ParaCrawl

paracrawl.eu

Kenneth Heafield, University of Edinburgh

neural.mt

Co-financed by the European Union
Connecting Europe Facility

ParaCrawl: crawl the web for parallel corpora

All 26	EU + EEA official languages
+3	Spanish co-official languages
1–640	Millon words per language
510,482	Websites
1+	Petabyte of compressed web pages

Parallel Corpus Size

Language	Words	Language	Words
French	640,273,938	Finnish	54,984,783
German	502,903,379	Romanian	49,494,227
Spanish	491,951,545	Slovak	35,247,648
Italian	308,244,744	Hungarian	32,151,740
Portuguese	171,495,357	Bulgarian	28,243,306
Russian	157,061,045	Croatian	23,531,438
Dutch	143,294,712	Slovenian	19,915,661
Polish	94,612,131	Lithuanian	19,471,370
Swedish	79,278,861	Estonian	15,633,491
Czech	75,316,848	Irish	15,473,067
Danish	67,200,201	Latvian	15,058,052
Greek	57,752,932	Maltese	3,884,509

Words on English side, after filtering

Improving Quality

From	To	ParaCrawl BLEU Gain Release 1	Release 4
English	Finnish	+0.0	+1.2
Finnish	English	+2.5	+4.6
English	Latvian	+0.7	+1.9
Latvian	English	+0.9	+2.5
English	Romanian	+0.6	+1.3
Romanian	English	+2.4	+4.0
English	Czech	-1.4	-0.1
Czech	English	+0.6	+1.1
English	German	-3.2	+1.2
German	English	-1.0	+3.1

Gains relative to WMT data without ParaCrawl.

Actually 3 Projects

Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2017–14/3/2019

Broader Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2018–14/9/2020

Continued Provision of Web-Scale Parallel Corpora for Official European Languages
1/10/2019–31/9/2021

Actually 3 Projects + Patents Coming Soon

Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2017–14/3/2019

Broader Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2018–14/9/2020

Continued Provision of Web-Scale Parallel Corpora for Official European Languages
1/10/2019–31/9/2021

EuroPat: Unleashing European Patent Translations
15/9/2019–14/9/2021

Actually 3 Projects + Patents Coming Soon

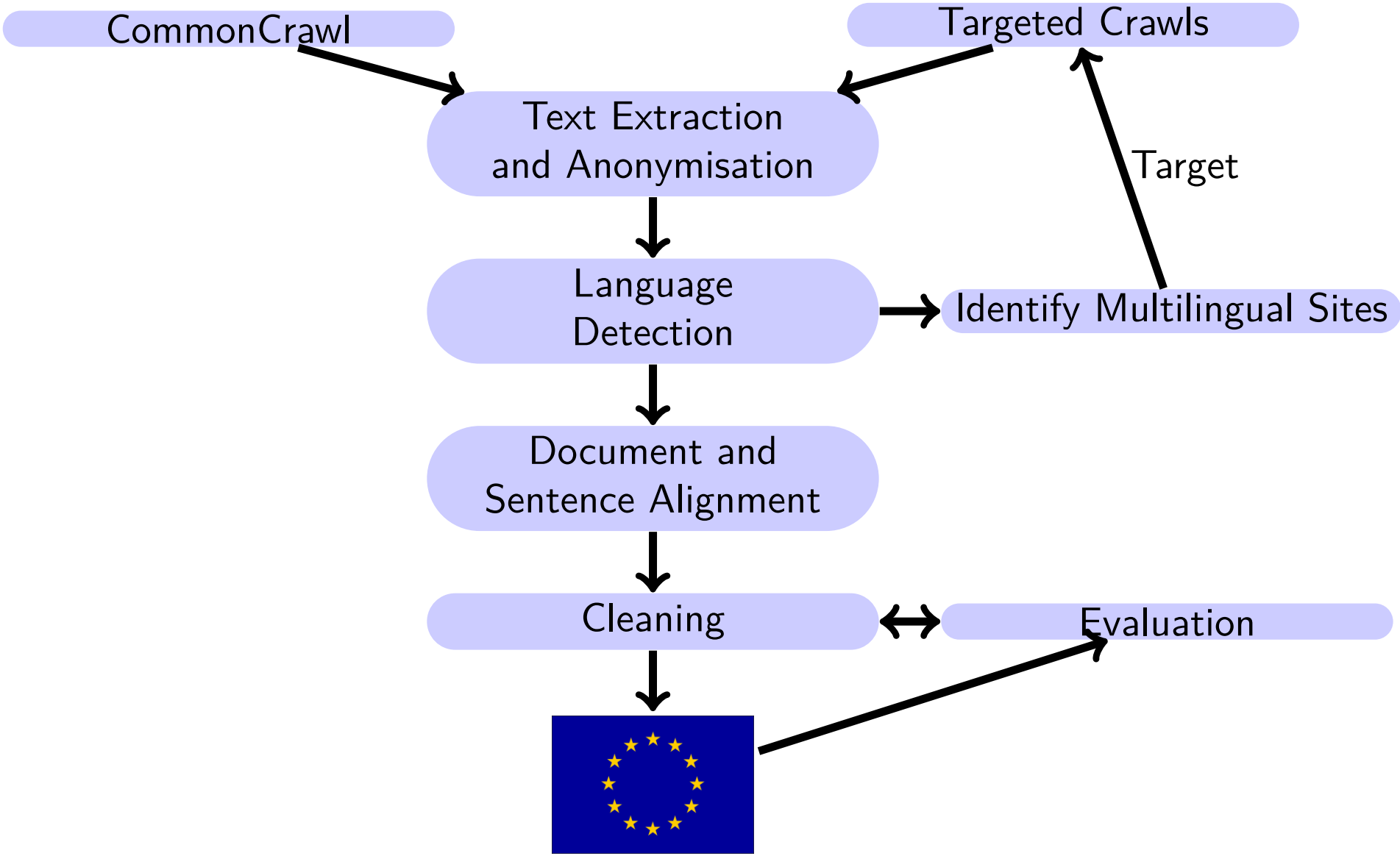
Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2017–14/3/2019

Broader Provision of Web-Scale Parallel Corpora for Official European Languages
15/9/2018–14/9/2020

Continued Provision of Web-Scale Parallel Corpora for Official European Languages
1/10/2019–31/9/2021

EuroPat: Unleashing European Patent Translations
15/9/2019–14/9/2021

More? Provision of Web-Scale Parallel Corpora for Official European Languages
Apply by 14/5/2019?


CommonCrawl
Targeted Crawls
Text Extraction and Anonymisation
Target
Language Detection
Identify Multilingual Sites
Document and Sentence Alignment
Cleaning
Evaluation

Finding sites

- Download Common Crawl
- Classify language on each page
- Pick sites with a mix of both languages

Finding sites

- Download Common Crawl
- Classify language on each page
- Pick sites with a mix of both languages

Problem: limited to sites in CommonCrawl.
Plan: get a petabyte of the Internet Archive.

Not Translated: `wordpress.com`

Blog hosting site

$\Longrightarrow$ multilingual, but few translations.

We blacklist large untranslated sites.

Language classification

Say you're looking for isiXhosa translations:

English	Do you have pets?
isiXhosa	Unazo izilwanaya zasekhaya?

Language classification

Say you're looking for isiXhosa translations:

English	Do you have pets?
isiXhosa	Unazo izilwanaya zasekhaya?

isiXhosa occurs 0.000008x as often as English on the web.
This is lower than error rate in language classification.

$\Longrightarrow$ Most of the "isiXhosa" was actually baseball statistics.

$\Longrightarrow$ Sometimes we need to build our own classifiers.

Matching

Match pages, then match their sentences.
Translate everything to English, do fuzzy matches.

Boilerplate: `santander.co.uk`

"Santander UK plc. Registered Office: 2 Triton Square, Regent's Place, London, NW1 3AN, United Kingdom. Registered Number 2294747. Registered in England and Wales. www.santander.co.uk. Telephone 0800 389 7000. Calls may be recorded or monitored. Authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority. Our Financial Services Register number is 106054. You can check this on the Financial Services Register by visiting the FCA's website www.fca.org.uk/register. Santander and the flame logo are registered trademarks."

$\Longrightarrow$ Match pages on boilerplate.

$\Longrightarrow$ Learn to translate boilerplate really well.

We use `boilerpipe` which tries to throw it out.

Templates: booking.com

"Solo travelers in particular like the location – they rated it 9.5 for a one-person stay."
"Les voyageurs individuels apprécient particulièrement l'emplacement de cet établissement. Ils lui donnent la note de 9,5 pour un séjour en solo."

"Solo travelers in particular like the location – they rated it 8.9 for a one-person stay."
"Les voyageurs individuels apprécient particulièrement l'emplacement de cet établissement. Ils lui donnent la note de 8,9 pour un séjour en solo."

Corpus of repetitive sentences is less useful.
$\Longrightarrow$ Diversity cleaning.

Cleaning

- Supervised classifier trained on 50k good, 50k bad sentences
- Test set *attempts* to have consistent cut-off across languages
- Pattern-based filtering

Shared Task on Corpus Filtering

Common techniques from 2018 Conference on MT:

- More aggressive language model filtering
- Score from translation systems, both directions
- Remove near-duplicates on source and target (not translated)

We will be implementing these

Copyright

Remember: 510,482 websites.
Crawls follow `robots.txt`

Crawler leaves contact information.
A few sites have asked to be removed and we have.

Coming 2020, help using the temporary exemption:
We post links to parallel sentences.
Software for end users to recrawl.

Conclusion

ParaCrawl provides:

- Broad coverage
- Very large corpora
- Demonstrated quality gains.