Shared Tasks at WMT 2021: Multilingual Low-Resource Translation for Indo-European Languages

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session V
23rd November 2021

Session V

Outline

1. WMT 2021, General View
   - The News Task
   - (Human) Evaluation
   - Best Performing Systems
2. Multilingual LR Translation for Indo-European Languages
   - Setting and Organisation
   - (Human) Evaluation
   - Best Performing Systems
3. Large-Scale Multilingual Machine Translation
   - Settings and Evaluation
   - DeltaLM

WMT 2021, General View

WMT, 15 Years of MT Evaluation Campaigns

EMNLP 2021
SIXTH CONFERENCE ON
MACHINE TRANSLATION (WMT21)

November 10-11, 2021
Punta Cana (Dominican Republic) and Online

Home

[HOME] [SCHEDULE] [RESULTS]
TRANSLATION TASKS: [NEWS] [SIMILAR LANGUAGES] [BIOMEDICAL] [EUROPEAN LOW RES MULTILINGUAL] [LARGE-SCALE MULTILINGUAL] [TRIANGULAR MT] [EFFICIENCY] [TERMINOLOGY] [UNSUP AND VERY LOW RES] [LIFELONG LEARNING]
EVALUATION TASKS: [QUALITY ESTIMATION] [METRICS]
OTHER TASKS: [AUTOMATIC POST-EDITING]

This conference builds on a series of annual workshops and conferences on statistical machine translation, going back to 2006:

- the NAACL-2006 Workshop on Statistical Machine Translation,
- the ACL-2007 Workshop on Statistical Machine Translation,
- the ACL-2008 Workshop on Statistical Machine Translation,
- the EACL-2009 Workshop on Statistical Machine Translation,
- the ACL-2010 Workshop on Statistical Machine Translation
- the EMNLP-2011 Workshop on Statistical Machine Translation,
- the NAACL-2012 Workshop on Statistical Machine Translation,
- the ACL-2013 Workshop on Statistical Machine Translation,
- the ACL-2014 Workshop on Statistical Machine Translation,
- the EMNLP-2015 Workshop on Statistical Machine Translation,
- the First Conference on Machine Translation (at ACL-2016),
- the Second Conference on Machine Translation (at EMNLP-2017),
- the Third Conference on Machine Translation (at EMNLP-2018).
- the Fourth Conference on Machine Translation (at ACL-2019).
- the Sixth Conference on Machine Translation (at EMNLP-2020).

NAACL 2006 WORKSHOP
ON STATISTICAL MACHINE TRANSLATION

Shared Task: Exploiting Parallel Texts for Statistical Machine Translation

June 8 and 9, 2006, in conjunction with NAACL 2006 in New York City

[HOME] | [PROGRAM] | [PROCEEDINGS] | [SHARED TASK] | [BASELINE SYSTEM] | [RESULTS]

The shared task of the workshop is to build a probabilistic phrase translation table for phrase-based statistical machine translation (SMT). Evaluation is translation quality on an unseen test set. We provide a parallel corpus as training data (with word alignment), a baseline statistical machine translation system, and additional resources. Participants may augment this system or use their own system.

Goals

The goals of staging this shared task are:

- get reference performance numbers in a large-scale translation task for European languages
- pose special challenges with word order (German-English) and translating from English into foreign languages
- offer interested parties a (relatively) smooth start with hands-on experience in state-of-the-art statistical machine translation methods
- create publicly available data for machine translation and machine translation evaluation

We hope that both beginners and established research groups will participate in this task.

Task Description

We provide training data for three European language pairs, and a common framework (including a language model and a basline system). The task is to improve methods to build a phrase translation table (e.g. by better word alignment, phrase extraction, phrase scoring), augment the system otherwiese (e.g. by preprocessing), or build entirely new translation systems.

The participants' system is used to translate a test set of unseen sentences in the source language. The translation quality is measured by the BLEU score, which measures overlap with a reference translation, and manual evaluation. Participants agree to contribute to the manual evaluation about eight hours of work.

To have a common framework that allows for comparable results, and also to lower the barrier to entry, we provide

- a fixed training set

WMT 2021, General View

Lots of Shared Tasks Related to LR-MT!

Shared Task	Papers	LR?
News	22	▶
Quality Estimation	12	▼
Metrics	8	▼
Terminology	8	▶
Similar Languages	7	▶
Multilingual Large Scale	7	▲
Unsupervised LR-MT	6	▲
Multilingual LR-MT	5	▲
Triangular MT	5	▶
Biomedical	5	▼
Efficiency	4	▼
APE	2	▼
Test Suites	1	▼

The News Task

Language Pairs Involved:

- English to/from Chinese
- English to/from Czech
- English to/from German
- English to/from Hausa
- English to/from Icelandic
- English to/from Japanese
- English to/from Russian
- French to/from German
- Hindi to/from Bengali
- Zulu to/from Xhosa

The News Task

Language Pairs Involved:

- English to/from Chinese
- English to/from Czech
- English to/from German
- **English to/from Hausa**
- English to/from Icelandic
- English to/from Japanese
- English to/from Russian
- French to/from German
- Hindi to/from Bengali
- **Zulu to/from Xhosa**

The News Task

Evaluation Campaign

- A truly large scale evaluation campaign (manual only)
- Reference-based direct assessments (into-English, crowdsourced)
  - 589 turker accounts (1,078 did not pass quality control!)
    488,396 translation assessment scores
- Source direct assessments + Contrastive (out-of-English)
  - 303,627 assessments + 64,031 assessments from constrastive
  - different conclusions!
- Document level introduce in the last editions
- Accurate quality control of crowdsourcing
- 1000-2000 assessment/system

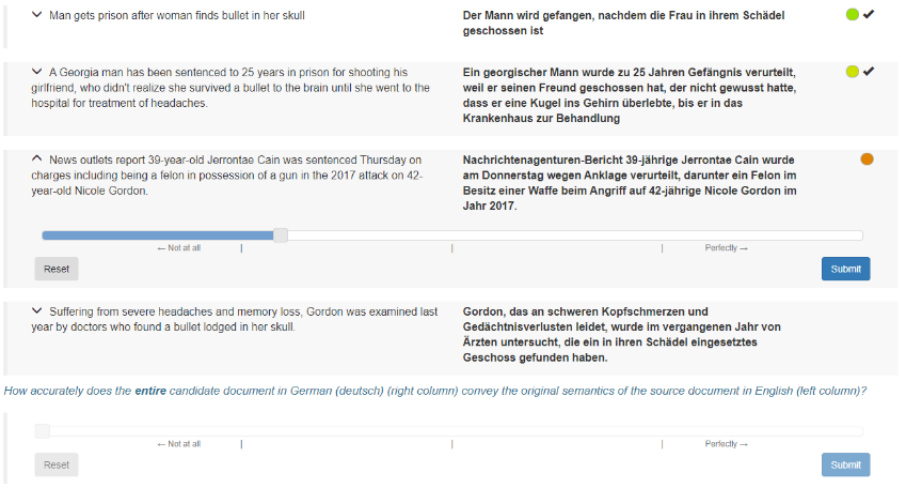
The News Task

Evaluation Campaign

- A truly large scale evaluation campaign (manual only)
- Reference-based direct assessments (into-English, crowdsourced)
  - 589 turker accounts (1,078 did not pass quality control!)
    488,396 translation assessment scores
- Source direct assessments + Contrastive (out-of-English)
  - 303,627 assessments + 64,031 assessments from constrastive
  - different conclusions!
- Document level introduce in the last editions
- Accurate quality control of crowdsourcing
- 1000-2000 assessment/system

The News Task

Evaluation Campaign (modified image from the Findings paper)



The News Task

Out-of-English Evaluation

Findings paper, out-of-English tables (page 19)

Findings paper, into-English tables (page 15)

The News Task

Some Selected Systems

The News Task

FACEBOOK-AI (Tran et al., 2021)

- Multilingual systems any-to-English, and English-to-any (128k BPE)
- large scale backtranslation
  1.3B parallel sentences + BT: 2.8B 2en; 1.0B en2
- in-domain finetuning
- ensembling
- noisy channel re-ranking
- scaling dense transformer (up to 4.7B parameters)
- sparse mixture of experts (up to 52B parameters)

The News Task

e-TRANSLATION (Oravecz et al., 2021)

- focus on data selection and filtering: heuristic rules and with a LM built from NewsCrawl data
- additional tagged, back-translated data from news corpora
  en–de: 32M parallel sentences + 226M for BT; fr–de: 14M + 15M for BT; cs–en: 45M + 88M for BT;
- vocabulary en–de: 36k BPE; fr–de: 30k BPE; cs–en: 36k BPE;
- parallel data is upsampled to a 1:1 ratio
- finetuning on a top subset of original parallel data ranked by the monolingual news LM and then fine-tuned further on previous years' test sets

The News Task

UEDIN (Chen et al., 2021; Pal et al., 2021)

BN-HI pre-trained on back-translated data, and fine-tuned on parallel data

- fine-tuning on in-domain data (*n*-gram matching, TF-IDF sim, and LM scoring with the validation set)
- ensemble of finetuned transformers

EN-DE rule-based and dual conditional cross-entropy filtering

- trained parallel and backtranslated data, and further trained on parallel sentences only
- finetuned on previous WMT sets and ensembled

The News Task

Shallow Summary

- High quality translation for most of the pairs
- No new architectures
- Clever ways to increase the capacity of transformers
- Clever ways to clean and augment the data
- **Ok, but mostly no low-resourced...**

The News Task

Shallow Summary

- High quality translation for most of the pairs
- No new architectures
- Clever ways to increase the capacity of transformers
- Clever ways to clean and augment the data
- **Ok, but mostly no low-resourced...**

Session V

Outline

1. WMT 2021, General View
   - The News Task
   - (Human) Evaluation
   - Best Performing Systems
2. Multilingual LR Translation for Indo-European Languages
   - Setting and Organisation
   - (Human) Evaluation
   - Best Performing Systems
3. Large-Scale Multilingual Machine Translation
   - Settings and Evaluation
   - DeltaLM

Multilingual LR Translation for Indo-European Languages

The Task

European Language Resource Coordination
Connecting Europe Facility

LT Bridge

Generalitat de Catalunya
Departament de Cultura

Thank you to all the DFKI people involved at any of the levels!

Kay, Leonie, Josef, Andrea, Corinna, Stephan, Eileen, Stefania...

Multilingual LR Translation for Indo-European Languages

The Subtasks

europeana

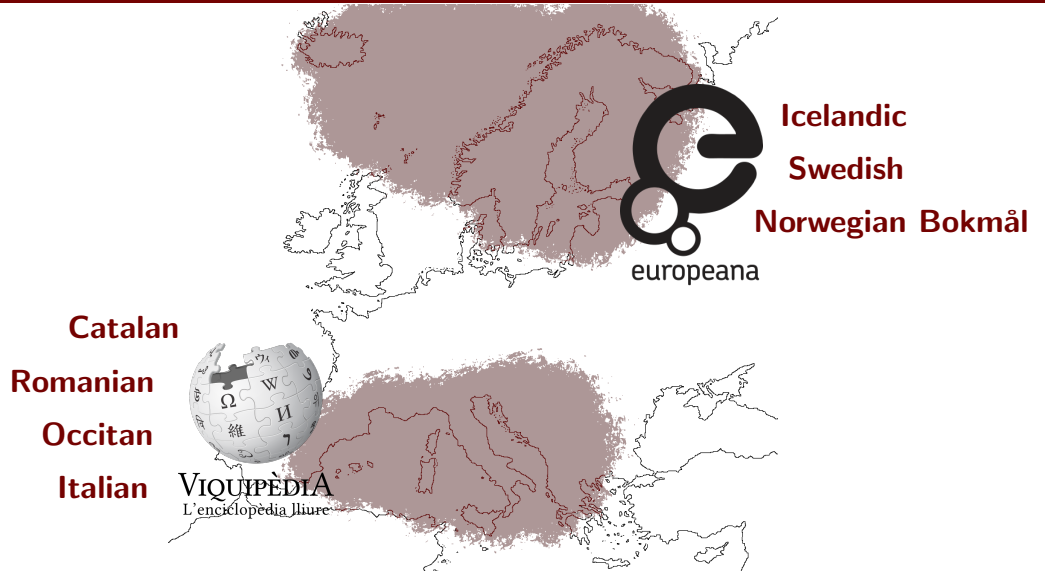
VIQUIPÈDIA
L'enciclopèdia lliure

Multilingual LR Translation for Indo-European Languages

Two Subtasks, two Indo-European Families

Multilingual LR Translation for Indo-European Languages

Two Subtasks, two Indo-European Families



Multilingual LR Translation for Indo-European Languages

Task 1: Translation of Descriptions of Archaeological Sites (is→{sv,nb})

europeana collections

Collections Explore Exhibitions Blog

Fornminjar

Landslag

Lýsing: Lækjar er fyrst getið um miðja 14. öld en fornleifar benda til að þar hafi verið búið allt frá fyrstu tíð. Búskap var hætt á jörðinni 1958 og er jörðin nú hluti af landi Ennis og hún nytjuð þaðan (Byggðasaga Skagafjarðar V. Bindi. Bls. 269-270). Engar byggingar eru lengur uppistandandi en bæjarhóllinn er enn vel greinanlegur. Túngarður sést að hluta vestan og norðan bæjarhólsins og eru tvær tóftir við vesturhluta hans. Um 20-50m beint norður af bæjarhólnum eru óglöggar, fornlegar minjar, skálatóft og fleiri byggingaleifar. Borkjarnar voru teknir í tóftir og túngarð auk þess sem könnunarskurður var grafinn í skálatóftina. Nánari lýsing: Lækur í Viðvíkursveit. Réttartóft austan skálatóftarinnar. Horft er til suðurs.

Multilingual LR Translation for Indo-European Languages

Task 1: Translation of Stories (nb→{is,sv})

europeana collections

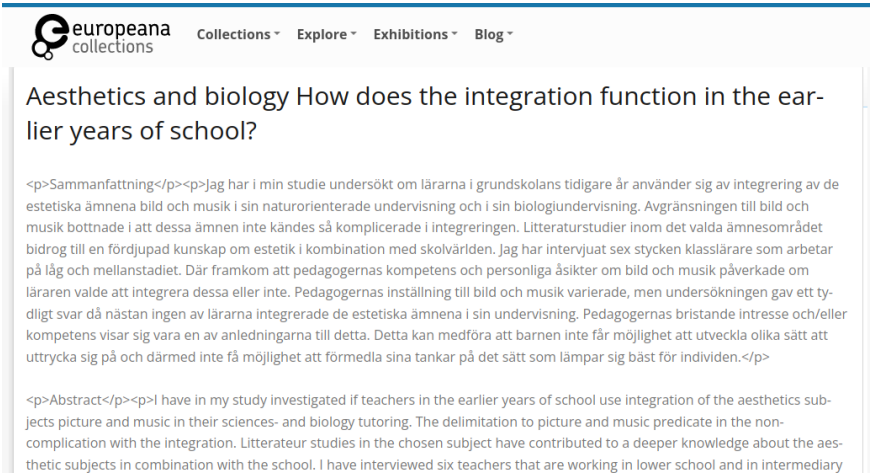
Collections Explore Exhibitions Blog

St. Hansfeiring ved Fjellbrudammen

St. Hans-feiring ved Fjellbrudammen har en lang tradisjon. Når dette ble en samlingplass midtsommerkvelden vet vi ikke, men fra mellomkrigstida og fram til rundt 1960 var det årviss feiring her. Da ble dammen gjerdet inn, fordi den hørte med til byens drikkevannskilde. Under krigen ble tradisjonen brutt, men behørig tatt opp igjen i 1945.
 <p>I løpet av de 40 årene årene som er gått siden feiringa tok slutt, vokste og grodde det til, slik at den grønne vollen er blitt helt ugjenkjennelig. Vemodig for oss som hadde opplevd mange trivelige St.Hans-aftener på Fjellbrua. På denne plassen var det buskapen til Moldegård samlet seg før de ble tatt ned til sommerfjøset for å bli melket. Sommerfjøset lå ned på Fjøsbakken vest for Vardevegen, like før der bommen står i dag. I forbindelse med markaplanen så ble klausuleringsbestemmelsene endret. Gjerdet rundt Fjellbrudammen ble tatt vekk og nå var muligheten til å gjenskape noe av det som var og tilrettelegge for bruk. Med stor dugnadsinnsats og i samarbeid med Molde kommune

Multilingual LR Translation for Indo-European Languages

Task 1: Translation of Theses' Abstracts (sv→{is,nb})



Multilingual LR Translation for Indo-European Languages

Task 1: Datasets Statistics

	Validation				Test			
	Docs.	**Sents.**	**SrcToks**	**TgtToks**	**Docs.**	**Sents.**	**SrcToks**	**TgtToks**
is2nb	26	467	6,096	6,932	24	563	8,256	9,301
is2sv	26	467	6,096	6,611	24	563	8,256	8,819
nb2is	19	502	7,673	7,495	16	540	9,218	8,867
nb2sv	19	502	7,673	7,499	16	540	9,218	8,804
sv2is	43	516	9,097	9,524	44	547	9,642	9,733
sv2nb	43	516	9,097	9,232	44	547	9,642	9,787

- collected around 1,000 sentences per language, different domain per lang.
- translated professional by translators (2 rounds)
- 6 translation directions

Multilingual LR Translation for Indo-European Languages

Task 2: Translation of Articles (ca)

https://ca.wikipedia.org/wiki/Calçot

Sense sessió iniciada Discussió per aquest IP Contribucions Crea un compte Inicia la sessió

Pàgina Discussió Mostra Modifica Mostra l'historial Més Cerca a Viquipèdia

VIQUIPÈDIA
L'enciclopèdia lliure

Portada
Article a l'atzar
Articles de qualitat

Comunitat
Portal viquipedista
Agenda d'actes
Canvis recents
La taverna
Contacte
Xat
Donatius
Ajuda

Eines
Què hi enllaça
Canvis relacionats

Calçot

Els **calçots** o **ceballots** són una varietat de cebes tendres, cebes poc bulboses i més suaus que es calcen a mesura que creixen, i que es mengen escalivats.


Calçots preparats

Sobre l'origen del calçot hi ha diverses versions però la més coneguda és la que atribueix a en Xat de Benaiges, un camperol que va viure a Valls a la fi del segle XIX, la invenció d'aquest cultiu.[1] En Xat de Benaiges va posar un parell de brots de cebes al foc i va descobrir per atzar un plat que a la primeria del segle XX ja havia esdevingut habitual en moltes llars de Valls.

De llavors ençà el consum dels calçots o calçotada ha esdevingut una festa gastronòmica coneguda arreu, especialment a Catalunya. Els calçots es mengen habitualment durant els mesos de gener, febrer i març, es couen amb llenya provinent de sarments, rabasses o rabassons de vinya i es consumeixen acompanyats d'una salsa típica, la dita *salvitxada*, que és una salsa de tipus romesco.

No fou fins a mitjan segle passat que el calçot es va fer conèixer fora de l'Alt Camp, particularment gràcies la divulgació feta per la *colla artística i humorística l'Olla*. La festa del calçot de Valls, que se celebra el darrer

Multilingual LR Translation for Indo-European Languages

Task 2: Datasets Statistics

	Validation				Test			
	Docs.	**Sents.**	**SrcToks**	**TgtToks**	**Docs.**	**Sents.**	**SrcToks**	**TgtToks**
ca2it	41	1,269	30,363	29,725	42	1,743	38,868	37,649
ca2oc	41	1,269	30,363	30,184	42	1,743	38,868	38,662
ca2ro	41	1,269	30,363	29,842	42	1,743	38,868	37,379

- collected around 3,000 sentences in Catalan
- translated professional by translators (2 rounds)
- 3 translation directions

Multilingual LR Translation for Indo-European Languages

Shared Task Challenges

C1 Multilinguality

C2 Limited data but related languages

C3 Specific vocabulary (cultural heritage, NEs)

C4 Document-level translation

Multilingual LR Translation for Indo-European Languages

Training Corpora, Slightly Constrained

We want SotA system, but still comparable among them. So, we allowed

- Parallel corpora (sentence and doc aligned)
- Monolingual corpora
- (Multiligual) pre-trained embeddings
- Wordnets with open license, BabelNet
- Multilingual home-made lexicons from Wikimedia

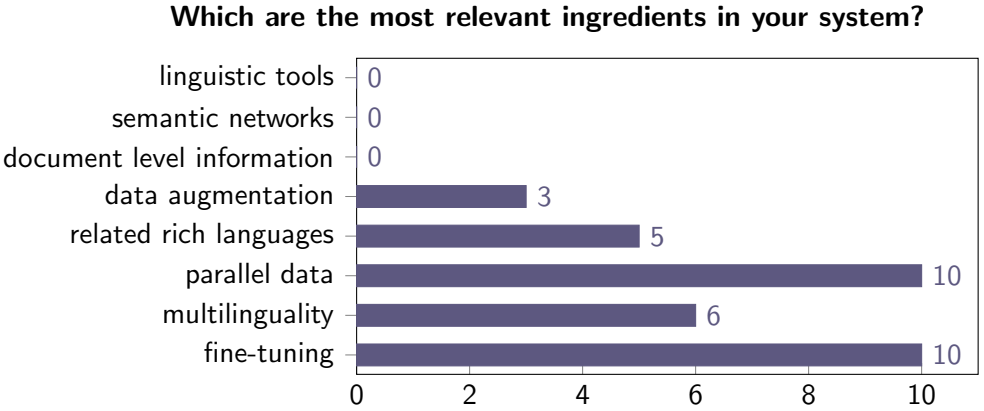
Multilingual LR Translation for Indo-European Languages

Training Corpora: Home-made Multiligual Lexicons

	Wikidata		Wikipedia		Wiktionary
	all	**cleaner**	**all**	**cleaner**	**all**
is2nb/nb2is	1,141,891	–	–	–	3,304/6,552
is2sv/sv2is	1,149,894	–	–	–	15,369/17,321
nb2sv/sv2nb	2,648,493	–	–	–	9,390/7,124
is-nb-sv	1,139,493	23,574	–	–	–
ca2it/it2ca	3,072,380	–	323,055	–	18,684/19,050
ca2oc/oc2ca	1,300,979	–	71,854	–	3,999/3,538
ca2ro/ro2ca	1,608,860	–	123,215	–	11,990/12,034
it2oc/oc2it	1,285,771	–	75,542	–	7,225/6,332
it2ro/ro2it	4,547,649	–	215,296	–	20,898/20,442
ro2oc/oc2ro	1,230,752	–	64,800	–	4,586/4,350
ca-it-ro	1,579,345	123,543	117,543	97,484	–

Multilingual LR Translation for Indo-European Languages

What did Participants Use?



Multilingual LR Translation for Indo-European Languages

Which Shared Task Challenges did the Participants Tackle?

C1 **Multilinguality**

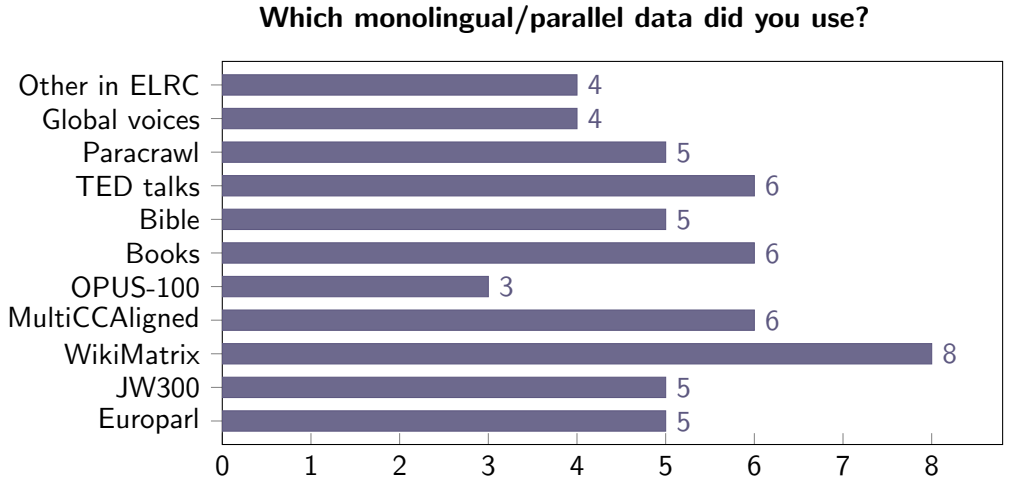
C2 **Limited data but related languages**

C3 Specific vocabulary (cultural heritage, NEs)

C4 Document-level translation

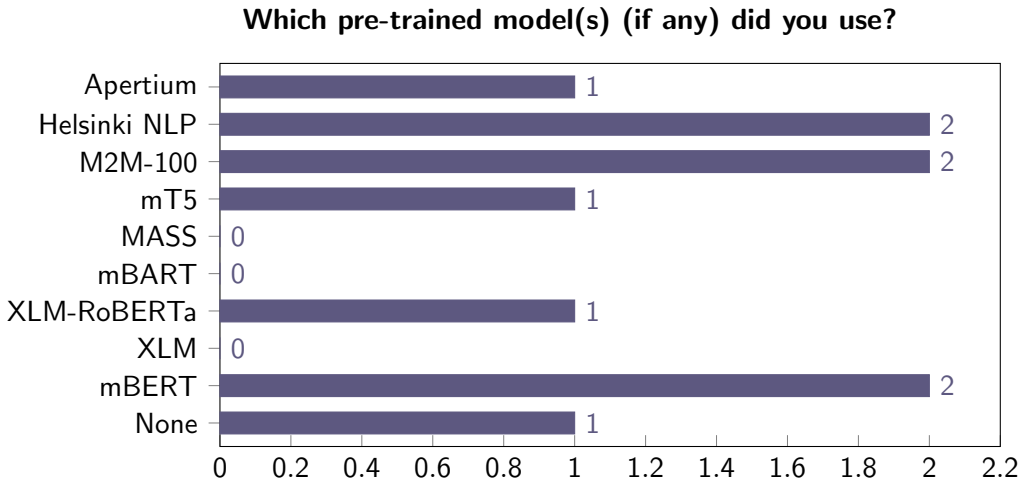
Multilingual LR Translation for Indo-European Languages

Corpora



Multilingual LR Translation for Indo-European Languages

Pre-Trained Models



Multilingual LR Translation for Indo-European Languages

Performance & Evaluation

- 11 submissions divided between the 2 tasks
- 2 baselines (M2M-100, mT5 finetuned)
- Automatic evaluation as average ranking of 5 metrics (official!)
  BLEU, TER, chrF, BertScore and COMET

Multilingual LR Translation for Indo-European Languages

Performance & Evaluation

- Automatic evaluation as average ranking of 5 metrics (official!)
  BLEU, TER, chrF, BertScore and COMET
- Manual evaluation (14 raters) of selected pairs: direct assessments with document context (DAs) using Appraise
  - reference DA for Swedish (nb2sv and is2sv)
  - source DA for ca2it and ca2oc
  - term accuracy for ca2it and ca2oc (source DA)
- High correlation between automatic and manual evaluation

Multilingual LR Translation for Indo-European Languages

Performance & Evaluation

- Automatic evaluation as average ranking of 5 metrics (official!)
  BLEU, TER, chrF, BertScore and COMET
- Manual evaluation (14 raters) of selected pairs: direct assessments with document context (DAs) using Appraise
  - reference DA for Swedish (nb2sv and is2sv)
  - source DA for ca2it and ca2oc
  - term accuracy for ca2it and ca2oc (source DA)
- High correlation between automatic and manual evaluation

Multilingual LR Translation for Indo-European Languages

Automatic Evaluation, Task 1: North Germanic Languages

	Average Ranking	BLEU	TER	chrF	COMET	BertScore
M2M-100 (baseline)	1.0±0.0	31.5	0.54	0.55	0.399	0.862
EdinSaar-Contrastive	2.2±0.4	27.1	0.57	0.54	0.283	0.856
EdinSaar-Primary	2.8±0.4	27.5	0.58	0.52	0.276	0.849
UBCNLP-Primary	4.0±0.0	24.9	0.60	0.50	0.076	0.847
UBCNLP-Contrastive	5.0±0.0	24.0	0.61	0.49	-0.068	0.837
mT5-devFinetuned (baseline)	6.0±0.0	18.5	0.78	0.42	-0.102	0.810

- High agreement between metrics
- Congrats M2M (Facebook)!

Multilingual LR Translation for Indo-European Languages

Automatic Evaluation, Task 1: North Germanic Languages

	Average Ranking	BLEU	TER	chrF	COMET	BertScore
M2M-100 (baseline)	1.0±0.0	31.5	0.54	0.55	0.399	0.862
EdinSaar-Contrastive	2.2±0.4	27.1	0.57	0.54	0.283	0.856
EdinSaar-Primary	2.8±0.4	27.5	0.58	0.52	0.276	0.849
UBCNLP-Primary	4.0±0.0	24.9	0.60	0.50	0.076	0.847
UBCNLP-Contrastive	5.0±0.0	24.0	0.61	0.49	-0.068	0.837
mT5-devFinetuned (baseline)	6.0±0.0	18.5	0.78	0.42	-0.102	0.810

- High agreement between metrics
- Congrats M2M (Facebook)!

Multilingual LR Translation for Indo-European Languages

Automatic Evaluation, Task 2: Romance Languages

	Average Ranking	BLEU	TER	chrF	COMET	BertScore
CUNI-Primary	1.2±0.4	50.1	0.401	0.694	0.566	0.901
CUNI-Contrastive	1.6±0.5	49.5	0.404	0.693	0.569	0.901
TenTrans-Contrastive	3.0±0.0	43.5	0.460	0.670	0.444	0.894
TenTrans-Primary	3.8±0.4	43.3	0.462	0.668	0.442	0.894
BSC-Primary	5.0±0.7	41.3	0.402	0.647	0.363	0.884
M2M-100 (baseline)	5.8±0.4	40.0	0.478	0.634	0.414	0.878
UBCNLP-Primary	7.2±0.4	35.4	0.528	0.588	0.007	0.854
mT5-devFinetuned (baseline)	8.0±0.7	29.3	0.592	0.553	0.059	0.850
UBCNLP-Contrastive	8.6±0.5	28.5	0.591	0.529	-0.374	0.825

- Congrats CUNI!

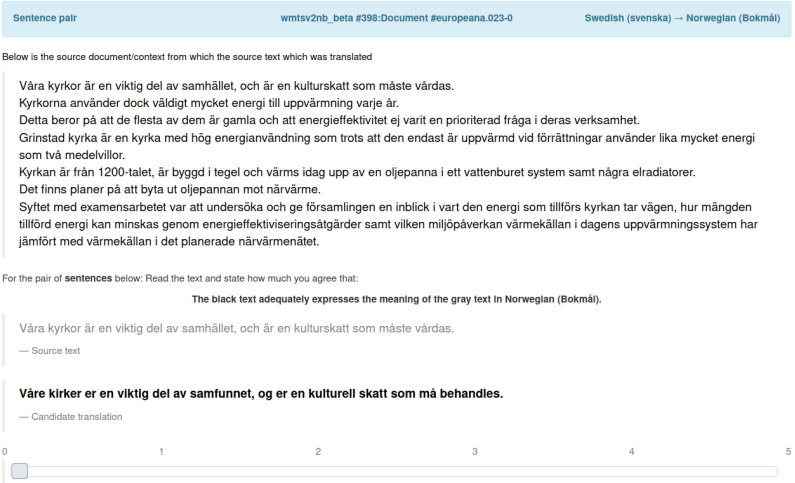
Multilingual LR Translation for Indo-European Languages

Manual Evaluation

- Similar but diffent to the News Evaluation Campaign
- We have less annotators, still need statistical significance in the results
- Manual evaluation (14 raters) of selected pairs: direct assessments with document context (DAs) on 100 sentences using Appraise
  - reference DA for Swedish (nb2sv and is2sv)
  - source DA for ca2it and ca2oc
  - term accuracy for ca2it and ca2oc (source DA)

Multilingual LR Translation for Indo-European Languages

Manual Evaluation: Customising Appraise



Multilingual LR Translation for Indo-European Languages

Manual Evaluation, Task 1: North Germanic Languages

	nb2sv		**is2sv**	
System	***z*-score**	**raw**	***z*-score**	**raw**
M2M-100	0.7±0.6	4.2±0.8	0.1±1.0	2.0±1.1
EdinSaar	0.2±0.7	3.6±1.1	-0.1±0.8	1.9±1.0
UBCNLP	0.2±0.8	3.5±1.2	-0.4±1.0	1.6±1.1
mT5-dFT	-1.2±0.7	1.5±1.1	0.4±1.1	2.4±1.2

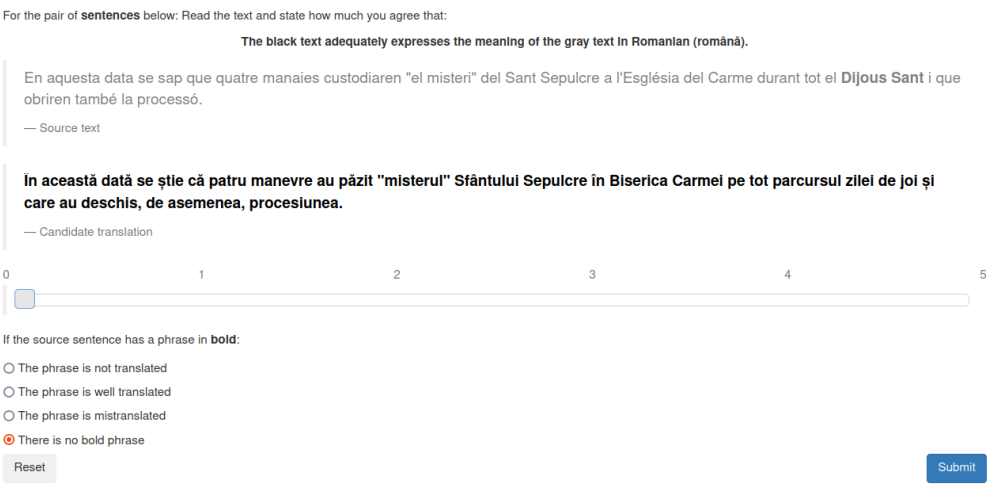
Multilingual LR Translation for Indo-European Languages

Manual Evaluation, Task 2: Romance Languages

	ca2it		ca2oc	
System	*z*-score	raw	*z*-score	raw
HUMAN	0.8±0.4	4.8±0.6	0.8±0.7	4.0±1.0
CUNI	0.5±0.7	4.4±0.9	0.5±0.8	3.6±1.1
M2M-100	0.4±0.7	4.2±1.0	-0.7±0.8	2.0±1.0
TenTrans	0.0±0.8	3.8±1.1	0.3±0.8	3.4±1.2
BSC	-0.1±0.8	3.7±1.1	0.3±0.9	3.4±1.2
UBCNLP	-0.5±1.0	3.1±1.3	0.0±0.9	3.0±1.2
mT5-dFT	-1.2±0.9	2.3±1.2	-1.0±0.7	1.7±0.9

Multilingual LR Translation for Indo-European Languages

Manual Evaluation: Customising Appraise for Term Accuracy



Multilingual LR Translation for Indo-European Languages

Term Accuracy

What's a term?

Plaça del Mercadal, segle XV, segle XIX i XX, la Casa Pinyol, Festes de Maig, Rambla de Badalona, la Cremada, la Segona República, Josep Maria Cuyàs, Baró de Maldà, 11 de maig de 1940, Francesc de Paula Giró i Prat, Aristeus antennatus, Productes de l'Empordà, 400 metres, mitjan segle XX, Canyó de Palamós, Confraria de Pescadors de Palamós, finals del segle XIX, Xat de Benaiges, començaments del segle XX, "salvitxada", la calçotada, Alt Camp, Congrés de Cultura Catalana, Valls, Concurs de salsa de la "calçotada", Fogueres de Sant Antoni, Nadal, Sant Antoni, Química Orgànica, Universitat de Barcelona, Junta d'Energia Nuclear, Universitat de Chicago, Universitat de València, Física Teòrica, Mecànica Teòrica, Premi d'Investigació Ramón y Cajal, Manaies de Girona, any 1751, Dijous Sant, Setmana Santa, segles xviii i xix, 1851, mitjans de segle XIX, finals del XVIII, port del Masnou, dos quilòmetres i mig, Club Nàutic del Masnou, Creu Roja, festival Ple de Riure, Masnou, N-II, Premià de Mar, any 2019, platja d'Ocata, Michelin, Ferran Adrià, El Celler de Can Roca, Can Fabes

Multilingual LR Translation for Indo-European Languages

Term Accuracy

- We sum the votes from all the raters per class and count the majority class, ties are discarded

	ca2it				ca2oc			
System	**well**	**mis**	**no**	Σ	**well**	**mis**	**no**	Σ
HUMAN	53	0	3	56	40	0	2	42
CUNI	39	3	5	47	30	7	1	38
M2M-100	33	2	6	41	26	9	0	35
TenTrans	37	0	9	46	32	4	1	37
BSC	27	7	5	39	33	4	0	37
UBCNLP	29	16	1	46	19	1	0	20
mT5-dFT	20	17	10	47	25	11	4	40

Multilingual LR Translation for Indo-European Languages

Term Accuracy: Observations

- The baselines (no in-domain training) have the largest number of mistranslations
- Translation quality 2it $>$ 2oc, but more mistranslations in 2it
  - sub-unit segmentation strategy?
- Multi-word named entities where one of the words has been literally translated and the others have not
- A number (specially centuries) is translated by another one

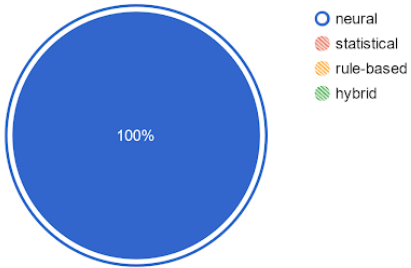
**Till now we only know the name of the systems,
but what are they addressing?**

Multilingual LR Translation for Indo-European Languages

Systems Characteristics

The architecture is best described as

10 responses

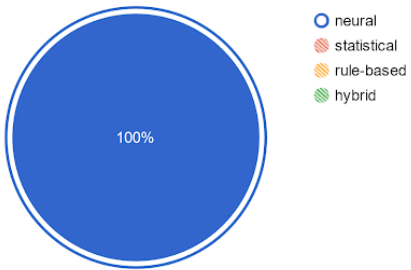


Multilingual LR Translation for Indo-European Languages

Systems Characteristics

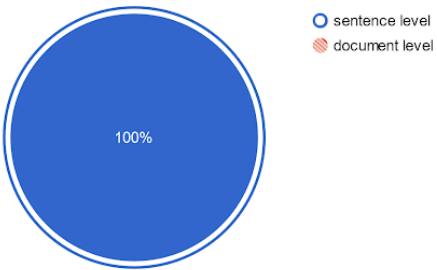
The architecture is best described as

10 responses



The system is best described as

10 responses

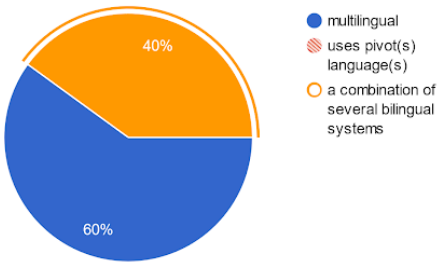


Multilingual LR Translation for Indo-European Languages

Systems Characteristics

The system is best described as

10 responses

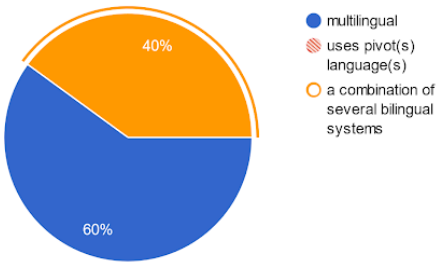


Multilingual LR Translation for Indo-European Languages

Systems Characteristics

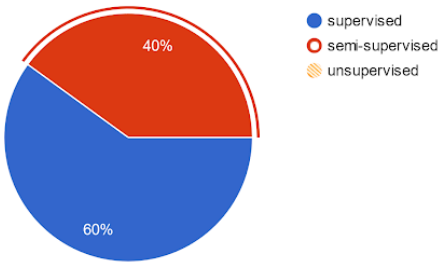
The system is best described as

10 responses



The system is best described as

10 responses

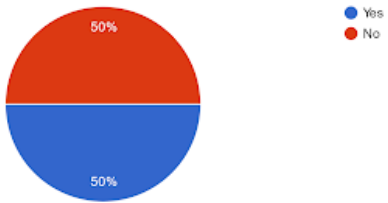


Multilingual LR Translation for Indo-European Languages

Systems Characteristics

Did you use data from the allowed rich languages (Danish, German or English for the Germanic family and Spanish, French, Portuguese or English for the Romance one)?

10 responses



Multilingual LR Translation for Indo-European Languages

Some Selected Systems

Tencent 腾讯

Multilingual LR Translation for Indo-European Languages

CUNI (Jon et al., 2021)

- Multilingual supervised machine translation model (primary) enriched with backtranslated data (contrastive)
- 41 M original parallel sentences including all language pairs in the task plus French and English
- Exploration of various subword granularities
- Phonemic representation of texts added via multi-task learning
- Character-level rescoring on the translations *n*-best lists for Catalan–Occitan

Multilingual LR Translation for Indo-European Languages

CUNI (Jon et al., 2021)

- Multilingual supervised machine translation model (**primary**) enriched with backtranslated data (**contrastive**)
- 41 M original parallel sentences including all language pairs in the task plus French and English
- Exploration of various subword granularities
- Phonemic representation of texts added via multi-task learning
- Character-level rescoring on the translations *n*-best lists for Catalan–Occitan

Multilingual LR Translation for Indo-European Languages

TenTrans (Yang et al., 2021)

- 8-to-4 multilingual model with Catalan–Italian–Romanian–Occitan as the target side and Spanish, French, Portuguese and English on the source side.
- In-domain finetuning (data selected using a domain classifier trained with multilingual BERT)
- Knowledge transfer: knowledge distillation of the M2M 1.2B model previously finetuned on the languages of the task
- Primary: ensemble of the in-domain multilingual and the distilled M2M
- Contrastive: adds a multilingual base model enriched with backtranslations to the ensemble and pivot-based methods to augment the training corpus

Multilingual LR Translation for Indo-European Languages

TenTrans (Yang et al., 2021)

- 8-to-4 multilingual model with Catalan–Italian–Romanian–Occitan as the target side and Spanish, French, Portuguese and English on the source side.
- In-domain finetuning (data selected using a domain classifier trained with multilingual BERT)
- Knowledge transfer: knowledge distillation of the M2M 1.2B model previously finetuned on the languages of the task
- Primary: ensemble of the in-domain multilingual and the distilled M2M
- **Contrastive**: adds a multilingual base model enriched with backtranslations to the ensemble and pivot-based methods to augment the training corpus

Multilingual LR Translation for Indo-European Languages

Conclusions

- Systems used direct neural translation, multilingual or bilingual, no translations done through a pivot language
- Multilingual systems trained with additional corpora with the related rich languages as source gave the best performance
- Data augmentation via backtranslations has been beneficial for all the systems
- Few improvements by selecting data close to the domain of the validation set, but the in-domain adaptation was not decisive to win the shared task
- Rankings would change if one only considers the most distant language within a family (Romanian and Icelandic)

Session V

Outline

1. WMT 2021, General View
   - The News Task
   - (Human) Evaluation
   - Best Performing Systems
2. Multilingual LR Translation for Indo-European Languages
   - Setting and Organisation
   - (Human) Evaluation
   - Best Performing Systems
3. Large-Scale Multilingual Machine Translation
   - Settings and Evaluation
   - DeltaLM

Large-Scale Multilingual Machine Translation

Track Details

Small Track #1: 5 Central/East European languages, 30 directions: Croatian, Hungarian, Estonian, Serbian, Macedonian, English

Small Track #2: 5 South East Asian languages, 30 directions: Javanese, Indonesian, Malay, Tagalog, Tamil, English

Large Track: All Languages, to and from English

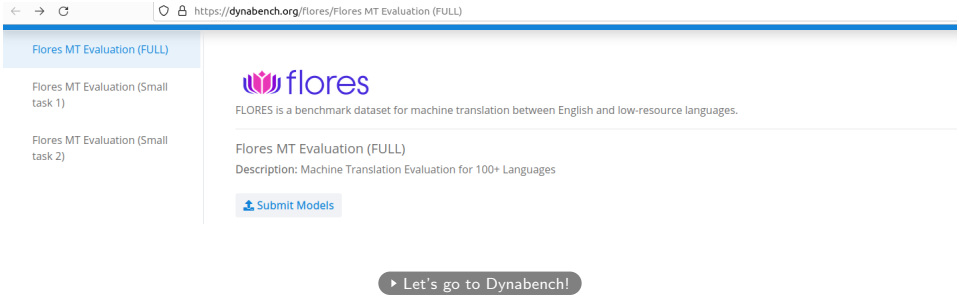
Large-Scale Multilingual Machine Translation

Large Track Languages

- Afrikaans
- Amharic
- Arabic
- Armenian
- Assamese
- Asturian
- Azerbaijani
- Belarusian
- Bengali
- Bosnian
- Bulgarian
- Burmese
- Catalan
- Cebuano
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Filipino (Tagalog)
- Finnish
- French
- Fula
- Galician
- Ganda
- Georgian
- German
- Greek
- Gujarati
- Hausa
- Hebrew
- Hindi
- Hungarian
- Icelandic
- Igbo
- Indonesian
- Irish
- Italian
- Japanese
- Javanese
- Kabuverdianu
- Kamba
- Kannada
- Kazakh
- Khmer
- Korean
- Kyrgyz
- Lao
- Latvian
- Lingala
- Lithuanian
- Luo
- Luxembourgish
- Macedonian
- Malay
- Malayalam
- Maltese
- Maori
- Marathi
- Mongolian
- Nepali
- Northern Sotho
- Norwegian
- Nyanja
- Occitan
- Oriya
- Oromo
- Pashto
- Persian
- Polish
- Portuguese
- Punjabi
- Romanian
- Russian
- Serbian
- Shona
- Sindhi
- Slovak
- Slovenian
- Somali
- Sorani Kurdish
- Spanish
- Swahili
- Swedish
- Tajik
- Tamil
- Telugu
- Thai
- Turkish
- Ukrainian
- Umbundu
- Urdu
- Uzbek
- Vietnamese
- Welsh
- Wolof
- Xhosa
- Yoruba
- Zulu

Large-Scale Multilingual Machine Translation

Dynabench Evaluation Platform



Large-Scale Multilingual Machine Translation

High-Quality Translations

LANGUAGE-PAIR LEADERBOARD

Dataset

Source Language	Target Language	Model	BLEU Score
Afrikaans (afr)	English (eng)	DeltaLM+Zcode	60.86
Welsh (cym)	English (eng)	DeltaLM+Zcode	60.05
English (eng)	Welsh (cym)	DeltaLM+Zcode	58.37
English (eng)	Maltese (mlt)	DeltaLM+Zcode	57.98
Maltese (mlt)	English (eng)	DeltaLM+Zcode	57.96
Swedish (swe)	English (eng)	DeltaLM+Zcode	52.63
Danish (dan)	English (eng)	DeltaLM+Zcode	52.40
Portuguese (Brazil) (por)	English (eng)	DeltaLM+Zcode	51.29
Welsh (cym)	Maltese (mlt)	DeltaLM+Zcode	50.15
Afrikaans (afr)	Maltese (mlt)	DeltaLM+Zcode	49.74

Page 1 of 1010

Previous Next

Large-Scale Multilingual Machine Translation

Low-Quality Translations

LANGUAGE-PAIR LEADERBOARD

Dataset

Source Language	Target Language	Model	BLEU Score
Lingala (lin)	Fula (ful)	DeltaLM+Zcode	1.41
Burmese (mya)	Kabuverdianu (kea)	DeltaLM+Zcode	1.42
Thai (tha)	Umbundu (umb)	DeltaLM+Zcode	1.42
Igbo (ibo)	Fula (ful)	DeltaLM+Zcode	1.42
Umbundu (umb)	Khmer (khm)	m2m-124-175m	1.43
Galician (glg)	Fula (ful)	m2m-124-175m	1.43
Estonian (est)	Fula (ful)	DeltaLM+Zcode	1.43
Luo (luo)	Khmer (khm)	615m	1.43
Hebrew (heb)	Umbundu (umb)	DeltaLM+Zcode	1.43
Catalan (cat)	Fula (ful)	m2m-124-175m	1.44

Page 1 of 1010

Previous Next

Large-Scale Multilingual Machine Translation

Maltese

LANGUAGE-PAIR LEADERBOARD

Dataset

Source Language	Target Language	Model	BLEU Score
Maltese (mlt)	English (eng)	DeltaLM+Zcode	57.96
Maltese (mlt)	Welsh (cym)	DeltaLM+Zcode	48.49
Maltese (mlt)	Portuguese (Brazil) (por)	DeltaLM+Zcode	42.38
Maltese (mlt)	French (fra)	DeltaLM+Zcode	41.44
Maltese (mlt)	Danish (dan)	DeltaLM+Zcode	39.68
Maltese (mlt)	Indonesian (ind)	DeltaLM+Zcode	39.29
Maltese (mlt)	Swedish (swe)	DeltaLM+Zcode	39.01
Maltese (mlt)	Irish (gle)	DeltaLM+Zcode	38.95
Maltese (mlt)	Malay (msa)	DeltaLM+Zcode	38.85
Maltese (mlt)	Afrikaans (afr)	DeltaLM+Zcode	38.23

Large-Scale Multilingual Machine Translation

Irish

LANGUAGE-PAIR LEADERBOARD

Dataset

Source Language	Target Language	Model	BLEU Score
Irish (gle)	English (eng)	DeltaLM+Zcode	45.61
Irish (gle)	Maltese (mlt)	DeltaLM+Zcode	43.70
Irish (gle)	Welsh (cym)	DeltaLM+Zcode	42.22
Irish (gle)	Portuguese (Brazil) (por)	DeltaLM+Zcode	35.36
Irish (gle)	French (fra)	DeltaLM+Zcode	34.43
Irish (gle)	Danish (dan)	DeltaLM+Zcode	33.77
Irish (gle)	Indonesian (ind)	DeltaLM+Zcode	32.90
Irish (gle)	Swedish (swe)	DeltaLM+Zcode	32.89
Irish (gle)	Malay (msa)	DeltaLM+Zcode	32.84
Irish (gle)	Afrikaans (afr)	DeltaLM+Zcode	32.32

Page 1 of 10

Previous Next

Large-Scale Multilingual Machine Translation

Catalan

LANGUAGE-PAIR LEADERBOARD

Dataset

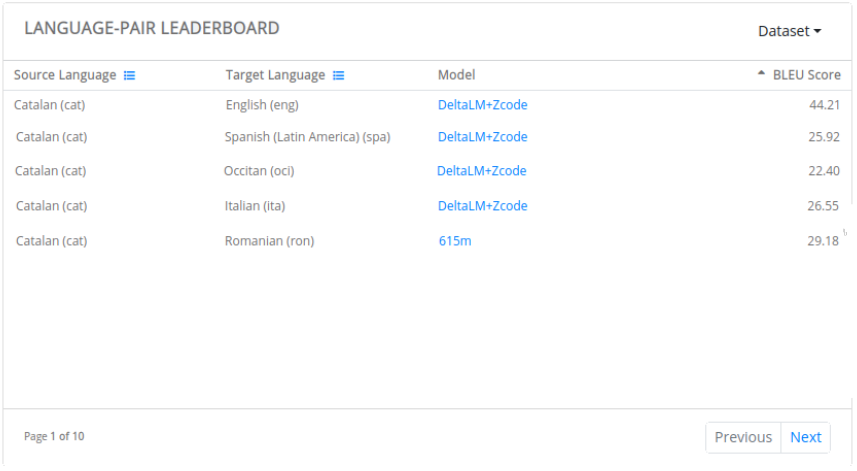
Source Language	Target Language	Model	BLEU Score
Catalan (cat)	English (eng)	DeltaLM+Zcode	44.21
Catalan (cat)	Maltese (mlt)	DeltaLM+Zcode	42.19
Catalan (cat)	Welsh (cym)	DeltaLM+Zcode	38.89
Catalan (cat)	French (fra)	DeltaLM+Zcode	36.48
Catalan (cat)	Portuguese (Brazil) (por)	DeltaLM+Zcode	36.15
Catalan (cat)	Danish (dan)	DeltaLM+Zcode	33.97
Catalan (cat)	Galician (glg)	615m	33.33
Catalan (cat)	Irish (gle)	DeltaLM+Zcode	33.07
Catalan (cat)	Swedish (swe)	DeltaLM+Zcode	32.80
Catalan (cat)	Bulgarian (bul)	DeltaLM+Zcode	32.04

Page 1 of 10

Previous Next

Large-Scale Multilingual Machine Translation

Catalan, because I can easily Interpret...



LANGUAGE-PAIR LEADERBOARD

Dataset

Source Language	Target Language	Model	BLEU Score
Catalan (cat)	English (eng)	DeltaLM+Zcode	44.21
Catalan (cat)	Spanish (Latin America) (spa)	DeltaLM+Zcode	25.92
Catalan (cat)	Occitan (oci)	DeltaLM+Zcode	22.40
Catalan (cat)	Italian (ita)	DeltaLM+Zcode	26.55
Catalan (cat)	Romanian (ron)	615m	29.18

Page 1 of 10

Previous Next

Large-Scale Multilingual Machine Translation

Catalan, because I can easily Interpret...

LANGUAGE-PAIR LEADERBOARD

Our task

Dataset

Source Language	Target Language	Model	M2M	CUNI	BLEU Score
Catalan (cat)	English (eng)	DeltaLM+Zcode			44.21
Catalan (cat)	Spanish (Latin America) (spa)	DeltaLM+Zcode			25.92
Catalan (cat)	Occitan (oci)	DeltaLM+Zcode	40.2	67.1	22.40
Catalan (cat)	Italian (ita)	DeltaLM+Zcode	46.6	49.5	26.55
Catalan (cat)	Romanian (ron)	615m	33.1	31.8	29.18

Page 1 of 10

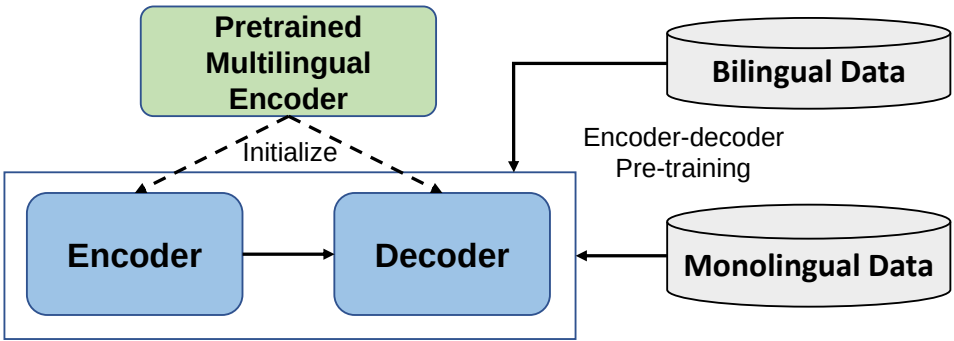
Previous Next

Main System Characteristics (from the findings paper)

- Combination of parallel, back-translated and noisy-parallel data
- Based on the pre-trained DeltaLM$_{LARGE}$ (next slides only if soon enough!)
- Mixture of direct and pivoted translation to improve the performance of individual directions
- Progressive learning: starts with a smaller architecture, noisier training data, and later changes to improve performance

Large-Scale Multilingual Machine Translation

DeltaLM: Basic Idea



DeltaLM

Basic Idea (Ma et al., 2021 —still in arXiv)

- "The decoder as the task layer of off-the-shelf pre-trained encoders"
- Encoder and the decoder are initialised with the pre-trained multilingual encoder
- Pre-train ΔLM with both monolingual data and bilingual data in a self-supervised way

DeltaLM

Basic Idea (Ma et al., 2021 —still in arXiv)

- "The decoder as the task layer of off-the-shelf pre-trained encoders"
- Encoder and the decoder are initialised with the pre-trained multilingual encoder
  - **How to initialise a decoder with an encoder??**
- Pre-train ΔLM with both monolingual data and bilingual data in a self-supervised way
  - **What's an appropriate pre-training task??**

DeltaLM

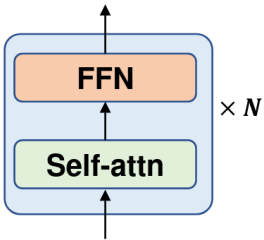
We could Use a Recap Section Now

1. Encoder and decoder modules in the transformer
   Session 3

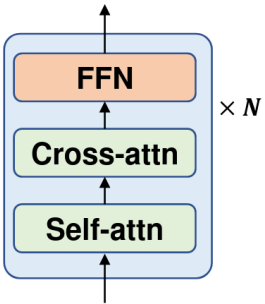
2. Pre-trained models for language generation
   (mBART example)
   Session 4

DeltaLM

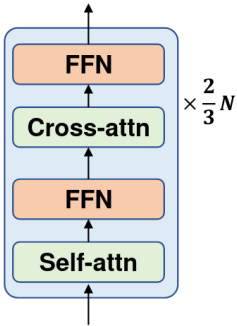
Interleaved Decoder



(a) Vanilla encoder



(b) Vanilla decoder



(c) Interleaved decoder

Any Encoder could be Used to Initialise, ΔLM Uses InfoXLM$_{BASE}$

InfoXLM (Chi et al., NAACL 2021)

- 12 layers and 768 hidden states
- Training with large-scale monolingual data and bilingual data
- Tasks: masked language model, translation language model, and cross-lingual contrast objectives
- Shared vocabulary of 250,000 tokens based on the SentencePiece
- By the way... InfoXLM is initialised with XLM-R (550M params)

DeltaLM

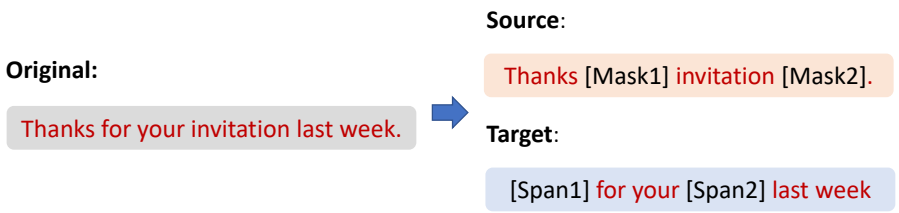
Architecture Characteristics

DeltaLM (Ma et al., 2021)

- 24 encoder layers, 12 interleaved decoder layers and 1024 hidden states (360M params)
- Training with large-scale monolingual data and bilingual data
- Tasks: span corruption and translation span corruption
- Shared vocabulary of 250,000 tokens based on the SentencePiece
- Initialised with InfoXLM which is initialised with XLM-R (550M params)

DeltaLM

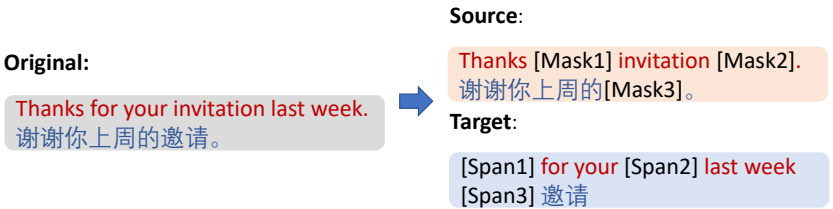
Pre-training Tasks: Span Corruption



- Introduced in mT5
- Data: large-scale multilingual corpora in 100 languages (6TB combination of CC100, CC-Net, and Wikipedia)

DeltaLM

Pre-training Tasks: Translation Span Corruption



- Introduced in mT6
- Data: concatenate two parallel sentences as the input for 77 languages (88GB of bilingual data from CCAligned and OPUS)

DeltaLM

A Comment on the Infrastructure

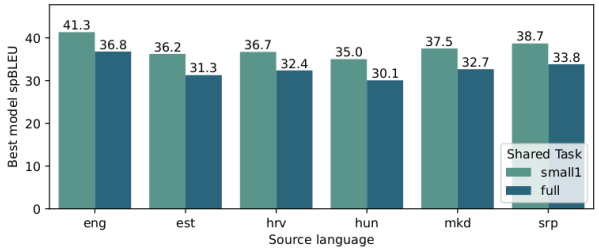
- Microsoft's submission trained on 64 NVIDIA V100 or 32 A100 GPUs
- It takes 1 week to train ΔLM with 32 V100 GPUs
- InfoXLM training
- 1.5 Million updates on 500 32GB Nvidia V100 GPUs for XML-R

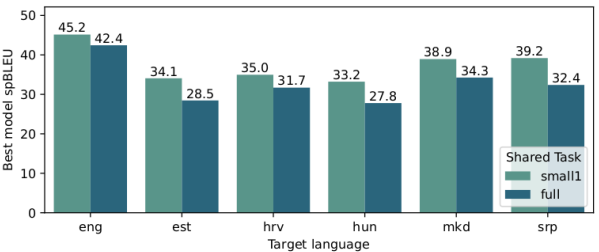
Main System Characteristics (from the findings paper)

- Combination of parallel, back-translated and noisy-parallel data
- Based on the pre-trained DeltaLM
- Mixture of direct and pivoted translation to improve the performance of individual directions
- Progressive learning: starts with a smaller architecture, noisier training data, and later changes to improve performance

Large-Scale Multilingual Machine Translation

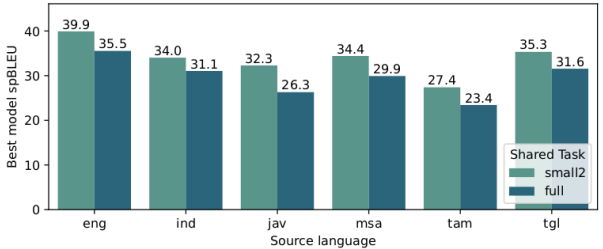
Comparison on spBLEU vs. the Degree of Multilingualism (FB slide)

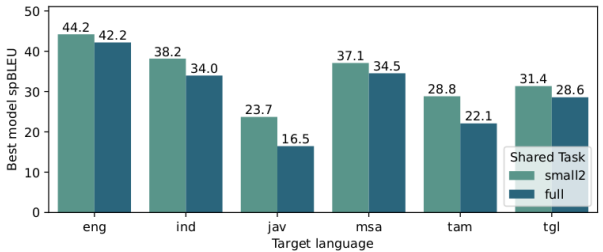




Large-Scale Multilingual Machine Translation

Comparison on spBLEU vs. the Degree of Multilingualism (FB slide)

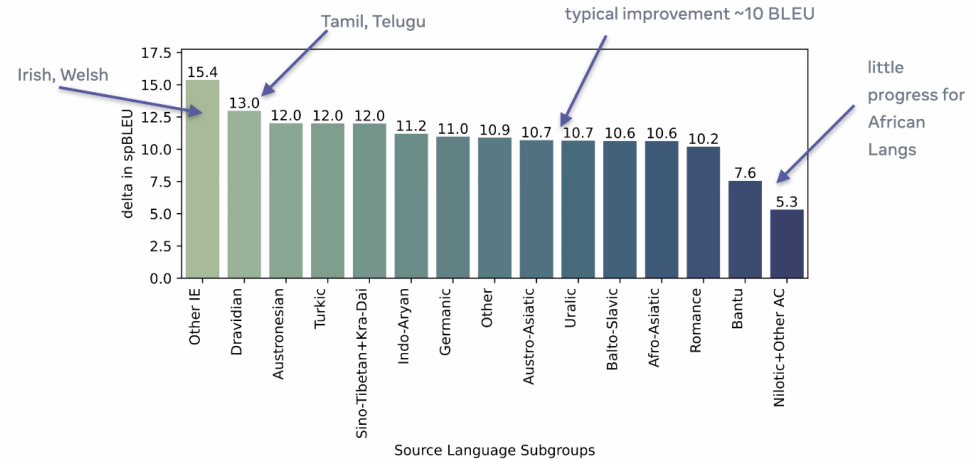




Large-Scale Multilingual Machine Translation

Improvement in the last 2 Years, from Facebook to Microsoft (FB slide)

Analysis of deltas per lang family (source)



Thanks! And...

see you soon, hopefully not virtually!

Thanks for listening!

Comments, questions & complaints to cristinae@dfki.de

Shared Tasks at WMT 2021: Multilingual Low-Resource Translation for Indo-European Languages

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session V
23rd November 2021