Using CL-ESA for Improving Ontology Translation

Kartik Asooja, Jorge Gracia, Nitish Aggarwal, Asunción Gómez-Pérez
OEG, UPM, Madrid
DERI, Galway, Ireland

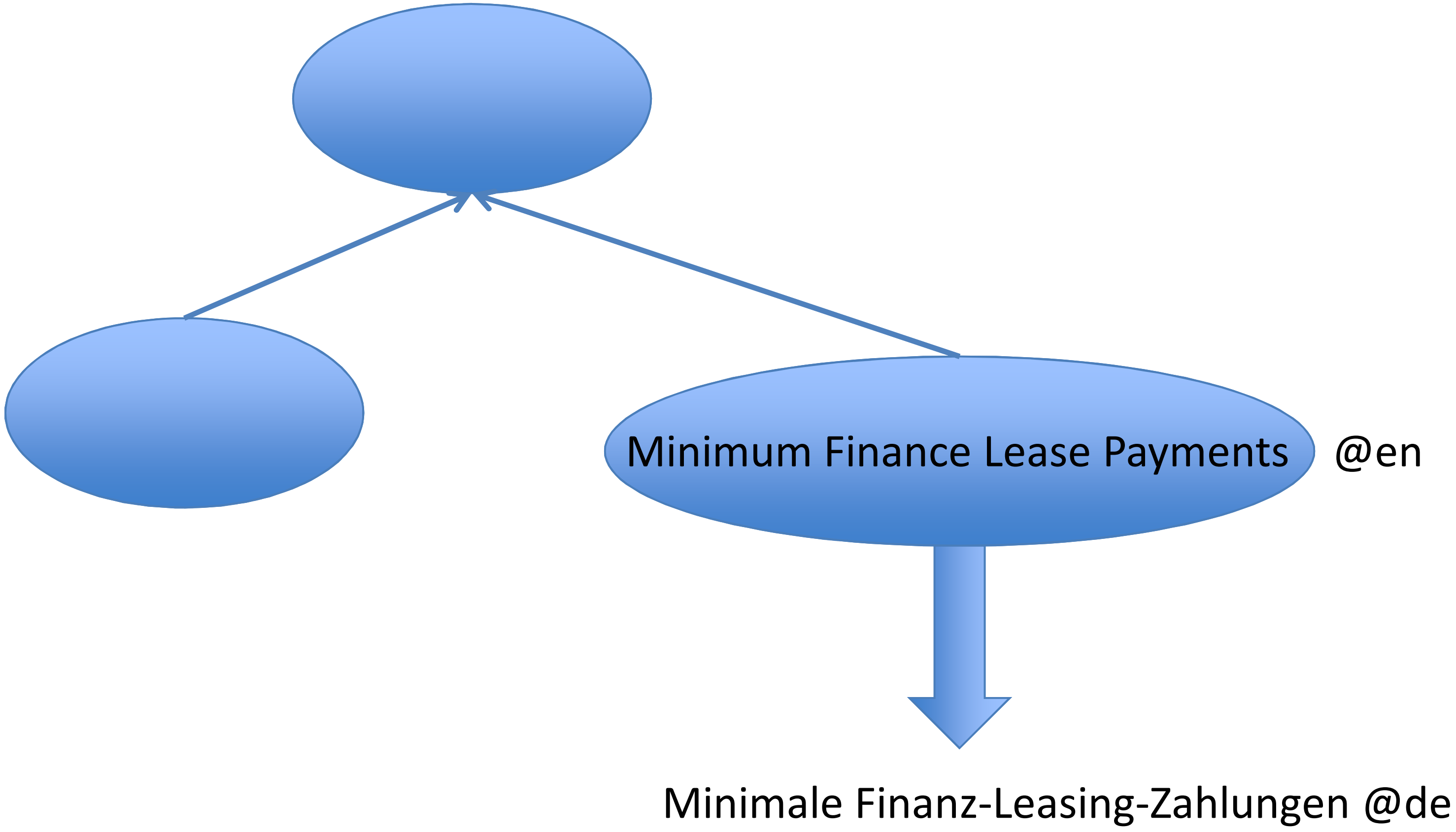
Monnet is supported
by the European Union
under Grant No. 248458

Overview

- Introduction
- Problem
- Motivation
- CL-ESA
- CL-ESA in SMT
- Evaluation
- Conclusion

Introduction

Goal: Ontology Localization : Adapt an ontology to a concrete language and culture community.


Minimum Finance Lease Payments
@en
Minimale Finanz-Leasing-Zahlungen @de

Problem

Special Characteristics of Ontology Translation compared to sentence-based MT

- Shorter texts (labels) => More ambiguity
- But: labels embedded in ontological context => less ambiguity

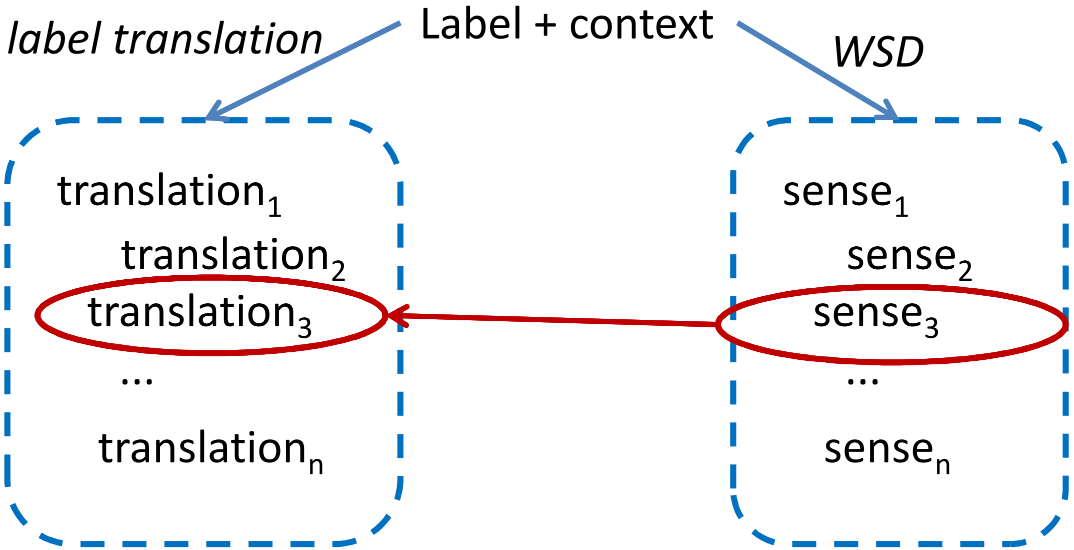
Can exploiting ontological context help in improving translation of an ontology ?

Our Goal: Exploit ontological context to improve Ontology translation following SMT approach.

Motivation

Parallelism between two tasks:

ontology label translation ↔ word sense disambiguation (WSD)

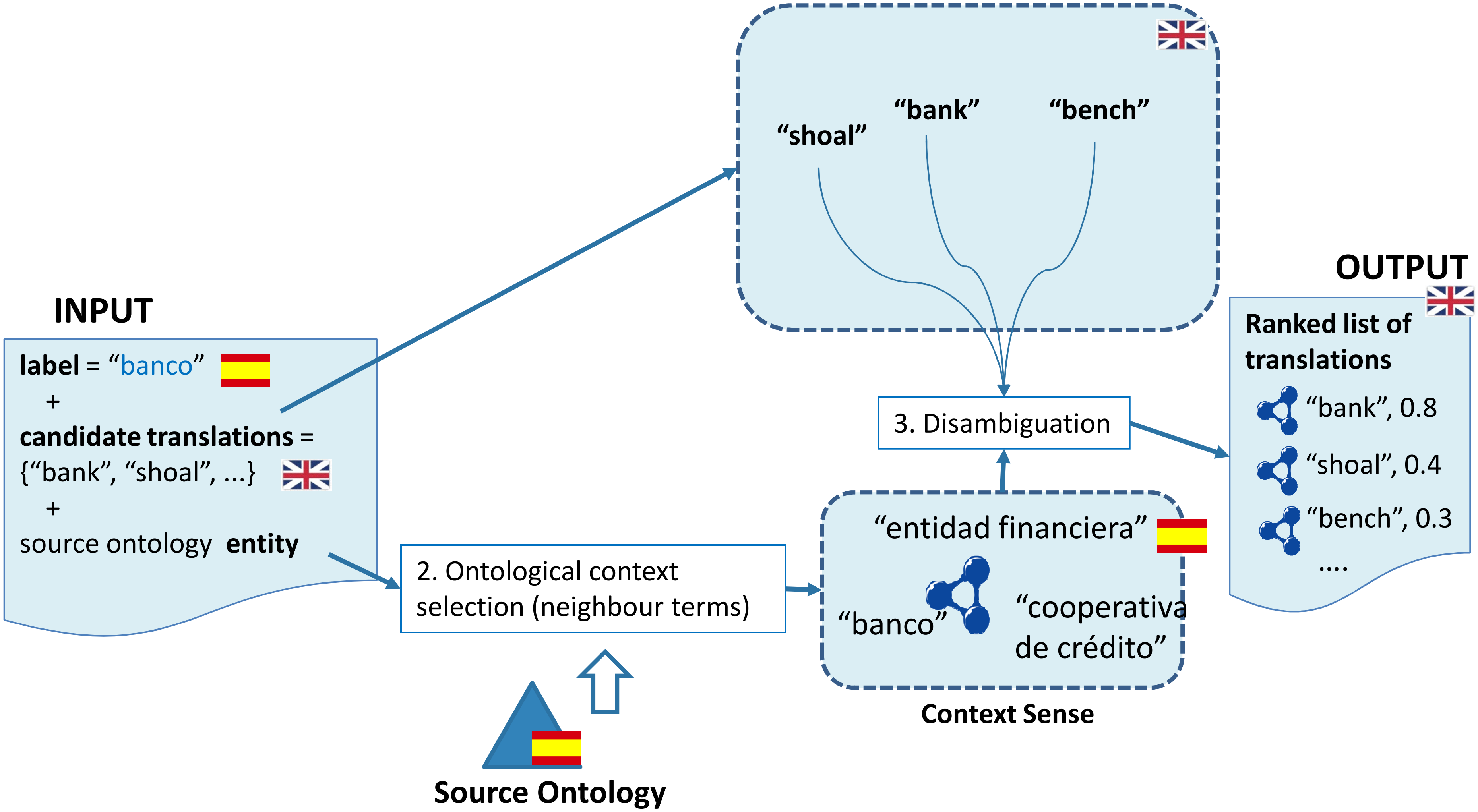


WSD has been shown to be beneficial for SMT[1,2]

1. Apidianaki M. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. EACL 2009

2. Carpuat M. and Wu D. Context-dependent phrasal translation lexicons for statistical machine translation. In Proceedings of MT Summit XI 2007

Motivation



INPUT

label = "banco" 🇪🇸

+

candidate translations = {"bank", "shoal", ...} 🇬🇧

+

source ontology **entity**

"shoal" "bank" "bench" 🇬🇧

2. Ontological context selection (neighbour terms)

Source Ontology 🇪🇸

"entidad financiera" 🇪🇸

"banco"

"cooperativa de crédito"

Context Sense

3. Disambiguation

OUTPUT 🇬🇧

Ranked list of translations

- "bank", 0.8
- "shoal", 0.4
- "bench", 0.3
- ....

CLESA

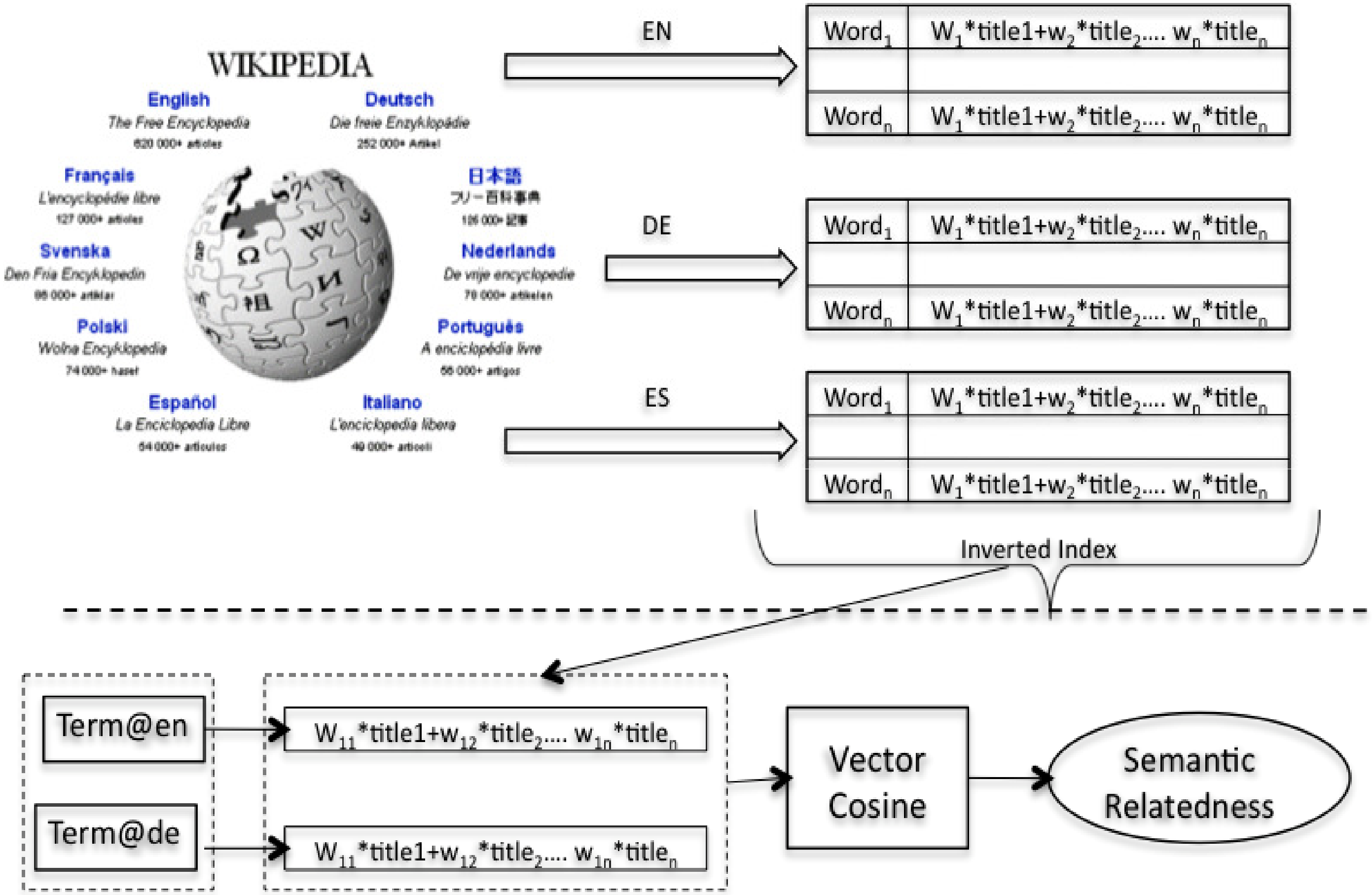
- Cross Lingual Explicit Semantic Analysis (CL-ESA)[1,2] can be used to disambiguate the phrases / translation candidates, given the ontological context.

- ESA calculates the semantic similarity between two texts by comparing the *distribution of their usages* under different explicit defined concepts.

- Wikipedia is commonly used for the implementation.

1. Potthast et. al. *A wikipedia-based multilingual retrieval model* [2008]
2. Sorg et. al. *Cross-lingual Information Retrieval with ESA* [2008]

CLESA



EN

Word$_1$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

Word$_n$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

DE

Word$_1$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

Word$_n$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

ES

Word$_1$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

Word$_n$ W_1*title1+w_2*title$_2$.... w_n*title$_n$

Inverted Index

Term@en

Term@de

W_{11}*title1+w_{12}*title$_2$.... w_{1n}*title$_n$

W_{11}*title1+w_{12}*title$_2$.... w_{1n}*title$_n$

Vector Cosine

Semantic Relatedness

Why CLESA ?

- Intuitive and simple model

- CL-ESA has been shown to perform better than the latent concept based semantic models like LSA, LDA for some tasks like Cross Lingual Information Retrieval (CLIR) [1]

[1] Cimiano et. al. *Explicit Versus Latent Concept Models for CLIR* [2009]

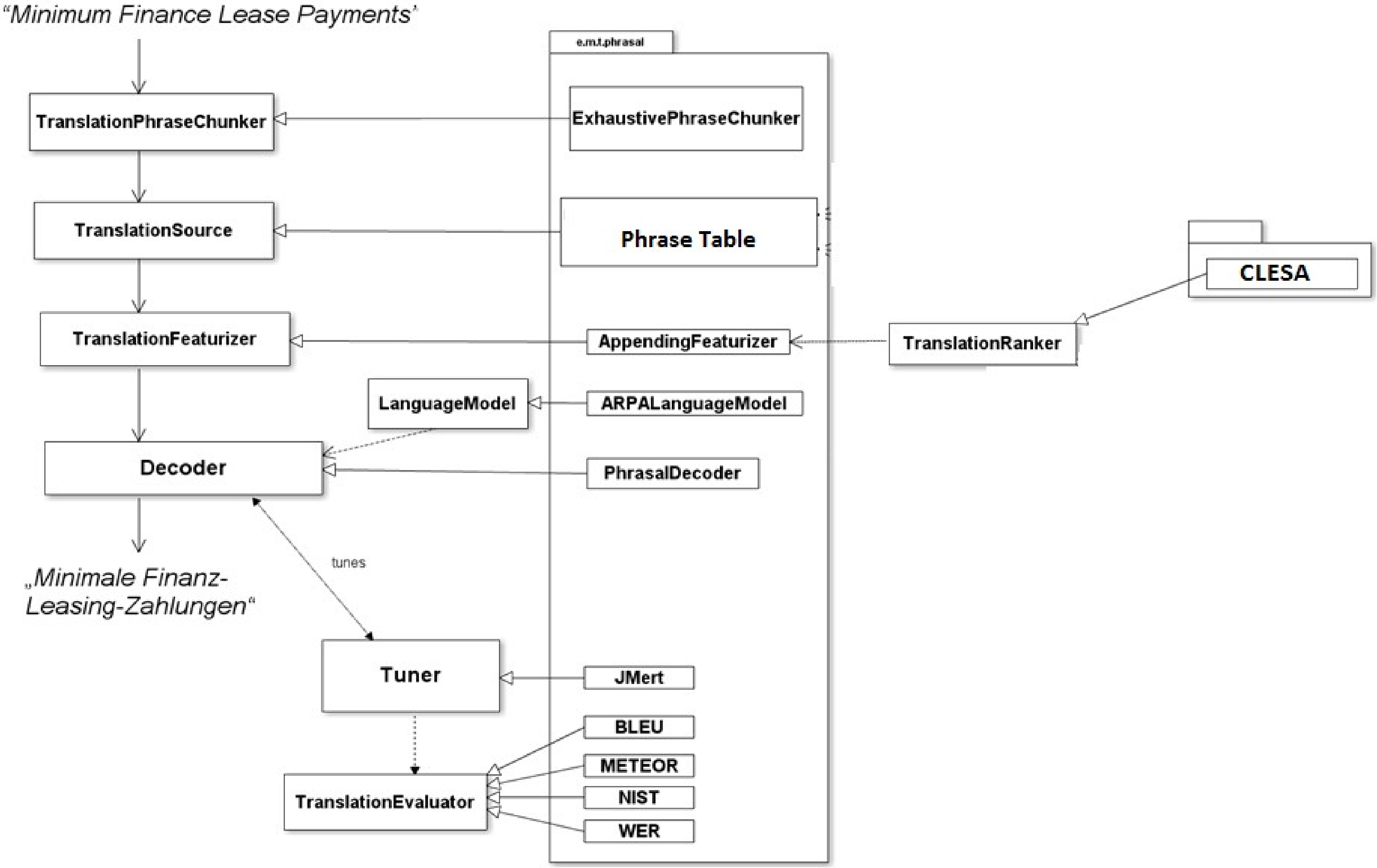
CLESA in SMT

$$argmax_{tgt} P(tgt|src) = argmax_{tgt} P(src|tgt) \; P_{LangModel}(tgt)$$

$$argmax_{tgt} P(tgt|src,O) = argmax_{tgt} P_{Translation}(src|tgt) P_{LangModel}(tgt) P_{Semantic}(tgt|O)$$

$P_{Semantic}(tgt|O)$ is the score given from CL-ESA to every translation candidate given the ontology

Monnet Architecture



"Minimum Finance Lease Payments'

TranslationPhraseChunker

TranslationSource

TranslationFeaturizer

LanguageModel

Decoder

„Minimale Finanz-Leasing-Zahlungen"

tunes

Tuner

TranslationEvaluator

e.m.t.phrasal

ExhaustivePhraseChunker

Phrase Table

AppendingFeaturizer

ARPALanguageModel

PhrasalDecoder

JMert

BLEU

METEOR

NIST

WER

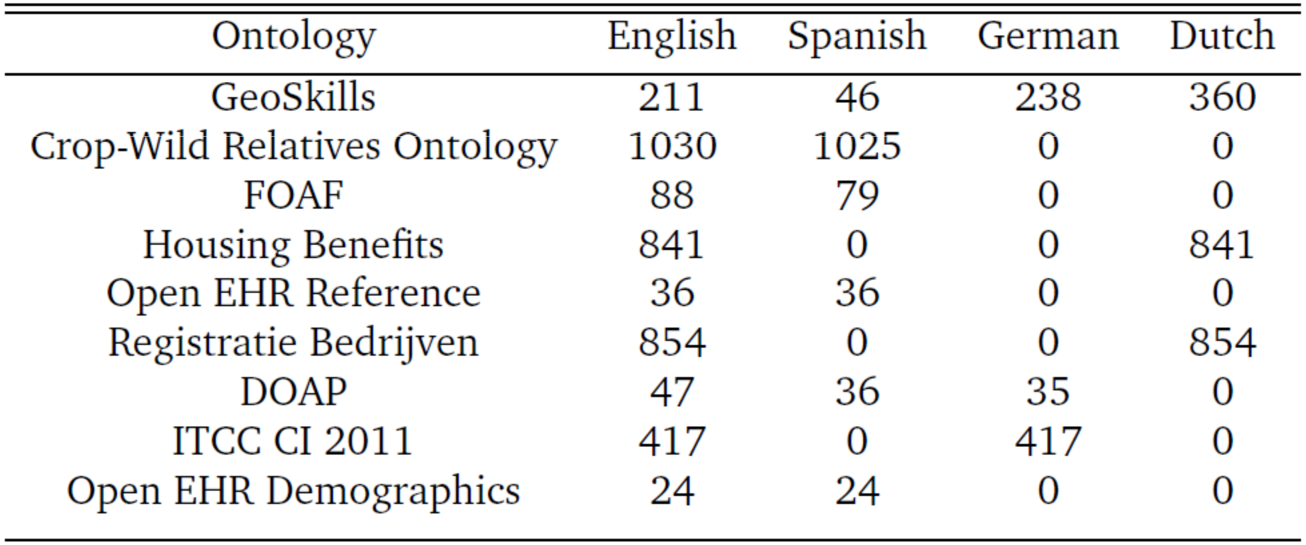
TranslationRanker

CLESA

Evaluation

Several multilingual ontologies were used for evaluation.

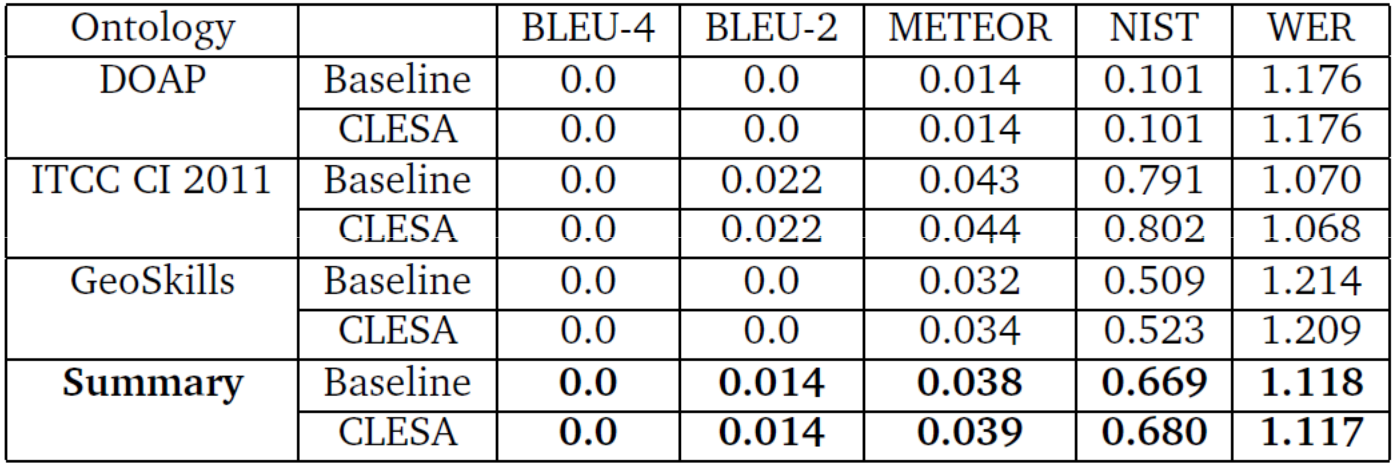
Automatic evaluation was done using standard MT metrics.



Ontology	English	Spanish	German	Dutch
GeoSkills	211	46	238	360
Crop-Wild Relatives Ontology	1030	1025	0	0
FOAF	88	79	0	0
Housing Benefits	841	0	0	841
Open EHR Reference	36	36	0	0
Registratie Bedrijven	854	0	0	854
DOAP	47	36	35	0
ITCC CI 2011	417	0	417	0
Open EHR Demographics	24	24	0	0

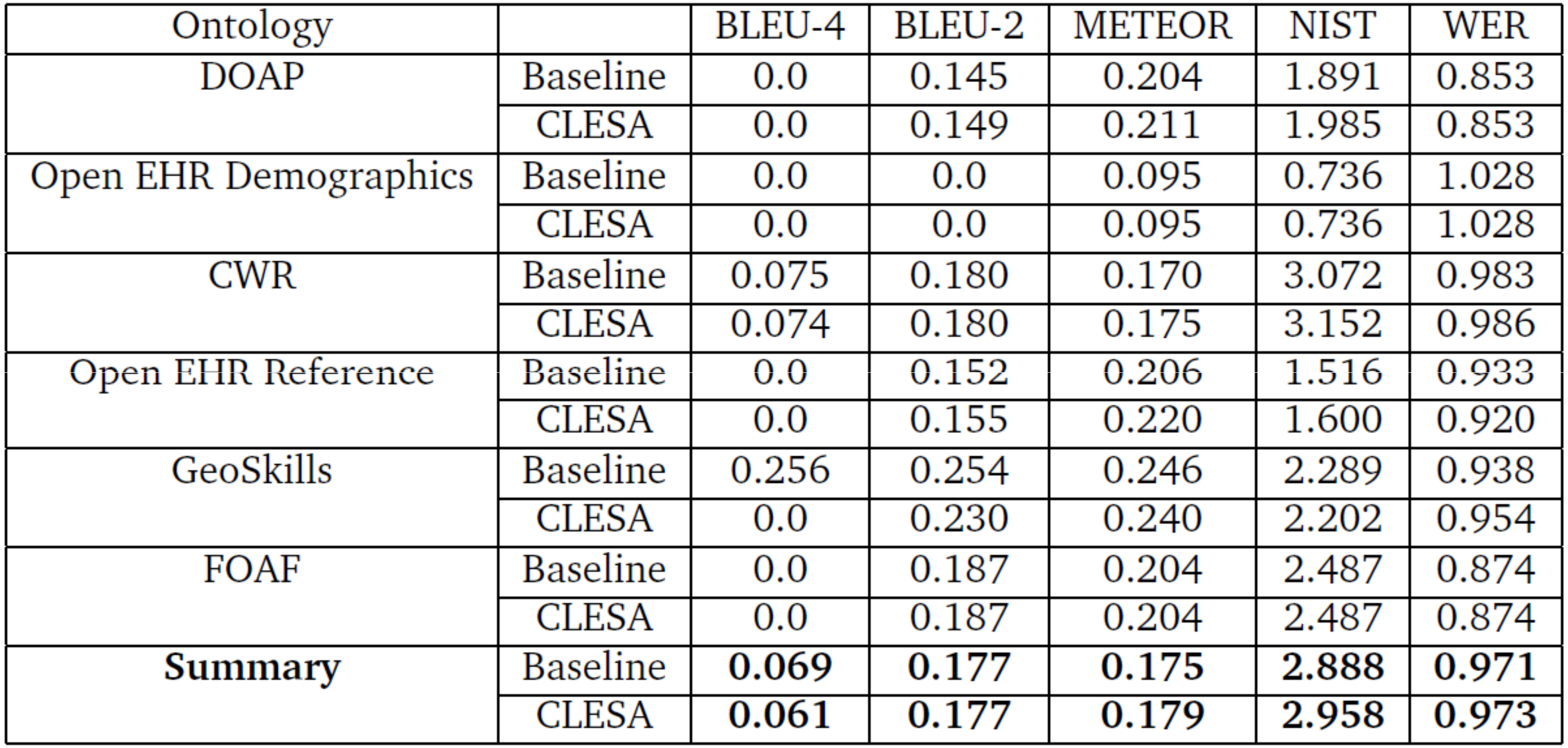
Ontologies used for evaluation

Evaluation: English - German



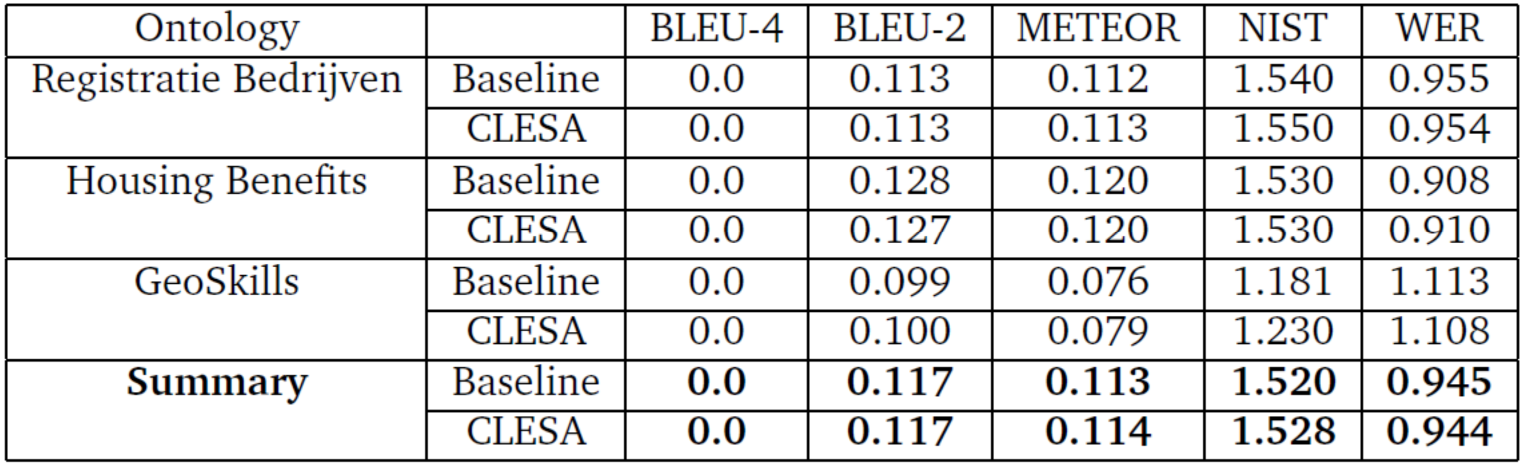
Ontology		BLEU-4	BLEU-2	METEOR	NIST	WER
DOAP	Baseline	0.0	0.0	0.014	0.101	1.176
	CLESA	0.0	0.0	0.014	0.101	1.176
ITCC CI 2011	Baseline	0.0	0.022	0.043	0.791	1.070
	CLESA	0.0	0.022	0.044	0.802	1.068
GeoSkills	Baseline	0.0	0.0	0.032	0.509	1.214
	CLESA	0.0	0.0	0.034	0.523	1.209
Summary	Baseline	**0.0**	**0.014**	**0.038**	**0.669**	**1.118**
	CLESA	**0.0**	**0.014**	**0.039**	**0.680**	**1.117**

Evaluation : English to Spanish



Ontology		BLEU-4	BLEU-2	METEOR	NIST	WER
DOAP	Baseline	0.0	0.145	0.204	1.891	0.853
	CLESA	0.0	0.149	0.211	1.985	0.853
Open EHR Demographics	Baseline	0.0	0.0	0.095	0.736	1.028
	CLESA	0.0	0.0	0.095	0.736	1.028
CWR	Baseline	0.075	0.180	0.170	3.072	0.983
	CLESA	0.074	0.180	0.175	3.152	0.986
Open EHR Reference	Baseline	0.0	0.152	0.206	1.516	0.933
	CLESA	0.0	0.155	0.220	1.600	0.920
GeoSkills	Baseline	0.256	0.254	0.246	2.289	0.938
	CLESA	0.0	0.230	0.240	2.202	0.954
FOAF	Baseline	0.0	0.187	0.204	2.487	0.874
	CLESA	0.0	0.187	0.204	2.487	0.874
Summary	Baseline	**0.069**	**0.177**	**0.175**	**2.888**	**0.971**
	CLESA	**0.061**	**0.177**	**0.179**	**2.958**	**0.973**

Evaluation : English to Dutch



Ontology		BLEU-4	BLEU-2	METEOR	NIST	WER
Registratie Bedrijven	Baseline	0.0	0.113	0.112	1.540	0.955
	CLESA	0.0	0.113	0.113	1.550	0.954
Housing Benefits	Baseline	0.0	0.128	0.120	1.530	0.908
	CLESA	0.0	0.127	0.120	1.530	0.910
GeoSkills	Baseline	0.0	0.099	0.076	1.181	1.113
	CLESA	0.0	0.100	0.079	1.230	1.108
Summary	Baseline	**0.0**	**0.117**	**0.113**	**1.520**	**0.945**
	CLESA	**0.0**	**0.117**	**0.114**	**1.528**	**0.944**

Conclusion

Metric scores are quite low, ***out of vocabulary*** could be the reason, qualitative analysis also required

Considering ontological context makes a ***slight*** improvement, thus proving it could be beneficial, more investigation required

Thanks !

References

1. **Cimiano et. al.** ***Explicit Versus Latent Concept Models for CLIR*** **[2009]**

2. **Potthast et. al.** ***A wikipedia-based multilingual retrieval model*** **[2008]**

3. Sorg et. al. ***Cross-lingual Information Retrieval with ESA*** **[2008]**