Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation

Pavel Pecina,
Antonio Toral, Andy Way

Dublin City University
Dublin, Ireland

Vassilis Papavassiliou
Prokopis Prokopidis, Maria Giagkou

Institute for Language & Speech Processing, Athens, Greece

In this talk ...

... we will show how you can adapt your SMT system to any domain of your interest by crawling domain-specific texts from the web (using existing tools only) ...

... in the example:

- ***Moses*** (the SMT system)
- ***Europarl*** (the general domain data source)
- ***Environment, Labour Legislation*** (the adaptation domains)
- ***English*** ↔ ***French, English*** ↔ ***Greek*** (the translation directions)

... and in the context of the **PANACEA project**

PANACEA

- FP7 STREP Project, number 24606
- Platform for the Automatic, Normalized, Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies
- A webservice-based production line that automates the stages involved in the acquisition, production, updating and maintenance of the Language Resources required by MT and other Language Technologies
- Project partners:

Outline

1. Motivation
2. Domain adaptation in SMT
3. Monolingual data acquisition
4. Parallel data acquisition
5. Experiments and results
6. Conclusions

Motivation

- SMT system is not guaranteed to perform optimally if the data for training and testing are not identically (and independently) distributed

- Main problems:
  - vocabulary coverage (domain-specific terminology)
  - divergence in style and genre (special vocabulary and grammar)

- All training, development, and test data should be:
  - from the same domain
  - of the same genre and style

Domain adaptation in SMT

- Domain-specific data (monolingual and parallel) usually not available in large enough amounts to train a system of a sufficient quality
- Even small amounts of such data can be used to adapt a general system to a particular domain:
  - Monolingual data → better language models
  - Small parallel data → better parameter tuning
  - Larger parallel data → better translation model
- Three principles:
  - using in-domain development data for parameter optimization
  - merging training data from general and specific domain and training new models
  - training domain-specific models and using them together with the general-domain models in the log-linear framework

Monolingual data acquisition process overview

1) web crawling

2) text normalization

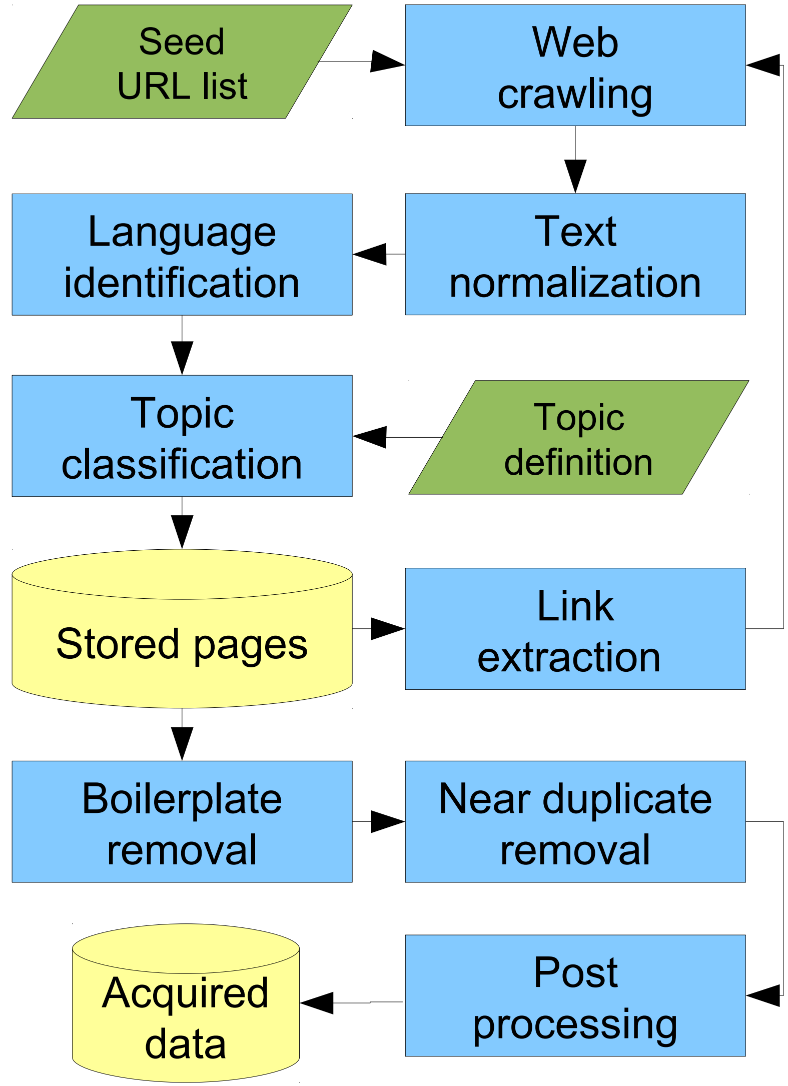
3) language identification

4) topic classification

5) document cleaning

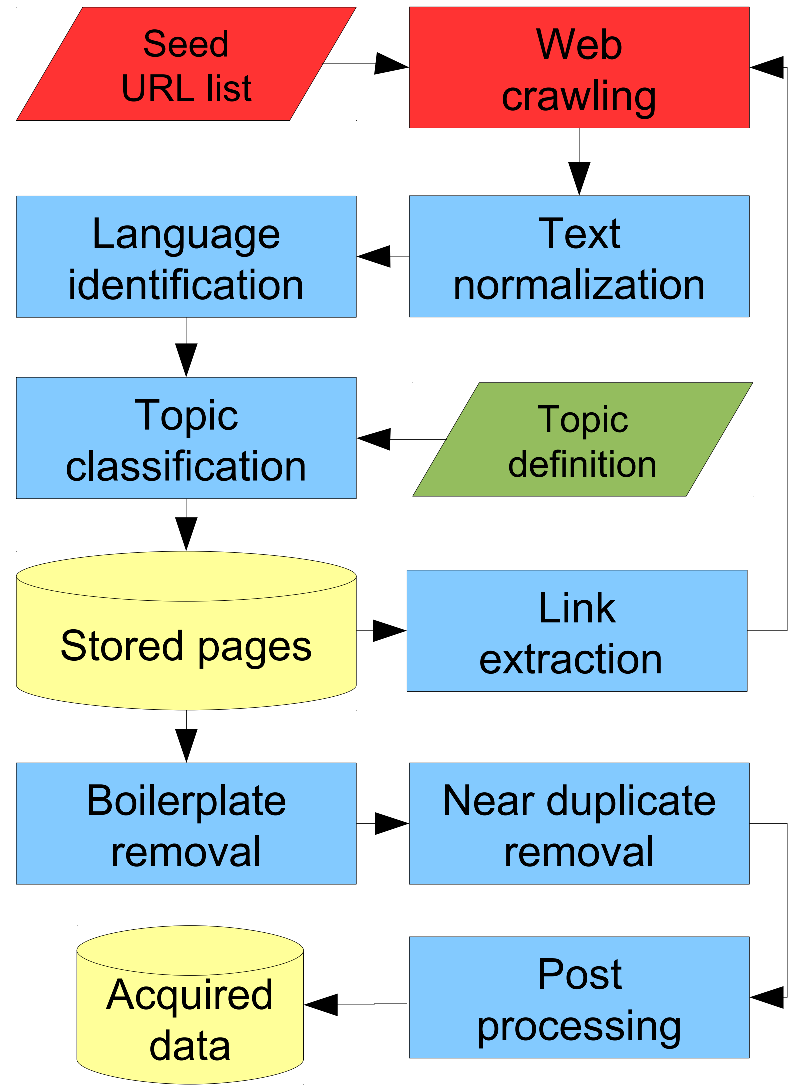
6) near-duplicate detection

7) post-processing



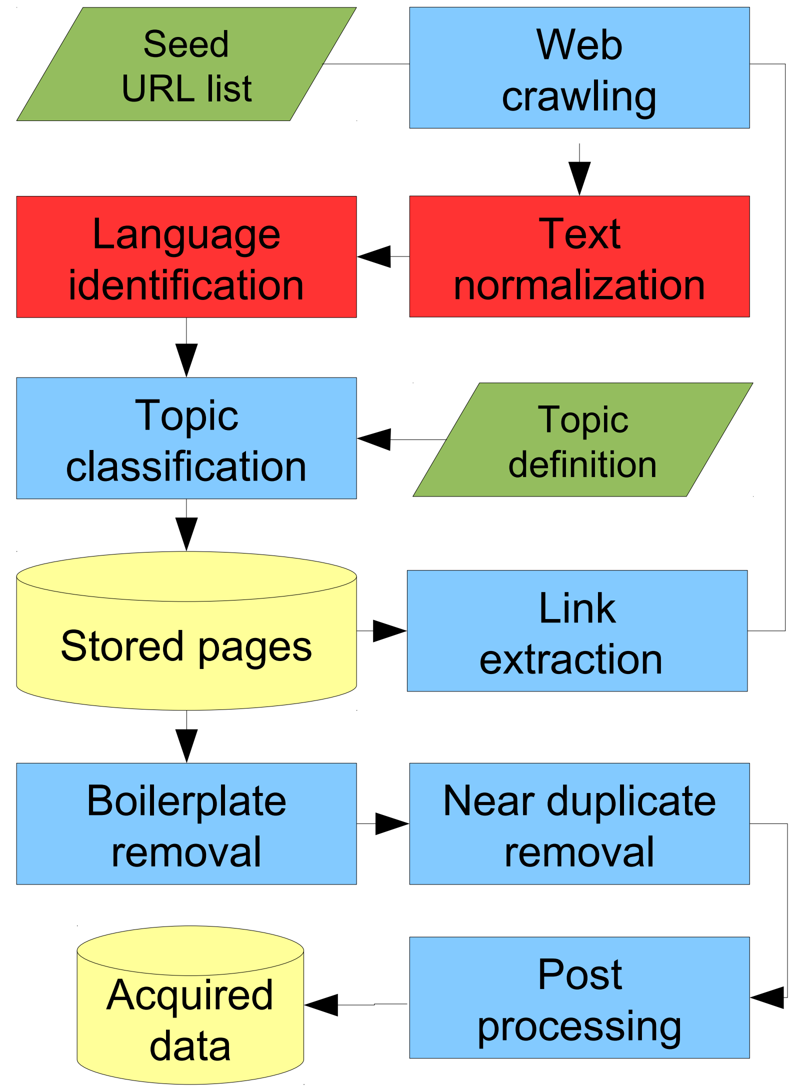
Domain-focused web crawling

- Based on an adapted **Combine** crawler (Ardö and Golub, 2007) interacting with a text to topic classifier
- Crawler's URL queue initialized with a **seed list** of URLs relevant to the targeted domain
- URL seed list sources:
  - the Open Directory Project (www.dmoz.org)
  - a web search engine queried for random tuples of domain-relevant terms



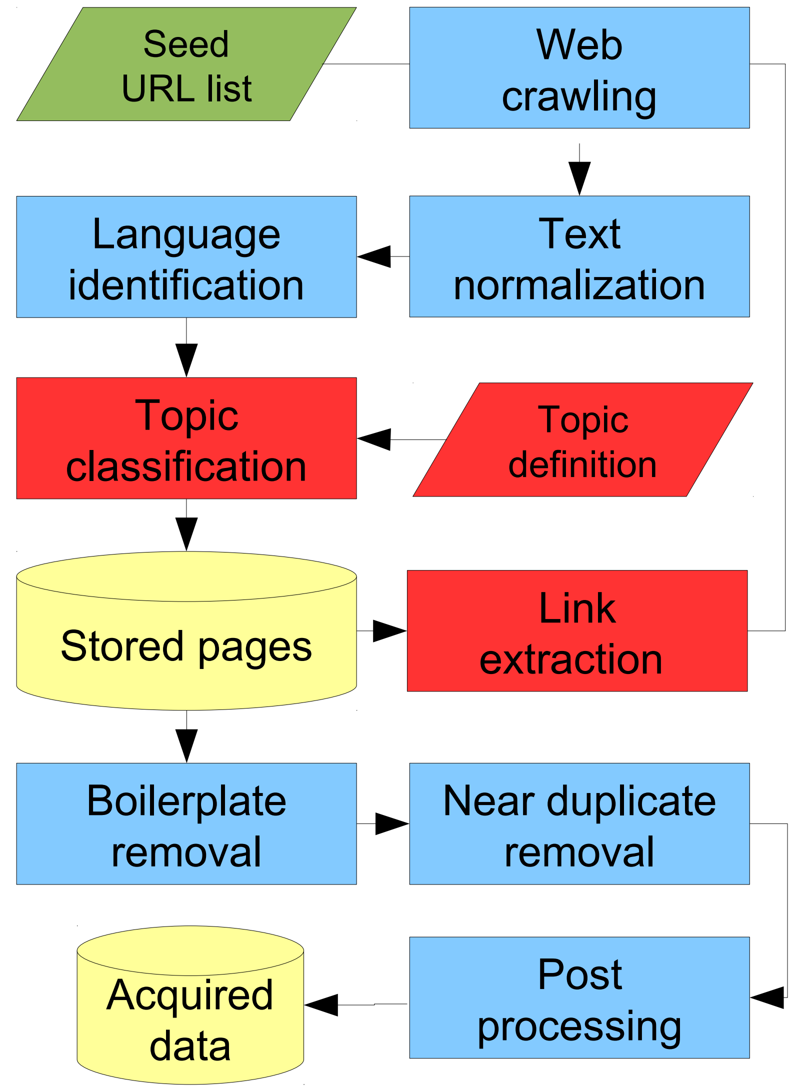
Text normalization and Lang ID

- **Text normalization**
  - file format detection (only HTML considered)
  - encoding identification and UTF-8 conversion
- **Language identification**
  - Lingua::Identify tool based on character n-grams
  - documents not in the targeted language discarded



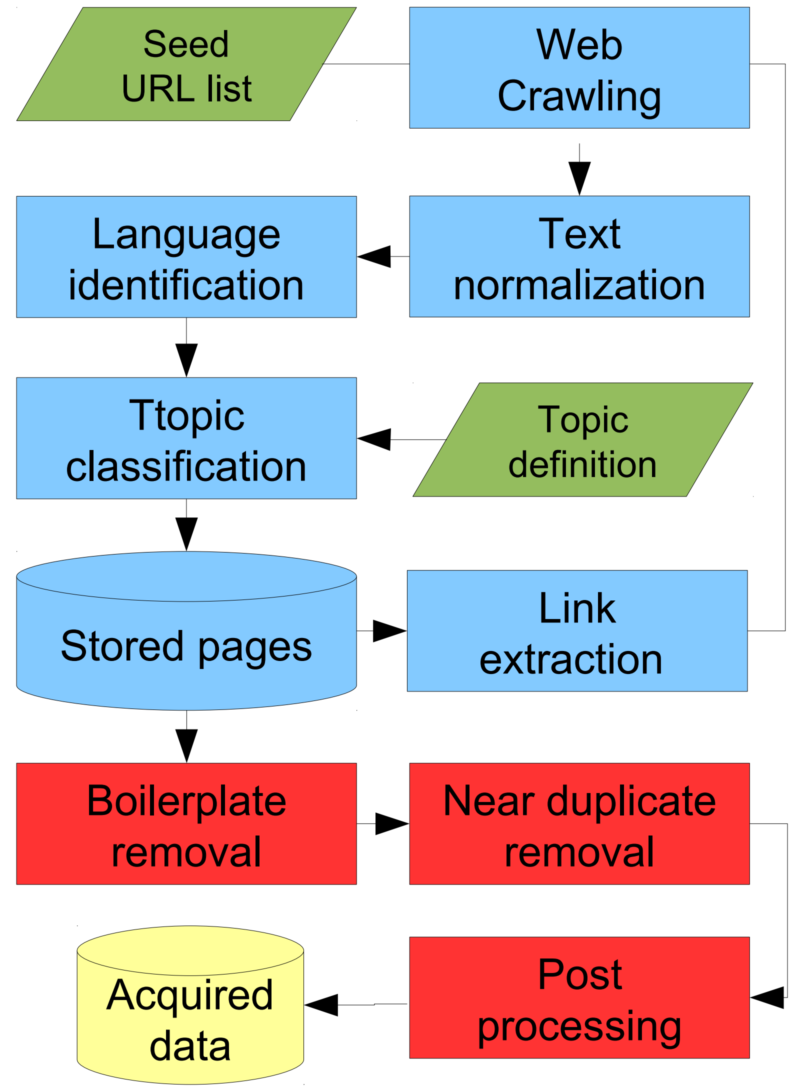
Topic classification and filtering

- Each topic defined by a list of (weighted) terms extracted from the **Eurovoc** multilingual thesaurus
- Example:
  100: air pollution = pollution_ENV
  100: biodiversity = natural_ENV
  100: climate change = natural_ENV
- Based on the terms found and their weights, each document is classified as relevant or discarded
- Links are extracted from relevant pages and moved to the crawler's queue



Monolingual data acquisition

- **Boilerplate removal**
  - headers, footers, menus, ads, etc. removed with the **Boilerpipe** tool (Kohlschütter et al., 2010)
- **Near duplicate removal**
  - very similar webapges detected by applying the **SpotSigs** algorithm (Theobald et al., 2008)
- **Postprocessing**
  - tokenization, sentence boundary identification by **Europarl tools**



Web-crawled monolingual data details and evaluation

- Documents from bilingual web sites excluded and used for acquisition of parallel data
- **Evaluation:** a sample of the pages classified by two human judges as in-domain or out-of-domain

lang	dom	sites	docs	sents	tokens	voc	new voc	accuracy
English	ENV	146	505	53,529	1,386,835	33,400	10,276	92.9
	LAB	150	461	43,599	1,223,697	25,183	6,674	91.6
French	ENV	106	543	31,956	1,196,456	36,097	9,485	95.7
	LAB	64	839	35,343	1,217,945	23,456	5,756	98.1
Greek	ENV	112	524	37,957	1,158,980	55,360	17,986	97.4
	LAB	117	481	34,610	1,102,354	52,887	16,850	88.1

Parallel data acquisition process overview

1) - 6) monolingual data crawling

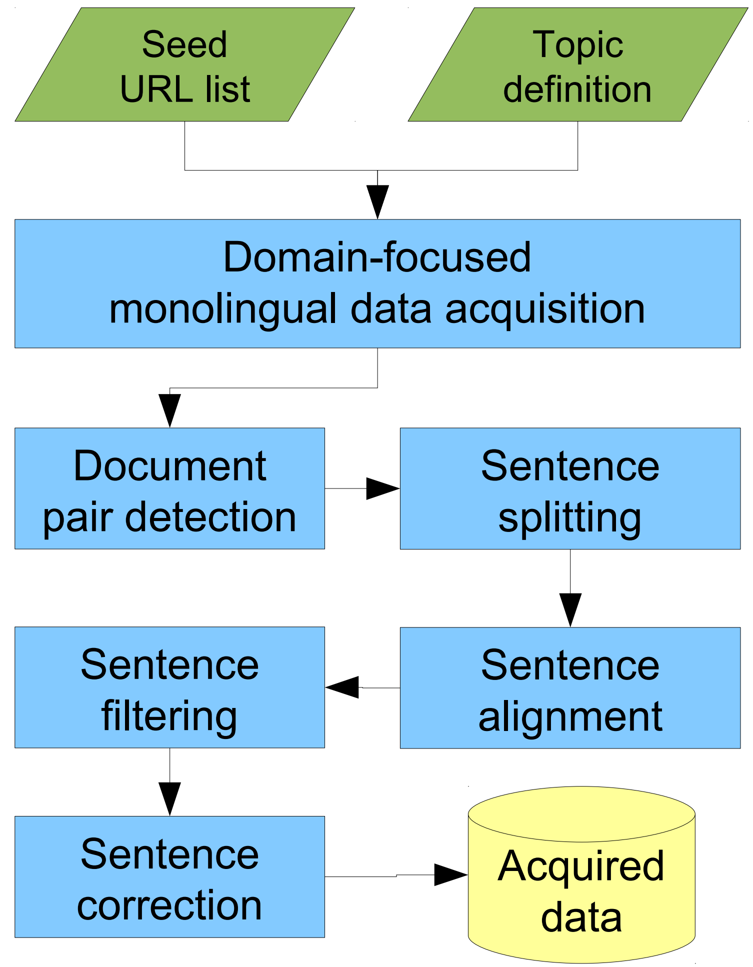
7) Document pairs detection

8) Sentence splitting

9) Sentence alignment

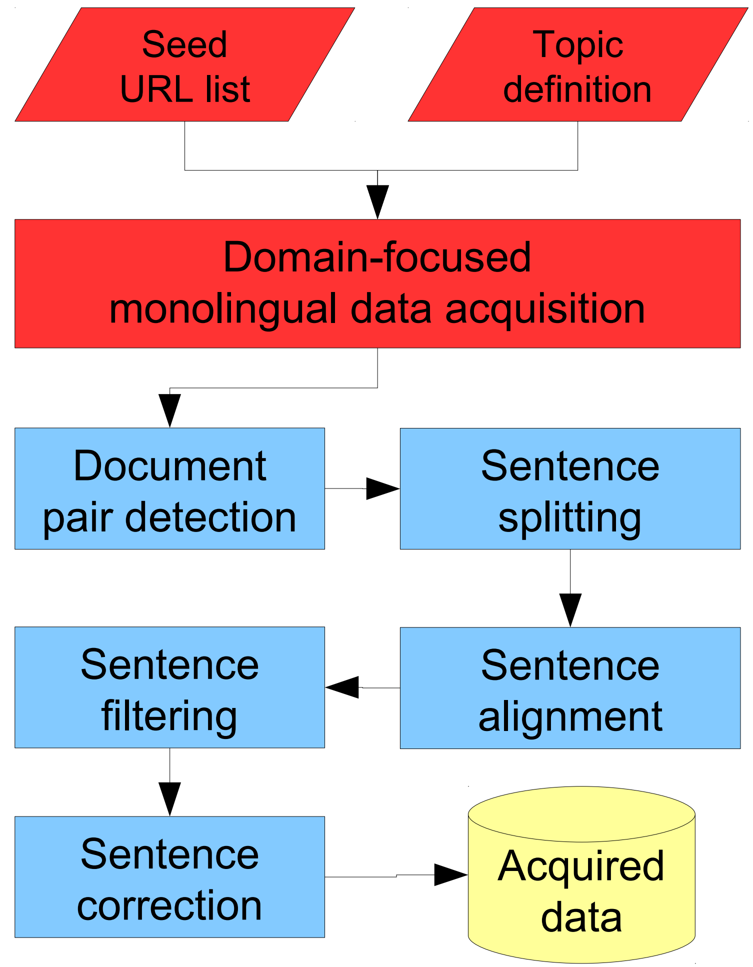
10) Sentence filtering

11) Sentence correction (manual)



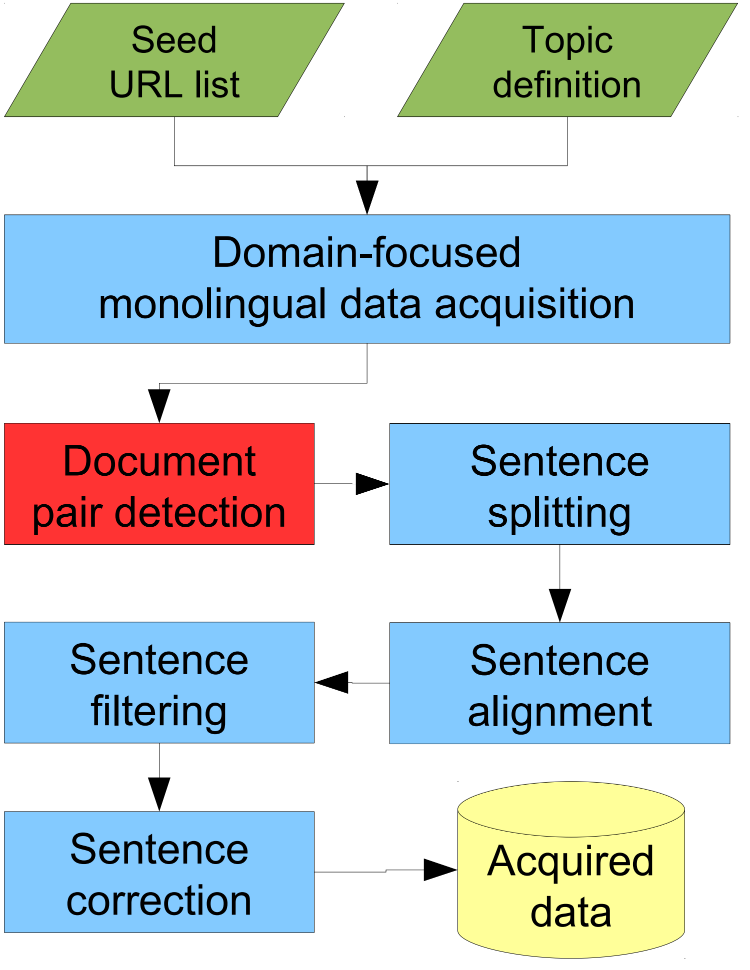
Parallel data crawling

- **Seed URL list** – relevant web sites in targeted pairs of languages, identified from the pool of web sites collected during the phase of monolingual data acquisition
- **Topic definition** – union of the topic definitions in the two targeted languages used for monolingual data acquisition
- The crawler constrained to follow internal links only



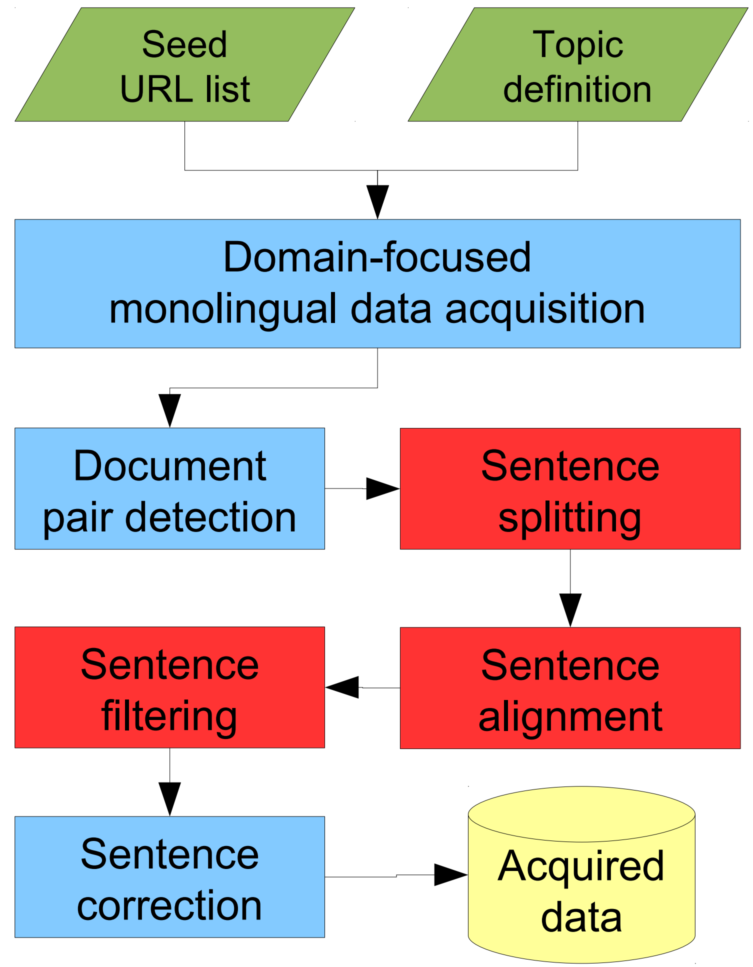
Parallel document detection

- Candidate parallel document identified by **Bitextor** (Esplà-Gomis and Forcada, 2010)
- It decides which documents could be considered translations of each other, based on the similarities of the HTML structures of the candidate parallel documents
- It also identifies pairs of paragraphs from which parallel sentences are then extracted



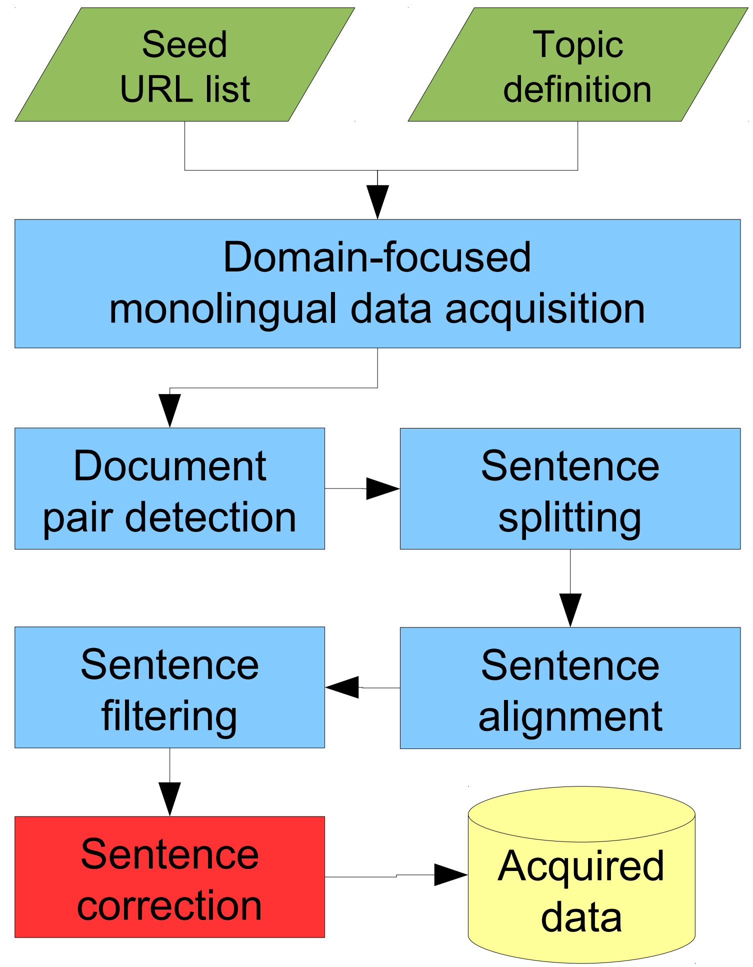
Parallel sentence processing

- Sentence splitting (and tokenization) with **Europarl** tools
- Sentence alignment with **Hunalign** (initial dictionary extracted from Europarl)
- Sentence filtering based on the **Hunalign** alignment score (threshold set after manual analysis of the results)



Parallel sentence correction

- Performed in order to create reliable **development** and **test sets** for each language pair and domain
- Low-cost procedure
- A sample of the filtered sentence pairs checked and corrected by native speakers



Parallel sentence correction details

- Two native speakers (one for each language pair) were instructed to check that:
  - sentence pairs belonged to the right domain
  - sentences within a sentence pair were equivalent in terms of content
  - translation quality is sufficient and correct the sentence pairs (if needed)
- Observations:
  - 55% accurate translations
  - 35% needed only minor corrections
  - 3–4% would require major corrections
  - 4–5% misaligned and would have to be translated completely
  - 3–4% from a different domain
- Results:
  - Only sentences requiring minor corrections had to be corrected (the remaining ones were discarded)

Web-crawled parallel data details

langage pair	dom	sites	docs	sents all /	filtered /	sampled /	corrected
English -- French	ENV	6	559	16,487	13,840	3,600	3,392
	LAB	4	900	33,326	23,861	3,600	3,411
English -- Greek	ENV	6	151	4,543	3,735	3,600	3,000
	LAB	4	125	3,094	2,707	2,700	2,506

- For each language pair and domain we obtained 2,000 sentence pairs for testing and 500-1,000 sentence pairs for parameter tuning 5-10 times cheaper than translating the data from scratch.

Experiments

- Evaluation performed in 8 scenarios:
  - 2 adaptation domains
  - 4 language pairs
  - both translation directions
- Four systems evaluated in each scenario:

  v0) **Out-of-domain** traning and development data (*baseline*)

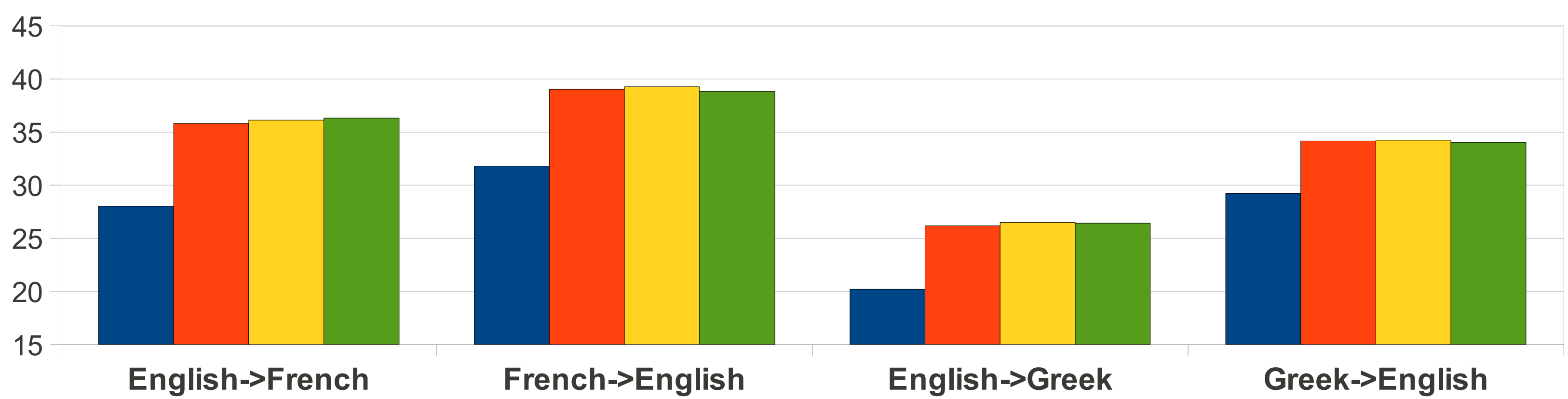
  v1) **Parallel** in-domain data for **parameter optimization**

  v2) **Monolingual** in-domain data for **language modelling:**
    - in-domain and general-domain data **merged in one model**

  v3) **Monolingual** in-domain data for **language modelling:**
    - in-domain and general domain data in **separate models**

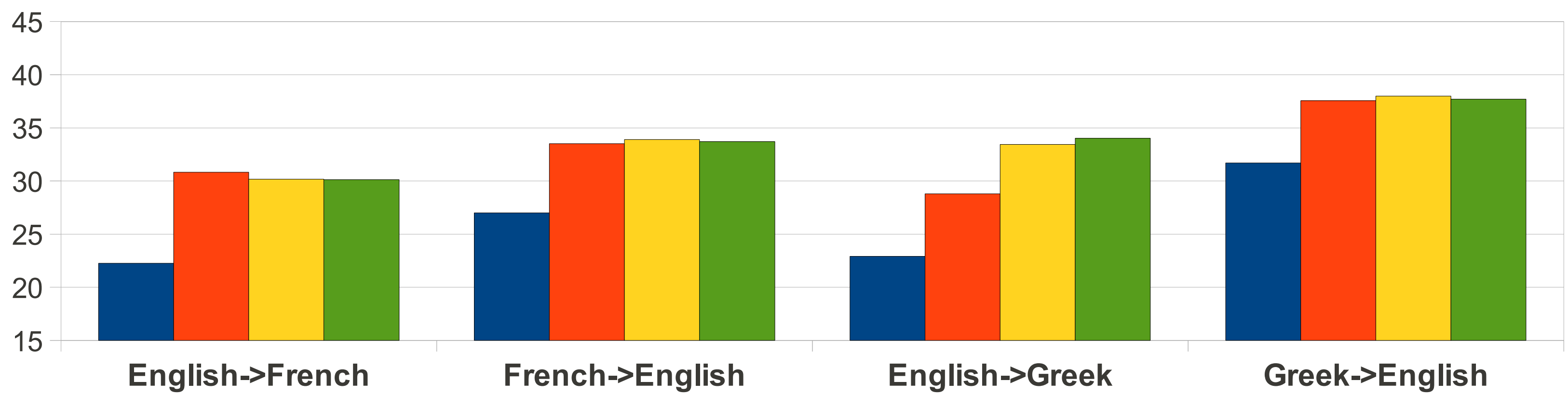
BLEU results: Natural Environment

sys	English→French	French→English	English→Greek	Greek→English
v0	28.03	31.79	20.20	29.23
v1	35.81 (27.76%)	39.04 (22.81%)	26.18 (29.60%)	34.16 (16.87%)
v2	36.13 (28.90%)	39.27 (23.53%)	26.50 (31.19%)	34.24 (17.14%)
v3	36.32 (29.58%)	38.84 (22.18%)	26.41 (30.74%)	34.15 (16.83%)



BLEU results: Labour Legislation

sys	English→French	French→English	English→Greek	Greek→English
v0	22.26	27.00	22.29	31.71
v1	30.84 (38.54%)	33.52 (24.15%)	28.79 (25.61%)	37.55 (18.42%)
v2	30.18 (35.58%)	33.91 (25.59%)	33.43 (45.86%)	38.00 (19.84%)
v3	30.12 (35.31%)	33.72 (24.89%)	34.03 (48.47%)	37.70 (18.89%)



Conclusions

- First steps towards domain adaptation of SMT based on data obtained by domain-focused web crawling
- Two types of web-crawled language resources tested (*in-domain parallel development data, in-domain monolingual training data*)
- The effect of using in-domain development data for parameter optimization is very substantial: 16–48% relative improvement
- The impact of using in-domain monolingual data for language modelling not confirmed (high OOV rate), which can be minimized only by improving the coverage of the translation models

Future work

- Crawling more parallel data and enhancing the translation models
- Overview of the PANACEA milestones:

milestone	parallel data		monolingual data		date
	source domain	annotation	source domain	annotation	
Test data	in-domain	–	–	–	t12
Baseline	general	–	general	plain	t12
System 1	general	–	general + in-domain	plain	t14
System 2	general + in-domain	morphology	general + in-domain	morphology	t22
System 3	general + in-domain	morphology + syntax	general + in-domain	morphology + syntax	t30

Thank you!

Questions?