Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A Case Study

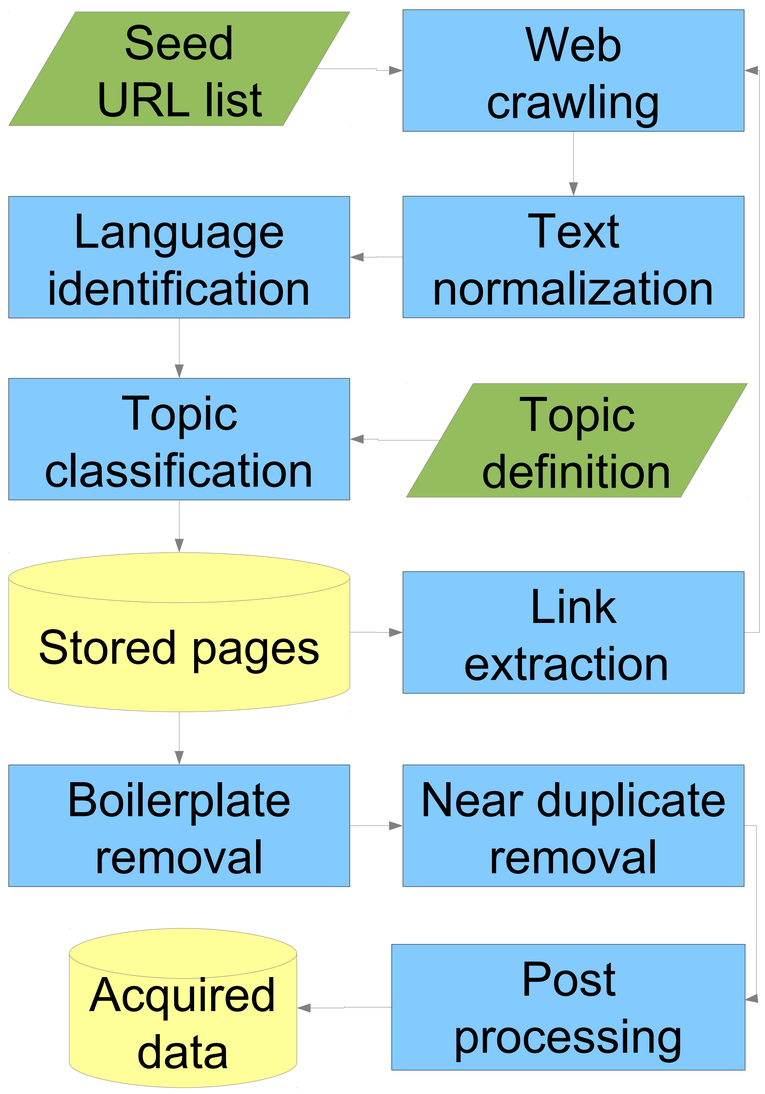
Pavel Pecina
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

Antonio Toral, Josef van Genabith
School of Computing
Dublin City University
Ireland

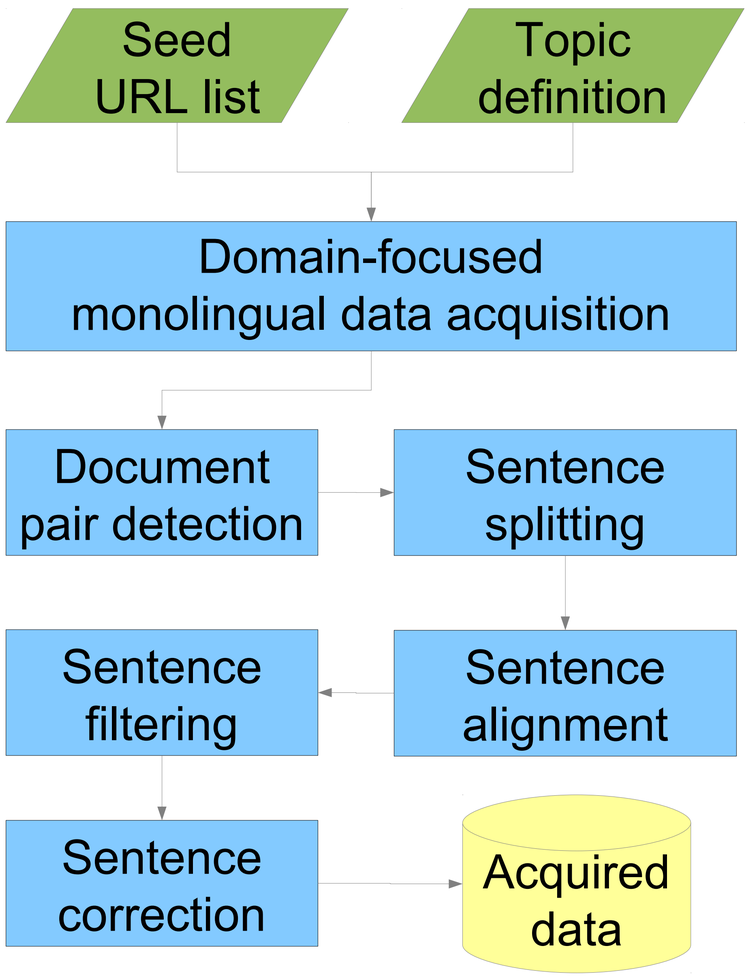
Vassilis Papavassiliou, Prokopis Prokopidis
Institute for Language and Speech Processing
Athena RIC
Athens, Greece

We tackle the problem of **domain adaptation** of **Statistical Machine Translation** by exploiting **domain-specific data** acquired by **domain-focused web-crawling**. We design and evaluate a procedure for **automatic acquisition** of **monolingual and parallel data** and their exploitation for **training**, **tuning**, and **testing** in a phrase-based Statistical Machine Translation system. We present a strategy for using such resources depending on their availability and quantity supported by results of a large-scale evaluation on the domains of **Natural Environment** and **Labour Legislation** and two language pairs: **English-French**, **English-Greek,** both directions. The average observed **increase of BLEU** is substantial at **49.5%** relative.

Monolingual Data Acquisition



Parallel Data Acquisition



Data

Domains: general (Europarl), environment, labour legislation

	domain	train	dev	test
EN-FR	gen	1.725.096	2.000	2.000
	env	10.240	1.392	2.000
	lab	20.261	1.411	2.000
EN-EL	gen	964.242	2.000	2.000
	env	9.653	1.000	2.000
	lab	7.064	506	2.000

Sentence pairs for each domain and data type

System Description

Preprocessing: Europarl tools (tokenisation, lowercasing)

LM: SRILM toolkit, interpolated 5-gram, Kneser-Ney discounting, trained on target sides of trai0ning data

Word alignment: GIZA++

Translation/reordering models: Moses, max length of phrases 7, parameters: distance, orientation-bidirectional-fe

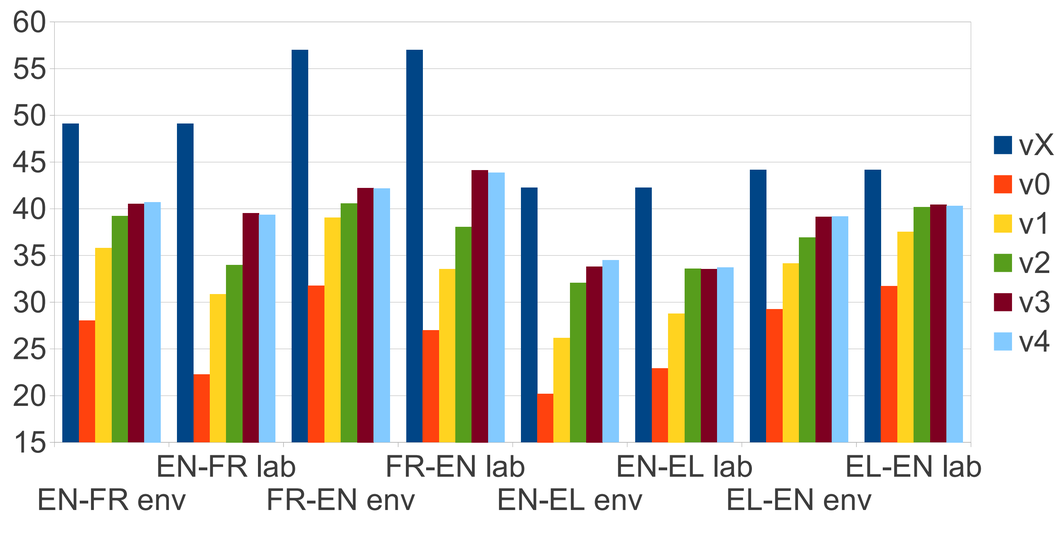
Decoder: Moses

Tuning: MERT on dev sets

Postprocessing: Moses recaser trained on target sides of training data, detokenisation by Europarl tools

sys	train	lm	dev	test
vX	gen	gen	gen	gen
v0	gen	gen	gen	spec
v1	gen	gen	spec	spec
v2	gen	gen+spec	spec	spec
v3	gen+spec	spec	spec	spec
v4	gen+spec	gen+spec	spec	spec

Results



	vX	v0	v1	v2	v3	v4
EN-FR env	49,12	28,03	35,81	39,23	40,53	40,72
EN-FR lab	49,12	22,26	30,84	34,00	39,55	39,35
FR-EN env	57,00	31,79	39,04	40,57	42,23	42,17
FR-EN lab	57,00	27,00	33,52	38,07	44,14	43,85
EN-EL env	42,24	20,20	26,18	32,06	33,83	34,50
EN-EL lab	42,24	22,92	28,79	33,59	33,54	33,71
EL-EN env	44,15	29,23	34,15	36,93	39,13	39,18
EL-EN lab	44,15	31,71	37,55	40,17	40,44	40,33

Results (BLEU) for each language pair, domain and system

Findings

Web-crawled data successfully used to adapt SMT to new domains → avg **improvement 49.5%** (BLEU)

Observations:

- Small amounts of in-domain parallel data more important than large amounts of in-domain monolingual data
- **500-1,000 sentence pairs** used as dev → **25%** improvement
- Additional parallel data (**7,000—20,000** as train) → **extra 25%**
- If parallel data not available, general-domain system can benefit from additional in-domain **monolingual data**, but **large amounts necessary** to obtain a moderate improvement

PANACEA website
http://panacea-lr.eu

SEVENTH FRAMEWORK PROGRAMME