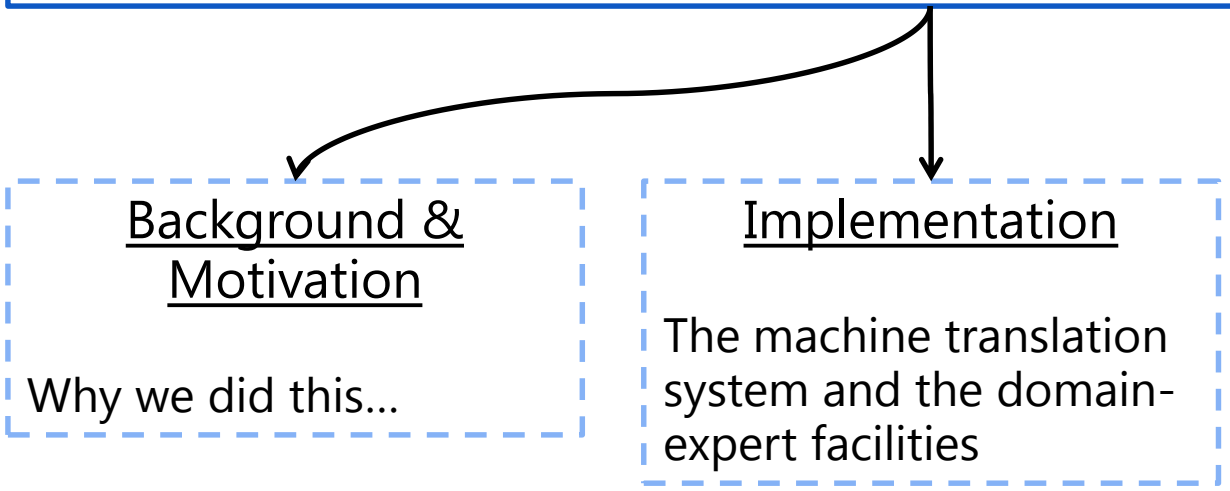
Background & Motivation

Why we did this…

Presentation's main points

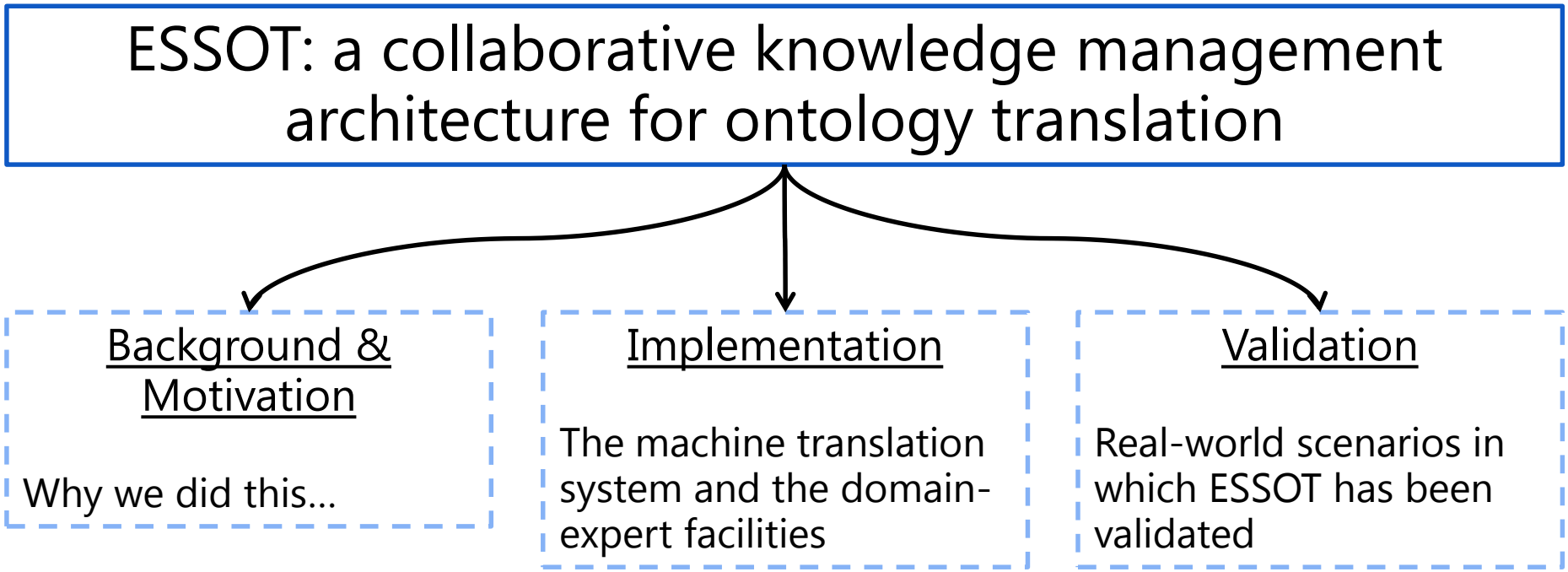
- Our Goal:
  to demonstrate why a domain-aware machine translation system can help domain experts in translating ontologies

ESSOT: a collaborative knowledge management architecture for ontology translation

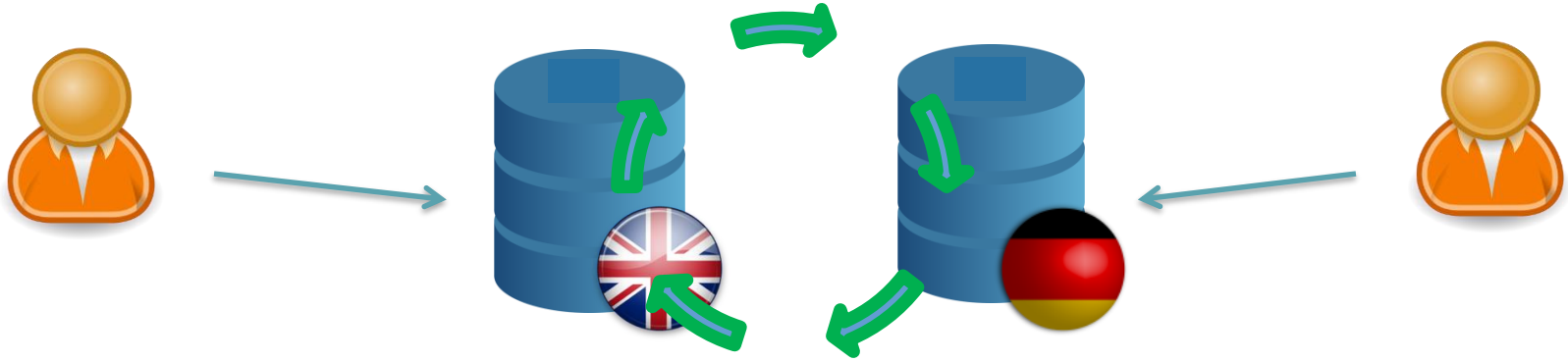


Presentation's main points

- Our Goal:
  to demonstrate why a domain-aware machine translation system can help domain experts in translating ontologies

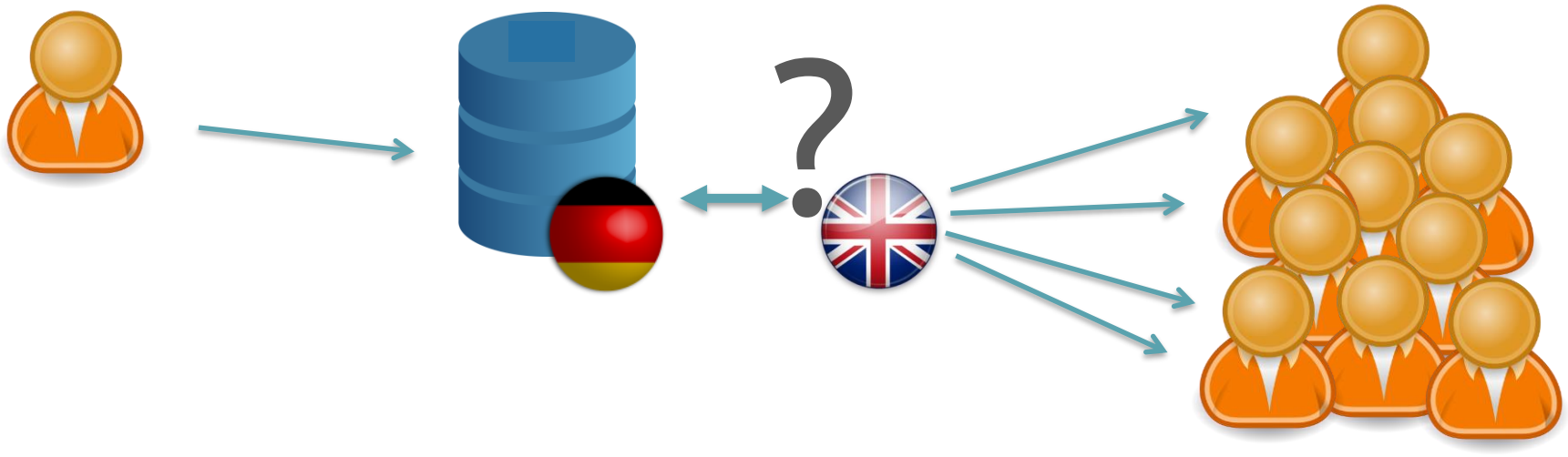


Motivation 1 – Breaking language barriers



- Ontologies created by people do not know other languages than their mother tongue
- Collaborative work and alignment in a multi-lingual scenario.
  - end-user applications

Motivation 2 – Knowledge Dissemination



- Artefacts are created in specific languages due to policy constraints

Translation Task: Documents vs Ontologies

Example: ‘”vessels” → Gefäß or Schiff

Translation Task: Documents vs Ontologies

Example: '"vessels" → Gefäß or Schiff

DOCUMENT

- Context 1 (Gefäß): The blood **vessels** are the part of the circulatory system that transports blood throughout the human body. There are three major types of blood **vessels**

- Context 2 (Schiff): Ships are typically large ocean-going **vessels**; whereas boats are smaller, and typically travel most often on inland or coastal waters.

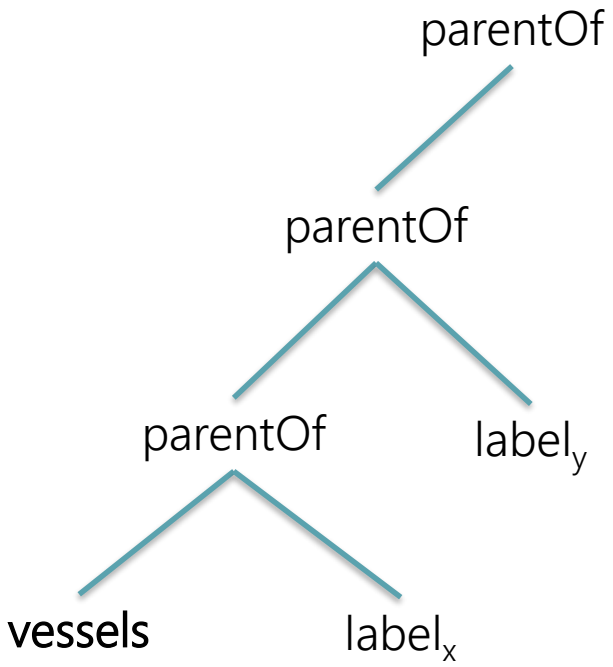
Translation Task: Documents vs Ontologies

Example: '"vessels" → Gefäß or Schiff

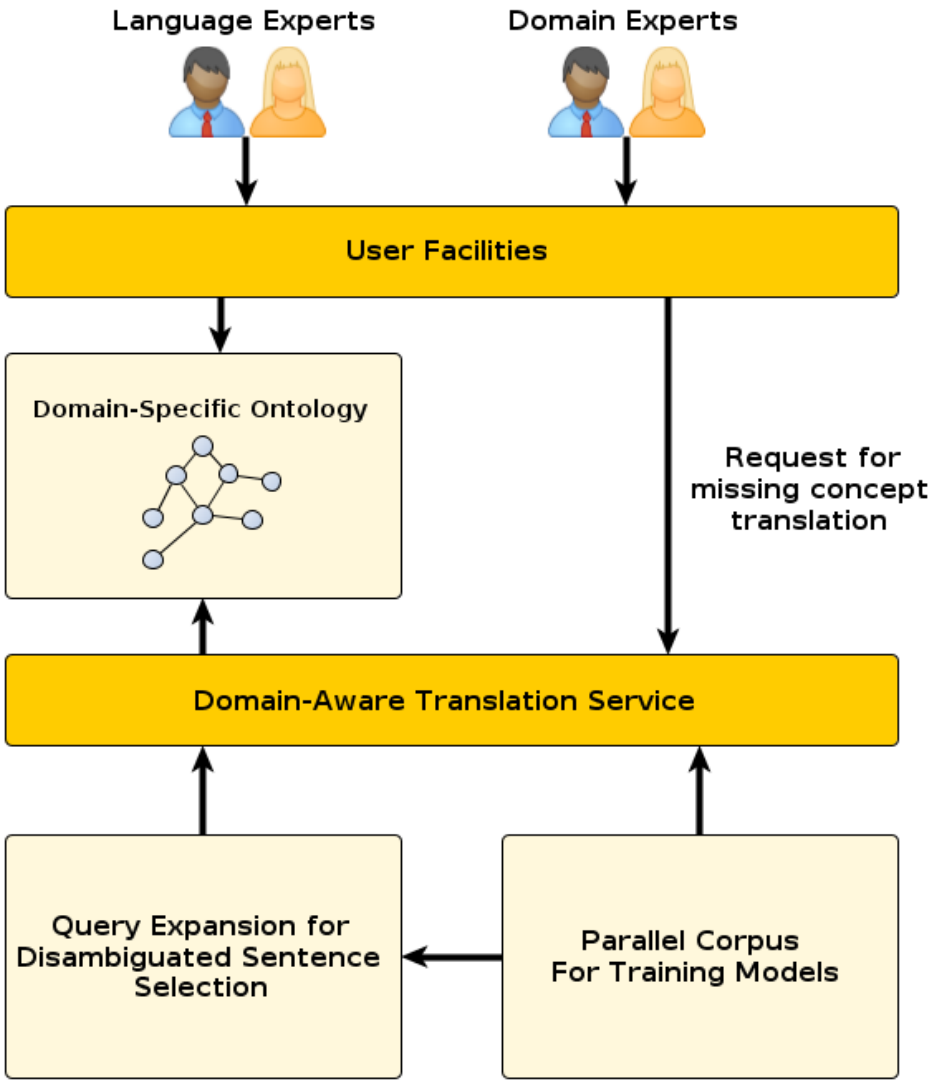
DOCUMENT

- Context 1 (Gefäß): The blood **vessels** are the part of the circulatory system that transports blood throughout the human body. There are three major types of blood **vessels**
- Context 2 (Schiff): Ships are typically large ocean-going **vessels**; whereas boats are smaller, and typically travel most often on inland or coastal waters.

ONTOLOGY



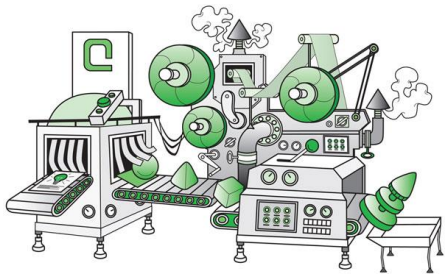
ESSOT Architecture



Construction of the Translation Model

Construction of the Translation Model

Construction of the Translation Model

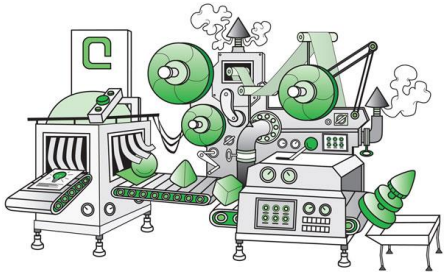


Word2Vec Machinery

Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Construction of the Translation Model

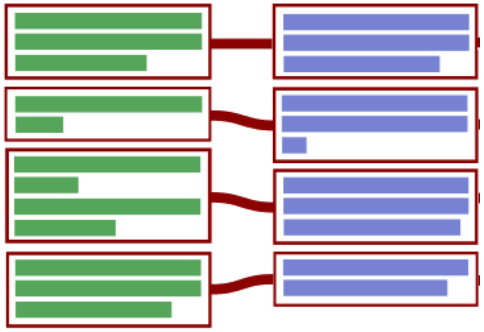


Word2Vec Machinery

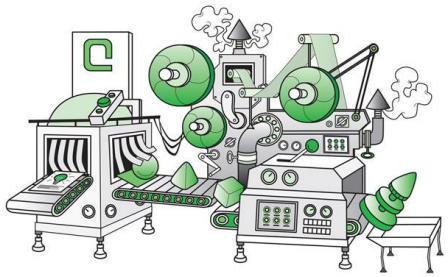
Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



Construction of the Translation Model

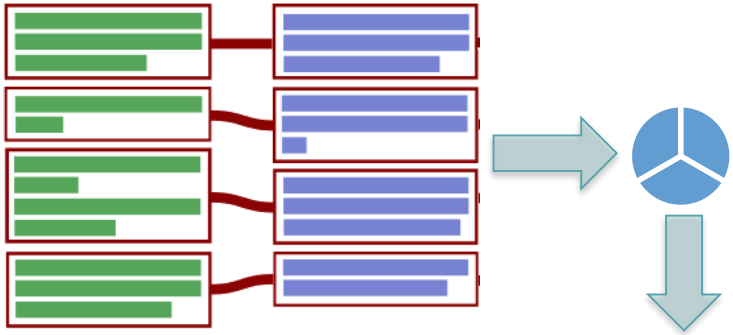


Word2Vec Machinery

Multi-dimensional Matrix

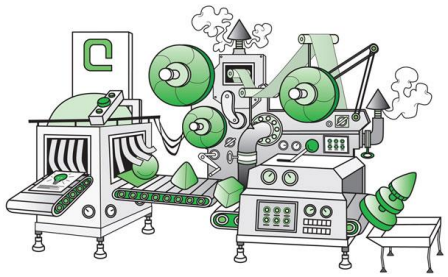
vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



Generic Translation Model

Construction of the Translation Model

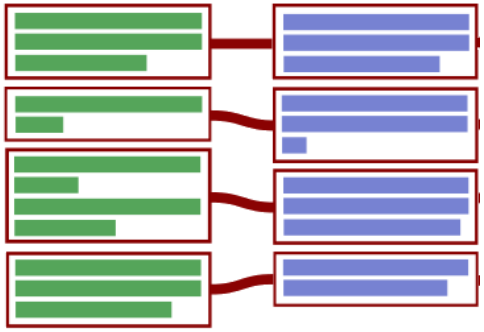


Word2Vec Machinery

Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



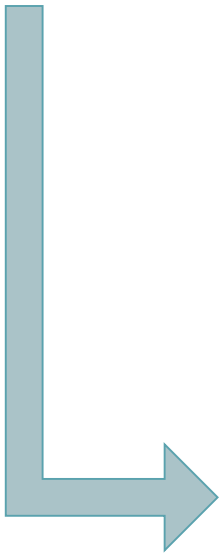
Generic Translation Model

n-best sentences

Online Domain Adaptation

Online Domain Adaptation

Ontology



vessel

blood	0.539
medical	0.512
body	0.493
vein	0.490
disease	0.487
biomedical	0.481

Online Domain Adaptation

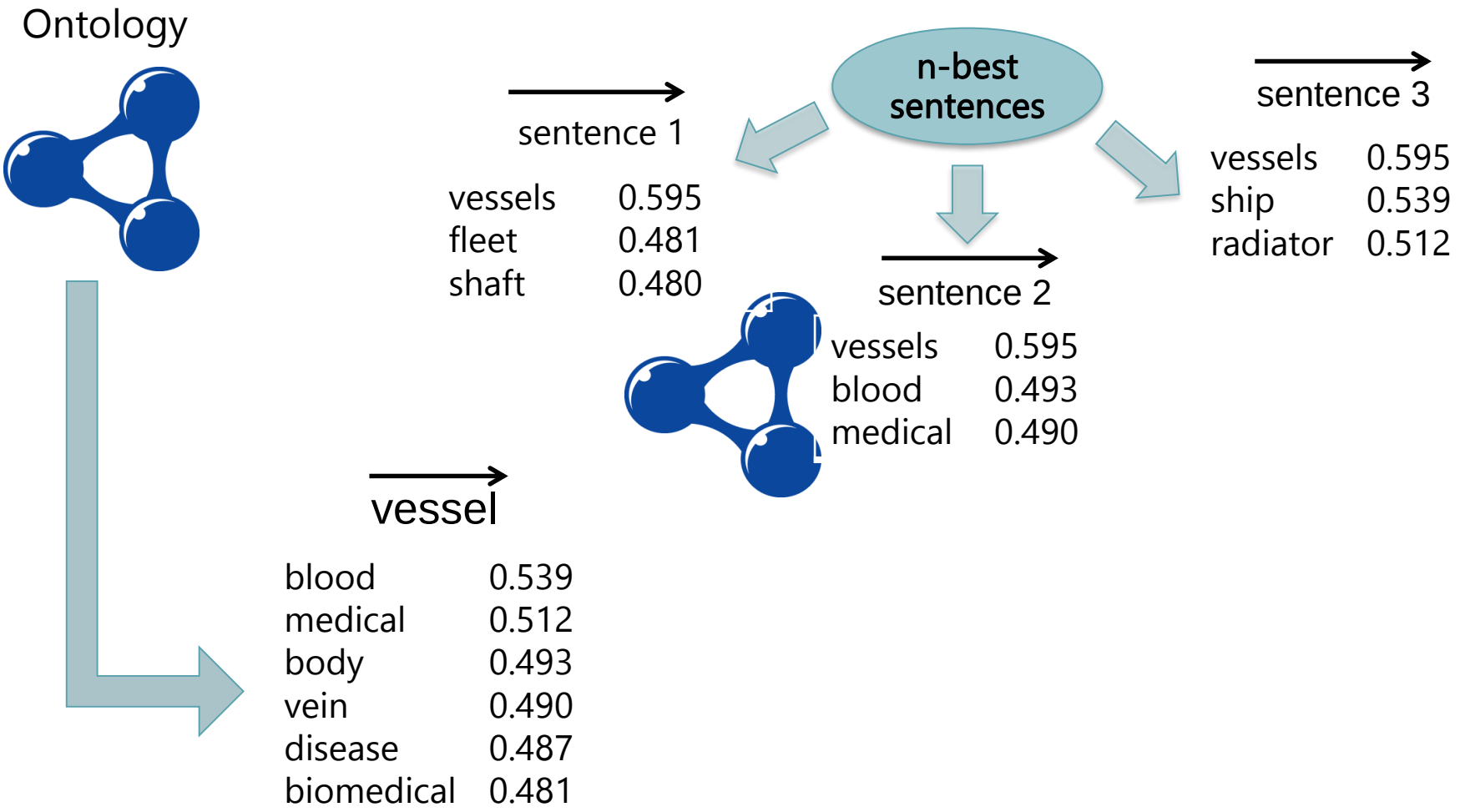
Ontology

n-best sentences

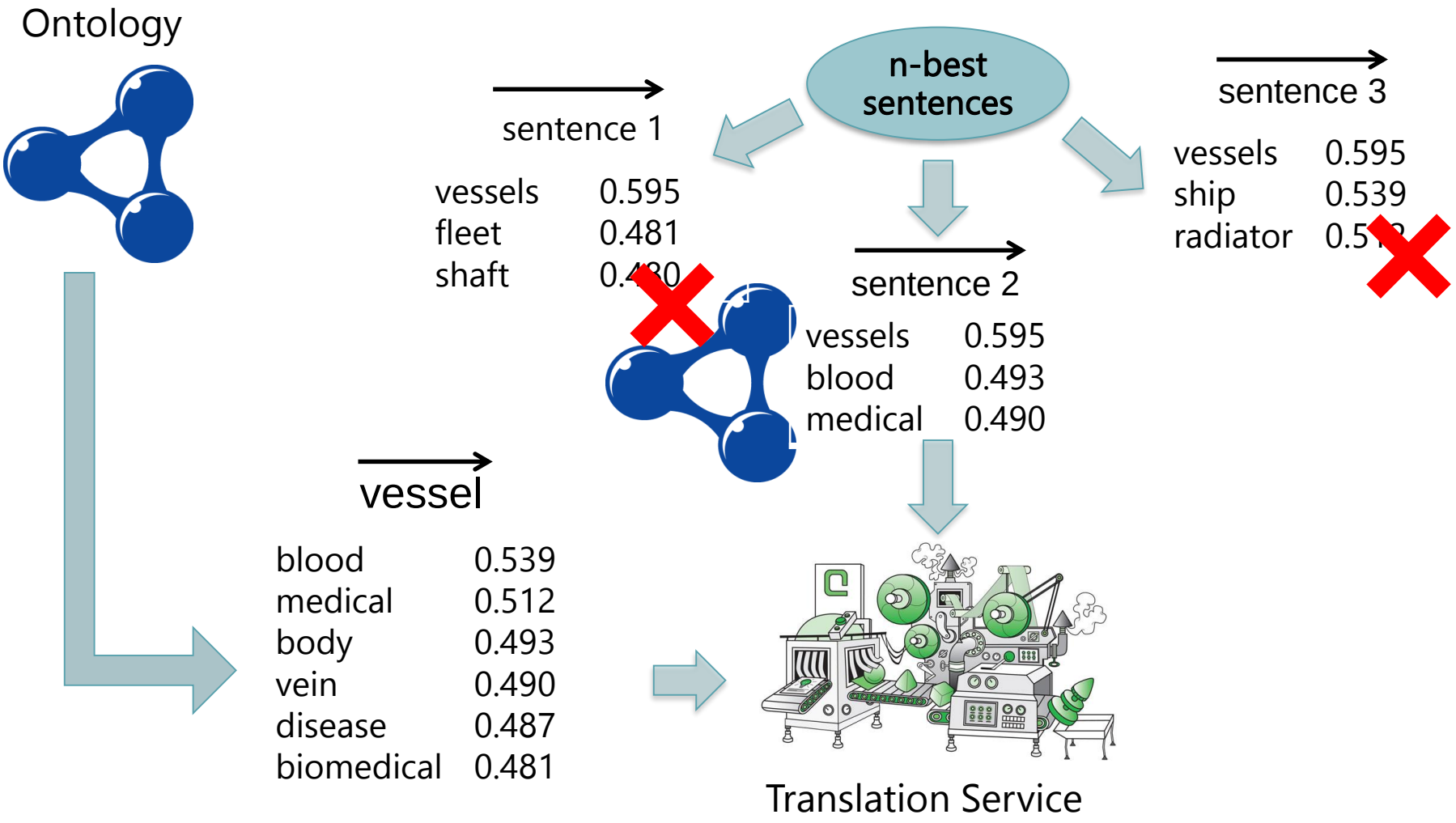
vessel

blood	0.539
medical	0.512
body	0.493
vein	0.490
disease	0.487
biomedical	0.481

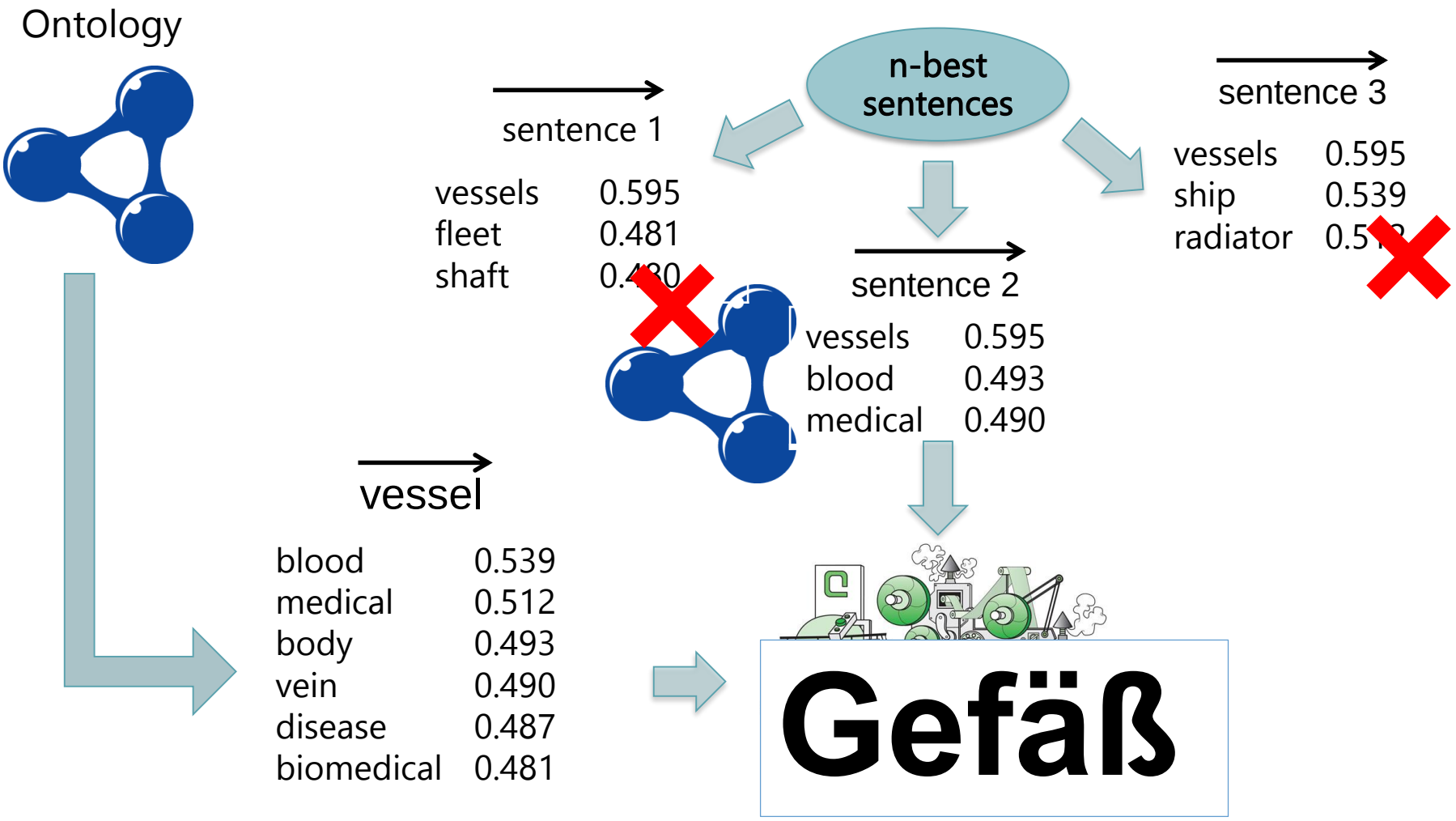
Online Domain Adaptation



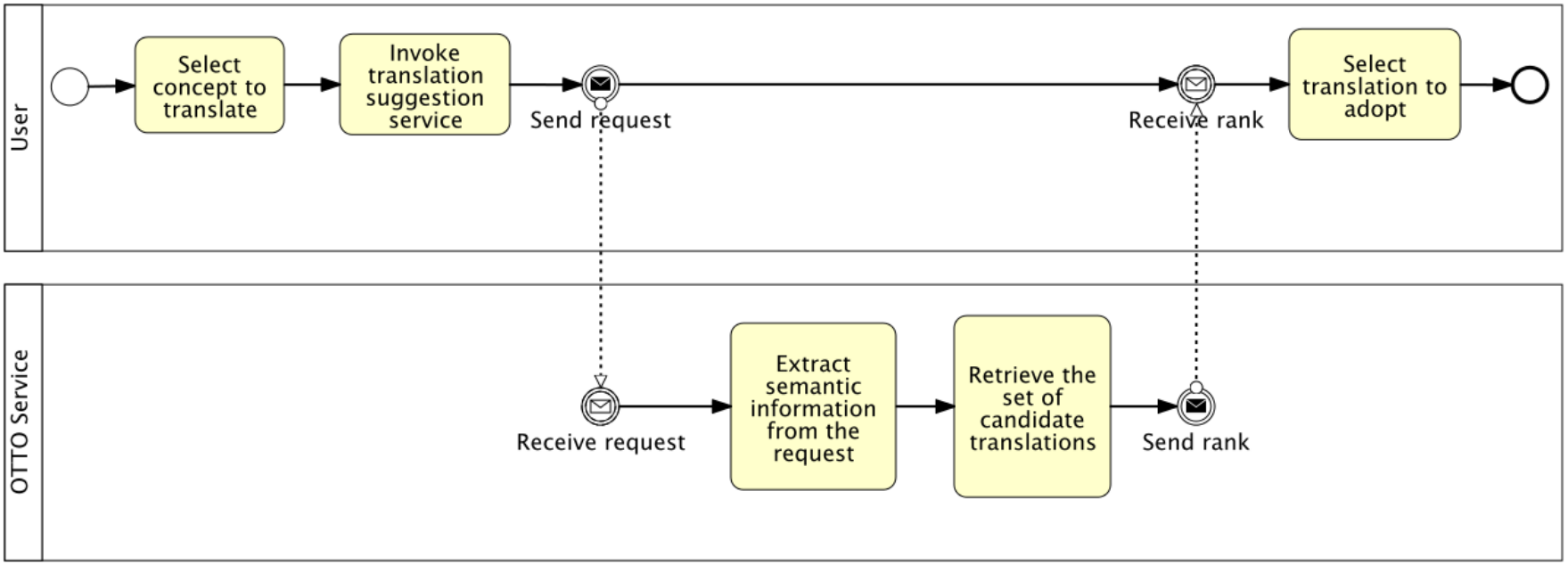
Online Domain Adaptation



Online Domain Adaptation

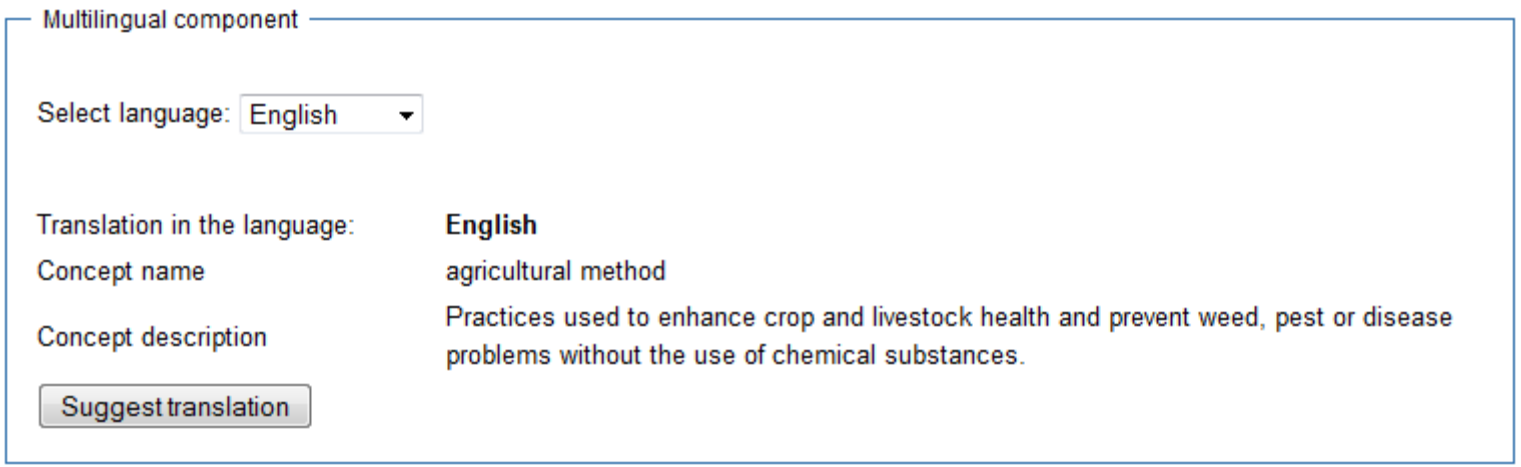


Translation Request Pipeline

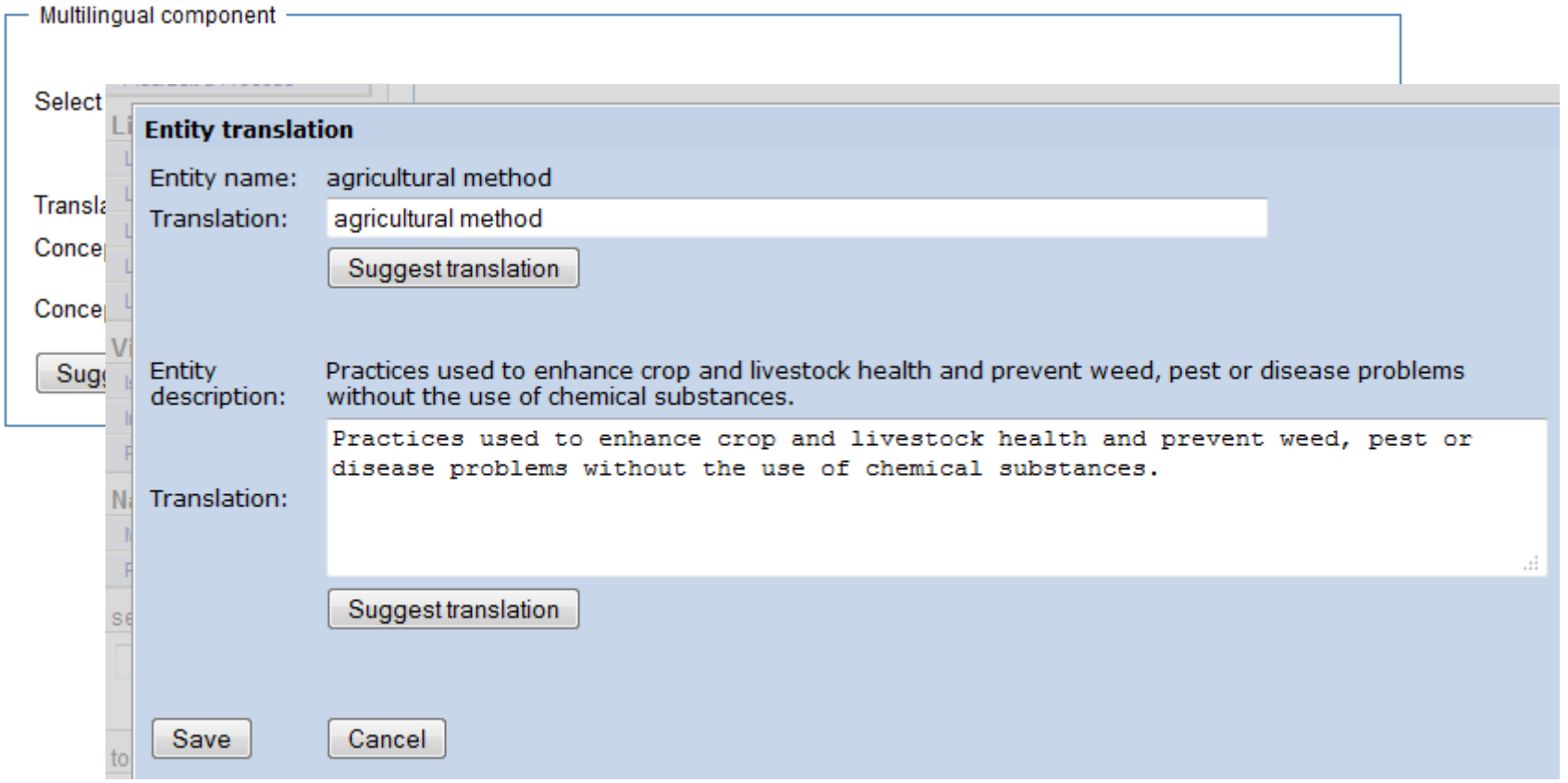


ESSOT: User Interfaces

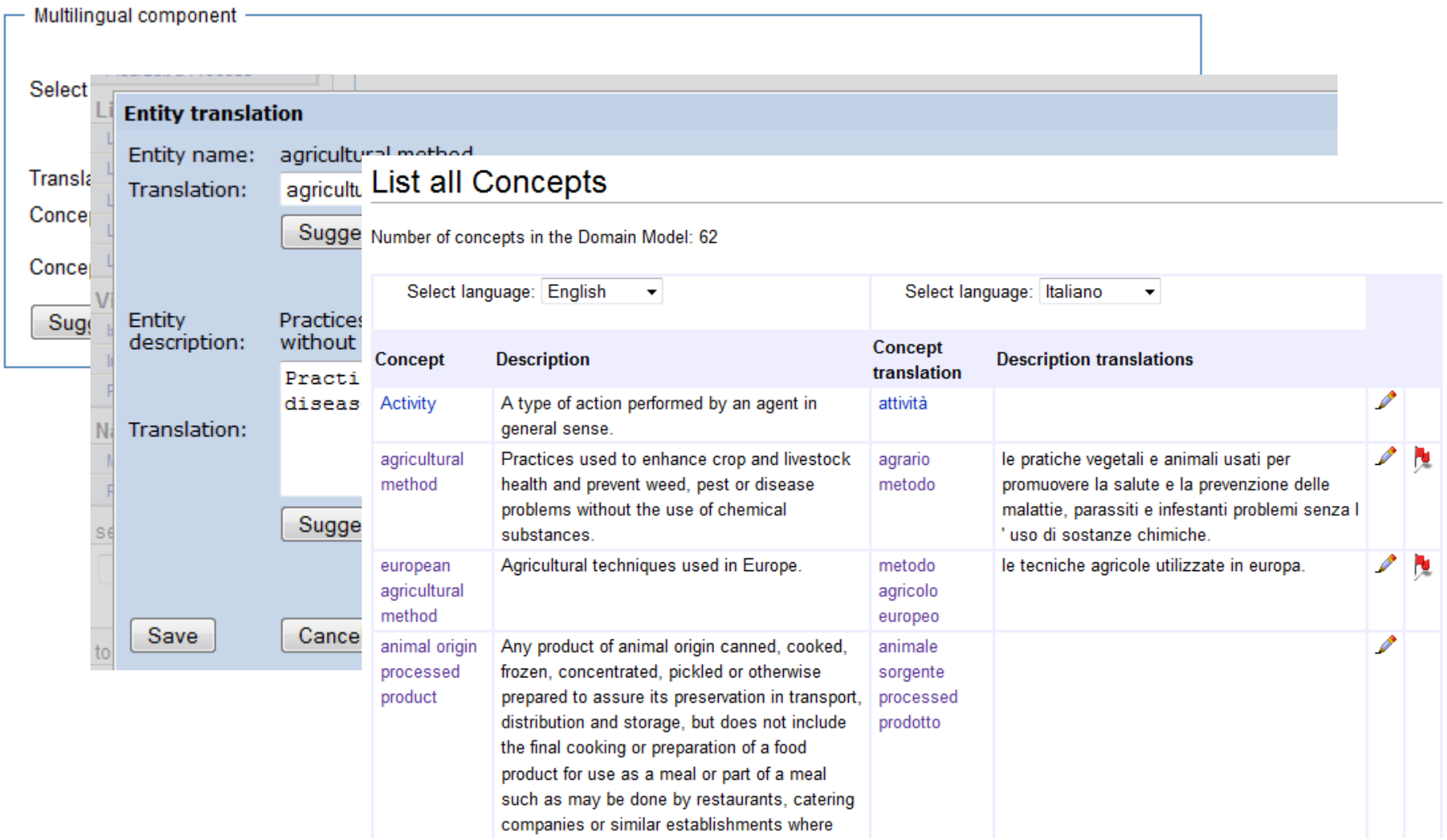
ESSOT: User Interfaces



ESSOT: User Interfaces



ESSOT: User Interfaces



ESSOT: The OTTO Translation Service

OTTO OnTology TranslatiOn System

Home | About | Team | API | Other Projects

Upload the ontology to be translated:

Choose File | No file chosen

or provide the internet address to it:

Enter ontology address

Submit Data

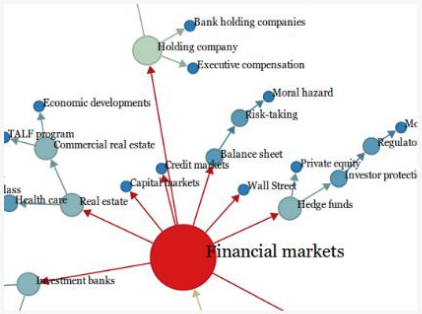
E-mail (optional):

The extraction and translation of an ontology may take several minutes (or hours). The provided optional email will be used to inform the user of the start and end of the translation process, as well the web address where the data will be stored. After the translation process, the provided email will be deleted from the system.

This service is brought to you by
http://nlp.insight-centre.org/

The OTTO Translation System was developed to enhance English monolingual ontologies in RDF (n-triples) or Owl with multilingual lexical knowledge. Currently, OTTO translates from English into German, Spanish, Italian, Irish, Slovenian and Czech.

Example of an RDF representation in the financial domain (see rdf/n-triples file).



RDF file was generated by Saffron.

http://server1.nlp.insight-centre.org/otto

Example

Input


```
{
  {
    "label2translate": "vessels",
    "concept_context":
      [
        "blood",
        "medical",
        "disease",
        "biomedical",
      ],
    "translate2":"de"
}
```


Output


```
{
  "possible_translations":
  {
    "blutgefäßen":-15.8438,
    "gefäßen":-2.4100,
    "halsgefäße":-2.6682
    .....
  },
  "time":"24 wallclock secs",
  "source_label":"vessels",
  "best":"gefäßen"
}
```


http://server1.nlp.insight-centre.org/otto/rest_service.html

Evaluation Procedure

- **Qualitative evaluation**:
  We collected subjective judgments from the language experts involved in the evaluation of the tool on general usability of the components and to provide feedback for future improvements.
  - **Research Questions 1:** Does the proposed system provide an effective support, in terms of the quality of suggested translations, to the management of multilingual ontologies?
  - **Research Questions 2:** Do the MoKi functionalities provide an effective support to the collaborative management of a multilingual ontology?

- **Quantitative Evaluation**:
  We collected objective measures concerning the effectiveness of the translations suggested by the embedded machine translation service.
  - 3 metrics: BLEU, METEOR, TER
  - 3 language pairs: EN←→DE, EN←→IT, EN←→ES
  - 6 ontologies:
    - Organic.Lingua, Presto
    - TheSoz, Geoskill, DOAP, STW

Qualitative Evaluation Results

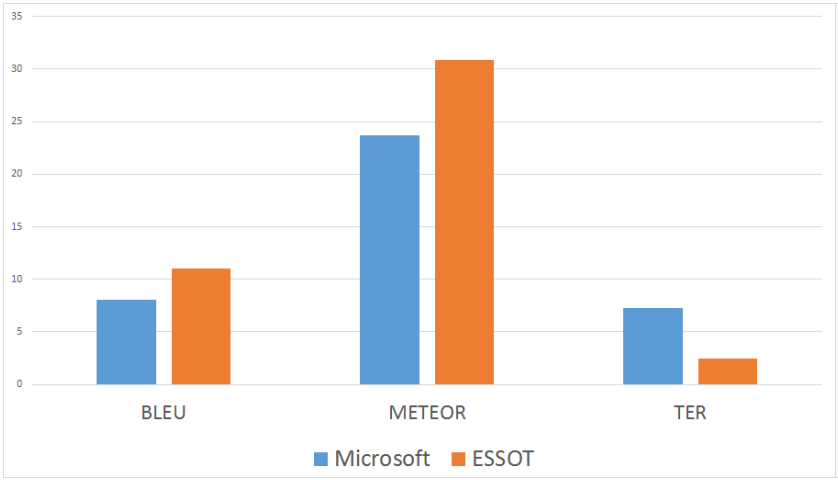
- Language Experts view
  - Pros: Easy to use for managing translations (9)
  - Cons: Usable interface for showing concept translations (3)

- Approval and Discussion
  - Pros: Pending approvals give a clear situation about concept status (4)
  - Cons: Discussion masks are not very useful and the approval process might be improved (8)

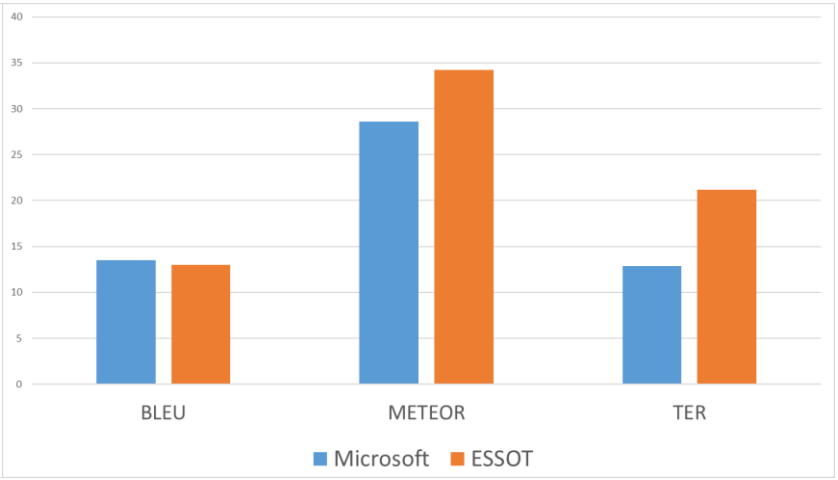
- Quick Translation feature
  - Pros: Best facility for translating concepts (8)
  - Cons: Improvable interface design (3)

Quantitative Evaluation Results

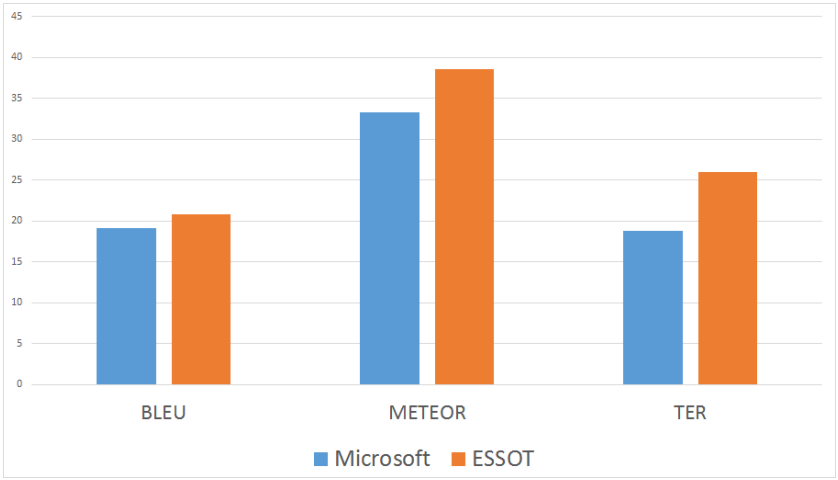
English ←→ German



English ←→ Italian



English ←→ Spanish



- General robustness of the approach
- EN←→DE: wrong translations needed many editing operations
- Computational time to improve
  - avg. 4.1 seconds per request

Conclusions

- ESSOT integrates the OTTO domain-adaptable semantic translation service and the MoKi collaborative knowledge management tool in an effective pipeline for multilingual ontology management.
- ESSOT provides helpful suggestions for performing ontology translation (RQ1) and the provided interfaces are usable and useful for supporting the language experts in the translation activity (RQ2)
- The translation quantitative evaluation shows significant improvements over the Microsoft Translation system
- Future work:
  - to extend this strategy to other problems (e.g. ontology enrichment);
  - improve the approval management task;
  - improve the service efficiency.

SEE YOU LATER AT THE DEMO SESSION !!!!!

It’s time for questions…

Mauro Dragoni
Fondazione Bruno Kessler

https://shell.fbk.eu/index.php/Mauro_Dragoni
dragoni@fbk.eu

Translation Examples

Ontology Label	Microsoft	OTTO	Gold standard
driver	**Treiber**	**Fahrer**	Fahrer
stroke	**Strich**	Schlaganfall	Schlaganfall
demonstration	**Vorführung**	**Demonstration**	Demonstration
race	**Rennen**	**Rasse**	Rasse
fine	**feine**	**Geldbuße**	Geldbuße
bonapartism	**Bonapartismus**	bonapartism	**Bonapartismus**
shamanism	Schamanismus	shamanism	Schamanismus

ESSOT Training Data

Corpus Name	Domain	→ German	→ Italian	→ Spanish
JRC-Acquis	legislative	√		√
DGT	legislative		√	√
ECB*	financial	√	√	√
EMEA*	medical	√	√	√
Europarl	proceedings of the EP	√	√	√
Gnome*	information technology		√	√
KDE*	information technology	√	√	√
OpenOffice*	information technology.	√		
PHP*	information technology	√		
OpenSubtitles13*	translated movie subtitles	√	√	√
TED	talk transcriptions		√	√
MultiUn	United Nation Documents	√		
Wikipedia Titles	terminology expressions		√	√
	Number of parallel sentences:	10,222,259	23,441,275	37,583,385

The Organic.Lingua and Presto Scenarios

- The Organic.Lingua project:
  multilingual services for discovery, retrieval, exploitation and extension of content related to Organic Agriculture and AgroEcology, particularly:
  - find resources in languages different from the ones in which the query has been formulated;
  - manage meta-data information for resources in different languages;
  - contribute to evolving content, e.g. providing services for content generation.

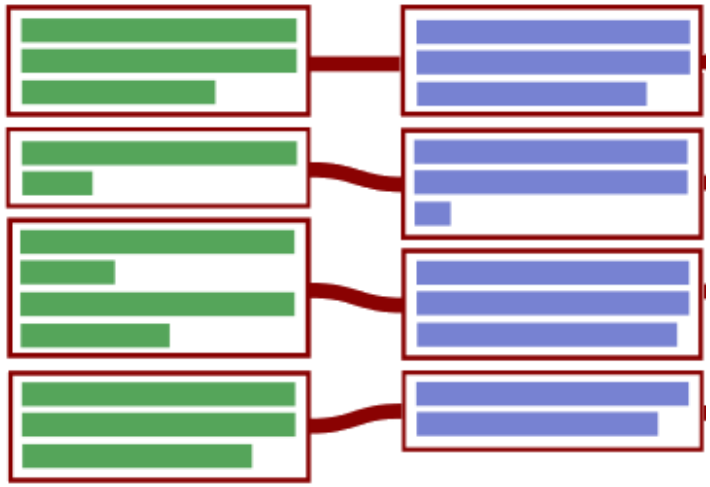
- The PRESTO project:
  the creation of a system for the customization of serious games scenarios based on virtual reality. Ontologies are used for:
  - manage interoperability between different 3D libraries;
  - support developers in developing behaviors of virtual agents.

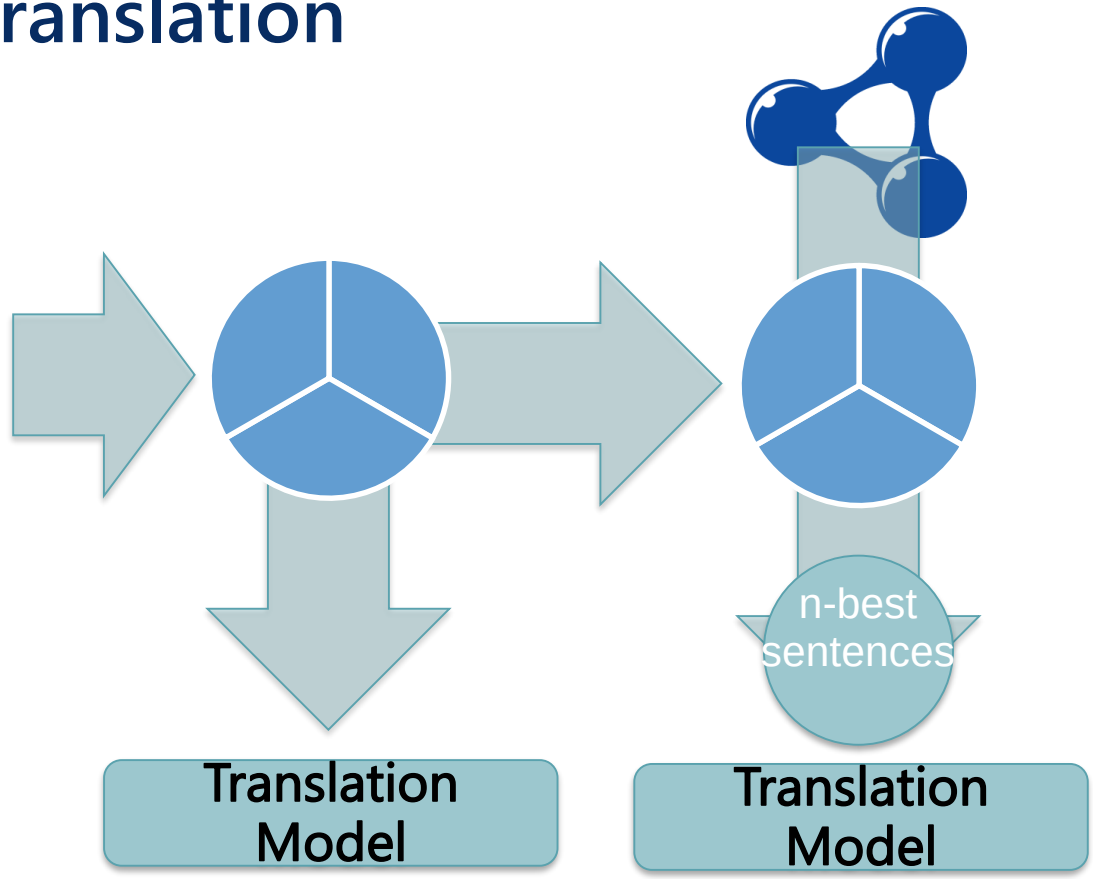
Construction of the Translation Model

- Word2Vec for building semantic vectors of words
  - Wikipedia pages used as a single huge textual file for building word vectors.
- Parallel corpora used for building a generic SMT model
  - Moses Toolkit has been used.
- Word vectors generated by W2V are used for filtering parallel sentences by detecting the ones that are more relevant.
- When a translation is requested:
  - the context of the concept is extracted from the ontology, packed and sent to the service;
  - vectors associated with each label are aggregated in order to generate a unique vector describing the full concept context; the same is done for the sentences containing the label to translate;
  - The aggregated vector is compared with the vectors of the relevant sentences;
  - Closest sentence is chosen for extracting the concept translation.

Statistical Machine Translation for Ontology Translation

Parallel Corpora





Probabilistic Model

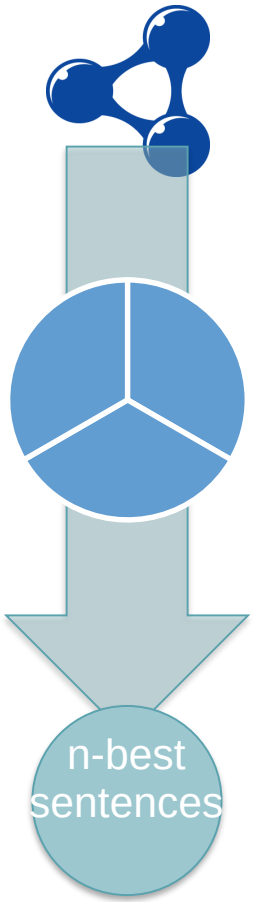
$$arg\max p(f|e)p(e)$$

Problem Space of Ontology Translation within SMT

	Out-of-Vocabulary (OOV) Problem	Out-of-Domain Translation
Source text:	acroscopic	vessels
Reference translation: (gold standard)	akroskopisch	Gefäße
Target transl. (generic):	acroscopic	Schiffe(1)
Target transl. (specific):	acroscopic	Gefäße

(1) Google Translate

Domain-specific data selection from available resources



Ontology to be translated

labels:

- *Administrative Act*
- *Insurance*
- *Equity*

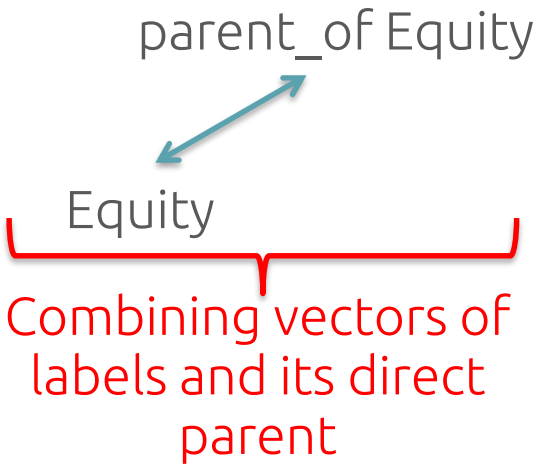
Parallel Resources

source sentences:

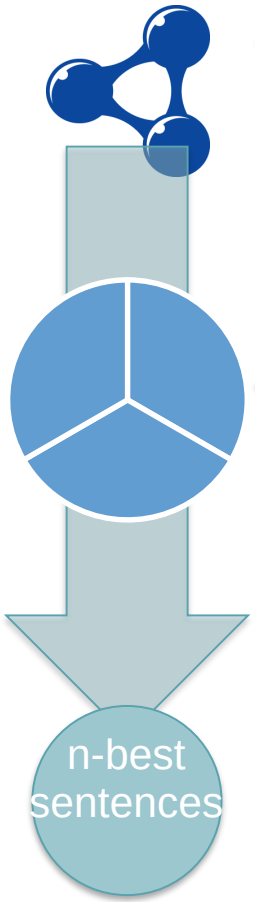
- *sentence1*
- *sentence2*
- *sentence3*

Word2Vec

- two-layer neural network, trained on large amount of text (Wikipedia)
- produces a distributed representation of words



Domain-specific data selection from available resources



intestinal infectious diseases *(ICD label)*

- infection (0.882381)
- gastroenteritis (0.879408)
- disease (0.874399)
- …

ataxia and anemia could be infectious *(k-sentence)*

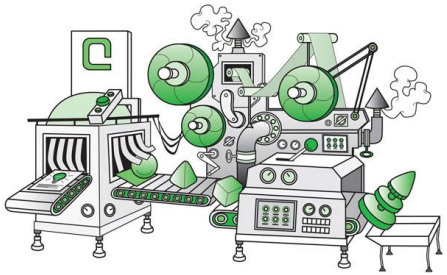
- infection (0.743936)
- encephalitis (0.721071)
- syndrome (0.711319)
- …

COSINE SIMILARITY between related words from sentences and ontology labels

Online Domain Adaptation

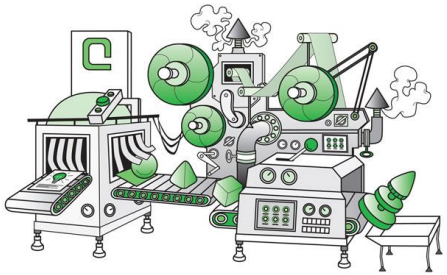
Online Domain Adaptation

Online Domain Adaptation



Word2Vec Machinery

Online Domain Adaptation

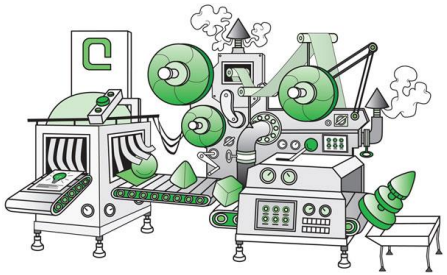


Word2Vec Machinery

Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Online Domain Adaptation

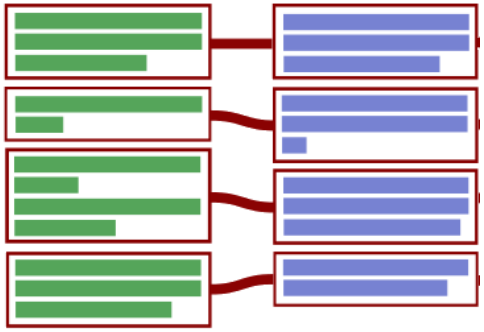


Word2Vec Machinery

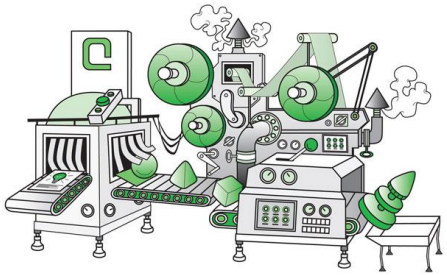
Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



Online Domain Adaptation

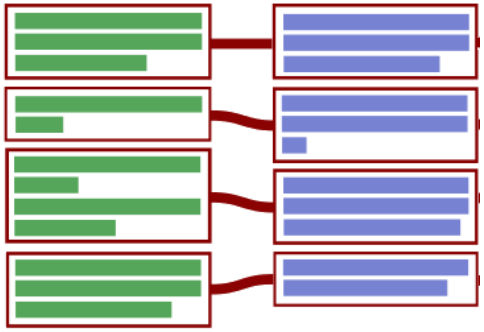


Word2Vec Machinery

Multi-dimensional Matrix

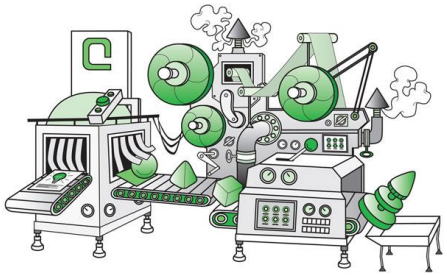
vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



Generic Translation Model

Online Domain Adaptation

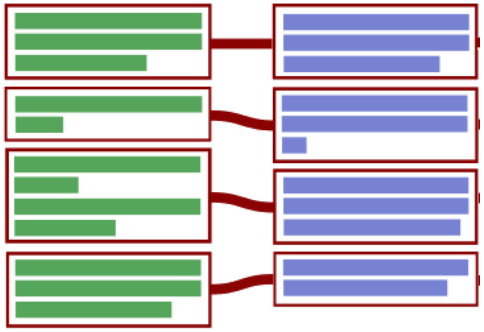


Word2Vec Machinery

Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

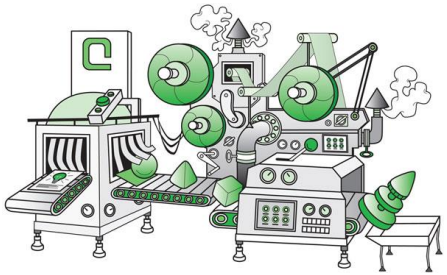
Parallel Corpora



Generic Translation Model

n-best sentences

Online Domain Adaptation

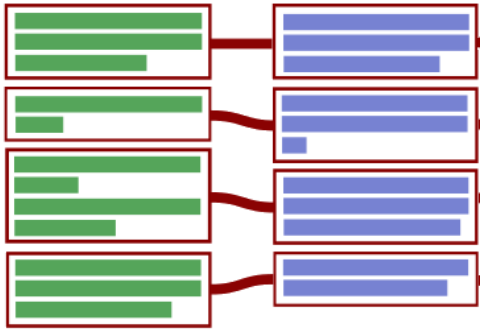


Word2Vec Machinery

Multi-dimensional Matrix

vessels	0.53956	0.2232 ...
ship	0.48783	0.6540 ...
radiator	0.51249	0.6511 ...
cruisers	0.59540	0.1213 ...
sank	0.48096	0.5415 ...
Wreckage	0.49036	0.5165 ...
fleet	0.48104	0.5646 ...
Shaft	0.16541	0.6544 ...

Parallel Corpora



Generic Translation Model

n-best sentences

Domain-specific data selection from available resources

... serious inflammation of the **vessels** ...
... **vessels** inflammation which may result in very rare ...
... arteries or of the blood **vessels** supplying the brain ...
... puncture of non-compressible **vessels** within the last month
...

Statistical Machine Translation System