MT Adaptation for Under-Resourced Domains – What Works and What Not

Mārcis Pinnis and Raivis Skadiņš

{marcis.pinnis|raivis.skadins}@tilde.lv

Tilde, Latvia

HLT 2012 / Tartu, Estonia / 04.10.2012.

TaaS ACCURAT

Overview

- The Aims of the paper
- Initial baseline SMT system
- Process chain overview
  - Acquisition of initial bi-lingual terminology
  - Collection of Comparable Corpora
  - Extraction of bi-lingual terminology
  - SMT system adaptation
- Big baseline and SMT system adaptation results
- Conclusion

The Aims

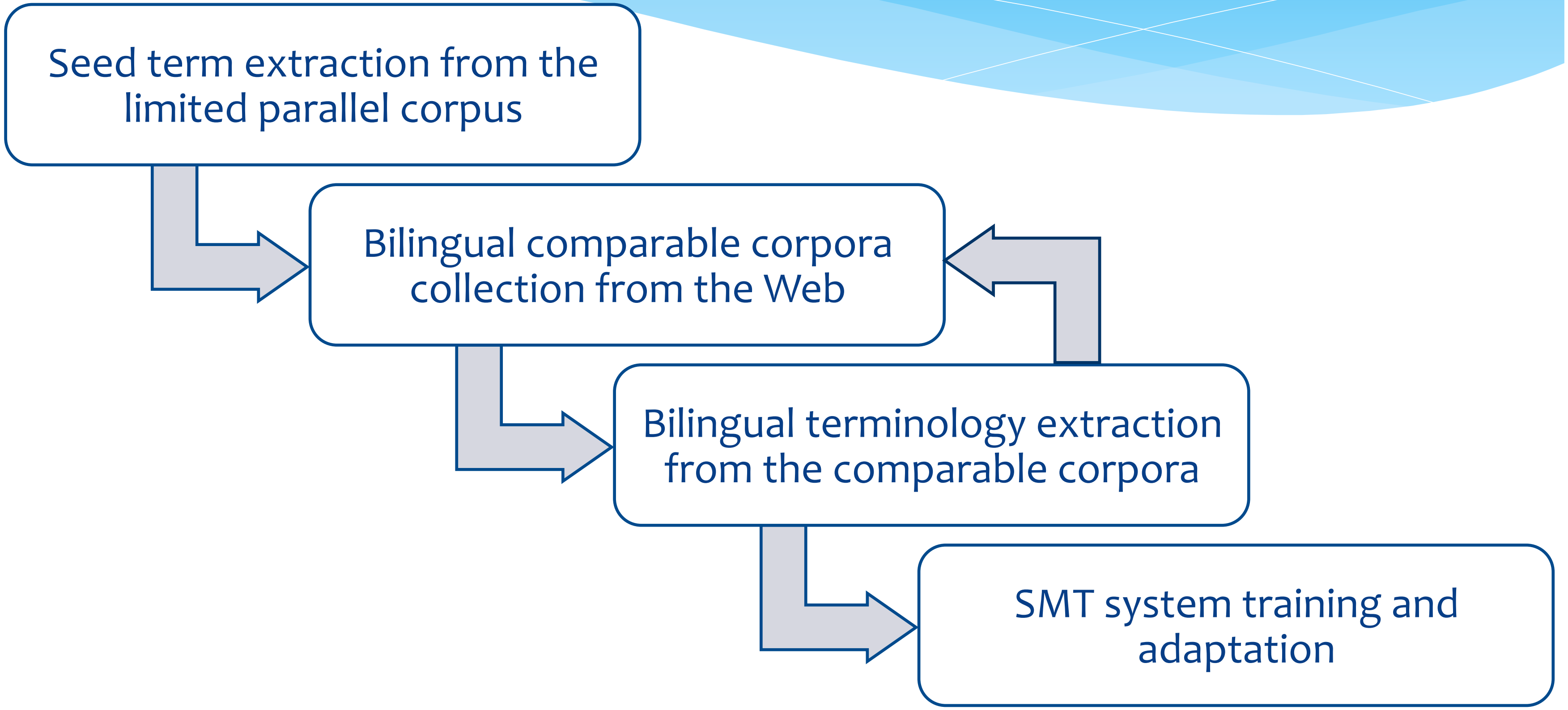
- To find methods for **SMT system adaptation** with a limited in-domain parallel corpus (or limited in-domain terminology)
- To use the Web for **in-domain corpora acquisition** that can be used in the SMT system adaptation process
- To show how general out-of-domain SMT systems can be tailored using **data extracted from in-domain comparable corpora**
- To start with very **limited in-domain parallel corpus** (~2700 sentence pairs)

Baseline SMT System

- English-Latvian translation direction
- Target domain – automotive texts
- Trained on a publicly available corpus – DGT-TM (2007)
  - 804,501 unique parallel sentence pairs
  - 791,144 unique Latvian sentences
- Tuned with MERT on 1,745 in-domain sentence pairs
- Evaluated on 872 in-domain sentence pairs
- Trained on the Let's MT! platform

Case sensitive	BLEU	NIST	TER	METEOR
No	10.97	3.9355	89.75	0.1724
Yes	10.31	3.7953	90.40	0.1301

Process Chain Overview



Seed term extraction from the limited parallel corpus

Bilingual comparable corpora collection from the Web

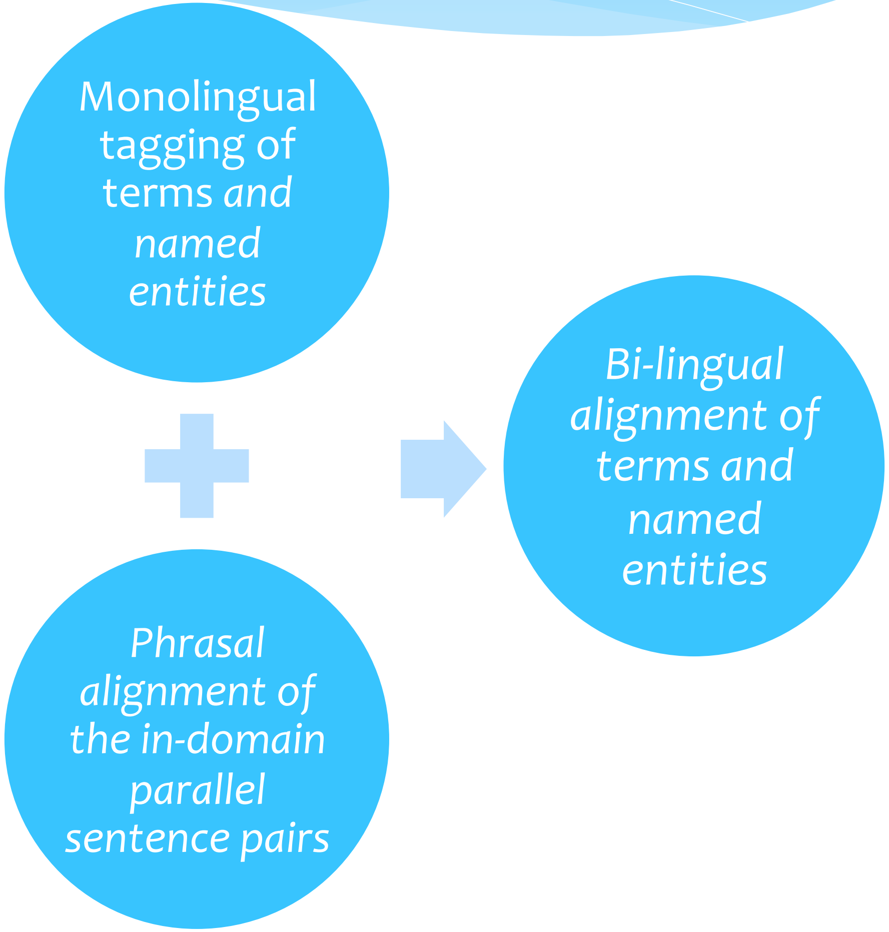
Bilingual terminology extraction from the comparable corpora

SMT system training and adaptation

- Steps 2 and 3 can be repeated in an iterative manner in order to bootstrap bilingual in-domain terminology

Initial Extraction and Alignment of Terms and Named Entities

- To find domain specific documents on the Web **we require seed terms** (to start crawling)
- The seed terms are extracted from the available parallel data
- ***Tilde's Wrapper System for CollTerm*** (*TWSC*) is used for monolingual term tagging
- ***TildeNER*** and ***OpenNLP*** are used for Latvian and English named entity recognition respectively
- ***Moses*** is used for phrasal alignment



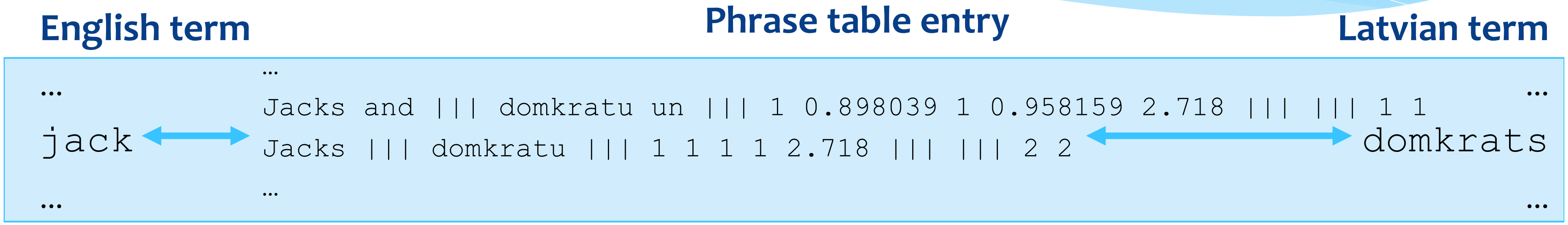
Monolingual tagging of terms *and named entities*

Phrasal alignment of the in-domain parallel sentence pairs

Bi-lingual alignment of terms and *named entities*

Bi-lingual Alignment of Terms and Named Entities

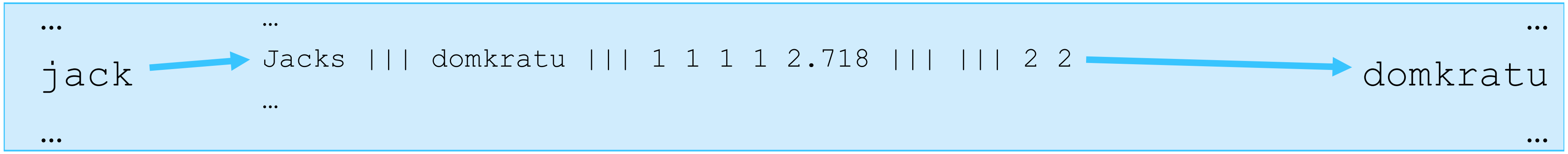
- Complete alignment



English term **Phrase table entry** **Latvian term**

```
…           …                                                                          …
            Jacks and ||| domkratu un ||| 1 0.898039 1 0.958159 2.718 ||| ||| 1 1
jack  ⟷     Jacks ||| domkratu ||| 1 1 1 1 2.718 ||| ||| 2 2           ⟷        domkrats
            …
…                                                                                      …
```

- Partial alignment



```
…           …                                                                          …
jack  →     Jacks ||| domkratu ||| 1 1 1 1 2.718 ||| ||| 2 2           →        domkratu
            …
…                                                                                      …
```

- To find inflected variants , **words in phrases are stemmed**
- With this process **542 unique English** and **786 unique Latvian** term and named entity phrases from the monolingually tagged corpora were **aligned in 783 pairs.**

Non-specific Phrase Filter

- Not all aligned phrases are **specific enough** for crawling of a **domain specific corpus**
- Therefore, we filter the phrases using reference corpus statistics

$$R(p_{src},p_{trg})=min\left(\sum_{i=1}^{|p_{src}|} IDF_{src}(p_{src}(i)),\sum_{j=1}^{|p_{trg}|} IDF_{trg}(p_{trg}(j))\right)$$

- 614 phrase pairs remained after the filtering step

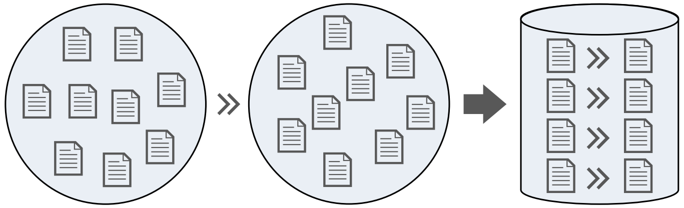
Comparable Corpora Collection

- For Web corpora crawling **55 English and 14 Latvian in-domain seed URLs** were manually collected
- A 48 hour focussed monolingual Web crawl was performed using the 614 bilingual phrases as seed terms and the collected URLs
- For crawling we use the ***Focussed Monolingual Crawler*** (***FMC***)

Language	Unique Documents	Unique Sentences	Tokens in Unique Sentences
English	34,540	1,481,331	20,134,075
Latvian	6,155	271,327	4,290,213

Document Alignment

- To minimise search space for bilingual term extraction the **monolingual corpora were aligned in document level** with a comparability metrics tool (*DictMetric*)
- 81,373 document pairs remained after filtering TOP 5 pairs for each Latvian as well as English document



Language	Unique Documents	Unique Sentences	Tokens in Unique Sentences
English	24,124	1,114,609	15,660,911
Latvian	5,461	247,846	3,939,921

Extraction of Term Pairs from Comparable Corpus

- Both monolingual corpora of the aligned comparable corpus are monolingually tagged with ***TWSC***
- This step extracts only terms (named entities are not considered)
- Terms in aligned documents are mapped using ***TerminologyAligner*** (***TEA***)
- TEA extracted **369 in-domain term pairs** (using a configuration that achieves precision of more than 90%)

TaaS ACCURAT

SMT System Adaptation
In-domain Language Model

- We start our adaptation experiments by adding an **in-domain language model** trained on the monolingual in-domain Latvian corpus (247,846 sentences) that was collected with ***FMC***
- We also test the system's performance by using only the in-domain language model

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Baseline	10.97	10.31	3.9355	3.7953	89.75	90.40	0.1724	0.1301
Int_LM	**11.30**	**10.61**	**3.9606**	**3.8190**	89.74	90.34	**0.1736**	**0.1312**
In-domain_ LM_only	11.16	10.52	3.9447	3.8074	**89.31**	**89.92**	0.1726	0.1305

SMT System Adaptation
Added In-domain Terminology

- In the next experiments
  we add to the general parallel corpora
  in-domain terminology translations; The following sets of bilingual terms are added:
  - 610 term pairs from the tuning data
  - 369 term pairs extracted from the Web
  - 6,767 unique in-domain terms from EuroTermBank

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Int_LM	11.30	10.61	3.9606	3.8190	89.74	90.34	0.1736	0.1312
Int_LM+T_Terms	12.93	12.12	4.2243	4.0598	88.58	89.32	0.1861	0.1418
Int_LM+T&CC_Terms	13.50	12.65	4.2927	4.1105	88.86	89.70	0.1878	0.1443
Int_LM+ETB_Terms	11.26	10.52	3.9456	3.7882	89.43	90.04	0.1737	0.1290

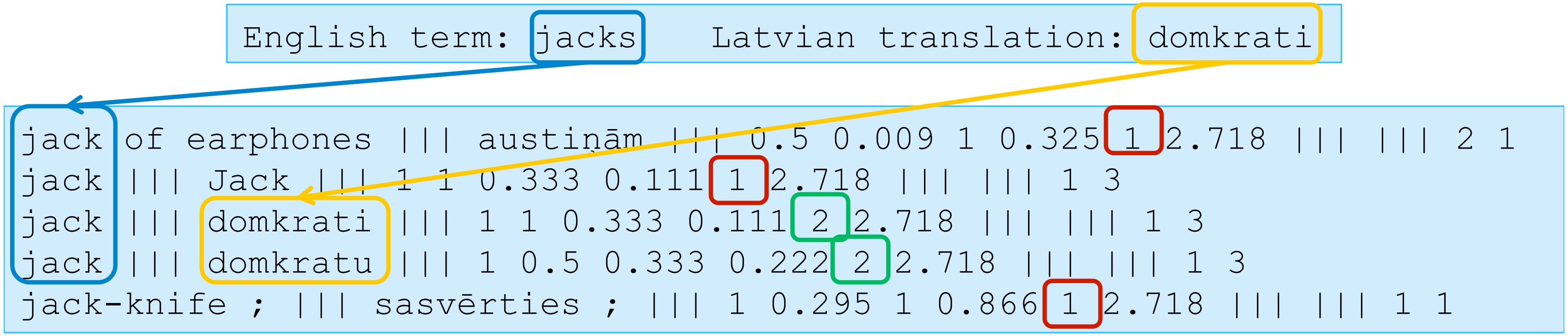
SMT System Adaptation
Added Pseudo-parallel Sentence Pairs

- In the next experiments we extracted 6,718 and 678 unique **pseudo-parallel sentence pairs** with LEXACC using two parallelism confidence score thresholds 0.51 and 0.35 respectively; the pairs were added to the SMT system's parallel data before training

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Int_LM	11.30	10.61	3.9606	3.8190	89.74	90.34	0.1736	0.1312
Int_LM+LEXACC_0.35	10.75	10.09	3.7935	3.6682	90.31	90.86	0.1646	0.1229
Int_LM+LEXACC_0.51	11.08	10.28	3.9132	3.7709	90.23	90.78	0.1706	0.1286

Term-aware Phrase Table

- To prefer in-domain terminology usage, we raise the weight of in-domain term translations in the phrase table by adding a new feature to the Moses phrase table



```
English term: jacks     Latvian translation: domkrati
```

```
jack of earphones ||| austiņām ||| 0.5 0.009 1 0.325 1 2.718 ||| ||| 2 1
jack ||| Jack ||| 1 1 0.333 0.111 1 2.718 ||| ||| 1 3
jack ||| domkrati ||| 1 1 0.333 0.111 2 2.718 ||| ||| 1 3
jack ||| domkratu ||| 1 0.5 0.333 0.222 2 2.718 ||| ||| 1 3
jack-knife ; ||| sasvērties ; ||| 1 0.295 1 0.866 1 2.718 ||| ||| 1 1
```

- Phrases containing bilingual terminology for the new feature receive the value 2
- Phrases not containing bilingual terminology – 1

SMT System Adaptation
Term-aware Phrase Table

- We modified the phrase table of the SMT systems containing previously added in-domain terminology
- The systems were re-tuned with MERT

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Int_LM+T_Terms	12.93	12.12	4.2243	4.0598	88.58	**89.32**	0.1861	0.1418
Int_LM +T&CC_Terms	13.50	12.65	4.2927	4.1105	88.86	89.70	0.1878	0.1443
Int_LM+T_Terms +6th	13.19	12.36	4.2657	4.0962	88.84	89.62	0.1876	0.1439
Int_LM +T&CC_Terms +6th	**13.61**	**12.78**	**4.3514**	**4.1747**	**88.54**	**89.32**	**0.1920**	**0.1469**

Big System Evaluation

- To validate the method consistency on larger corpora we trained a new system consisting of:
  - 5,363,043 parallel sentence pairs
  - 33,270,743 monolingual Latvian sentences
- For improved systems the setup is as before

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Big_Baseline	15.85	15.00	4.8448	4.6934	73.80	75.12	0.2098	0.1651
Big_Int_LM+T&CC_Terms	17.24	16.12	5.0020	4.8278	72.16	73.59	0.2163	0.1717
Big_Int_LM+T&CC_Terms+6th	18.21	17.08	5.1476	4.9626	70.22	71.62	0.2191	0.1747

Conclusion

- We presented **techniques for SMT domain adaptation** utilizing:
  - **bilingual terminology**
  - **bilingual comparable corpora** collected from the Web

- **Integration of terminology** within SMT systems even with simple techniques can achieve an SMT system quality improvement of up to **23.1%** over the baseline system

- **Term-aware phrase tables** can further boost the quality up to **24.1%** over the baseline system

TaaS ACCURAT

Thank you!

ACCURAT

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 248347.

TaaS

The research within the project TaaS leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 296312

ESF
EIROPAS SOCIĀLAIS FONDS
EIROPAS SAVIENĪBA
LATVIJAS UNIVERSITATE
ANNO 1919
IEGULDĪJUMS TAVĀ NĀKOTNĒ

This work has been supported by the European Social Fund within the project «Support for Doctoral Studies at University of Latvia».

TaaS

SEVENTH FRAMEWORK PROGRAMME