The OSCAR Project

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Improving Data Quality in Multilingual Heterogeneous Web-Based Corpora

DFKI Projektbüro Berlin
Speech and Language Technology
Dr. Pedro Ortiz Suarez
portizs.eu
pedro.ortiz@dfki.de

0. The OSCAR Project

The OSCAR Project

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- Open Source
- Data-Focused
- Web-Based
- Community Driven
- Text-based 👀
- Highly-optimized Software

- **The OSCAR project** is an Open Source project **aiming** to provide high-quality web based data for **Machine Learning** applications in **as many languages as possible**.
- From the beginning the OSCAR project has aimed to **increase** the amount of **data available** for **mid- to low-resource languages**, with the intend to make **latest developments in ML** available to **as many people as possible**.
- For the moment the project has focused on textual data, but is **looking to expand** to other domains in the future and reach and help new communities.
- We try to provide high performance data pipelines, usable even in **low-resource infrastructures**.

1. The History of the OSCAR Project

1.1 Pedro Does a Ph.D.

BASNUM

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

RAQUETTE. ſ. f. Eſpece de pallette pour jouër à la paume, & au volant. Elle eſt faite d'un treillis de cordes de boyaux (dont les unes s'appellent *montans*, & les autres *travers*) fort tenduës ſur un tour de bois qui a un manche de mediocre longueur. Un de ſes côtez s'appelle les *droits*, & l'autre les *nœuds*. Paſquier a remarqué qu'anciennement on ne joüoit point à la paume avec des *raquettes*: c'étoit avec la paume de la main; & de là il conjecture qu'eſt venu le nom de jeu de *paume*. On n'avoit inventé les *raquettes* qu'un peu avant le temps de Paſquier, à ce qu'il dit.

Menage derive ce mot de *retiquetta*, diminutif de *retis*, *reticus* & *reticulum*.

On dit proverbialement pour ſe mocquer d'un homme qui ſe vante de pluſieurs choſes qu'il n'a pas faites, C'eſt un grand caſſeur de *raquettes*.

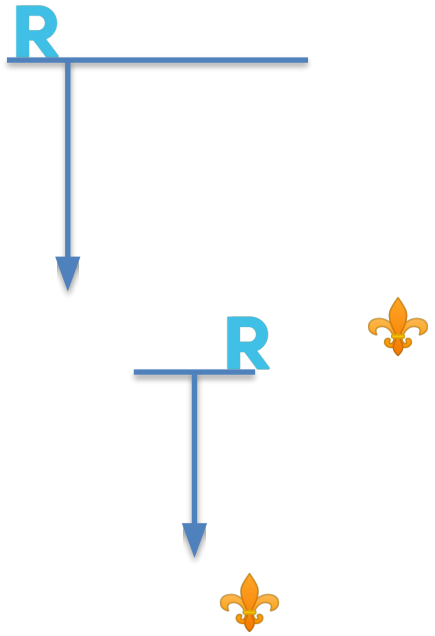
RAQUETTE, ſe dit auſſi d'une certaine machine que les Sauvages de Canada attachent à leurs pieds pour marcher plus commodément ſur la neige, & qui eſt faite à-peu-près en forme de *raquette* à jouër.

RAQUETTE, ſe dit auſſi d'une eſpece de figuier d'Inde qui croît aux Iles Antilles: c'eſt cette eſpece que Mr. Tournefort appelle *opuntia vulgò herbariorum*, J. BAUH. Ses feuilles ſont épaiſſes, longues, quelquefois larges comme une raquette, d'où vient que les François lui ont donné ce nom. Voyez FIGUIER D'INDE.

```
<div n="RAQ">
    <entry xml:lang="fre" xml:id="Raquette">
        <sense><form><orth rendition="#uc">raquette.</orth><gramGrp><pos expand="Substantif">s.</pos>
        <gen expand="Feminin">f.</gen></gramGrp></form>  <def>Espece de pallette pour jouër
        à la paume, &amp;amp; au volant.</def> <note>Elle est faite d'iun treillis de
        cordes de boyaux édont les unes s'appellent montans
        &amp;amp; les autres travers J fort tendues sur un tour de bois
        qui a un manche de mediocre longueur.</note> <xr corresp="#droits #noeuds">Un de ses cô-
        tea s'appelle les droits, &amp;amp; l'autre les nœuds</xr>. <etym><bibl><author ref="#Pasquier_ISN0000000010907010">Pasquier</author></bibl>
        a remarqué qu'anciennement on ne joüoit point à la
        paume avec des raquettes : c'étoit avec la paume de la
        main , &amp;amp; de là il conjectue qu'est venu le nom de jeu
        de pame.</etym> <note>On n'avoit inventé les raquettes qu'un peu
        avant le temps de <bibl><author ref="#Pasquier_ISN0000000010907010">Pasquier</author></bibl>, à ce qu'il dit.</note>


        <etym><bibl><author ref="#Men_ISN0000000080815971">Menage</author></bibl> derive ce mot de
        <lang rendition="#i" xml:lang="lat">retiquetta</lang>, diminutif de retis,
        reticus &amp;amp; reticulum.</etym>
        </sense>

        <sense><usg>On dit proverbialement</usg> <gloss>pour se mocquer d'un homme qui
        se vante de plusieurs choses qu'il n'a pas faites,</gloss> <seg type="Proverbe">C'est
        un grand casseur de raquettes.</seg></sense>


        <sense><form><orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une certaine machine que
        les Sauvages de Canada attachent à leurs pieds pour
        marcher plus commodément sur la neige, &amp;amp; qui est
        faite à-peu-près en forme de raquette à jouër.</def></sense>


        <sense><form><orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une especе de figuier d'In-
        de qui croît aux Iles Antilles :</def> <note>c'est cette espece que
        Mr. Tournefort appelle opuntia vule herbariorum,
        J. Bauh. Ses feuilles sont épaisses, longues, quel-
        quefois larges comme une raquette, d'où vient que les
        François lui ont donné ce nom.</note> <xr>Voyez <ref corresp="#Ficuier">Ficuier
        d'Inde.</ref></xr></sense>
    </entry>
</div>
```

Plan

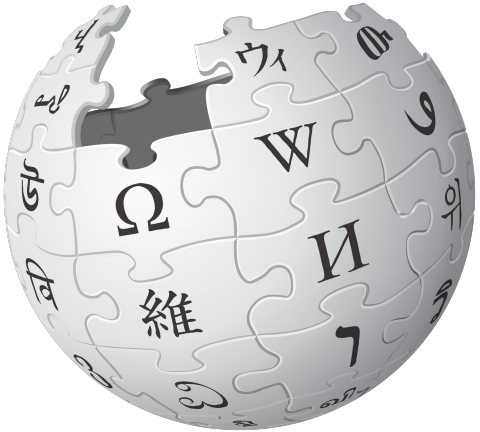
- Put together a big Contemporary French corpus
- Train Contemporary French language model
- Put together a small Historical French corpus
- Continue the training of the contemporary French language model with this corpus.



Available Resources

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

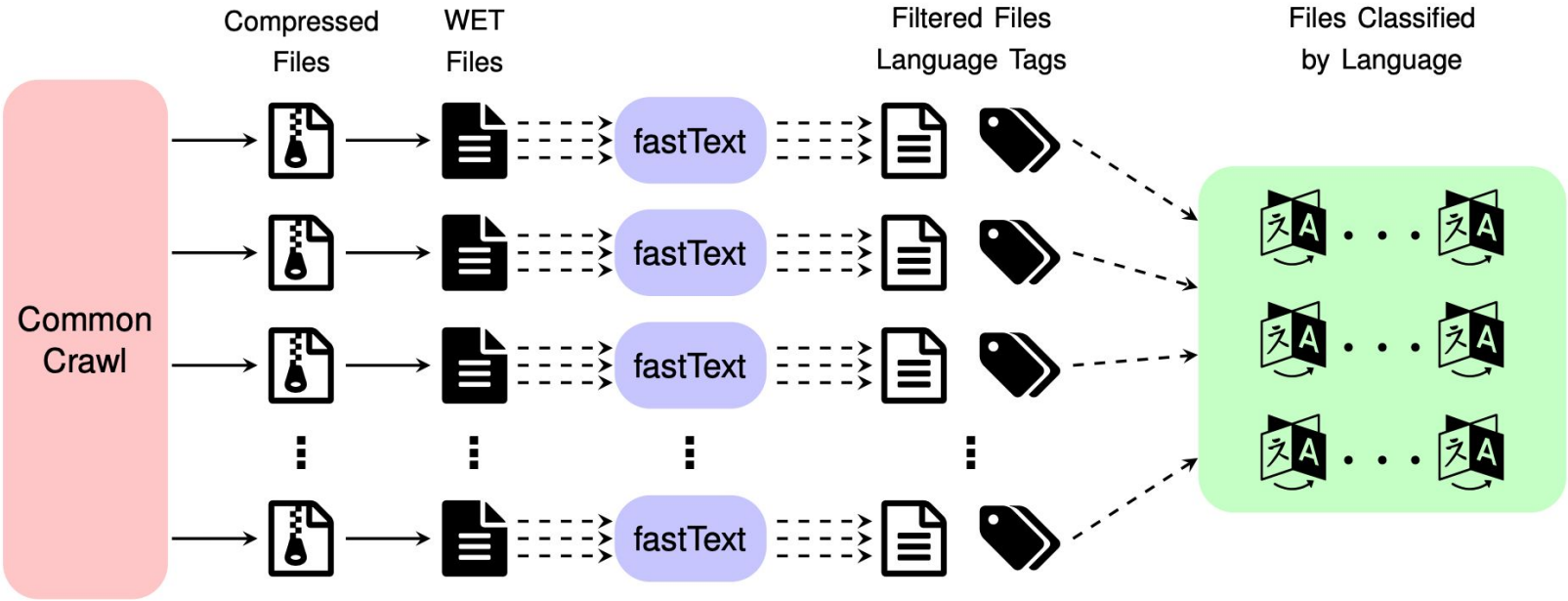
- Wikipedia
  - Too Small
  - Too Regular
- Frwac
  - Many encoding Problems
  - Old
  - Still too small
- Common Crawl?
  - YES!



1.2 The First OSCAR Corpus or OSCAR 2019

(Grave et al., 2018)

Goclassy


Compressed Files
WET Files
Filtered Files Language Tags
Files Classified by Language
Common Crawl
fastText
fastText
fastText
fastText

OSCAR 2019

OSCAR 2019

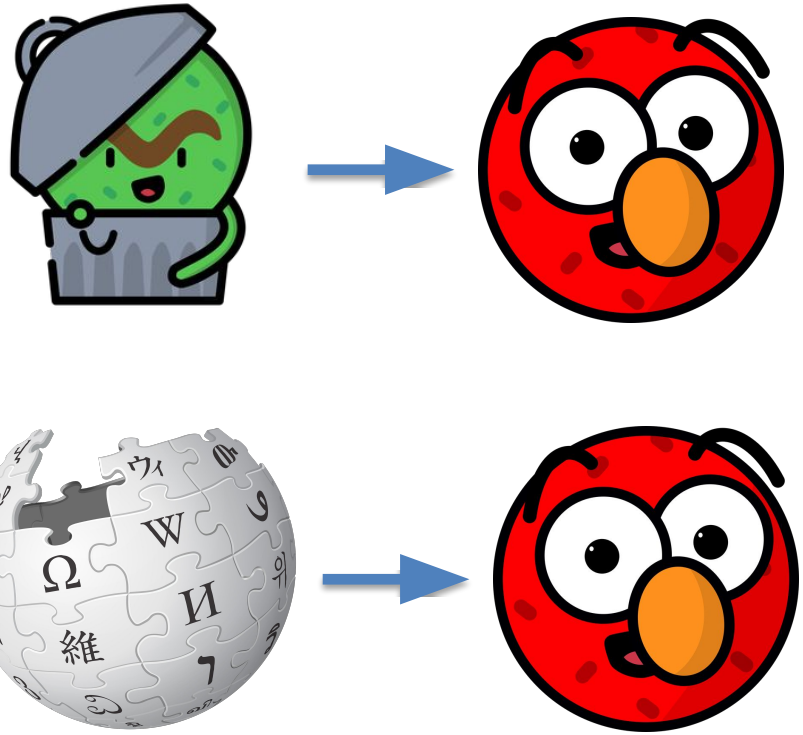
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Language	Size		Words	
	Orig	Dedup	Orig	Dedup
English	2.3T	1.2T	418,187,793,408	215,841,256,971
Russian	1.2T	568G	92,522,407,837	46,692,691,520
Spanish	278G	149G	47,545,122,279	25,928,290,729
French	282G	138G	46,896,036,417	23,206,776,649
German	308G	145G	44,878,908,446	21,529,164,172
Italian	137G	69G	22,248,707,341	11,250,012,896
Portuguese	124G	64G	20,641,903,898	10,751,156,918
Chinese	508G	249G	14,986,424,850	6,350,215,113
Japanese	216G	106G	4,962,979,182	1,123,067,063
Polish	109G	47G	15,277,255,137	6,708,709,674
Total OSCAR	6.3T	3.2T	844,315,434,723	425,651,344,234

2. Is The Data Any Good?

2.1 First Automatic Evaluation

Is OSCAR Too Noisy?



Results on UD Treebanks

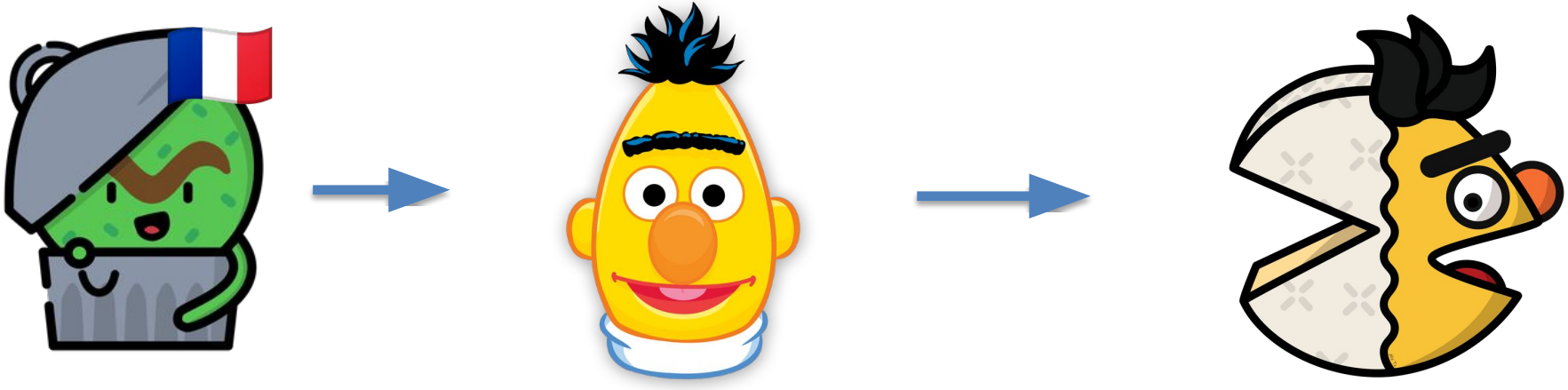
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Treebank	Model	UPOS	UAS	LAS
Bulgarian BTB	UDify	98.89	95.54	92.40
	UDPipe 2.0	98.98	93.38	90.35
	+mBERT	99.20	95.34	92.62
	+ELMo$_{Wikipedia}$	99.17	94.93	92.05
	+ELMo$_{OSCAR}$	**99.40**	**96.01**	**93.56**
Catalan-AnCora	UDify	98.89	94.25	92.33
	UDPipe 2.0	98.88	93.22	91.06
	+mBERT	**99.06**	**94.49**	92.74
	+ELMo$_{Wikipedia}$	99.05	93.99	92.24
	+ELMo$_{OSCAR}$	**99.06**	**94.49**	**92.88**
Danish-DDT	UDify	97.50	87.76	84.50
	UDPipe 2.0	97.78	86.88	84.31
	+mBERT	98.21	89.32	87.24
	+ELMo$_{Wikipedia}$	98.45	89.05	86.92
	+ELMo$_{OSCAR}$	**98.62**	**89.84**	**87.95**

Treebank	Model	UPOS	UAS	LAS
Finnish-FTB	UDify	93.80	86.37	81.40
	UDPipe 2.0	96.65	90.68	87.89
	+mBERT	96.97	91.68	89.02
	+ELMo$_{Wikipedia}$	97.27	92.05	89.62
	+ELMo$_{OSCAR}$	**98.13**	**93.81**	**92.02**
Finnish-TDT	UDify	94.43	86.42	82.03
	UDPipe 2.0	97.45	89.88	87.46
	+mBERT	97.57	91.66	89.49
	+ELMo$_{Wikipedia}$	97.65	91.60	89.34
	+ELMo$_{OSCAR}$	**98.36**	**93.54**	**91.77**
Indonesian-GSD	UDify	93.36	86.45	80.10
	UDPipe 2.0	93.69	85.31	78.99
	+mBERT	94.09	86.47	80.40
	+ELMo$_{Wikipedia}$	93.94	86.16	80.10
	+ELMo$_{OSCAR}$	**94.12**	**86.49**	**80.59**

2.2 Evaluating in a High Resource Language

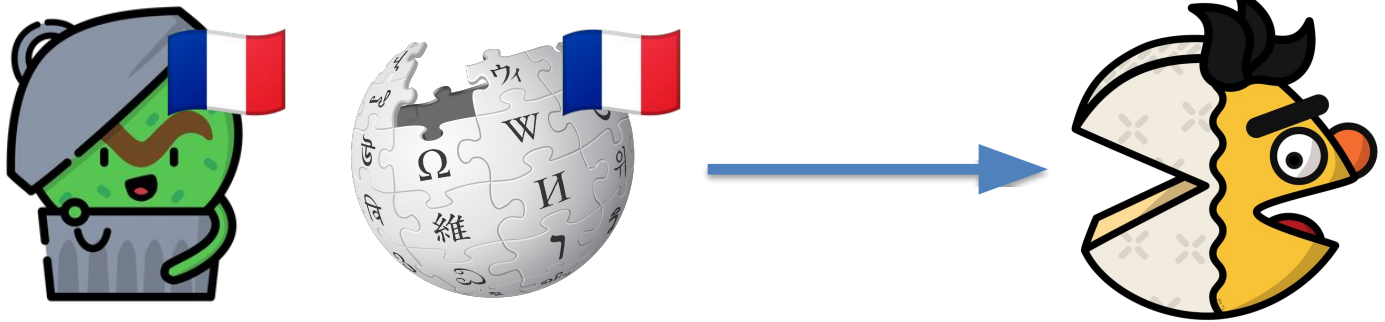
CamemBERT



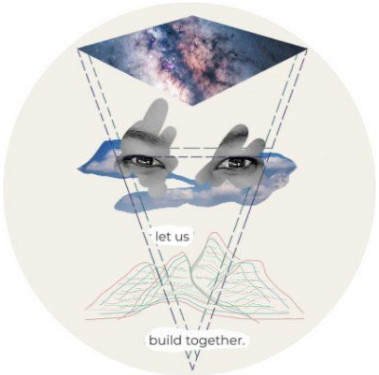
OSCAR vs Wikipedia

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		Average		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55
Embeddings (with UDPipe Future (tagging, parsing) or LSTM+CRF (NER))													
Wiki	4GB	98.09	92.31	98.74	93.55	96.24	78.91	95.78	89.79	97.21	88.64	91.23	-
CCNet	4GB	**98.22**	**92.93**	99.12	94.65	97.17	**82.61**	**96.74**	89.95	97.81	90.04	**92.30**	-
OSCAR	4GB	98.21	92.77	99.12	**94.92**	97.20	82.47	**96.74**	**90.05**	**97.82**	**90.05**	91.90	-
OSCAR	138GB	98.18	92.77	**99.14**	94.24	**97.26**	82.44	96.52	89.89	97.77	89.84	91.83	-



3. Quality at Glance: The First Human OSCAR Evaluation

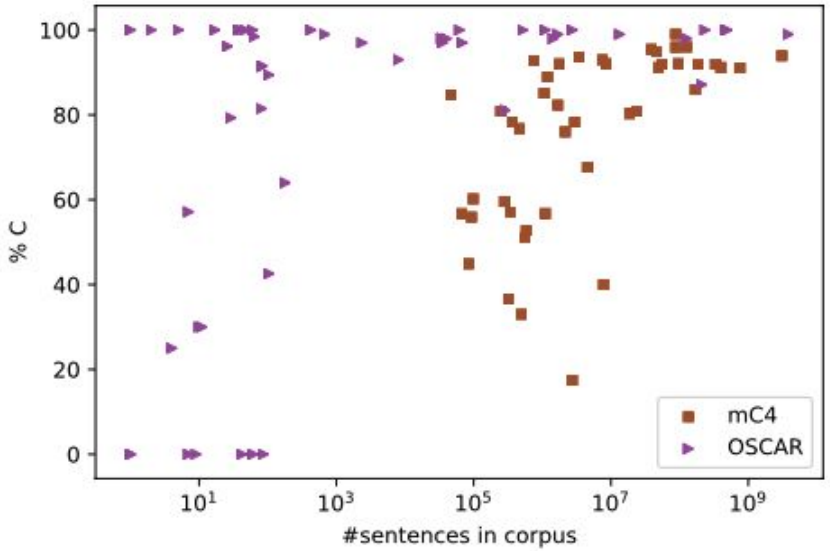


Results

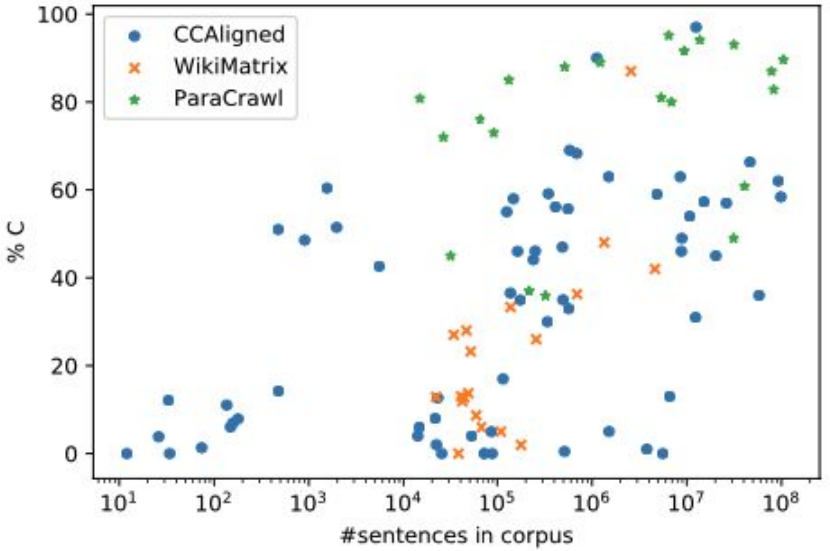
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

		Parallel			Monolingual	
		CCAligned	ParaCrawl v7.1	WikiMatrix	OSCAR	mC4
	#langs audited / total	65 / 119	21 / 38	20 / 78	51 / 166	48 / 108
	%langs audited	54.62%	55.26%	25.64%	30.72%	44.44%
	#sents audited / total	8037 / 907M	2214 / 521M	1997 / 95M	3517 / 8.4B	5314 / 8.5B
	%sents audited	0.00089%	0.00043%	0.00211%	0.00004%	0.00006%
macro	C	29.25%	76.14%	23.74%	87.21%	72.40%
macro	X	29.46%	19.17%	68.18%	-	-
macro	WL	9.44%	3.43%	6.08%	6.26%	15.98%
macro	NL	31.42%	1.13%	1.60%	6.54%	11.40%
macro	offensive	0.01%	0.00%	0.00%	0.14%	0.06%
macro	porn	5.30%	0.63%	0.00%	0.48%	0.36%
micro	C	53.52%	83.00%	50.58%	98.72%	92.66%
micro	X	32.25%	15.27%	47.10%	-	-
micro	WL	3.60%	1.04%	1.35%	0.52%	2.33%
micro	NL	10.53%	0.69%	0.94%	0.75%	5.01%
micro	offensive	0.00%	0.00%	0.00%	0.18%	0.03%
micro	porn	2.86%	0.33%	0.00%	1.63%	0.08%
	#langs =0% C	7	0	1	7	0
	#langs <50% C	44	4	19	11	9
	#langs >50% NL	13	0	0	7	1
	#langs >50% WL	1	0	0	3	4

Results



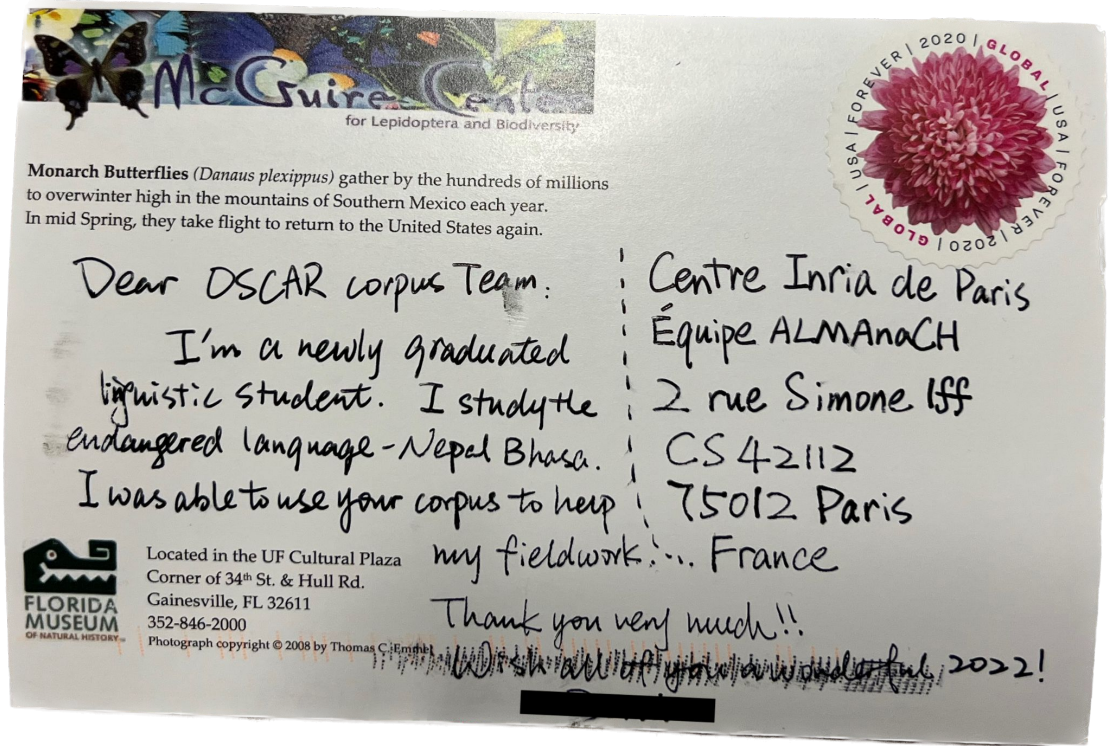
(a) Monolingual corpora



(b) Parallel corpora

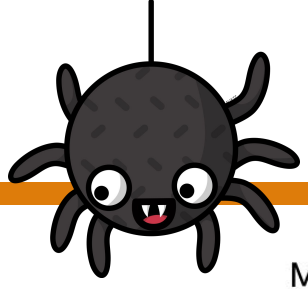
3.1 Nice Feedback from Users

Postcards!

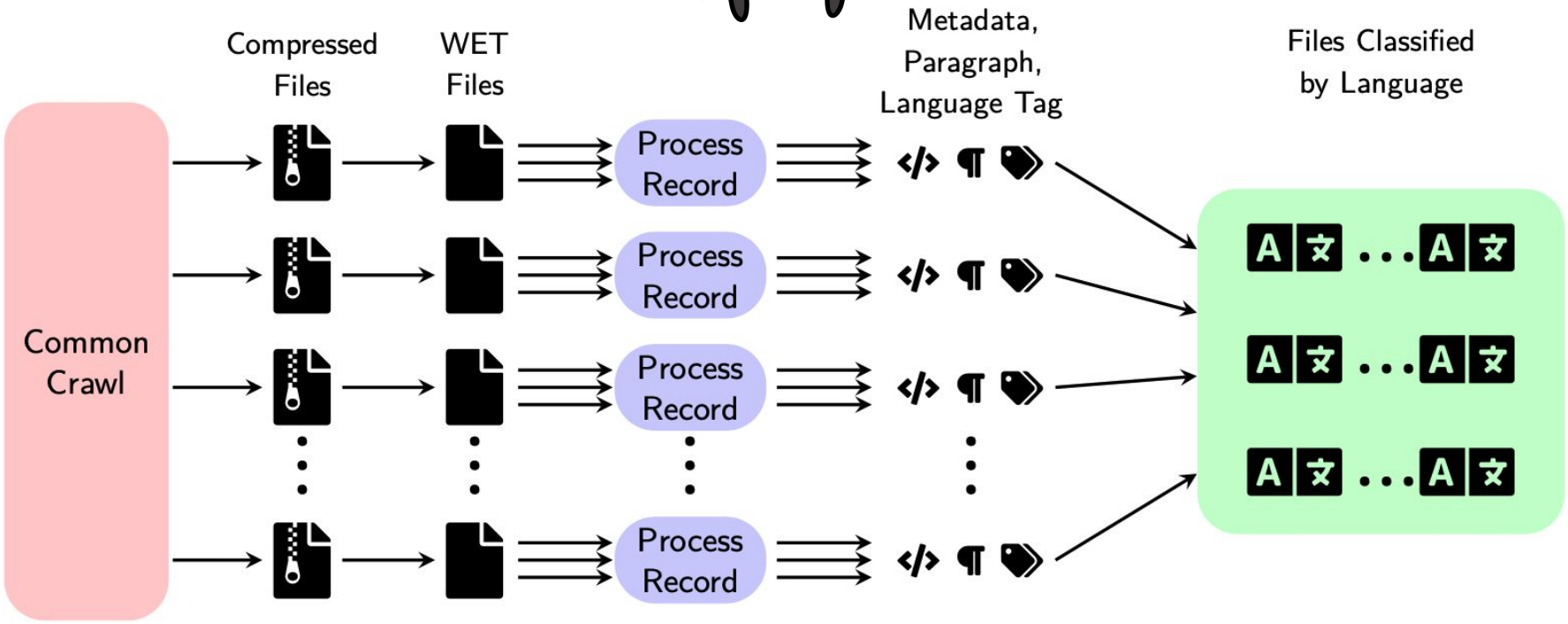

McGuire Center
for Lepidoptera and Biodiversity
Monarch Butterflies (Danaus plexippus) gather by the hundreds of millions to overwinter high in the mountains of Southern Mexico each year. In mid Spring, they take flight to return to the United States again.
USA | FOREVER | 2020 | GLOBAL
Dear OSCAR corpus Team:
I'm a newly graduated linguistic student. I study the endangered language - Nepal Bhasa. I was able to use your corpus to help my fieldwork. Thank you very much!!
Centre Inria de Paris
Équipe ALMAnaCH
2 rue Simone Iff
CS 42112
75012 Paris
France
Located in the UF Cultural Plaza
Corner of 34th St. & Hull Rd.
Gainesville, FL 32611
352-846-2000
FLORIDA MUSEUM OF NATURAL HISTORY
Photograph copyright © 2008 by Thomas C. Emmel

4. Ungoliant: A New Pipeline, two New OSCARS

Ungoliant

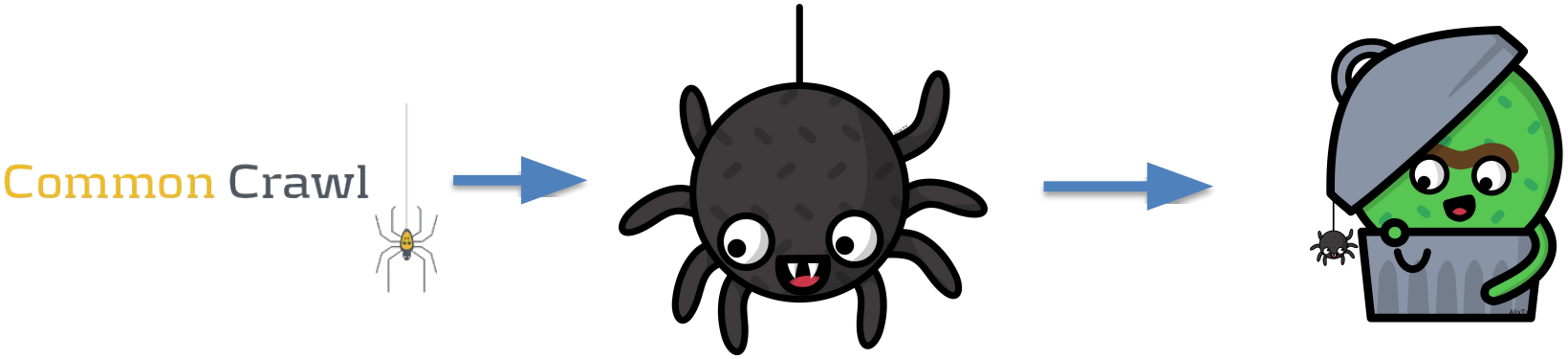


Deutsches Forschungszentrum für Künstliche Intelligenz GmbH



OSCAR 21.09 and OSCAR 22.01

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH



The OSCAR Project

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- OSCAR 21.09
  - Dump of 02.21
  - Added Metadata
- OSCAR 22.01
  - Dump of 12.21
  - Moved to Document Oriented Language Classification
  - Added Annotations

- **tiny**: The document has a low (<5) number of lines.
- **short_sentences**: The document has a high number (>50%) of short lines (<400 bytes)
- **header**: The document has a high number of short lines at its head, suggesting the presence of low quality content.
- **footer**: The document has a high number of short lines at its tail, suggesting the presence of low quality content.
- **noisy**: The document has a high percentage of punctuation (>50%)
- **adult**: The document contains adult content. This annotation uses a blocklist and labels a tiny part of the corpus: It does not catch most of the adult content.

5. Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning

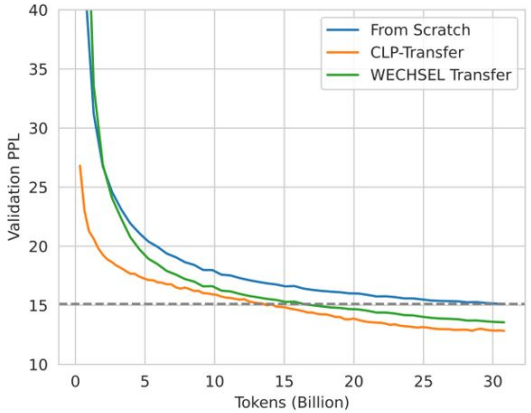
Cross-lingual & Progressive Transfer Learning

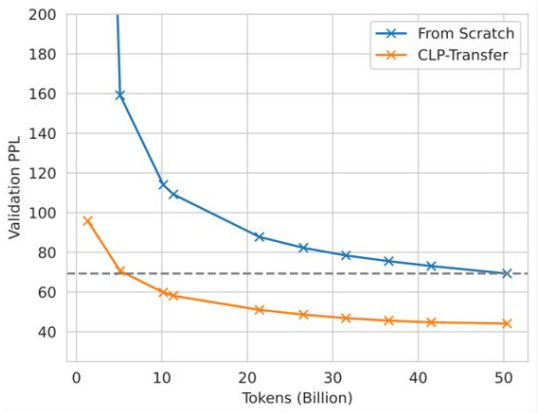
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- The “**Open**” in OpenGPT-X stands also for accessibility in terms of data and compute requirements for training LLMs.
- More LLMs are made publicly available (BLOOM, OPT, ...) that we can exploit to train our own LLM **more resource-efficiently**.
- Goal: Train a *large* model in a *target* language (e.g., a large German model).
- **CLP transfer learning.** Instead of training a model from scratch with randomly initialized weights, we recycle weights from pretrained models:
  - Cross-lingual: Transfer a *large* model in a *source* language (e.g., English) to our *target* language.
  - Progressive: Transfer a *small* model in our *target* language to the *large* model size (can be trained with fewer resources or is publicly available).

Cross-lingual & Progressive Transfer Learning

- We train two monolingual German language with CLP:
  - GPT2-XL (1.5B parameters, English)
  - BLOOM (7.1B parameters, multilingual - no German)
- CLP outperforms sole cross-lingual transfer (WECHSEL) and reduces the training effort compared to from scratch by up to **80% for BLOOM** (50% for GPT2-XL).





Cross-lingual & Progressive Transfer Learning

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Pretrained models: hf.co/malteos/bloom-6b4-clp-german

Source code: github.com/malteos/clp-transfer

Model demo: opengptx.dfki.de

Arxiv preprint: arxiv.org/abs/2301.09626

6. What's Next for OSCAR?

6.1. How to filter Unwanted data

KenLM Models and Other Filtering Strategies

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- CCNet introduced perplexity-based filtration with **KenLM models** (Kneser-Ney Language Model). However these models where trained on Wikipedia, which is **not very representative**.
- These exact same models were used in other initiatives such as BigScience, confined with lists of **flagged words** (which can give a high number of false positives).
- Other strategies include using **closed class words lists**, but spelling is not necessarily consistent in some languages.

6.2. What About KenLM Models Used Differently

Training KenLM Models with Adult-tagged data

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Approach	Model	Macro F1 on Test	Model Prediction on 1GB OSCAR	
			Prediction Speed (s)	% of Harmful Content Predicted
1st approach	distilBERT	91%	23529.01	78.7%
	FastText	89%	41.66	74.8%
2nd approach	FastText	91%	44.26	65.4%
3rd approach	Perplexity_4.22	94%	~50	0.49%
3rd approach	Perplexity_5.31	98%	~50	0.79%
3rd approach	Perplexity_13.51	99%	~50	1.01%

Table 3: Models performance (1st approach: 1 GB OSCAR Data; 2nd & 3rd approach: Mixed Data from OSCAR & Other Sources)

Training KenLM Models with Adult-tagged data

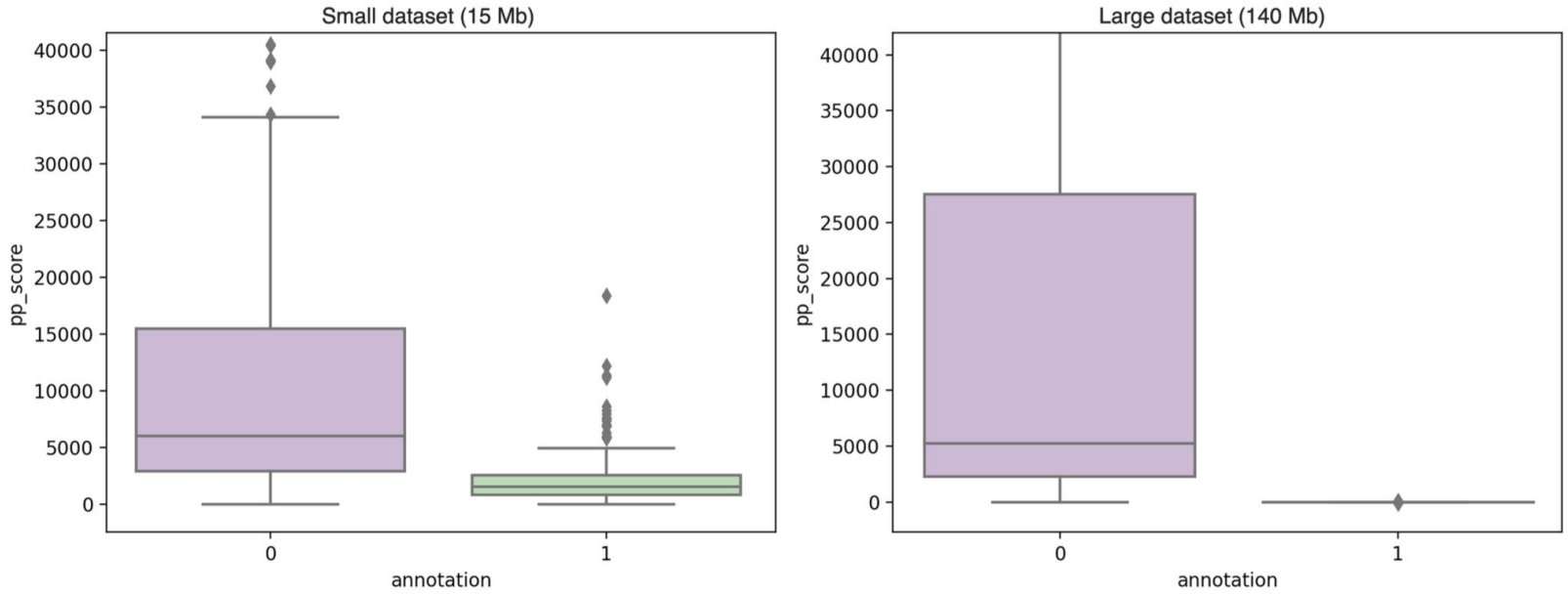


Figure 1: Perplexity Score Distribution

Training KenLM Models with Adult-tagged data

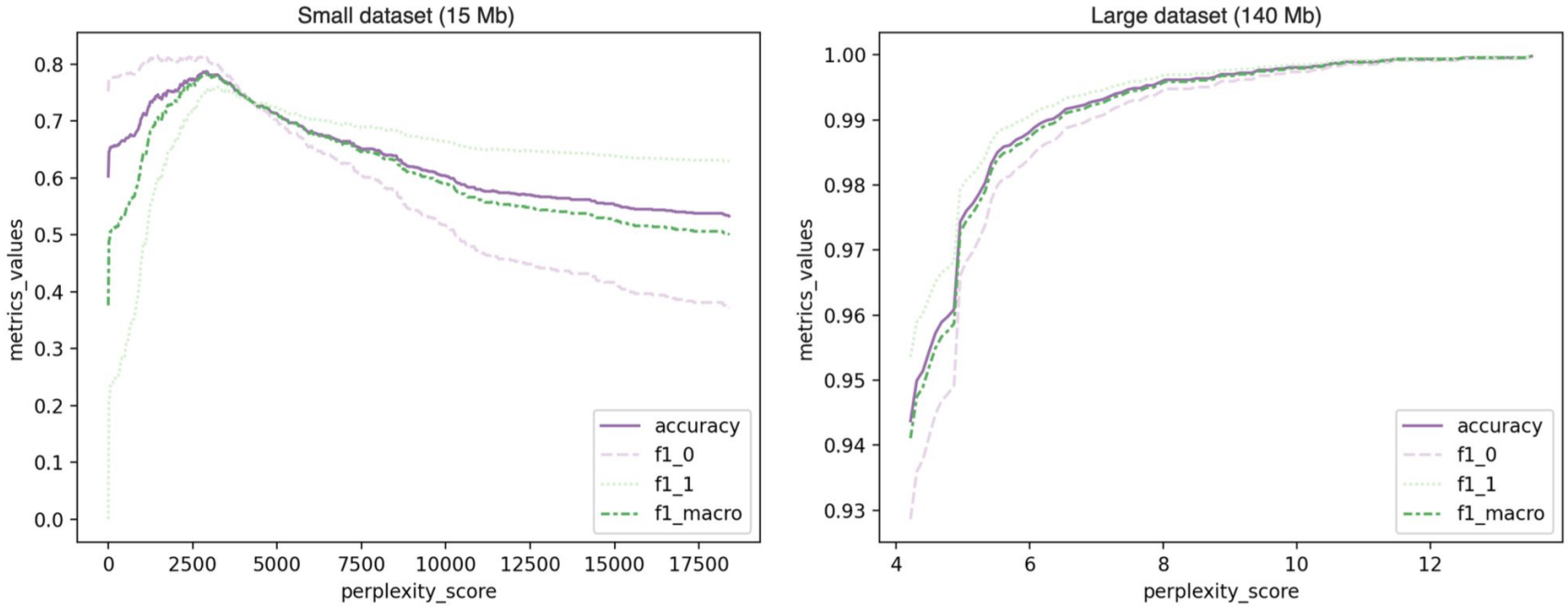
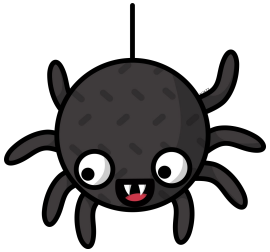
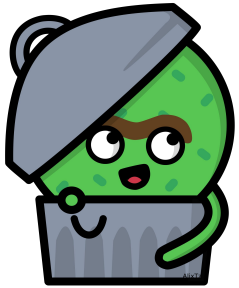


Figure 2: Classification Performance by Threshold on the Validation Set

6.3. OSCAR 23.01

OSCAR 23.01





- More fine-grained categories for annotations (No KenLM).
- Language targeted and human curated block-lists
- Locality-sensitive hashing for removing near document duplicates.
- Perplexity-based adult content annotations for at least 73 languages.
- Dual Month Common Crawl Dump.
- https://oscar-prive.huma-num.fr/2301/
- Email: contact@oscar-project.org

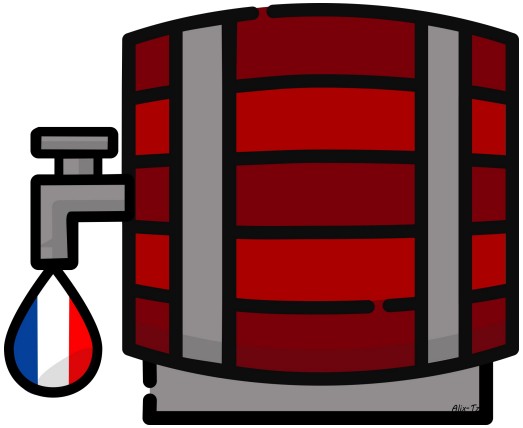
7. Should I Train my Model on OSCAR alone?

NO

I mean…
Probably not

7.1 Comparing OSCAR to a Reference Corpus

CaBeRnet : A Balanced Corpus for Contemporary French



CaBeRnet Sub-set	Tokens	Unique Forms	TTR
Oral	122 864 888	291 744	0.0024
Popular	131 444 017	458 521	0.0035
News	132 708 943	462 971	0.0035
Fiction	198 343 802	983 195	0.0050
Academic	126 431 211	1 433 663	0.0113
Total	711 792 861	2 558 513	0.0036

Following Biber's definition, *"**representativeness** refers to the extent to which a sample includes the full range of variability in a population"* (Biber, 1993, 244)

CaBeRnet in Parsing

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Model	GSD			Sequoia			Spoken			ParTUT		
	UPOS	UAS	LAS	UPOS	UAS	LAS	UPOS	UAS	LAS	UPOS	UAS	LAS
Baseline UDPipe Future	97.63	90.65	88.06	98.79	92.37	90.73	95.91	82.90	77.53	96.93	92.17	89.63
+ELMo$_{CBT}$	97.49	90.21	87.37	98.40	92.18	90.56	96.60	85.05	79.82	97.27	92.55	90.44
+ELMo$_{Wikipedia}$	97.92	92.13	89.77	99.22	94.28	92.97	97.28	85.61	80.79	**97.62**	94.01	91.78
+ELMo$_{CaBeRnet}$	97.87	92.02	89.62	99.33	94.42	93.14	***97.30***	85.39	80.63	97.43	94.02	91.86
+ELMo$_{OSCAR}$	97.85	92.41	90.05	99.30	94.43	93.25	97.10	85.83	**80.94**	97.47	**94.74**	**92.55**
+ELMo$_{OSCAR+CaBeRnet}$	**97.98**	**92.57**	**90.22**	**99.34**	**94.51**	**93.38**	97.24	**85.91**	80.93	97.58	94.47	92.05
State-of-the-art												
UDify	97.83	93.60	91.45	97.89	92.53	90.05	96.23	85.24	80.01	96.12	90.55	88.06
UDPipe Future + mBERT	97.98	92.55	90.31	*99.32*	94.88	93.81	97.23	*86.27*	*81.40*	*97.64*	94.51	92.47
CamemBERT	*98.19*	*94.82*	*92.47*	99.21	*95.56*	*94.39*	96.68	86.05	80.07	97.63	95.21	*92.90*

7.2. What happens with specific domain data?

D'AlemBERT and CamemBERT

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Original							Normalised or Contemporary						
Model	16	17	18	19	20	Avg	Model	16	17	18	19	20	Avg
Drama							*Drama*						
Pie Extended	*90.34*	*94.47*	*94.64*	-	-	*93.15*	Pie Extended	*93.69*	*95.75*	*95.61*	*95.03*	*93.71*	*94.76*
CamemBERT	87.06	89.01	90.92	-	-	89.00	CamemBERT	90.18	91.51	91.37	91.13	91.42	91.12
D'AlemBERT	**94.17**	**96.59**	**96.28**	-	-	**95.68**	D'AlemBERT	**96.25**	**96.97**	**96.80**	**96.25**	**95.00**	**96.25**
Varia							*Varia*						
Pie Extended	*89.85*	*93.44*	*95.98*	-	-	*93.09*	Pie Extended	*92.52*	*94.81*	*95.98*	*92.24*	*94.03*	*93.94*
CamemBERT	86.90	88.85	92.85	-	-	89.53	CamemBERT	89.79	90.69	93.06	90.54	89.78	93.94
D'AlemBERT	**93.86**	**95.73**	**96.95**	-	-	**95.51**	D'AlemBERT	**94.52**	**96.64**	**96.88**	**94.90**	**95.30**	**95.65**
Both							*Both*						
Pie Extended	*90.08*	*93.95*	*95.33*	-	-	*93.12*	Pie Extended	*93.08*	*95.28*	*95.80*	*93.65*	*93.87*	*94.35*
CamemBERT	86.98	88.93	91.89	-	-	89.27	CamemBERT	89.99	91.10	92.22	90.84	90.60	92.53
D'AlemBERT	**94.02**	**96.16**	**96.62**	-	-	**95.60**	D'AlemBERT	**95.39**	**96.81**	**96.84**	**95.58**	**95.15**	**95.95**

8. The Future of OSCAR

8.1 A First Glimpse into Language Identification

Heterogeneity is Hard!

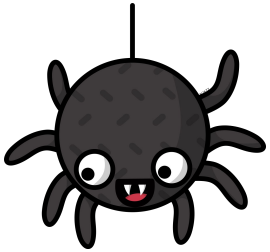
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

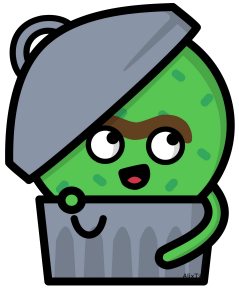
#LID classifier	dataset	corr.guesses	ratio	time
cld2-polyglot	dsl2014	11077/12600	0.879127	0.688s
cld2-python	dsl2014	11072/12600	0.87873	0.595s
cld3-python	dsl2014	10454/12600	0.829683	3.968s
fasttext -m lid.176.bin	dsl2014	11084/12600	0.879683	0.391s
fasttext -m lid.176.bin -l	dsl2014	11023/12600	0.874841	0.358s
fasttext -m lid.176.ftz	dsl2014	10678/12600	0.84746	0.570s
fasttext -m lid.218a.bin	dsl2014	11161/12600	0.885794	2.815s
fasttext -m lid.218a.bin -l	dsl2014	11186/12600	0.887778	2.800s
fasttext -m lid.218a.ftz	dsl2014	11154/12600	0.885238	7.031s
cld2-polyglot	europarl	20775/21000	0.989286	0.695s
cld2-python	europarl	20779/21000	0.989476	0.589s
cld3-python	europarl	20811/21000	0.991	4.300s
fasttext -m lid.176.bin	europarl	20804/21000	0.990667	0.410s
fasttext -m lid.176.bin -l	europarl	20805/21000	0.990714	0.408s
fasttext -m lid.176.ftz	europarl	20681/21000	0.98481	0.537s
fasttext -m lid.218a.bin	europarl	20966/21000	0.998381	3.225s
fasttext -m lid.218a.bin -l	europarl	20963/21000	0.998238	3.175s
fasttext -m lid.218a.ftz	europarl	20969/21000	0.998524	7.578s

#LID classifier	dataset	corr.guesses	ratio	time
cld2-polyglot	tatoeba	15415/19457	0.79226	0.451s
cld2-python	tatoeba	15417/19457	0.792363	0.312s
cld3-python	tatoeba	12250/19457	0.629593	2.096s
fasttext -m lid.176.bin -l	tatoeba	15108/19457	0.776481	0.149s
fasttext -m lid.176.bin	tatoeba	15317/19457	0.787223	0.169s
fasttext -m lid.176.ftz	tatoeba	13631/19457	0.70057	0.168s
fasttext -m lid.218a.bin -l	tatoeba	15697/19457	0.806753	1.663s
fasttext -m lid.218a.bin	tatoeba	15741/19457	0.809015	1.702s
fasttext -m lid.218a.ftz	tatoeba	15777/19457	0.810865	2.536s
cld2-polyglot	twitter	140240/187461	0.748102	5.972s
cld2-python	twitter	140858/187461	0.751399	3.752s
cld3-python	twitter	111628/187461	0.595473	26.457s
fasttext -m lid.176.bin -l	twitter	144679/187461	0.771782	2.279s
fasttext -m lid.176.bin	twitter	143463/187461	0.765295	2.372s
fasttext -m lid.176.ftz	twitter	139219/187461	0.742656	2.593s
fasttext -m lid.218a.bin -l	twitter	127447/187461	0.679859	19.502s
fasttext -m lid.218a.bin	twitter	121086/187461	0.645926	19.751s
fasttext -m lid.218a.ftz	twitter	120192/187461	0.641157	39.719s

8.2 Ideas for the Future

The Future of OSCAR





- Use multiple Common Crawl Dumps.
- Use HTML sources instead of text-extracts.
- Start a multimodal, multilingual OSCAR.
- Find new partners for data distribution.
- Add parallel corpora and code.
- Implement new data distribution formats.
- Work on data reconstruction.
- Work on data exploration tools.
- Find alternative sources.
- Contribute to Common Crawl.
- Work on community management and projects.
- Define a data-quality framework.
- Work on personal data identification and anonymization.

8.2 The OSCAR Community

A Growing Community

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Inria

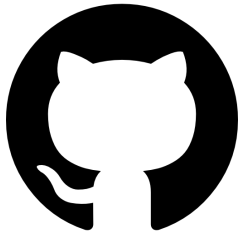
Common Crawl

IN OMNIBUS VERITAS
UNIVERSITÄT MANNHEIM

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Help us Improve OSCAR

- https://oscar-project.org/

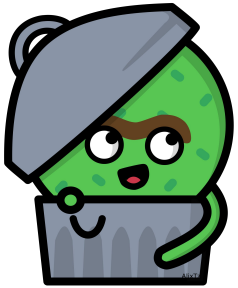


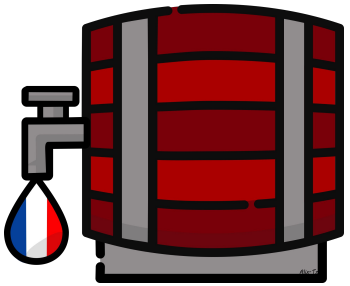
- https://github.com/oscar-project

- https://discord.com/invite/4JNg9FTar4

Thank You!

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH





Special thanks to Alix Chagué (https://alix-tz.github.io/en/index.html) for All The logos of my projects

Questions?