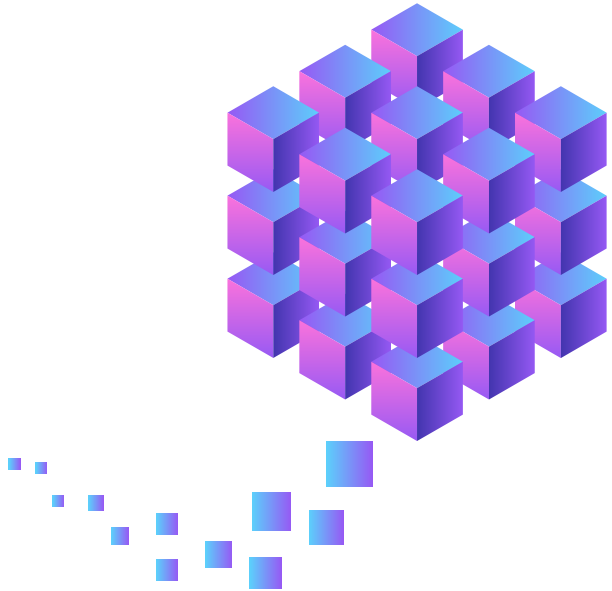
LREC 2022
Marseille

Quality versus Quantity: Building Catalan-English MT Resources



Ona de Gibert, Ksenia Kharitonova,
Blanca Calvo Figueras, Jordi Armengol-Estapé,
Maite Melero

24.06.2022

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Motivation

GOAL Build MT resources for the Catalan-English pair

Less-Resource

How can we **build** MT resources for less-resource pair?

Quality

How can we **evaluate** the quality of a given dataset?

Filtering

How can we **improve** the quality of a given dataset?

Related Work

Training Data

Large multilingual parallel corpora

- OPUS (Tiedemann, 2012)
- CCAligned (El-Kishky et al., 2020)
- Wikimatrix (Schwenk et al., 2019)
- ParaCrawl (Bañón et al., 2020)

Quality

Poor quality (Kreutzer et al. 2021)

Quality assessment

- Monolingual (Caswell et al., 2020)
- Parallel corpora (Kreutzer et al. 2021)

Parallel Corpus Filtering

Parallel corpus filtering Shared Task at WMT (Koehn et al., 2020)

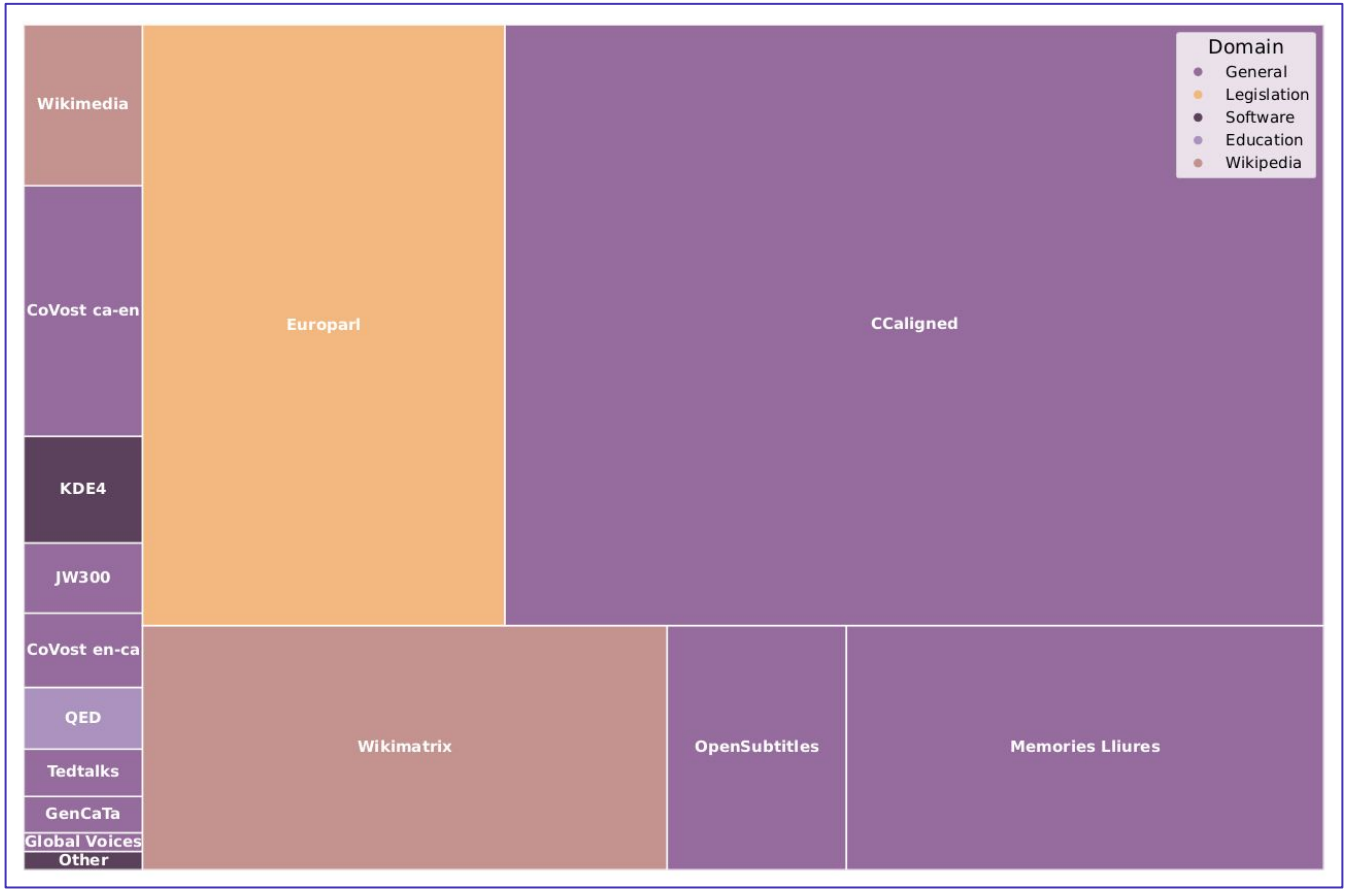
Filters based on:

- Heuristics
- Sentence embeddings
- Binary classifier

Step 1: Build

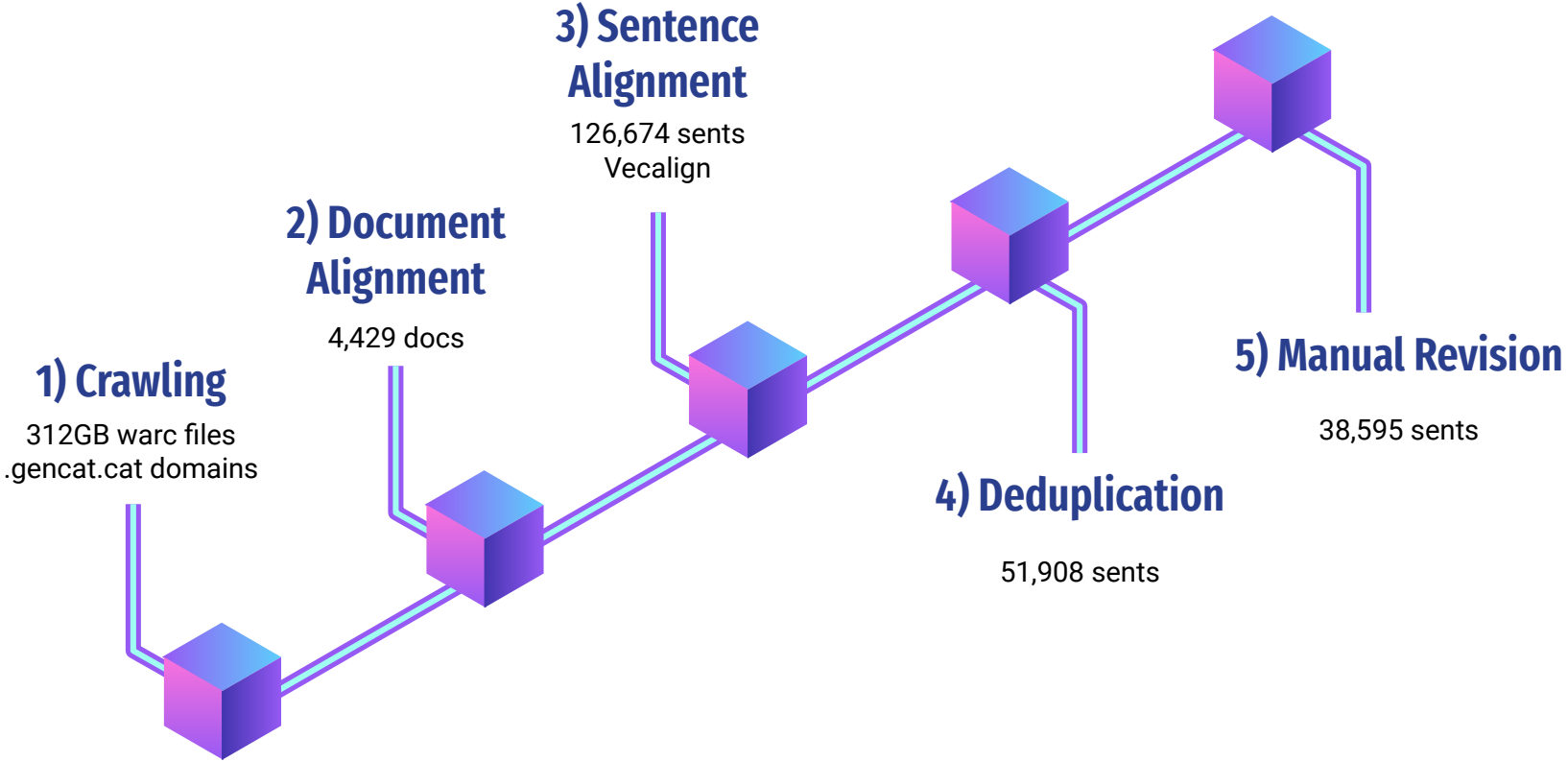
	Dataset	Sentences	Tokens	Tokens/Sent	Source	Domain
1	CCaligned	5,787,682	89,606,874	15.48	(El-Kishky et al., 2020)	General
2	COVID-19 Wikipedia	1,531	34,836	22.75	(Tiedemann, 2012)	Health
3	CoVost ca-en*	263,891	809,660	10.17	(Wang et al., 2020)	General
4	CoVost en-ca*	79,633	2,953,096	11.19	(Wang et al., 2020)	General
5	Eubookshop	3,746	82,067	21.91	(Tiedemann, 2012)	Legislation
6	Europarl	1,965,734	50,417,289	25.65	(Koehn, 2005)	Legislation
7	GEnCaTa*	38,595	858,385	22.24	New	General
8	Global Voices	21,342	438,032	20.52	(Tiedemann, 2012)	General
9	Gnome*	2,183	30,228	13.85	(Tiedemann, 2012)	Software
10	JW300	97,081	1,809,252	18.64	(Agić and Vulić, 2019)	General
11	KDE4*	144,153	1,450,631	10.06	(Tiedemann, 2012)	Software
12	Memories Lliures*	1,173,055	9,452,382	8.06	Softcatalà	Software
13	Open Subtitles	427,913	2,796,350	6.53	(Lison and Tiedemann, 2016)	General
14	Opus Books	4,580	73,416	16.03	(Tiedemann, 2012)	Narrative
15	QED*	69,823	1,058,003	15.15	(Abdelali et al., 2014)	Education
16	Tatoeba*	5,500	34,872	6.34	(Tiedemann, 2012)	General
17	Tedtalks	50,979	770,774	15.12	Softcatalà	General
18	Ubuntu	6,781	33,321	4.91	(Tiedemann, 2012)	Software
19	Wikimatrix	1,205,908	28,111,517	23.31	(Schwenk et al., 2019)	Wikipedia
20	Wikimedia*	208,073	5,761,409	27.69	(Tiedemann, 2012)	Wikipedia
	Total	11,558,183	196,582,394	15.78		

Step 1: Build


Domain
General
Legislation
Software
Education
Wikipedia
Wikimedia
CoVost ca-en
KDE4
JW300
CoVost en-ca
QED
Tedtalks
GenCaTa
Global Voices
Other
Europarl
CCaligned
Wikimatrix
OpenSubtitles
Memories Lliures

GEnCaTa

a High Quality Parallel Corpus



Step 2: Evaluate

We perform a **Human Audit** of 100 aligned segments/dataset → 2000 aligned segments

Error Taxonomy (Kreutzer et al. 2021)

- CC: Correct translation, natural sentence
- CS: Correct translation, but single word or short phrase
- CB: Correct translation, but boilerplate
- X: Incorrect translation
- WL: Wrong language
- NL: Not language

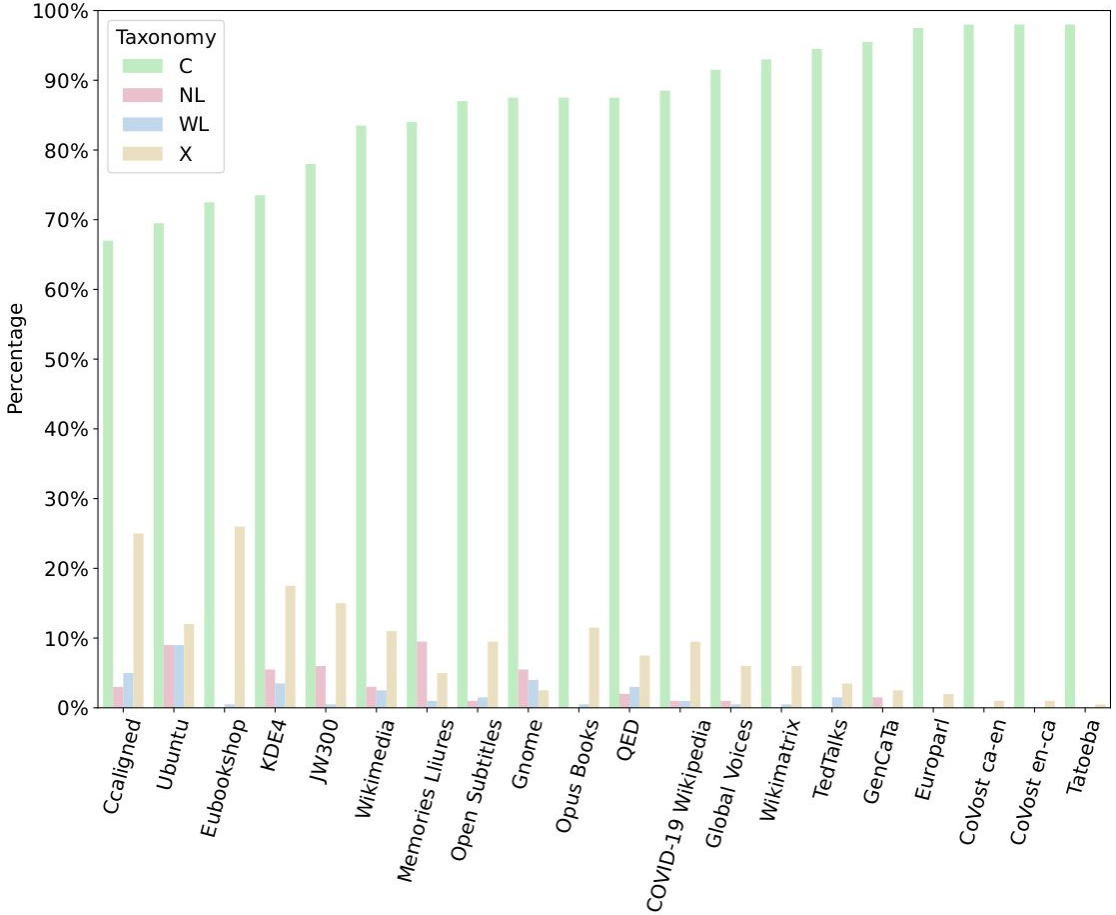
Step 2: Evaluate

We perform a **Human Audit** of 100 aligned segments/dataset → 2000 aligned segments

Error Taxonomy (Kreutzer et al. 2021)

- CC: Correct translation, natural sentence
  - *És al nord-est dels Plans de Puigmaçana, al sud-oest del Corral del Sastre. / It's to the northeast of the Puigmaçana Plains, southwest of the Corral del Sastre.*
- CS: Correct translation, but single word or short phrase
  - Silencia l'usuari / Silence User
- CB: Correct translation, but boilerplate
  - Molt bones festes i millor entrada d'any 2015 / Merry Christmas and a happy 2015
- X: Incorrect translation
  - Agraeixo a la meva esposa Marina, que s'ha quedat amb mi. / This did not include his half-sister, Margaret who was allowed to be with them.
- WL: Wrong language
  - Malditos bastardos, no saben nada de la historia de su propio pais / You little rascals know nothing about our own history.
- NL: Not language
  - @@image: 'figures/api_browser.png'; md5=7e3b2481bf743644470726421cb5afb1 / external ref='figures/api_browser.png' md5='47cae67d29c708139b9740d06fd2521e'

Human Audit Results



Step 3: Improve

Binary classifier

We fine-tune mBERT with the labeled GEnCaTa dataset

Label	Train	Valid	Test
Positive	23,897	7,490	7,489
Negative	8,011	2,510	2,511
Total	31,908	10,000	10,000

Table 3: Train, valid and test splits of the GEnCaTa dataset for parallel corpus filtering

Model	F1	Accuracy
mBERT-uncased	0.968	0.952
mBERT-cased	0.970	0.955

Table 4: Fine-tuning of mBERT results on the GEnCaTa dataset for parallel corpus filtering

Step 3: Improve

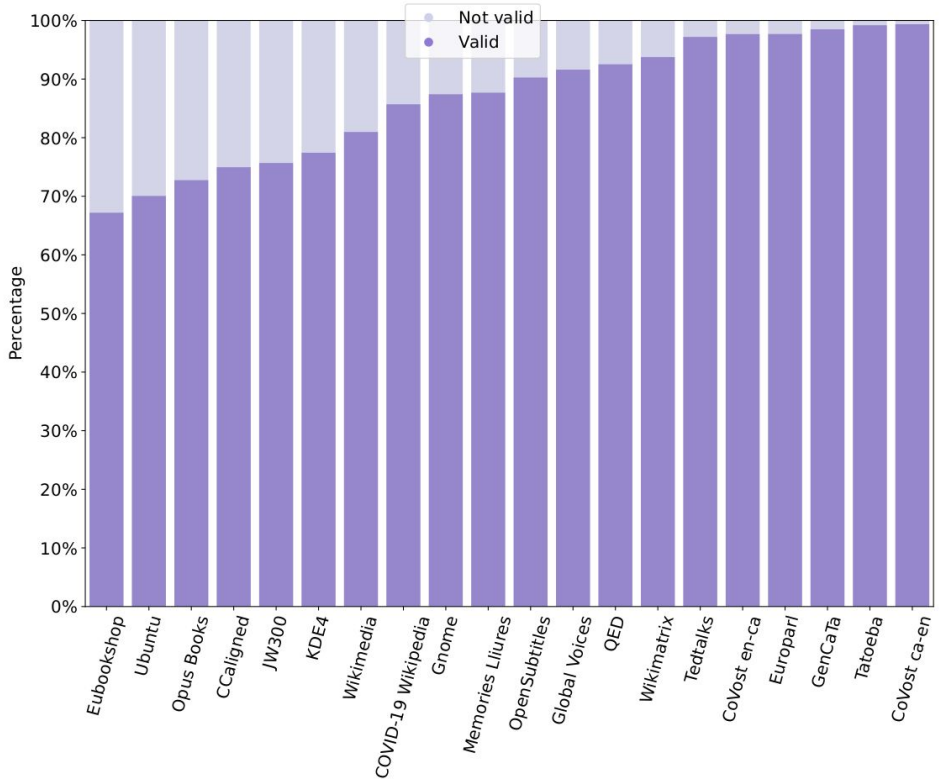
Binary classifier

We fine-tune mBERT with the labeled GEnCaTa dataset

Filtering

We filter all the compiled data with our binary classifier

0.89 Pearson with Human Audit



Step 3: Improve

Binary classifier

We fine-tune mBERT with the labeled GEnCaTa dataset

Filtering

We filter all the compiled data with our binary classifier

0.89 Pearson with Human Audit

MT systems

We train two MT systems:

- RAW: raw data
- FIL: filtered data

New evaluation resources!

Evaluation Resources

Existing test sets

- Flores-101 (Goyal et al., 2021)
- Catalan United Nations test set (Costa-jussà, 2020)
- WMT20 Biomedical Shared Task test set (Bawden et al., 2020)

New test sets

Dataset	Languages	Domain	Sent.	Tokens
CyberMT	ca, es, en	cybersecurity	1,715	33,050
TaCon	ca, es, en eu, ga	legislation	1,110	18,275
WMT13	ca, es, en, de, ru, fr, cs	newswire	3,000	59,340

Table 4: Language resources for MT evaluation. *Tokens* refers to Catalan tokens.

Parallel Corpus Filtering

Binary classifier

We fine-tune mBERT with the labeled GEnCaTa dataset

Filtering

We filter all the compiled data with our binary classifier

0.89 Pearson with Human Audit

MT systems

We train two MT systems:

- RAW: raw data
- FIL: filtered data

Direction	Test set	RAW	FIL	Increase
EN → CA	Cyber	40.2	43.1	1,7
	Flores-101	35.7	38.0	1,3
	TaCon	28.9	30.2	2,9
	WMT13	31.2	32.9	2,3
CA → EN	Cyber	47.4	49.5	1,9
	Flores-101	34.7	37.6	2,6
	TaCon	32.4	35.0	2,1
	WMT13	34.1	36.0	2,9

Table 5: sBLEU scores for MT evaluation

Zero-Shot Cross-lingual Transfer Learning

We apply our model to new language pairs:

- 3,000 sentences of the WMT as positive examples
- 3,000 negative examples generated synthetically
- All available language combinations: CA, CS, DE, EN ES, FR, RU

Zero-Shot Cross-lingual Transfer Learning

We apply our model to new language pairs:

- 3,000 sentences of the WMT as positive examples
- 3,000 negative examples generated synthetically
- All available language combinations: CA, CS, DE, EN ES, FR, RU

	Target						
Source	CA	CS	DE	EN	ES	FR	RU
CA	-	0.952	0.979	-	0.985	0.982	0.954
CS	0.947	-	0.976	0.987	0.948	0.940	0.972
DE	0.879	0.934	-	0.987	0.937	0.949	0.958
EN	-	0.894	0.961	-	0.938	0.957	0.925
ES	0.977	0.947	0.980	0.988	-	0.982	0.971
FR	0.960	0.916	0.979	0.988	0.967	-	0.964
RU	0.936	0.972	0.979	0.981	0.975	0.969	-

Conclusions & Future Work

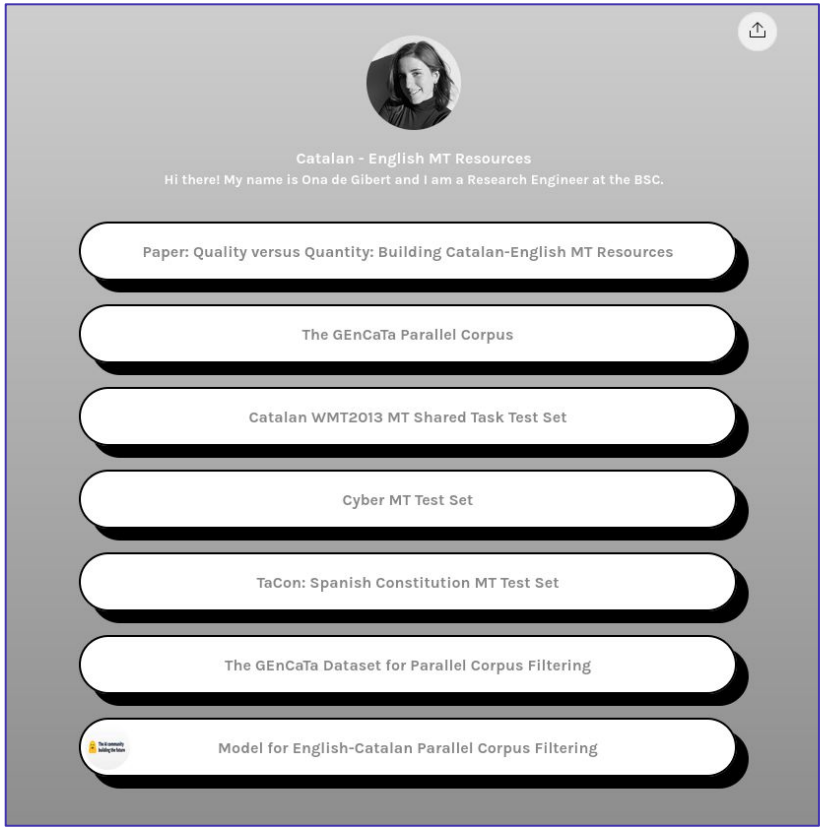
Quality assessment is worth!

- Investigate further the task of quality estimation of parallel corpora and its impact in the obtained MT engines.
- Conduct a more qualitative analysis of the output of the MT systems to gain linguistic insights from the results.
- Explore the transfer-learning capabilities of our model

Published Resources

- The GEnCaTa Parallel Corpus
- Catalan WMT2013 MT Shared Task Test Set
- Cyber MT Test Set
- TaCon: Spanish Constitution MT Test Set
- The GEnCaTa Dataset for Parallel Corpus Filtering
- Model for English-Catalan Parallel Corpus Filtering

Published Resources



Thank you!

You can find me at:

 ona.degibert@bsc.es

 linkedin.com/in/onadegibertbonet

 twitter.com/OnaDeGibert