

Web-Scale bitext mining and its impact on NMT

Kenneth Heafield, University of Edinburgh

paracrawl.eu

neural.mt



Co-financed by the European Union
Connecting Europe Facility



Microsoft



ParaCrawl: crawl the web for parallel corpora

All 24 EU official languages

1–640 Million words per language

510,482 Websites

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

Coming

Spanish–Basque

Spanish–Catalan

Spanish–Galician

English–Chinese

English–Icelandic

English–Norwegian (both forms)

English–Somali

Proposal under review:

Dutch–French

German–Polish

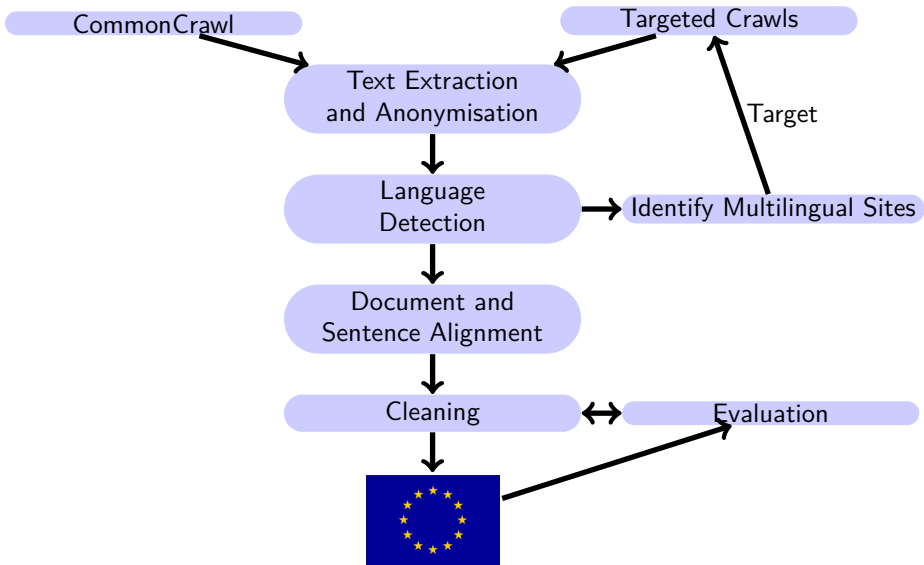
Quality Impact

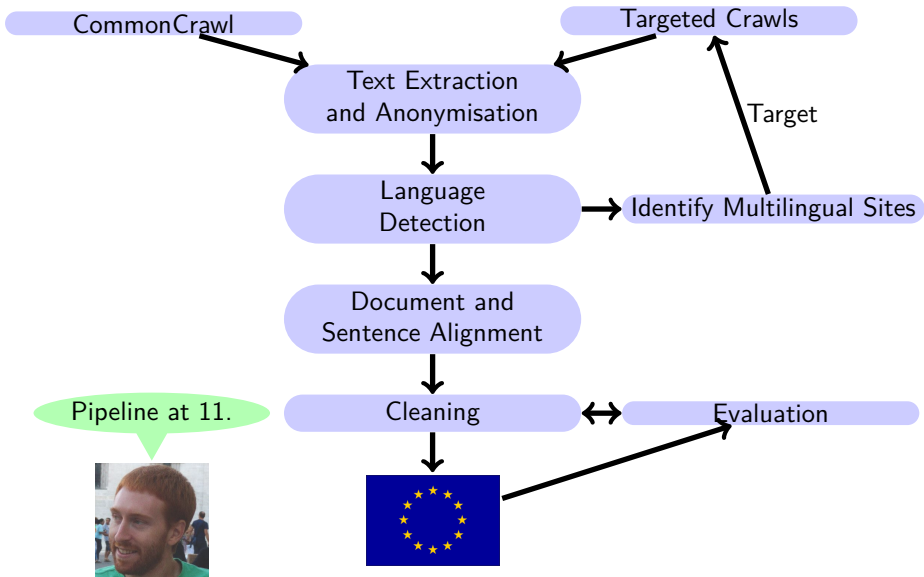
Pair	Baseline	+ParaCrawl	Gain
Czech→English	25.7	26.3	+0.6
Finnish→English	21.7	24.2	+2.5
German→English	29.7	31.4	+1.7
Latvian→English	15.6	16.4	+0.8
Romanian→English	29.2	32.4	+3.1

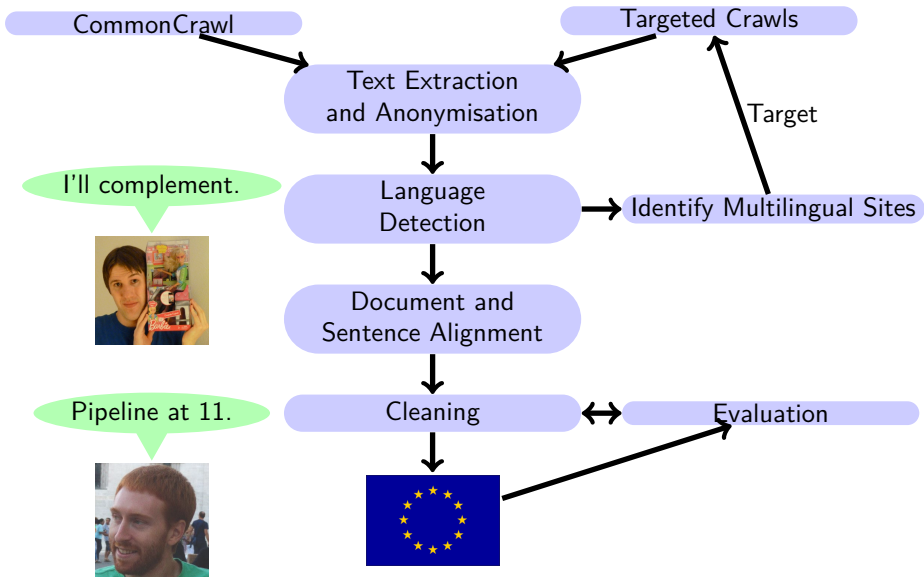
BLEU scores on 2017 Conference on Machine Translation test

Outline

- 1 The pipeline
- 2 Tips for using data
- 3 Legal







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.

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.

wordpress.com is multilingual, but has few translations

→ We have a small blacklist.

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.

Lower than classifier error rate, so mostly we get garbage:

“zojaexaTorquay, Devon,” “GT-001 / GT-100 Ver.2 MIDI”

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.

Lower than classifier error rate, so mostly we get garbage:

“zojaexaTorquay, Devon,” “GT-001 / GT-100 Ver.2 MIDI”

Low-resource languages need a second-pass language model filter.

Sorry Estonia, Ireland, Latvia, Malta, and South Africa.

Matching

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

Matching

Match pages, then match their sentences.

Translate everything to English, do fuzzy matches.

Boilerplate

On bankofamerica.com, most text is a legal disclaimer.

⇒ Match on disclaimer, not content.

⇒ Great at translating disclaimers, terrible at content.

We use boilerpipe to remove boilerplate
...but it's not perfect

Templates: the booking.com problem

“Solo travelers in particular like the location – they rated it 8.3 for a one-person stay.”

“Les voyageurs individuels apprécient particulièrement l'emplacement de cet établissement. Ils lui donnent la note de 8,3 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.”

Fuzzy matching can get numbers wrong.
MT learns weird repetitive sentences.

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

Outline

- ① The pipeline
- ② **Tips for using data**
- ③ Legal

Threshold by Quality

We provide quality scores in the TMX files.
Apply a threshold for better data.

ParaCrawl 1 German → English:

Filter	Δ BLEU
--------	---------------

60%	+1.7
-----	------

100%	-1
------	----

Weight ParaCrawl Lower

Neural MT makes it easy to weight sentences.
Just duplicate good data in the corpus!

We typically use 7x good, 1x ParaCrawl

Filter for relevance

ParaCrawl is broad, not your domain.

Train in-domain language model, filter the corpus.

We'll offer a ready-made package in ParaCrawl 2.

Languages are different

There's different amounts of data on the web
Different sources.
And you have different corpora

⇒ Sadly there isn't a one-size-fits-all filtering
But we are working to make scores comparable across languages.

Legal Issues

Personal Data

Copyright

Our stance

Follow robots.txt

Can ask to be removed. Only one person has.

We don't claim to own the web.

If we did something copyrightable, dedicate to public domain (CC0).

Delivery to eTranslation



Delivery to eTranslation

The first word of
copyright is copy.

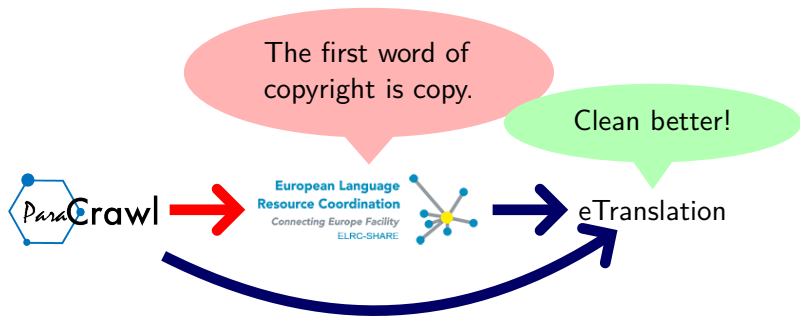


Delivery to eTranslation bypassing ELRC


The first word of
copyright is copy.





Delivery to eTranslation bypassing ELRC



ELRA sold webcrawl parallel corpora:

 [Browse Resources](#) [Information](#) Cart total [View cart](#) [Register](#) [Login](#)

Linguatools Webcrawl Parallel Corpus German-English 2015

 12  1





▸ View resource name in all available languages





ISLRN: 800-190-274-236-9

ID: **ELRA-W0091**

The corpus consists of 10 million German-English parallel sentences that were crawled from the internet between 10/2013 and 04/2015. The sentences were gathered from over 112,000 different hosts. An elaborate multi-step quality filtering was applied, including language identification filter, machine translation filter, grammaticality filter, etc. to get as clean data as possible. There are no duplicate sentence pairs, and there is no overlap with existing publicly available corpora like europarl, DGT-TM, etc. Web pages have been automatically categorized for subject area. The corpus is available in TMX and Moses format (encoding UTF-8). [Read Less](#)

▸ View resource description in French

MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	1000.00 € 	4800.00 € 
Licence: Commercial Use - ELRA VAR	4800.00 € 	4800.00 € 

NON MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	1200.00 € 	5000.00 € 
Licence: Commercial Use - ELRA VAR	5000.00 € 	5000.00 € 

ELRA took this down after my last talk.

Quotation in UK law

(1ZA) Copyright in a work is not infringed by the use of a quotation from the work (whether for criticism or review or otherwise) provided that—

(a) the work has been made available to the public,

(b) the use of the quotation is **fair dealing** with the work,

(c) the extent of the quotation is no more than is required by the specific purpose for which it is used, and

(d) the quotation is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).

National Library of the Netherlands

The Netherlands has no legal deposit law.

National library archives the web anyway.

Report on legal issues: tinyurl.com/ycn6daap

Conclusion: “adopted a pragmatic way to handle the copyright issues: the opt-out approach. This approach assumes implicit permission for web archiving.”

ELRC-endorsed Temporary Exemption

Legal Basis

InfoSoc directive: mandatory exemption for temporary acts of reproduction
“it is not impossible to organize the crawling process in such a way as to comply with the conditions of the exception” –ELRC report

Option

We provide URLs, sentence positions, and checksums (Forcada et al 2006)
eTranslation runs our script to download pages again.
Script creates model, automatically deletes web pages.
Ok to keep model “lawful use”

ELRC-endorsed Temporary Exemption

Legal Basis

InfoSoc directive: mandatory exemption for temporary acts of reproduction
“it is not impossible to organize the crawling process in such a way as to comply with the conditions of the exception” –ELRC report

Option

We provide URLs, sentence positions, and checksums (Forcada et al 2006)
eTranslation runs our script to download pages again.
Script creates model, automatically deletes web pages.
Ok to keep model “lawful use”

This doesn't appear in the ELRC report summaries!

Saner Approach: ROAM

Randomise Shuffle the sentences.

Omit Remove data. In Germany, “up to 15% of a work can be reproduced and communicated to the public” –ELRC

Anonymise Replace phone numbers, e-mail addresses, etc. with a constant.

Mix Jumble sentences from different sources.

Conclusion

ParaCrawl provides:

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