Searching the Web for Cross-lingual Parallel Data

Ahmed El-Kishky[†], Philipp Koehn[*], Holger Schwenk[†]

Facebook AI[†], Johns Hopkins University[*]
http://www.statmt.org/web-mining-tutorial/

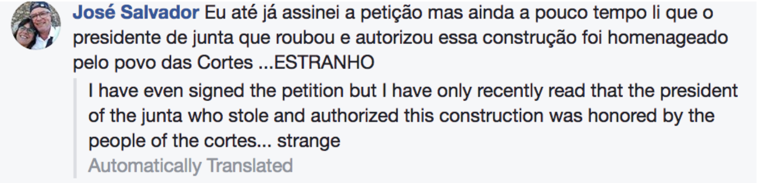
July 25, 2020

Overview

1. Background and Motivation
2. Multilingual Corpora and Web Crawling
3. Multilingual Representations
   - LASER
   - Evaluation
4. Parallel Document Retrieval
5. Local Sentence Alignment
6. Global Sentence Alignment
   - WikiMatrix
   - CCMatrix
   - WMT/TED
7. Parallel Sentence Filtering

Intro

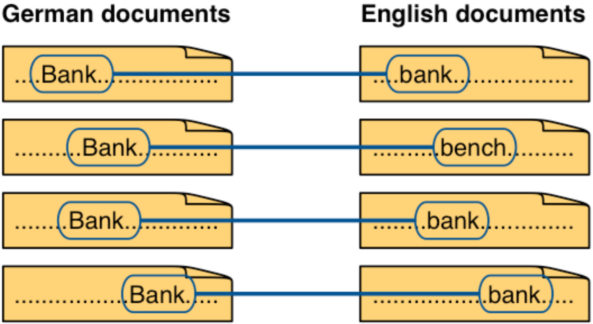
For example, machine translation.



José Salvador Eu até já assinei a petição mas ainda a pouco tempo li que o presidente de junta que roubou e autorizou essa construção foi homenageado pelo povo das Cortes ...ESTRANHO

> I have even signed the petition but I have only recently read that the president of the junta who stole and authorized this construction was honored by the people of the cortes... strange
>
> Automatically Translated

Learning from Data



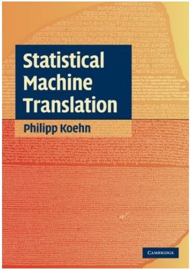
Needed: examples of translated sentences

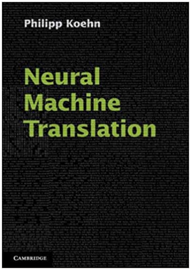
Data-Driven Machine Translation

- Given: parallel corpora
  (collections of translated sentences)

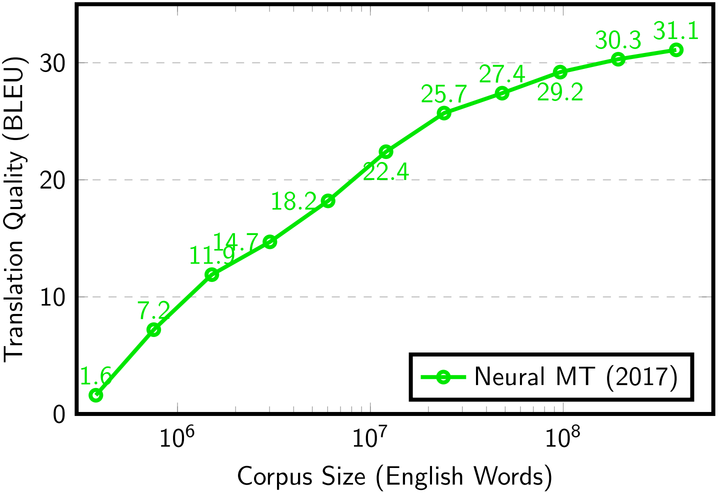
- Output: machine translation models

- Since ∼2000: statistical methods
- Since ∼2015: neural methods

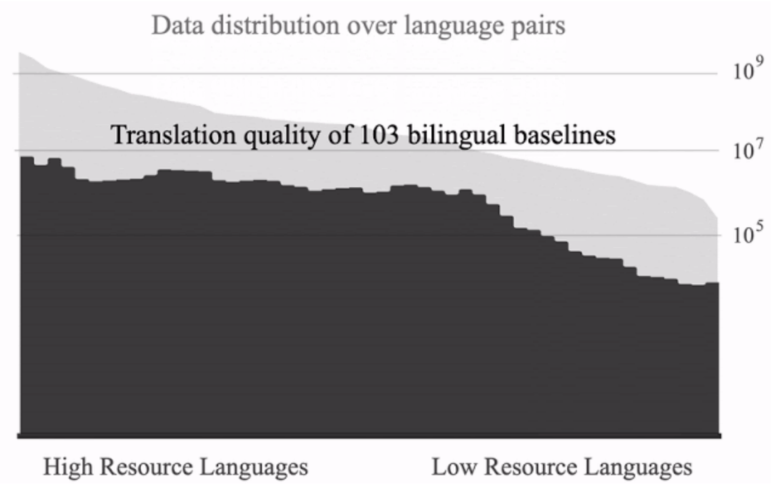




More Data is Better



More Data is Better



(from Google)

More Data is Better

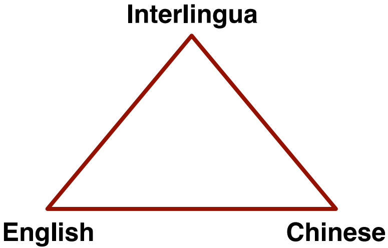
Don't think about algorithms, get more data!

If you want to think, think about getting more data!

Eric Brill, 2001

Towards Interlingua

Language-agnostic meaning representations.



Parallel corpora give us two corners of this triangle

Other Uses of Parallel Data

For example, multi-lingual hate speech detection.

- Annotate an English corpus
- Train a classifier
- But: use language-independent representations of input (trained on parallel data)

Naturally Occurring Data

- Translation is a common human activity
- Billion dollar industry that
  - localizes products and their documentation
  - makes information accessible in many languages
  - enables communication in multi-lingual organizations
  - translates books, TV shows, movies, ...
- We do not need to create this data.
- We just need to find it.

Large Pools of Data

- For instance, Europarl, 2005
  - well structured web site with clear mapping between translations
  - specialized scripts for crawling, text extraction, alignment
  - maintained structure: sessions, speakers, paragraphs

Europarl: A Parallel Corpus for Statistical Machine Translation
Koehn, MT Summit 2005

- Other efforts like this
  - Project Syndicate (”news commentary”)
  - Global Voices
  - EU Bookstore
  - United Nations
  - Acquis Communitaire

Commoncrawl

- The web on a hard drive
- Extraction pipeline

  Dirt Cheap Web-Scale Parallel Text from the Common Crawl, Smith, Saint Amand, Plamada, Koehn, Callison-Burch, Lopez, ACL 2013

  - detect document pairs based on URL
  - use HTML structure to check document matches
  - extract text (in chunks indicated by HTML)
  - sentence alignment
  - sentence filtering
- Decent amounts: French (120m words), German (80m), ..., Pashto (200k)

CCMatrix

- Preview: we wil present methods to extract parallel sentences from CommonCrawl
- CCMatrix: largest collection of high quality mined bitexts
    - 4.5 billion parallel sentences in 39 languages
    - ”matrix”: aligned across all pairs, not just paired with English
- Extraction purely with retrieval over sentence embeddings

Paracrawl

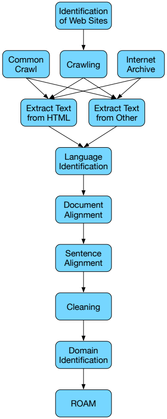
- Crawling the web for parallel data
  - funding from two Google grants (2014, 2016)
  - funding from the EU since 2017
- Currently in collaboration with Edinburgh, Alicante, Prompsit, TAUS, Omniscien Technology

⇒ Corpora with billions* of words for major languages

* amounts vary based on degree of filtering — for German–English, raw corpus has 4 billion sentence pairs, recommended corpus only 40 million deduplicated sentence pairs)

Processing Pipeline

- Identifying multi-lingual web sites
- Crawling
- Text extraction from HTML and PDF
- Document alignment
- Sentence alignment
- Sentence pair repair (Bifixer)
- Sentence pair filtering



Candidate Web Sites

- Extracted all text from CommonCrawl
  - Language ID on all of it

    N-gram Counts and Language Models from the Common Crawl, Buck, Heafield, Van Ooyen, LREC 2014

  ⇒ List of web sites with content in multiple languages
  CommonCrawl has 1.6 million domains with de-en data, 1.7 million for es-en, etc.

- Search for language name (”Chinese”) or flags (en.gif)
- For low resource languages, crawl all web sites with language content

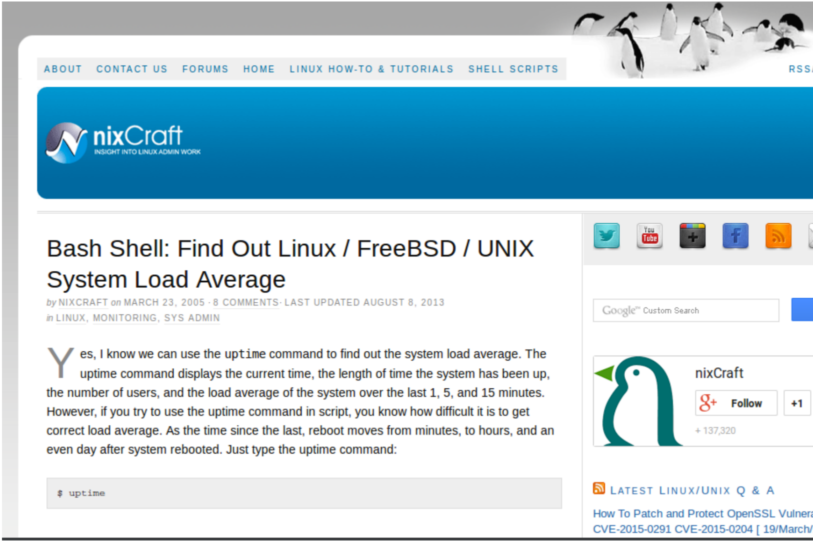
Crawling

- Several off-the-shelf tools available
  - HTTrack: multi-platform tool for crawling
  - Heritrix: Internet Archive's web crawler
  - Creepy: Python library with basic resources for crawling
  - Wget: popular Unix tool
- Many practical problems
  - large sites
  - protected content
  - interference with web server operations
  - robots.txt

Extract text

- Raw crawls: HTML, TXT, PDF, junk
- Converted into usable format, for each document
  - URL
  - language identification
  - raw HTML (base64)
  - extracted text (base64)
- Special challenges by formats such as PDF

A Web Page



Bash Shell: Find Out Linux / FreeBSD / UNIX System Load Average

by NIXCRAFT *on* MARCH 23, 2005 · 8 COMMENTS · LAST UPDATED AUGUST 8, 2013
in LINUX, MONITORING, SYS ADMIN

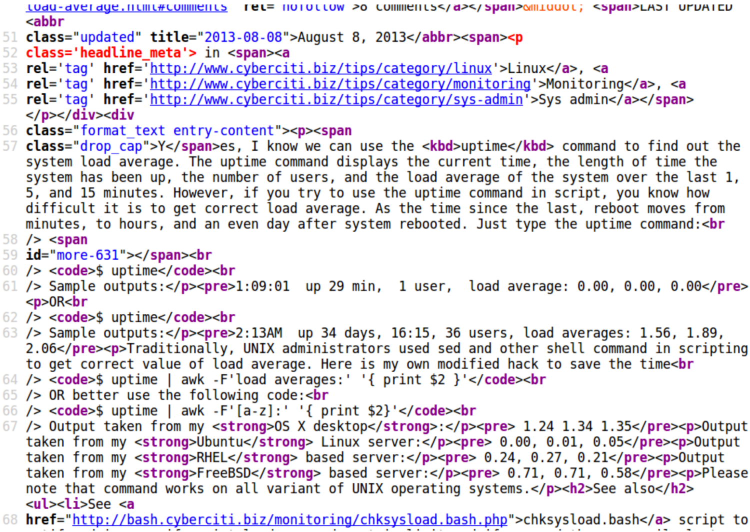
Yes, I know we can use the `uptime` command to find out the system load average. The uptime command displays the current time, the length of time the system has been up, the number of users, and the load average of the system over the last 1, 5, and 15 minutes. However, if you try to use the uptime command in script, you know how difficult it is to get correct load average. As the time since the last, reboot moves from minutes, to hours, and an even day after system rebooted. Just type the uptime command:

```
$ uptime
```

LATEST LINUX/UNIX Q & A

How To Patch and Protect OpenSSL Vulnera
CVE-2015-0291 CVE-2015-0204 [19/March/

HTML Source



```
load-average.html#comments' rel='nofollow'>8 comments</a></span>&middot; <span>LAST UPDATED
<abbr
class="updated" title="2013-08-08">August 8, 2013</abbr><span><p
class='headline_meta'> in <span><a
rel='tag' href='http://www.cyberciti.biz/tips/category/linux'>Linux</a>, <a
rel='tag' href='http://www.cyberciti.biz/tips/category/monitoring'>Monitoring</a>, <a
rel='tag' href='http://www.cyberciti.biz/tips/category/sys-admin'>Sys admin</a></span>
</p></div><div
class="format_text entry-content"><p><span
class="drop_cap">Y</span>es, I know we can use the <kbd>uptime</kbd> command to find out the
system load average. The uptime command displays the current time, the length of time the
system has been up, the number of users, and the load average of the system over the last 1,
5, and 15 minutes. However, if you try to use the uptime command in script, you know how
difficult it is to get correct load average. As the time since the last, reboot moves from
minutes, to hours, and an even day after system rebooted. Just type the uptime command:<br
/> <span
id="more-631"></span><br
/> <code>$ uptime</code><br
/> Sample outputs:</p><pre>1:09:01  up 29 min,  1 user,  load average: 0.00, 0.00, 0.00</pre>
<p>OR<br
/> <code>$ uptime</code><br
/> Sample outputs:</p><pre>2:13AM  up 34 days, 16:15, 36 users, load averages: 1.56, 1.89,
2.06</pre><p>Traditionally, UNIX administrators used sed and other shell command in scripting
to get correct value of load average. Here is my own modified hack to save the time<br
/> <code>$ uptime | awk -F'load averages:' '{ print $2 }'</code><br
/> OR better use the following code:<br
/> <code>$ uptime | awk -F'[a-z]:' '{ print $2}'</code><br
/> Output taken from my <strong>OS X desktop</strong>:</p><pre> 1.24 1.34 1.35</pre><p>Output
taken from my <strong>Ubuntu</strong> Linux server:</p><pre> 0.00, 0.01, 0.05</pre><p>Output
taken from my <strong>RHEL</strong> based server:</p><pre> 0.24, 0.27, 0.21</pre><p>Output
taken from my <strong>FreeBSD</strong> based server:</p><pre> 0.71, 0.71, 0.58</pre><p>Please
note that command works on all variant of UNIX operating systems.</p><h2>See also</h2>
<ul><li>See <a
href="http://bash.cyberciti.biz/monitoring/chksysload.bash.php">chksysload.bash</a> script to
```

Method 1: Strip Tags

```
LAST UPDATED August 8, 2013 in Linux , Monitoring , Sys
admin Y es, I know we can use the uptime command to find
out the system load average. The uptime command displays
the current time, the length of time the system has been
up, the number of users, and the load average of the
system over the last 1, 5, and 15 minutes. However, if
you try to use the uptime command in script, you know how
difficult it is to get correct load average. As the time
since the last, reboot moves from minutes, to hours, and
an even day after system rebooted. Just type the uptime
command: $ uptime Sample outputs: 1:09:01 up 29 min, 1
user, load average: 0.00, 0.00, 0.00
```

Method 2: HTML Parser

```
LAST UPDATED August 8, 2013

in Linux, Monitoring, Sys admin

Y

es, I know we can use the uptime command to find out the system
load average. The uptime command displays the current time, the
length of time the system has been up, the number of users, and
the load average of the system over the last 1, 5, and 15
minutes. However, if you try to use the uptime command in
script, you know how difficult it is to get correct load
average. As the time since the last, reboot moves from minutes,
to hours, and an even day after system rebooted. Just type the
uptime command:

$ uptime

Sample outputs: 1:09:01 up 29 min, 1 user, load average: 0.00,
0.00, 0.00
```

What Language?

Muitas intervenções alertaram para o facto de a política dos sucessivos governos PS, PSD e CDS, com cortes no financiamento das instituições do Ensino Superior e com a progressiva desresponsabilização do Estado das suas funções, ter conduzido a uma realidade de destruição da qualidade do Ensino Superior público.

Clues: Letter N-Grams

Muitas intervenções alertaram para o facto de a política dos sucessivos governos PS, PSD e CDS, com cortes no financiamento das instituições do Ensino Superior e com a progressiva desresponsabilização do Estado das suas funções, ter conduzido a uma realidade de destruição da qualidade do Ensino Superior público.

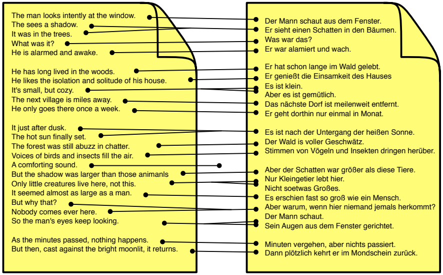
Align Documents

- Paracrawl method
  - translate foreign document into English
  - score based on n-gram matches
  - matching of URL
  - other features

Quick and Reliable Document Alignment with TF/IDF Cosine Distance, Buck and Koehn, WMT 2016

- Shared task WMT 2016
  - n-gram matches on (translated) documents powerful
  - only very recently more research on topic

Align Sentences



The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

He has long lived in the woods.
He likes the isolation and solitude of his house.
It's small, but cozy.
The next village is miles away.
He only goes there once a week.

It just after dusk.
The hot sun finally set.
The forest was still abuzz in chatter.
Voices of birds and insects fill the air.
A comforting sound.
But the shadow was larger than those animanls
Only little creatures live here, not this.
It seemed almost as large as a man.
But why that?
Nobody comes ever here.
So the man's eyes keep looking.

As the minutes passed, nothing happens.
But then, cast against the bright moonlit, it returns.

Der Mann schaut aus dem Fenster.
Er sieht einen Schatten in den Bäumen.
Was war das?
Er war alamiert und wach.

Er hat schon lange im Wald gelebt.
Er genießt die Einsamkeit des Hauses
Es ist klein.
Aber es ist gemütlich.
Das nächste Dorf ist meilenweit entfernt.
Er geht dorthin nur einmal in Monat.

Es ist nach der Untergang der heißen Sonne.
Der Wald is voller Geschwätz.
Stimmen von Vögeln und Insekten dringen herüber.

Aber der Schatten war größer als diese Tiere.
Nur Kleingetier lebt hier.
Nicht soetwas Großes.
Es erschien fast so groß wie ein Mensch.
Aber warum, wenn hier niemand jemals herkommt?
Der Mann schaut.
Sein Augen aus dem Fenster gerichtet.

Minuten vergehen, aber nichts passiert.
Dann plötzlich kehrt er im Mondschein zurück.

- Given: pair of documents
- Task: match sentence
- Allow 1-2 mappings etc.?
- Reordering of sentences?
- Several established tools (Hunalign, Bleualign, ...)

Corpus Cleaning

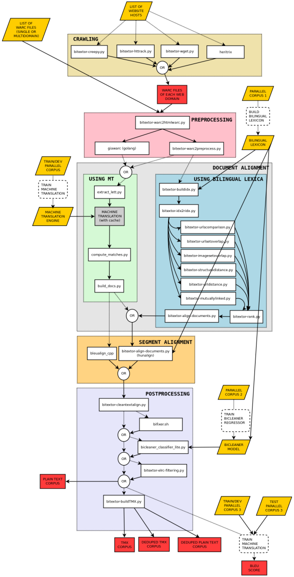
- Two objectives for clean corpus
- Fluency
  - well-formed language
- Adequacy
  - foreign and English sentence have same meaning, style, etc.
- Open question: what is harmful noise?
- Shared tasks at WMT 2018-2020

Open-Source Code: Bitextor

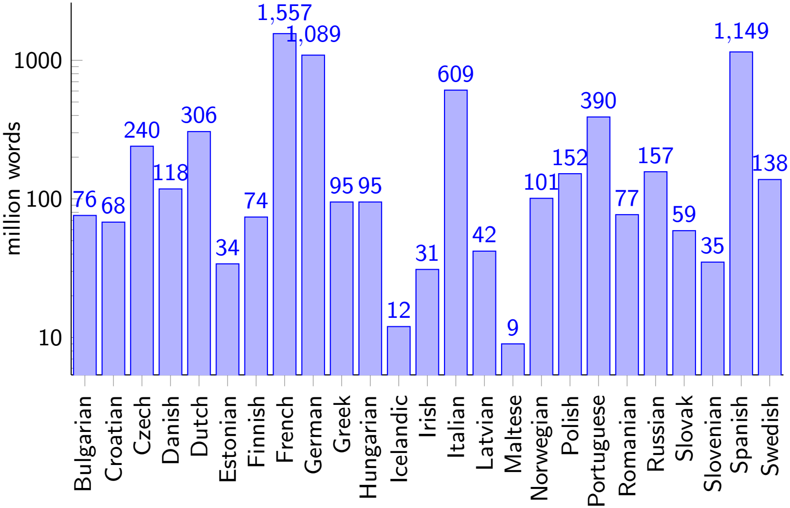
https://github.com/bitextor/bitextor

- Bitextor: Integrated tool to execute entire pipeline

- Pipeline management for distributed computation


LIST OF WEBSITE HOSTS
LIST OF WARC FILES (SINGLE OR MULTIDOMAIN)
CRAWLING
bitextor-creepy.py
bitextor-httrack.py
bitextor-wget.py
heritrix
OR
WARC FILES OF EACH WEB DOMAIN
PARALLEL CORPUS 1
BUILD BILINGUAL LEXICON
BILINGUAL LEXICON
PREPROCESSING
bitextor-warc2htmlwarc.py
giawarc (golang)
bitextor-warc2preprocess.py
TRAIN/DEV PARALLEL CORPUS
TRAIN MACHINE TRANSLATION
MACHINE TRANSLATION ENGINE
OR
DOCUMENT ALIGNMENT
USING MT
extract_lett.py
MACHINE TRANSLATION (with cache)
compute_matches.py
build_docs.py
USING BILINGUAL LEXICA
bitextor-buildidx.py
bitextor-idx2ridx.py
bitextor-urlscomparison.py
bitextor-urlsetoverlap.py
bitextor-imagesetoverlap.py
bitextor-structuredistance.py
bitextor-urldistance.py
bitextor-mutuallylinked.py
OR
bitextor-align-documents.py
bitextor-rank.py
SEGMENT ALIGNMENT
bleualign_cpp
bitextor-align-documents.py (hunalign)
OR
POSTPROCESSING
bitextor-cleantextalign.py
bifixer.sh
OR
bicleaner_classifier_lite.py
OR
bitextor-elrc-filtering.py
OR
PLAIN TEXT CORPUS
bitextor-buildTMX.py
TMX CORPUS
DEDUPED TMX CORPUS
DEDUPED PLAIN TEXT CORPUS
PARALLEL CORPUS 2
TRAIN BICLEANER REGRESSOR
BICLEANER MODEL
TRAIN/DEV PARALLEL CORPUS 3
TEST PARALLEL CORPUS 3
TRAIN MACHINE TRANSLATION
BLEU SCORE

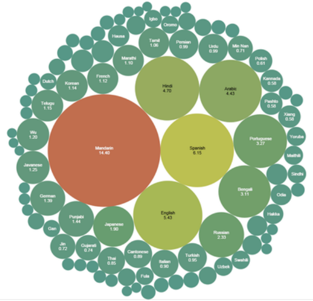
ParaCrawl Release 6



Multilingual Models

- 7 111 living languages

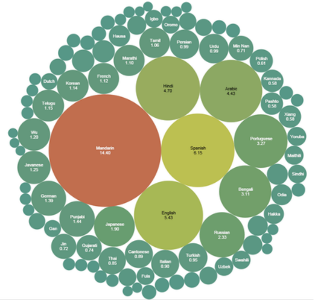
Native speakers



Multilingual Models

- 7 111 living languages
- 40% are endangered

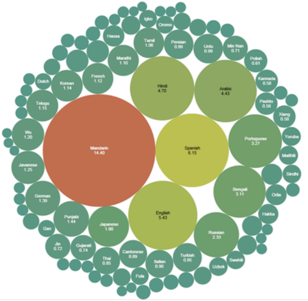
Native speakers



Multilingual Models

- 7 111 living languages
- 40% are endangered
- 23 languages account for half the population

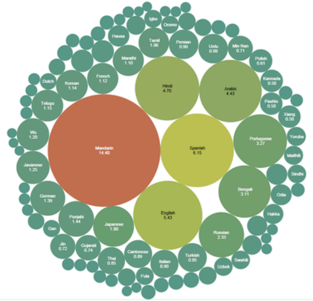
Native speakers



Multilingual Models

- 7 111 living languages
- 40% are endangered
- 23 languages account for half the population
- MT: < 100 languages

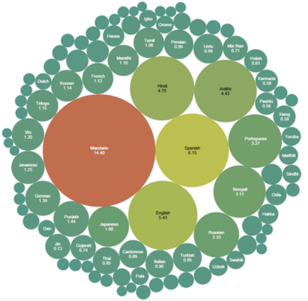
Native speakers



Multilingual Models

- 7 111 living languages
- 40% are endangered
- 23 languages account for half the population
- MT: $<$ 100 languages
- Almost all NLP applications are mostly English (classification, sentiment analysis or NLI, Q&A, dialog, ...)

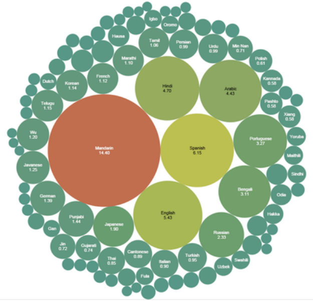
Native speakers



Multilingual Models

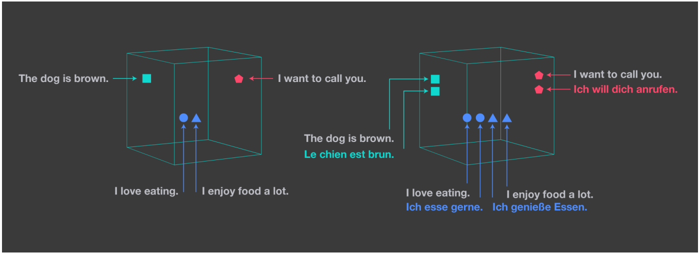
- 7 111 living languages
- 40% are endangered
- 23 languages account for half the population
- MT: < 100 languages
- Almost all NLP applications are mostly English (classification, sentiment analysis or NLI, Q&A, dialog, . . .)

Native speakers



⇒ Input in foreign language is translated into English

Multilingual Models

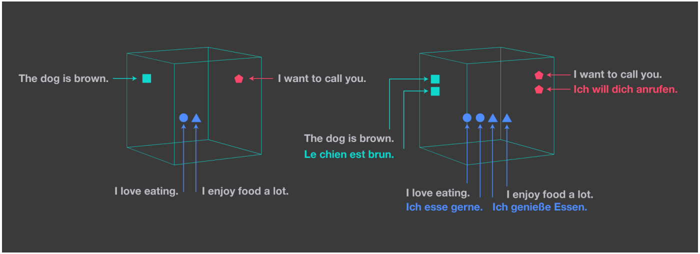


Motivation

- Try to embed sentences written in many languages into one joint space

  ⇒ cross-lingual transfer for various NLP applications
  - benefit of similarities among languages

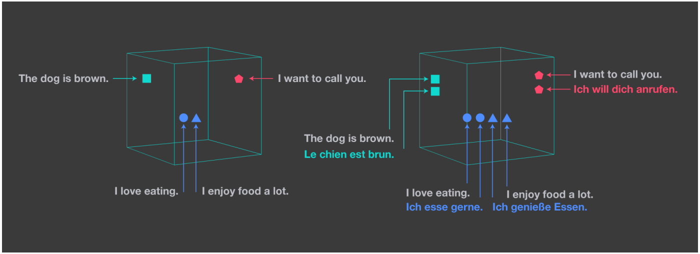
Multilingual Models



Motivation

- Try to embed sentences written in many languages into one joint space
  - ⇒ cross-lingual transfer for various NLP applications
  - benefit of similarities among languages
- This gives us a highly semantic representation

Multilingual Models

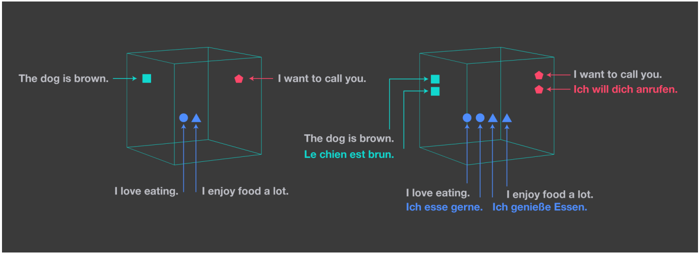


Motivation

- Try to embed sentences written in many languages into one joint space
  - ⇒ cross-lingual transfer for various NLP applications
  - benefit of similarities among languages
- This gives us a highly semantic representation

⇒ Sentences with similar meaning are close (mono- or cross-lingual)

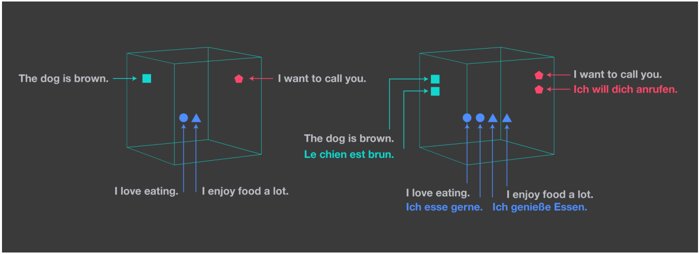
Multilingual Models



Applications:

- zero-shot transfer
- bitext mining and filtering
- large-scale similarity search
- paraphrasing
- data augmentation
- . . .

Multilingual Models



Applications:

- zero-shot transfer
- bitext mining and filtering
- large-scale similarity search
- paraphrasing
- data augmentation
- ...

Multilingual Models

Some approaches:

- MUSE
  - unsupervised multilingual word embeddings
- LASER
  - supervised multilingual **sentence** embeddings
- XLM
  - unsupervised multilingual sentence embeddings
- Sentence BERT
  - fine-tuned for linguistic similarity

⋮

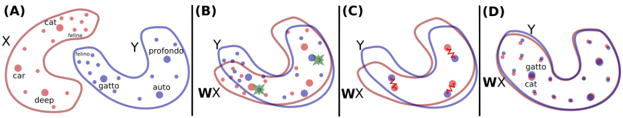
Multilingual Models

Some approaches:

- MUSE
  - unsupervised multilingual word embeddings
- LASER
  - supervised multilingual **sentence** embeddings
- XLM
  - unsupervised multilingual sentence embeddings
- Sentence BERT
  - fine-tuned for linguistic similarity

⋮

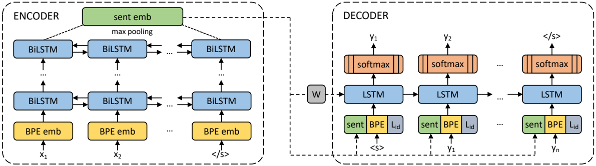
MUSE



Principle

- Learn multilingual word embeddings without any aligned data
- fastText embeddings aligned in a common space
  - learn transformation of space X to Y
- A. Conneau et al., *Word Translation Without Parallel Data*, ICLR'18
- `https://github.com/facebookresearch/MUSE`

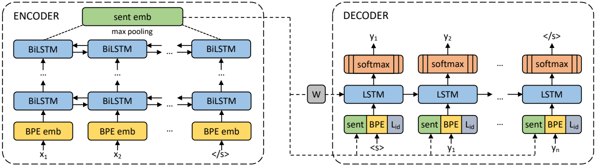
LASER: Architecture



Seq2seq approach with one joint encoder and decoder

- Based on `fairseq`
- Shared encoder and decoder for several languages
- No attention, but max-pooling
- Sentence representation is used at the input at each time step and to initialize decoder
- Also target language embedding

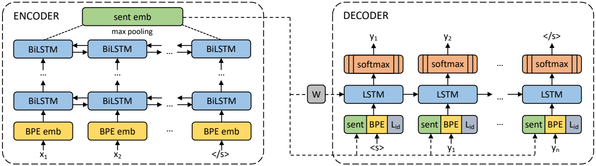
LASER: Architecture



Training strategies

- *N*:1 translation is enough to learn a joint embedding
- No explicit criterion to enforce joint embedding
  - ranking loss
  - GAN to predict language
  - . . .

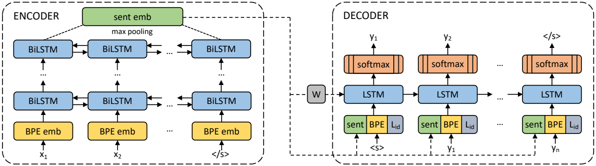
LASER: Architecture



Training strategies

- N:1 translation is enough to learn a joint embedding
- No explicit criterion to enforce joint embedding
  - ranking loss
  - GAN to predict language
  - ...
- But N:1 doesn't cover target language (English)

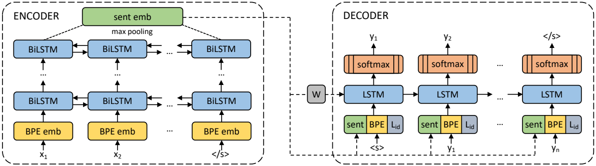
LASER: Architecture



Training strategies

- *N*:1 translation is enough to learn a joint embedding
- No explicit criterion to enforce joint embedding
  - ranking loss
  - GAN to predict language
  - . . .
- But *N*:1 doesn't cover target language (English)
- Limited success with (noisy) autoencoder

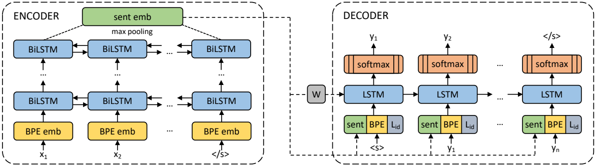
LASER: Architecture



Training strategies

- How to have a language at the input and output ?
  - in the past: $N \rightarrow$ (N-1)
- Two target languages are enough
  - English and Spanish
  - independently aligned
  - not all input languages need to be aligned to both
- Language pair is changed at each mini-batch
- Trained on 223M sentences of public bitexts

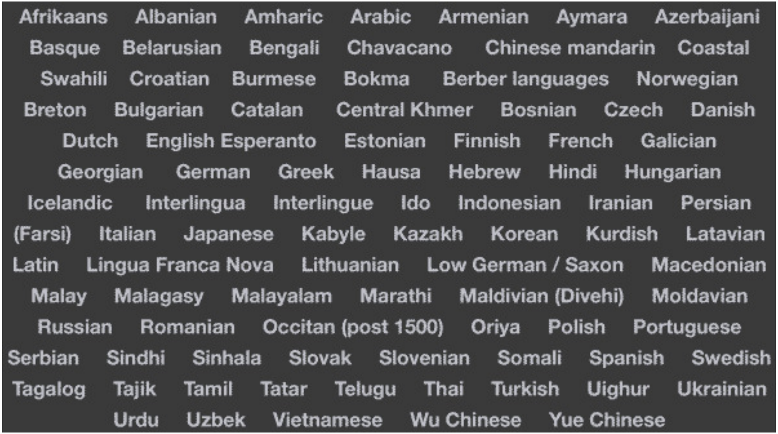
LASER: Architecture



Encoder

- 5 layer BiLSTM (depth helps !)
- No information on input (or target) language
- Shared BPE tokens, 50k BPE operations
- No pretraining of BPE embeddings
- The training procedure makes no assumption on the encoder:
  $\Rightarrow$ transformers, convolutional, . . .

LASER: Training languages



Afrikaans Albanian Amharic Arabic Armenian Aymara Azerbaijani
Basque Belarusian Bengali Chavacano Chinese mandarin Coastal
Swahili Croatian Burmese Bokma Berber languages Norwegian
Breton Bulgarian Catalan Central Khmer Bosnian Czech Danish
Dutch English Esperanto Estonian Finnish French Galician
Georgian German Greek Hausa Hebrew Hindi Hungarian
Icelandic Interlingua Interlingue Ido Indonesian Iranian Persian
(Farsi) Italian Japanese Kabyle Kazakh Korean Kurdish Latavian
Latin Lingua Franca Nova Lithuanian Low German / Saxon Macedonian
Malay Malagasy Malayalam Marathi Maldivian (Divehi) Moldavian
Russian Romanian Occitan (post 1500) Oriya Polish Portuguese
Serbian Sindhi Sinhala Slovak Slovenian Somali Spanish Swedish
Tagalog Tajik Tamil Tatar Telugu Thai Turkish Uighur Ukrainian
Urdu Uzbek Vietnamese Wu Chinese Yue Chinese

LASER: Training languages

22 different writing scripts:

Arabic	هناك العديد من اللغات في العالم.	Hebrew	ישנן שפות רבות בעולם.
Armenian	Աշխարհում շատ լեզուներ կան:	Kanjii	世界にはたくさんの言語があります。
Burmese	ကမ်ဘာပေါ်တွင်ဘာသာစကားများစွာရှိပါတယ်။	Khmer	មានភាសាជាច្រើននៅលើពិភពលោក។
Chinese	世界上有很多种语言。	Latin	There are many languages in the world.
Cyrillic	В мире много языков.	Malayalam	ലോകത്തിൽ അനേകം ഭാഷകൾ ഉണ്ട്.
Devanagari	दुनिया में कई भाषाएं हैं।	Persian	زبان های بسیاری در جهان وجود دارد.
Eastern-Nagari	বিশ্বের অনেক ভাষা আছে।	Sinhala	ලෝකයේ බොහෝ භාෂාවන් පවතී.
Ge'ez	በዓለም ውስጥ ብዙ ቋንቋዎች አሉ.	Tamil	உலகில் பல மொழிகள் உள்ளன.
Georgian	მსოფლიოში ბევრი ენაა.	Telugu	ప్రపంచంలో అనేక భాషలు ఉన్నాయి.
Greek	Υπάρχουν πολλές γλώσσες στον κόσμο.	Thaana	*No free translation for Maldivian (Dhivehi)*
Hangul	세계에는 많은 언어가 있습니다.	Thai	มีหลายภาษาในโลกนี้

- One single encoder can handle all these scripts
- All these sentences are close in the embedding space
- It is not necessary to specify the language or script
- Code-switching is also supported

LASER toolkit

Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Trans. Assoc. Comput. Linguistics 7: 597-610 (2019)

Well established in community, academia and industry

- `https://github.com/facebookresearch/LASER/`
- M. Artetxe and H. Schwenk, *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*, TACL'19 and arXiv'18
- Fast and easy to use (2000 sentences/sec)
- One model for many applications
- Current SOTA for filtering and mining bitexts

Applications of Multilingual Embeddings

- Zero-shot transfer in NLP
  - Use ML embeddings to train English NLP system

  ⇒ apply it to other languages without any modification
  - classification, NLI, QA, . . .

Applications of Multilingual Embeddings

- Zero-shot transfer in NLP
  - Use ML embeddings to train English NLP system

  ⇒ apply it to other languages without any modification
  - classification, NLI, QA, ...
- Bitexts mining and filtering
  - sentence similarity ∼ distance in joint space

Applications of Multilingual Embeddings

- Zero-shot transfer in NLP
  - Use ML embeddings to train English NLP system

  ⇒ apply it to other languages without any modification
  - classification, NLI, QA, . . .
- Bitexts mining and filtering
  - sentence similarity ∼ distance in joint space
- Large-scale similarity search
  - index many sentences, search for closest ones
  - paraphrasing, data augmentation, . . .

Applications of Multilingual Embeddings

- Zero-shot transfer in NLP
  - Use ML embeddings to train English NLP system

  ⇒ apply it to other languages without any modification
  - classification, NLI, QA, . . .
- Bitexts mining and filtering
  - sentence similarity ∼ distance in joint space
- Large-scale similarity search
  - index many sentences, search for closest ones
  - paraphrasing, data augmentation, . . .

We always use the same LASER sentence embeddings, no task-specific fine-tuning

XNLI: Cross-Lingual NLI

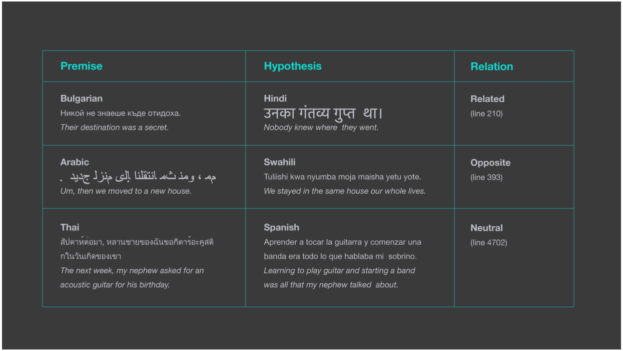
- Fixed LASER embeddings
- NLI classifier trained on English only

XNLI: Cross-Lingual NLI

- Fixed LASER embeddings
- NLI classifier trained on English only
- Zero-shot transfer to any language supported by LASER

XNLI: Cross-Lingual NLI

- Fixed LASER embeddings
- NLI classifier trained on English only
- Zero-shot transfer to any language supported by LASER
- We can arbitrarily combine sentence in any language



Premise	Hypothesis	Relation
Bulgarian Никой не знаеше къде отидоха. *Their destination was a secret.*	**Hindi** उनका गंतव्य गुप्त था। *Nobody knew where they went.*	**Related** (line 210)
Arabic مم ، ومنذ ثم انتقلنا إلى منزل جديد . *Um, then we moved to a new house.*	**Swahili** Tuliishi kwa nyumba moja maisha yetu yote. *We stayed in the same house our whole lives.*	**Opposite** (line 393)
Thai สัปดาห์ต่อมา, หลานชายของฉันขอกีตาร์อะคูสติกในวันเกิดของเขา *The next week, my nephew asked for an acoustic guitar for his birthday.*	**Spanish** Aprender a tocar la guitarra y comenzar una banda era todo lo que hablaba mi sobrino. *Learning to play guitar and starting a band was all that my nephew talked about.*	**Neutral** (line 4702)

Bitext Mining Approach

Margin criterion

Semantic similarity $\propto$ distance

$\Rightarrow$ mine parallel sentences

$$\begin{aligned}&\text{margin}(x, y)\\ &= \frac{\cos(x, y)}{\sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k}}\end{aligned} \tag{1}$$

(Artexe and Schwenk, arXiv Nov'18 and ACL'19)

Results for 93 Languages: BUCC

	TRAIN				TEST			
	de-en	fr-en	ru-en	zh-en	de-en	fr-en	ru-en	zh-en
Azpeita et '17	83.3	78.83	-	-	83.7	79.5	-	-
Grégoire&Langlais '17	-	20.7	-	-	-	20	-	-
Zhang & Zweigenbaum '17	-	-	-	43.48	-	-	-	45.13
Azpeita et al. '18	84.3	80.6	80.9	76.5	85.5	81.5	81.3	77.5
Bouamor & Sajad '18	-	75.2	-	-	-	76.0	-	-
Leong & Chao '18	-	-	-	58.5	-	-	-	56
Schwenk ACL'18	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Artetxe&Schwenk arXiv'18	94.8	91.9	90.9	91.0	95.6	92.9	92.0	**92.6**
Proposed method	**95.4**	**92.4**	**92.3**	**91.2**	**96.2**	**93.9**	**93.3**	92.3

- Significantly outperforms other systems of the BUCC eval

Results for 93 Languages: BUCC

	TRAIN				TEST			
	de-en	fr-en	ru-en	zh-en	de-en	fr-en	ru-en	zh-en
Azpeita et '17	83.3	78.83	-	-	83.7	79.5	-	-
Grégoire&Langlais '17	-	20.7	-	-	-	20	-	-
Zhang & Zweigenbaum '17	-	-	-	43.48	-	-	-	45.13
Azpeita et al. '18	84.3	80.6	80.9	76.5	85.5	81.5	81.3	77.5
Bouamor & Sajad '18	-	75.2	-	-	-	76.0	-	-
Leong & Chao '18	-	-	-	58.5	-	-	-	56
Schwenk ACL'18	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Artetxe&Schwenk arXiv'18	94.8	91.9	90.9	91.0	95.6	92.9	92.0	**92.6**
Proposed method	**95.4**	**92.4**	**92.3**	**91.2**	**96.2**	**93.9**	**93.3**	92.3

- Significantly outperforms other systems of the BUCC eval
- New system trained on 93 languages is better than dedicated system, limited to eval languages

Generalization to New Languages

System trained on the 21 languages of Europarl

	De-En	Fr-En
State-of-the-art	85.5	81.5
Our approach	95.6	92.9

Generalization to New Languages

System trained on the 21 languages of Europarl

	De-En	Fr-En	Ru-En
State-of-the-art	85.5	81.5	81.3
Our approach	95.6	92.9	62.0

- Good performance on Russian (precision=80%) **although Russian was not used during training**

Generalization to New Languages

System trained on the 21 languages of Europarl

	De-En	Fr-En	Ru-En
State-of-the-art	85.5	81.5	81.3
Our approach	95.6	92.9	62.0

- Good performance on Russian (precision=80%) **although Russian was not used during training**

⇒ Very promising to mine data for dialects and minority languages which are in the same family than a trained language, Gallician, Nepali, . . .

Results for 93 Languages: Bitext Filtering

WMT'19: Bitext filtering for low-resource conditions

- Filter very noisy Paracrawl crawled bitexts (40-60M)
- Evaluation by training SMT and NMT systems:
- Train: En/Ne (586k), En/Si (645k) + En/Hi (1.5M)

	Ne/En 1M		Ne/En 5M		Si/En 1M		Si/En 5M	
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
LASER	**4.21**	**6.88**	4.63	2.84	**4.27**	**6.39**	**4.94**	4.02
2nd	4.10	5.48	**4.74**	**3.43**	4.19	4.97	4.62	**4.44**

Results for 93 Languages: Bitext Filtering

WMT'19: Bitext filtering for low-resource conditions

- Filter very noisy Paracrawl crawled bitexts (40-60M)
- Evaluation by training SMT and NMT systems:
- Train: En/Ne (586k), En/Si (645k) + En/Hi (1.5M)

	Ne/En 1M		Ne/En 5M		Si/En 1M		Si/En 5M	
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
LASER	**4.21**	**6.88**	4.63	2.84	**4.27**	**6.39**	**4.94**	4.02
2nd	4.10	5.48	**4.74**	**3.43**	4.19	4.97	4.62	**4.44**

- Overall best results by significant margin (+25%)

Results for 93 Languages: Bitext Filtering

WMT'19: Bitext filtering for low-resource conditions

- Filter very noisy Paracrawl crawled bitexts (40-60M)
- Evaluation by training SMT and NMT systems:
- Train: En/Ne (586k), En/Si (645k) + En/Hi (1.5M)

	Ne/En 1M		Ne/En 5M		Si/En 1M		Si/En 5M	
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
LASER	**4.21**	**6.88**	4.63	2.84	**4.27**	**6.39**	**4.94**	4.02
2nd	4.10	5.48	**4.74**	**3.43**	4.19	4.97	4.62	**4.44**

- Overall best results by significant margin (+25%)
- Good filtering is more important for NMT than SMT

Results for 93 Languages: Bitext Filtering

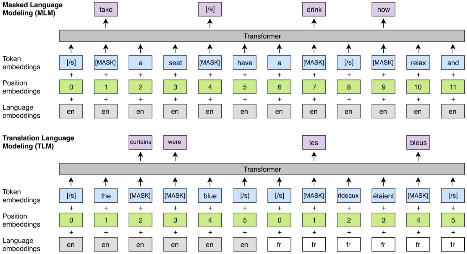
WMT'19: Bitext filtering for low-resource conditions

- Filter very noisy Paracrawl crawled bitexts (40-60M)
- Evaluation by training SMT and NMT systems:
- Train: En/Ne (586k), En/Si (645k) + En/Hi (1.5M)

	Ne/En 1M		Ne/En 5M		Si/En 1M		Si/En 5M	
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
LASER	**4.21**	**6.88**	4.63	2.84	**4.27**	**6.39**	**4.94**	4.02
2nd	4.10	5.48	**4.74**	**3.43**	4.19	4.97	4.62	**4.44**

- Overall best results by significant margin (+25%)
- Good filtering is more important for NMT than SMT
- Vishrav et al, *Low-Resource Corpus Filtering using Multilingual Sentence Embeddings*, WMT'19

XLM: Architecture


Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr

Multilingual extension of BERT

- Unsupervised: Masked LM training (MLM)
  - Joint BPE or SentencePiece vocabulary
- Supervised: Cross-Lingual LM training (CLM)
  - attention can attend words in either language which encourages alignment

XLM: Applications

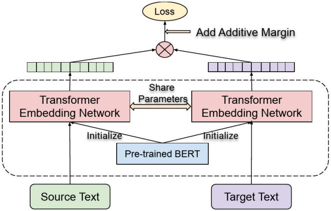
Results

- Trained unsupervised on 2.5 billon sentences of CC
- Supports 100 languages, very strong results on many English and cross-lingual tasks (GLUE, XNLI, . . .)
  - generally task specific fine-tuning
- G. Lample and A. Conneau,
  Cross-lingual Language Model Pretraining, NIPS'19
- A. Conneau et al.,
  Unsupervised Cross-lingual Representation Learning at Scale, ACL'20
- `https://github.com/facebookresearch/XLM`
- Application to similarity search requires some sort of fine-tuning

Sentence BERT

Recent research

- Several works aim in achieving transformer-based language agnostic sentence representations
- Feng et al., *Language-agnostic BERT Sentence Embedding*, arxiv Jul'20



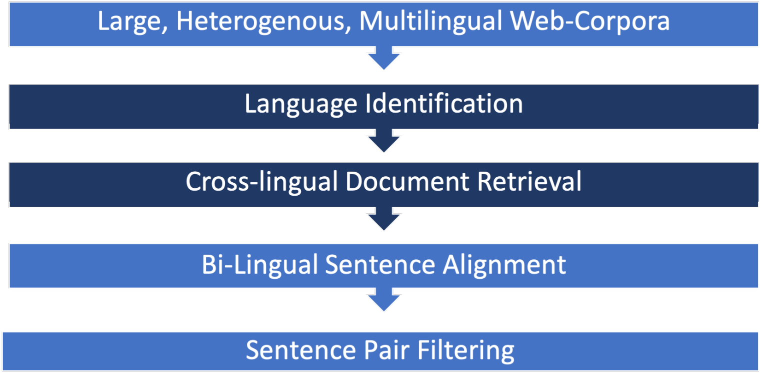
- Very Interesting results, but no comparision with margin-based LASER mining

parallel document retrieval

Cross-lingual Document Retrieval

Finding pairs of documents that are translations/near-translations of each other.

Cross-Lingual Document Retrieval



Large, Heterogenous, Multilingual Web-Corpora

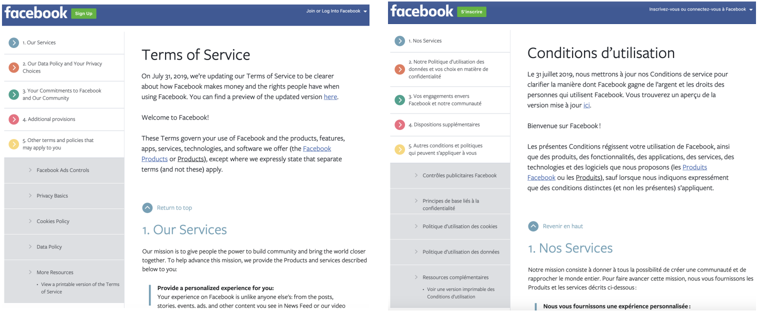
Language Identification

Cross-lingual Document Retrieval

Bi-Lingual Sentence Alignment

Sentence Pair Filtering

Cross-lingual Document Retrieval



1. Our Services
2. Our Data Policy and Your Privacy Choices
3. Your Commitments to Facebook and Our Community
4. Additional provisions
5. Other terms and policies that may apply to you

> Facebook Ads Controls

> Privacy Basics

> Cookies Policy

> Data Policy

> More Resources

- View a printable version of the Terms of Service

Terms of Service

On July 31, 2019, we're updating our Terms of Service to be clearer about how Facebook makes money and the rights people have when using Facebook. You can find a preview of the updated version here.

Welcome to Facebook!

These Terms govern your use of Facebook and the products, features, apps, services, technologies, and software we offer (the Facebook Products or Products), except where we expressly state that separate terms (and not these) apply.

Return to top

1. Our Services

Our mission is to give people the power to build community and bring the world closer together. To help advance this mission, we provide the Products and services described below to you:

> **Provide a personalized experience for you:**
> Your experience on Facebook is unlike anyone else's: from the posts, stories, events, ads, and other content you see in News Feed or our video

1. Nos Services
2. Notre Politique d'utilisation des données et vos choix en matière de confidentialité
3. Vos engagements envers Facebook et notre communauté
4. Dispositions supplémentaires
5. Autres conditions et politiques qui peuvent s'appliquer à vous

> Contrôles publicitaires Facebook

> Principes de base liés à la confidentialité

> Politique d'utilisation des cookies

> Politique d'utilisation des données

> Ressources complémentaires

- Voir une version imprimable des Conditions d'utilisation

Conditions d'utilisation

Le 31 juillet 2019, nous mettrons à jour nos Conditions de service pour clarifier la manière dont Facebook gagne de l'argent et les droits des personnes qui utilisent Facebook. Vous trouverez un aperçu de la version mise à jour ici.

Bienvenue sur Facebook !

Les présentes Conditions régissent votre utilisation de Facebook, ainsi que des produits, des fonctionnalités, des applications, des services, des technologies et des logiciels que nous proposons (les Produits Facebook ou les Produits), sauf lorsque nous indiquons expressément que des conditions distinctes (et non les présentes) s'appliquent.

Revenir en haut

1. Nos Services

Notre mission consiste à donner à tous la possibilité de créer une communauté et de rapprocher le monde entier. Pour faire avancer cette mission, nous vous fournissons les Produits et les services décrits ci-dessous :

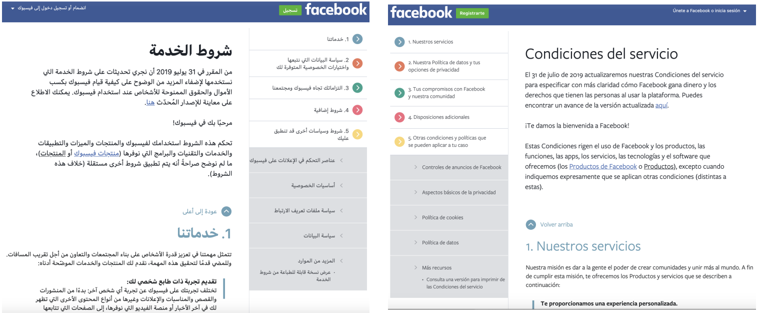
> **Nous vous fournissons une expérience personnalisée :**

(a) English Webpage

(b) French Webpage

Figure: Two web documents that are translations of each other.

Cross-lingual Document Retrieval



facebook تسجيل

انضمام أو تسجيل دخول إلى فيسبوك

1. خدماتنا

2. سياسة البيانات التي نتبعها واختيارات الخصوصية المتوفرة لك

3. التزاماتك تجاه فيسبوك ومجتمعنا

4. شروط إضافية

5. شروط وسياسات أخرى قد تنطبق عليك

عناصر التحكم في الإعلانات على فيسبوك

أساسيات الخصوصية

سياسة ملفات تعريف الارتباط

سياسة البيانات

المزيد من الموارد

- عرض نسخة قابلة للطباعة من شروط الخدمة

شروط الخدمة

من المقرر في 31 يوليو 2019 أن نجري تحديثات على شروط الخدمة التي نستخدمها لإضفاء المزيد من الوضوح على كيفية قيام فيسبوك بكسب الأموال والحقوق الممنوحة للأشخاص عند استخدام فيسبوك. يمكنك الاطلاع على معاينة للإصدار المُحدّث هنا.

مرحبًا بك في فيسبوك!

تحكم هذه الشروط استخدامك لفيسبوك والمنتجات والميزات والتطبيقات والخدمات والتقنيات والبرامج التي نوفرها (منتجات فيسبوك أو المنتجات)، ما لم نوضح صراحةً أنه يتم تطبيق شروط أخرى مستقلة (خلاف هذه الشروط).

عودة إلى أعلى

1. خدماتنا

تتمثل مهمتنا في تعزيز قدرة الأشخاص على بناء المجتمعات والتعاون من أجل تقريب المسافات. وللمضي قدمًا لتحقيق هذه المهمة، نقدم لك المنتجات والخدمات الموضّحة أدناه:

تقديم تجربة ذات طابع شخصي لك:
تختلف تجربتك على فيسبوك عن تجربة أي شخص آخر: بدءًا من المنشورات والقصص والمناسبات والإعلانات وغيرها من أنواع المحتوى الأخرى التي تظهر لك في آخر الأخبار أو منصة الفيديو التي نوفرها، إلى الصفحات التي تتابعها

facebook Registrarte

Únete a Facebook o inicia sesión

1. Nuestros servicios

2. Nuestra Política de datos y tus opciones de privacidad

3. Tus compromisos con Facebook y nuestra comunidad

4. Disposiciones adicionales

5. Otras condiciones y políticas que se pueden aplicar a tu caso

Controles de anuncios de Facebook

Aspectos básicos de la privacidad

Política de cookies

Política de datos

Más recursos

- Consulta una versión para imprimir de las Condiciones del servicio

Condiciones del servicio

El 31 de julio de 2019 actualizaremos nuestras Condiciones del servicio para especificar con más claridad cómo Facebook gana dinero y los derechos que tienen las personas al usar la plataforma. Puedes encontrar un avance de la versión actualizada aquí.

¡Te damos la bienvenida a Facebook!

Estas Condiciones rigen el uso de Facebook y los productos, las funciones, las apps, los servicios, las tecnologías y el software que ofrecemos (los Productos de Facebook o Productos), excepto cuando indiquemos expresamente que se aplican otras condiciones (distintas a estas).

Volver arriba

1. Nuestros servicios

Nuestra misión es dar a la gente el poder de crear comunidades y unir más al mundo. A fin de cumplir esta misión, te ofrecemos los Productos y servicios que se describen a continuación:

Te proporcionamos una experiencia personalizada.

(a) Arabic Webpage

(b) Spanish Webpage

Figure: Two web documents that are translations of each other.

Motivation

- Training data for information retrieval
  - Supervision for learning-to-rank
  - Supervision for retrieval
- Source of training data for learning multilingual representation
  - Cross-lingual word representations
  - Cross-lingual sentence representations
  - Cross-lingual document representation
- Source of training data for machine translation (BLEU goes up)
  - Mine parallel data for low-resource directions
  - Web parallel data covers a variety of domains

Objective

Given a corpus of web-documents, automatically identify pairs of documents that are translations of each other.

Objective

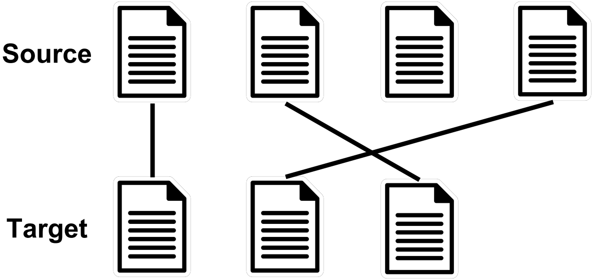


Figure: Documents are aligned 1-to-1 within each domain.

Evaluation

- Recall only i.e. what percentage of the test-set pairs is found
- 1-1 rule; every document can only occur in one pair.

CC-Aligned

CC-Aligned: A Massive Collection of Cross-Lingual Web Documents

Motivation

Creating a large cross-lingual parallel document dataset can be valuable

- High-quality multilingual dataset can be used to benchmark document alignment algorithms
- Parallel dataset can be used for supervision for cross-lingual representation
- A large parallel dataset can be mined for parallel sentences for NMT training

Insights

URL Signals for Parallel Web Documents

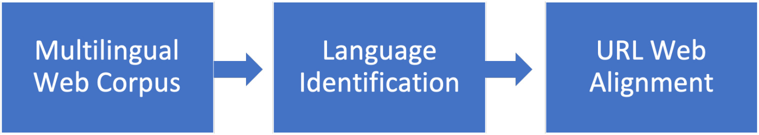
- URLs often contain language codes signifying the language a piece of web content is in
- URL structural information can be used as a signal for identifying parallel documents
  - `https://anonymizedURL.com`
  - `https://fr-fr.anonymizedURL.com`

Insights

Source URL	Target URL
eng.aaa.com	aaa.com
aaa.com/**en-gb**/b	aaa.com/**zh-cn**/b
aaa.com/**English**/b	aaa.com/**Yoruba**/b
aaa.com/b/**en**	aaa.com/b/**vi**
aaa.com/b/	**thai.**aaa.com/b/
aaa.com/b**&lang=english**	aaa.com/b**&lang=arabic**
aaa.com/b**?lang=en**	aaa.com/b**?lang=fr**
aaa.com/b	aaa.com/b**?lang=1**

Table: URL matching via language identifiers.

CCAligned Dataset



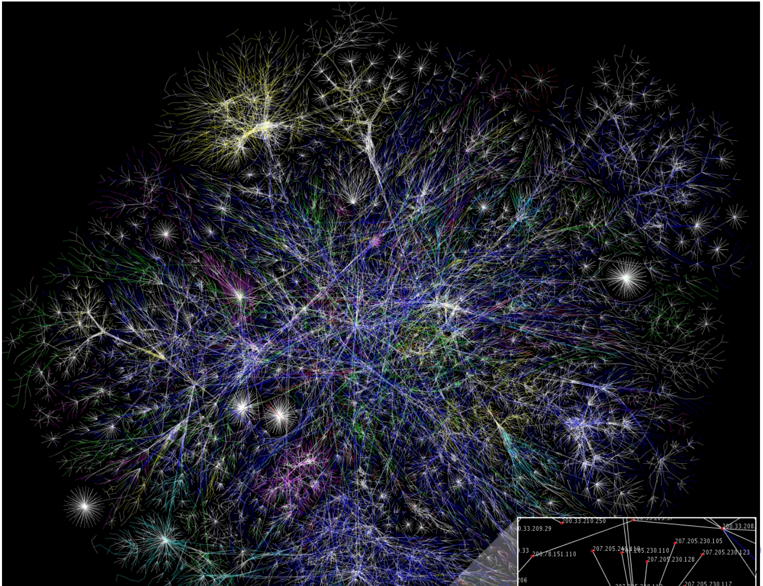
Multilingual Web Corpus

Language Identification

URL Web Alignment

Common Crawl Corpus

CommonCrawl Corpus: An Open Repository of Web Data



Common Crawl Corpus

CommonCrawl Corpus: An Open Repository of Web Data

- Text content
- Status information
- HTTP response code
- HTML title
- HTML meta tags
- RSS/Atom information
- All anchors/hyperlinks

Corpus Statistics

Corpus Statistics

- 68 CommonCrawl Snapshots (every month 2013-2020)
- Each snapshot contains over 2 billion web-documents
- 169.4 billion web documents
- 107.8 million distinct web-domainis

Preprocessing Steps

URL Normalization

- URL Normalization: removing the protocol and host name
  - `https://www.aaa.com` → `aaa.com`)
- Deduplicate based on normalized URL
  - URL that appears more than once, we select the instance that possesses the longest document content.
    1. document was deleted and gets shorter
    2. document is amended and gets longer.

De-duplication Corpus Statistics

- 169.4 billion documents → 29.6 billion
- 83% reduction from raw corpus
- 107.8 million distinct web-domains

Mining Parallel Documents

The next step is to mine parallel web documents.

1. De-duplicate CommonCrawl corpus
2. Perform language identification on each web-document.
3. Apply URL-Matching Heuristics

Mined Parallel Documents

Parallel Cross-lingual Documents

1. 364 million aligned documents
   - 100M with English
   - 264M without English
2. 4598 language pairs
   - 98 with English
   - 4500 without English

CCAligned Dataset Quality

Human annotators evaluated quality of the mined documents

	Language	**P**$_{maj}$	**K**α	**P**$_{adj}$
High	German	90.0	0.74	96.7
	Chinese	86.7	0.68	93.3
Mid	Arabic	83.3	0.72	90.0
	Romanian	76.7	0.50	96.7
Low	Estonian	83.3	0.68	90.0
	Burmese	86.7	0.88	100.0
	Avg	84.4	0.70	94.5

Dataset Analysis

Dataset Analysis

- High-precision collection of cross-lingual documents
- Dataset was constructed using **ONLY** URL-features
- Can one evaluate content-based alignment strategies on this dataset?

Content-based Alignment: Direct Embedding

Direct Embedding (DE) with LASER

- Embed the entire document using LASER embedding
- Each document d has its dense vector representation $\mathbf{v}_d$

Content-based Alignment: Sentence Average Embedding

Sentence Averaging (SA) with LASER

1. Decompose each document into sentences
2. Embed each sentence using LASER
3. document embedding by averaging the sentence vectors into a document vector $\mathbf{v}_d$

$$\mathbf{v}_d = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_{s_i} \tag{2}$$

Content-based Alignment: Weighted Sentence Average Embedding

Weighted Sentence Averaging (WSA) with LASER. Try common information retrieval tricks

1. Sentence Length (SL): Longer sentences more important than shorter
2. Inverse Document Frequency (IDF): More frequent sentences may be unimportant
3. SL-IDF: Combine both

Content-based Alignment: Weighted Sentence Average Embedding

Weighted Sentence Averaging (WSA) with LASER

$$\mathbf{v}_d = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_{s_i} \times \mathbf{v}_{s_i} \tag{3}$$

$$SL_{s_i} = \frac{|s_i|}{\sum_{s \in d} count(s) \times |s|} \tag{4}$$

$$IDF_{s_i} = \log \frac{N+1}{1 + |\{d \in D : s \in d\}|} \tag{5}$$

$$SLIDF_{s_i} = SL_{s_i} \times IDF_{s_i} \tag{6}$$

Scoring Function

Cross-lingual Document Similarity

- dense document representations for each document from the source and target sets
- score pairs to evaluate how semantically similar documents are
- given two documents *a* and *b*, compute their semantic similarity using a cosine similarity

$$sim(a,b) = \frac{\mathbf{v}_a \cdot \mathbf{v}_b}{||\mathbf{v}_a||\ ||\mathbf{v}_b||} \tag{7}$$

Competitive Matching

Ensuring the pairs are 1-to-1 (each aligned document is in at most one pair)

- each document in the source document set, D_s is paired with each document in the target set, D_t
- $D_s \times D_t$ scored pairs – a fully connected bipartite graph
- expected output assumes each page in the non-dominant language has a translated or comparable counterpart
- $min(|D_s|, |D_t|)$ expected number of aligned pairs
- Hungarian algorithm $\mathcal{O}(max(|D_s||D_t|)^3)$ ... intractable

Competitive Matching

Algorithm 1: Competitive Matching

```
Input: $P = \{(d_s, d_t) | d_s \in D_s, d_t \in D_t\}$
Output: $P' = \{(d_{s,i}, d_{t,i}), \ldots\} \subset P$
$scored \leftarrow \{(p, score(p))$ for $p \in P\}$
$sorted \leftarrow sort(scored)$ in descending order
aligned $\leftarrow \varnothing$
$S_s \leftarrow \varnothing$
$S_t \leftarrow \varnothing$
for $d_s, d_t \in sorted$ do
    if $d_s \notin S_s \wedge d_t \notin S_t$
    $aligned \leftarrow aligned \cup \{(d_s, d_t)\}$
    $S_s \leftarrow S_s \cup d_s$
    $S_t \leftarrow S_t \cup d_t$
end
return aligned
```

Alignment Results High-Resource

	Recall				
Language	**DE**	**SA**	**SL**	**IDF**	**SLIDF**
French	0.39	**0.84**	0.83	0.82	**0.84**
Spanish	0.34	0.53	0.55	**0.58**	0.57
Russian	0.06	0.48	0.50	**0.61**	0.60
German	0.52	0.74	**0.76**	0.74	**0.76**
Italian	0.22	0.54	0.55	0.55	**0.57**
Portuguese	0.17	0.36	0.39	0.33	**0.40**
Dutch	0.28	0.51	0.54	0.52	**0.56**
Indonesian	0.11	0.36	**0.48**	0.43	**0.48**
Polish	0.17	0.38	0.41	**0.44**	0.42
Turkish	0.12	0.30	0.34	**0.45**	0.41
Swedish	0.19	0.37	0.37	0.38	**0.39**
Danish	0.27	0.46	0.65	0.60	**0.67**
Czech	0.15	0.36	**0.41**	0.32	**0.41**
Bulgarian	0.07	0.34	0.37	0.40	**0.44**
Finnish	0.06	0.24	0.32	0.43	**0.44**
Norwegian	0.13	0.26	0.33	0.33	**0.38**
Macro-AVG	0.20	0.41	0.45	0.47	**0.49**

Alignment Results Mid-Resource

	Recall				
Language	**DE**	**SA**	**SL**	**IDF**	**SLIDF**
Romanian	0.15	0.39	0.40	0.40	**0.41**
Vietnamese	0.06	0.13	0.18	0.15	**0.23**
Ukrainian	0.05	0.49	0.70	0.70	**0.74**
Greek	0.05	0.22	0.24	**0.34**	0.30
Korean	0.06	0.49	0.47	0.49	**0.51**
Arabic	0.04	0.26	0.46	0.42	**0.51**
Croatian	0.16	0.32	0.36	0.34	**0.36**
Slovak	0.20	0.37	**0.44**	0.41	0.42
Thai	0.02	0.15	0.28	0.19	**0.35**
Hebrew	0.05	0.19	0.30	0.27	**0.33**
Hindi	0.04	0.03	0.33	0.28	**0.43**
Hungarian	0.15	0.41	0.39	0.39	**0.46**
Lithuanian	0.11	0.61	0.72	0.74	**0.80**
Slovenian	0.13	0.20	0.26	0.31	**0.33**
Farsi	0.06	0.22	0.37	0.40	**0.49**
Macro-AVG	0.09	0.28	0.39	0.39	**0.44**

Alignment Results Low-Resource

	Recall				
Language	**DE**	**SA**	**SL**	**IDF**	**SLIDF**
Estonian	0.28	0.57	0.62	0.58	**0.64**
Bengali	0.05	0.47	**0.59**	0.51	0.58
Albanian	0.23	0.56	0.60	0.57	**0.61**
Macedonian	0.02	0.16	**0.22**	0.19	0.08
Urdu	0.06	**0.29**	0.23	0.27	0.24
Serbian	0.06	0.46	**0.58**	0.47	0.56
Azerbaijani	0.08	0.27	0.28	**0.34**	0.27
Armenian	0.02	0.08	0.13	0.12	**0.17**
Belarusian	0.07	0.26	0.44	0.36	**0.51**
Georgian	0.06	0.18	0.23	**0.25**	**0.25**
Tamil	0.02	0.13	0.19	0.23	**0.34**
Marathi	0.02	0.13	**0.20**	0.10	0.16
Kazakh	0.05	0.16	0.24	0.25	**0.33**
Mongolian	0.03	0.01	0.05	0.10	**0.22**
Burmese	0.01	**0.35**	0.18	0.08	0.26
Bosnian	0.18	0.49	0.64	0.50	**0.65**
Macro-AVG	0.08	0.29	0.34	0.31	**0.37**

Downstream Mining

From the aligned documents, can do further sentence-level mining.

- From this dataset, mined **2.25** billion parallel sentences covering 4598 language pairs
- 950 million pairs are sentences paired with English sentences
- 1.3 billion pairs are non-English sentence pairs

Follow-up Research

From CCAligned aligned documents, there are many open research problems that can leverage this data

- Mine more, higher quality parallel sentences from the CCAligned documents
- Use CCAligned documents as supervision for supervised document alignment (mine parallel documents using high-recall method)
- Leverage parallel documents to learn cross-linigual document representations and cross-lingual document retrieval

Cross-lingual Sentence Mover's Distance

Massively Multilingual Document Alignment with Cross-lingual Sentence Mover's Distance

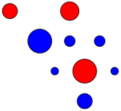
Motivation & Insight

- Motivation
  - Cross-lingual retrieval based on content is more general than using metadata (URL, timestamp, etc)
  - CCAligned is high-precision. For more training data (especially for low-resource direction, need a high-recall approach)
- Insight
  - Creating document level fixed representations may be destructive for variable-length documents.
  - Averaging sentence embedding places equal importance to all sentences
  - How well sentences match up between document pairs is a good signal for parallel documents

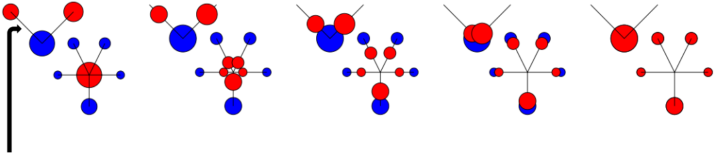
Earth Mover's Distance

- measure of the distance between two probability distributions over a region *D*
- For example: if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region D
  - the EMD is the minimum cost of turning one pile into the other
  - the cost is assumed to be amount of dirt moved times the distance by which it is moved

Earth Mover's Distance



- red distribution: “dirt”
- blue distribution: “holes”



The distance between points (ground distance) can be Euclidean distance, Manhattan...

Cross-lingual Sentence Mover's Distance

- Each document has a distribution over sentences
  - multinomial distribution - normalize bag of sentences (nBOS)
- Euclidean distance between source document sents and target document sents
  - Leverage LASER embeddings to compute Euclidean distances

Weighting Sentences Based on Importance

- XL-SMD requires a distribution over sentences for each document
- Each sentence has probability mass allocated to it.
- 4weighting schemes for each sentence investigated
  - Uniform weighting (each sentence equally weighted)
  - Sentence length (Longer sentences = more mass)
  - Inverse document frequency (IDF)
  - SL-IDF

Document Mass Normalization

Normalizing the mass to unit measure in both the source and target documents each each document has a legitimate distribution and the induced distance metric is valid.

$$d'_{A,i} = \frac{d_{A,i}}{\sum_{s \in A} d_{A,s}} \tag{8}$$

Optimal Transport

- $\Delta(i, j)$ is distance between the i_{th} and j_{th} sentences
- V denote vocab size (sentences within a document pair)
- $\Delta(i, j) = ||v_i - v_j||$

$$XLSMD(A, B) = \underset{T \geq 0}{min} \sum_{i=1}^{V} \sum_{j=1}^{V} T_{i,j} \times \Delta(i, j) \tag{9}$$

subject to:

$$\forall i \sum_{j=1}^{V} T_{i,j} = d_{A,i}$$

$$\forall j \sum_{i=1}^{V} T_{i,j} = d_{B,j}$$

Where $T \in R^{V \times V}$ is a nonnegative matrix, where each $T_{i,j}$ denotes how much of sentence i in document A is assigned to sentences j in document B, and constraints ensure the flow of a given sentence cannot exceed its allocated mass.

Greedy Mover's Distance

Solving the optimal transport problem is of cubic complexity and slow. Can it be approximated?

- find the two closest sentences and moves as much mass between the two sentences as possible
- the algorithm moves to the next two closest pairs
- terminates when all mass has been moved between the source and target document
- maintains mass constraints

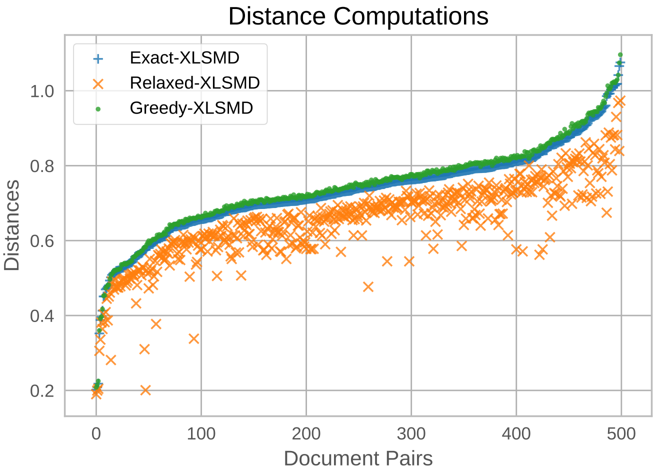
Approximation Performance

How does this approximation compare to the exact?

Method	Kendall-Tau	Recall	MAE	Runtime (s)
Exact-XLSMD	1.00	0.69	0.000	0.402
Relaxed-XLSMD	0.70	0.58	0.084	0.031
Greedy-XLSMD	0.98	0.69	0.010	0.107

Table: Comparing exact XLSMD computation to approximation schemes for computing XLSMD on 10 webdomains.

Approximate Distances



Approximate Performance

Which approximate computation works better?

Approximation	**Low**	**Mid**	**High**	**All**
Relaxed-XLSMD	0.44	0.43	0.50	0.46
Greedy-XLSMD	0.54	0.50	0.56	0.54

Table: Document alignment performance of fast methods for approximating the same variant of XLSMD.

Alignment Results High-Resource

	Recall					
Language	**DE**	**SA**	**SMD**	**SL**	**IDF**	**SLIDF**
French	0.39	0.84	0.81	0.84	0.83	**0.85**
Spanish	0.34	0.53	0.59	0.63	0.62	**0.64**
Russian	0.06	0.64	0.69	0.69	0.70	**0.71**
German	0.52	0.74	**0.78**	0.76	0.77	0.77
Italian	0.22	0.47	0.55	0.56	0.56	**0.59**
Portuguese	0.17	0.36	0.39	**0.41**	0.38	0.40
Dutch	0.28	0.49	0.54	0.54	0.54	**0.56**
Indonesian	0.11	0.47	0.49	0.52	0.51	**0.53**
Polish	0.17	0.38	0.45	0.45	**0.46**	**0.46**
Turkish	0.12	0.38	0.52	0.56	0.57	**0.59**
Swedish	0.19	0.40	0.44	0.44	**0.46**	0.45
Danish	0.27	0.62	0.63	**0.69**	0.65	**0.69**
Czech	0.15	0.40	0.43	**0.44**	**0.44**	0.43
Bulgarian	0.07	0.43	0.52	0.54	**0.55**	0.52
Finnish	0.06	0.47	0.51	0.51	**0.54**	0.52
Norwegian	0.13	0.33	0.37	0.39	**0.42**	0.41
AVG	0.20	0.50	0.54	0.56	0.56	**0.57**

Alignment Results Mid-Resource

	Recall					
Language	**DE**	**SA**	**SMD**	**SL**	**IDF**	**SLIDF**
Romanian	0.15	0.40	0.44	0.43	**0.45**	0.43
Vietnamese	0.06	0.28	0.29	0.29	0.29	**0.32**
Ukrainian	0.05	0.68	0.67	0.78	0.78	**0.82**
Greek	0.05	0.31	0.47	0.48	**0.49**	**0.49**
Korean	0.06	0.34	0.60	0.54	**0.61**	0.60
Arabic	0.04	0.32	0.63	0.59	**0.65**	0.61
Croatian	0.16	0.37	0.40	0.40	**0.41**	0.40
Slovak	0.20	0.41	0.46	**0.46**	**0.46**	0.44
Thai	0.02	0.19	0.41	0.33	**0.47**	0.41
Hebrew	0.05	0.18	0.39	**0.43**	0.41	0.41
Hindi	0.04	0.27	0.34	**0.54**	0.52	0.53
Hungarian	0.15	0.49	0.50	**0.54**	0.51	**0.54**
Lithuanian	0.11	0.73	0.79	0.79	**0.80**	**0.80**
Slovenian	0.13	0.33	0.34	0.35	**0.36**	**0.36**
Persian	0.06	0.32	0.56	0.57	0.53	**0.59**
AVG	0.09	0.37	0.49	0.50	**0.52**	**0.52**

Alignment Results Low-Resource

	Recall					
Language	**DE**	**SA**	**SMD**	**SL**	**IDF**	**SLIDF**
Estonian	0.28	0.52	0.69	0.66	**0.74**	0.72
Bengali	0.05	0.32	0.78	0.72	0.77	**0.79**
Albanian	0.23	0.56	**0.66**	0.65	0.65	**0.66**
Macedonian	0.02	0.33	0.32	0.36	**0.38**	0.33
Urdu	0.06	0.22	**0.60**	**0.60**	0.49	0.56
Serbian	0.06	0.59	**0.75**	0.74	0.74	0.71
Azerbaijani	0.08	0.34	0.74	0.74	**0.75**	0.74
Armenian	0.02	0.18	0.32	0.35	0.34	**0.38**
Belarusian	0.07	0.47	0.67	0.69	**0.73**	0.71
Georgian	0.06	0.24	0.46	**0.48**	0.45	0.45
Tamil	0.02	0.20	0.51	0.45	0.51	**0.53**
Marathi	0.02	0.11	0.43	**0.46**	0.33	0.39
Kazakh	0.05	0.31	0.44	**0.46**	0.45	0.45
Mongolian	0.03	0.13	0.18	0.22	0.21	**0.23**
Burmese	0.01	0.10	0.26	0.33	**0.46**	**0.46**
Bosnian	0.18	0.64	0.61	0.69	0.65	**0.72**
AVG	0.08	0.33	0.53	0.54	0.54	**0.55**

WMT 2016 Shared Task

WMT 2016 Shared Task

WMT 2016 Shared Task: Challenges

Big-ish websites

- E.g. cinedoc.org: 50k English, 50k French pages
- Makes 2.5B possible pairs
- Only allowed to pick 50k

Language detection unreliable

- Made sure test set can be found
- Some participants ran their own pipelines

WMT 2016 Shared Task: Challenges II

Near duplicates

- Removed pages when text was exactly the same
- www.taize.fr/fr article10921.html
- www.taize.fr/fr article10921.html?chooselang=1
- Almost identical

Submissions

- 11 participating groups
- 19 submissions
- Up to 95% recall (NovaLincs-URL-Coverage)

NovaLincs

- use a phrase table from a phrase-based statistical machine translation system to compute coverage scores
- based on the ratio of phrase pairs covered by a document pair.
- NOVALINCS-COVERAGE (88.6%)
- NOVALINCSCOVERAGE-URL (85.8%) coverage first then URL
- NOVALINCS-URL-COVERAGE (95.0%) URL first then coverage

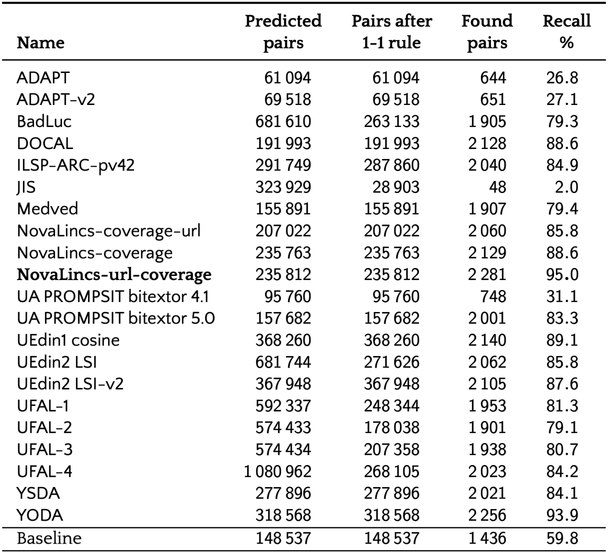
YODA

- uses the machine translation of the French document, and finds the English corresponding document based on bigram and 5-gram matches, assisted by a heuristics based on document length ratio
- YODA: (93.9%)

UEdin1

- uses cosine similarity between tf/idf weighted vectors, extracted by collecting n-grams from the English and machine translated French text. compare many hyperparameters such as weighting schemes and two pair selection algorithms.
- UEdin1: (89.1%)

Submission Results



Name	Predicted pairs	Pairs after 1-1 rule	Found pairs	Recall %
ADAPT	61 094	61 094	644	26.8
ADAPT-v2	69 518	69 518	651	27.1
BadLuc	681 610	263 133	1 905	79.3
DOCAL	191 993	191 993	2 128	88.6
ILSP-ARC-pv42	291 749	287 860	2 040	84.9
JIS	323 929	28 903	48	2.0
Medved	155 891	155 891	1 907	79.4
NovaLincs-coverage-url	207 022	207 022	2 060	85.8
NovaLincs-coverage	235 763	235 763	2 129	88.6
NovaLincs-url-coverage	235 812	235 812	2 281	95.0
UA PROMPSIT bitextor 4.1	95 760	95 760	748	31.1
UA PROMPSIT bitextor 5.0	157 682	157 682	2 001	83.3
UEdin1 cosine	368 260	368 260	2 140	89.1
UEdin2 LSI	681 744	271 626	2 062	85.8
UEdin2 LSI-v2	367 948	367 948	2 105	87.6
UFAL-1	592 337	248 344	1 953	81.3
UFAL-2	574 433	178 038	1 901	79.1
UFAL-3	574 434	207 358	1 938	80.7
UFAL-4	1 080 962	268 105	2 023	84.2
YSDA	277 896	277 896	2 021	84.1
YODA	318 568	318 568	2 256	93.9
Baseline	148 537	148 537	1 436	59.8

Figure: Documents are aligned 1-to-1 within each domain.

Submission Results: Fuzzy Matching

Allowing 5% edits between predicted and expected

Name	Pairs found	Δ	Recall	Δ	Rank	Δ
ADAPT	726	+82	30.2	+3.4	20	0
ADAPT-v2	733	+82	30.5	+3.4	19	0
BadLuc	2 062	+157	85.9	+6.5	13	+3
DOCAL	2 235	+107	93.1	+4.5	4	+1
ILSP-ARC-pv42	2 185	+145	91.0	+6.0	7	+2
JIS	48	0	2.0	0.0	21	0
Medved	1 986	+79	82.7	+3.3	15	0
NovaLincs-coverage-url	2 130	+70	88.7	+2.9	9	−1
NovaLincs-coverage	2 192	+63	91.3	+2.6	6	−2
NovaLincs-url-coverage	2 303	+22	95.9	+0.9	2	−1
UA PROMPSIT bitextor 4.1	775	+27	32.3	+1.1	18	0
UA PROMPSIT bitextor 5.0	2 117	+116	88.1	+4.8	10	+2
UEdin1 cosine	2 227	+87	92.7	+3.6	5	−2
UEdin2 LSI	2 146	+84	89.3	+3.5	8	−1
UEdin2 LSI-v2	2 281	+176	95.0	+7.3	3	+3
UFAL-1	2 060	+107	85.8	+4.5	14	−1
UFAL-2	1 954	+53	81.4	+2.2	17	0
UFAL-3	1 980	+42	82.4	+1.8	16	−2
UFAL-4	2 078	+55	86.5	+2.3	12	−2
YSDA	2 102	+81	87.5	+3.4	11	0
YODA	2 307	+51	96.0	+2.1	1	+1

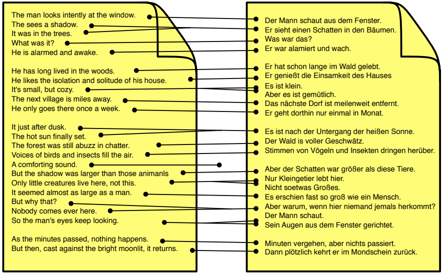
Figure: Documents are aligned 1-to-1 within each domain.

Shared Task Insights

- Machine translated text helpful
- Finding matching n-grams works well
- Big boost by combination with URL-matching baseline

sentence alignment

Sentence Alignment



The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

He has long lived in the woods.
He likes the isolation and solitude of his house.
It's small, but cozy.
The next village is miles away.
He only goes there once a week.

It just after dusk.
The hot sun finally set.
The forest was still abuzz in chatter.
Voices of birds and insects fill the air.
A comforting sound.
But the shadow was larger than those animanls
Only little creatures live here, not this.
It seemed almost as large as a man.
But why that?
Nobody comes ever here.
So the man's eyes keep looking.

As the minutes passed, nothing happens.
But then, cast against the bright moonlit, it returns.

Der Mann schaut aus dem Fenster.
Er sieht einen Schatten in den Bäumen.
Was war das?
Er war alamiert und wach.

Er hat schon lange im Wald gelebt.
Er genießt die Einsamkeit des Hauses
Es ist klein.
Aber es ist gemütlich.
Das nächste Dorf ist meilenweit entfernt.
Er geht dorthin nur einmal in Monat.

Es ist nach der Untergang der heißen Sonne.
Der Wald is voller Geschwätz.
Stimmen von Vögeln und Insekten dringen herüber.

Aber der Schatten war größer als diese Tiere.
Nur Kleingetier lebt hier.
Nicht soetwas Großes.
Es erschien fast so groß wie ein Mensch.
Aber warum, wenn hier niemand jemals herkommt?
Der Mann schaut.
Sein Augen aus dem Fenster gerichtet.

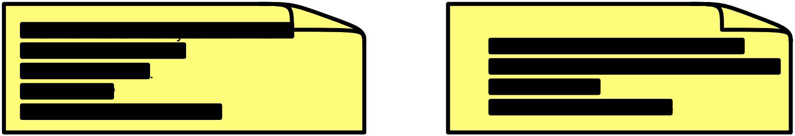
Minuten vergehen, aber nichts passiert.
Dann plötzlich kehrt er im Mondschein zurück.

"Local" alignment: limited to document pairs

- given: document pair
- output: matching sentence pairs

We also respect the order of sentences.

Church and Gale (1993)

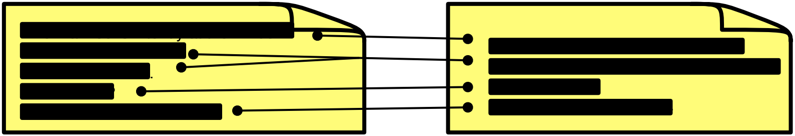


- Consider only the lengths of the sentences

$$\text{abs}(\log \frac{\text{length}_e}{\text{length}_f})$$

- Find the Viterbi path that with the best length ratios
- Additional cost factors for alignments other than 1-1

Church and Gale (1993)

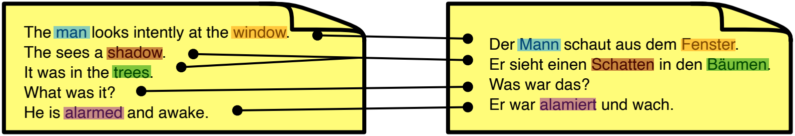


- Consider only the lengths of the sentences

$$\text{abs}(\log \frac{\text{length}_e}{\text{length}_f})$$

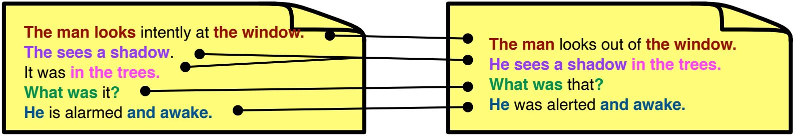
- Find the Viterbi path that with the best length ratios
- Additional cost factors for alignments other than 1-1

Use of Dictionaries


The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.
Der Mann schaut aus dem Fenster.
Er sieht einen Schatten in den Bäumen.
Was war das?
Er war alamiert und wach.

- Given a word translation dictionary man = Mann; window = Fenster; shaddow = Schatten; trees = Bäumen; alarmed = alamiert
- Find matching word pairs
- Score sentence pairs based on number of matches
- **Hunalign:** (Varga et al., 2005) tool using this feature
- **Gargantua:** unsupervised induction of translation dictionary

Translate and Match



The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

The man looks out of the window.
He sees a shadow in the trees.
What was that?
He was alerted and awake.

- Use machine translation to translate foreign sentences
- Match translation with English
- Use of standard machine translation metric to assess match: BLEU score
- **Bleualign** (Sennrich and Volk, 2010)

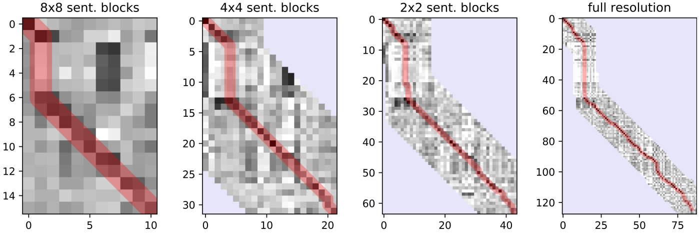
Use of Sentence Embeddings

- Multilingual sentence embeddings, e.g., LASER
- Sentences with similar meaning have similar embedding — **independent** of language
- Comparison based on Cosine distance

$$c(x,y) = \frac{(1-\cos(x,y))\ N(x)\ N(y)}{\sum_{s=1}^{S} 1-\cos(x,y_s) + \sum_{s=1}^{S} 1-\cos(x_s,y)}$$

- **Vecalign** (Thompson and Koehn, 2019)

Efficient Algorithm



- Complexity of alignment via dynamic programming: $O(n^2)$
- Coarse to fine algorithm: $O(n)$
  embedding for block = average of sentence embeddings

Evaluation: Text + Bild

Algorithm	O( )	P	R	F_1
Gargantua	N^2	0.48	0.54	0.51
Hunalign w/o lexicon	**N**	0.59	0.70	0.64
Hunalign w/ lexicon	**N**	0.61	0.73	0.66
Church and Gale	N^2	0.71	0.72	0.72
Moore	‡	0.86	0.71	0.78
Bleualign	N^2	0.83	0.78	0.81
Bleualign-NMT	N^2	0.85	0.83	0.84
Coverage-Based	N^2	0.85	0.84	0.85
Vecalign	**N**	**0.89**	**0.90**	**0.90**

‡O( ) is data dependent

Evaluation: Bible

	Verse-level F_1	
Languages	Vecalign	Hunalign
Afrikaans–Arabic	**0.863**	0.339
Afrikaans–Tagalog	**0.922**	0.775
Arabic–Norwegian	**0.787**	0.406
Arabic–Somali	**0.634**	0.067
Turkish–Somali	**0.533**	0.331
Norwegian–Somali	**0.697**	0.687
Somali–Afrikaans	**0.782**	0.738
Tagalog–Norwegian	**0.874**	0.764
Turkish–Afrikaans	**0.703**	0.401
Turkish–Tagalog	**0.647**	0.247

Evaluation: CommonCrawl

Language Pair	LASER-only	Vecalign (best setup)
English–Portuguese	31.5	32.9 (+1.4)
Portuguese–English	36.0	38.8 (+2.8)
English–Bulgarian	29.6	32.6 (+3.0)
Bulgarian–English	20.6	22.3 (+1.7)
English–Estonian	14.0	15.0 (+1.0)
English–Georgian	8.6	9.1 (+0.5)
English–Urdu	10.9	12.5 (+1.6)
English–Marathi	10.0	10.3 (+0.3)
English–Burmese	8.0	9.0 (+1.0)

Evaluation: Paracrawl

- Task: align sentences in document pairs (subset of ParaCrawl data)

Language	**Web Domains**	**Document Pairs**	**English Tokens**
German	21,806	17,109,018	10,788,923,009
Czech	12,179	6,661,650	4,089,806,440
Hungarian	5,560	2,770,432	1,504,698,348
Estonian	5,129	2,301,309	1,427,328,440
Maltese	933	303,198	134,232,546

Evaluation: Paracrawl

- Results: BLEU scores for best subset (English token count)

Language	**Hunalign**		**Vecalign**		**Bleualign**	
German	35.1	(100m)	**35.8**	**(150m)**	35.0	(100m)
Czech	21.0	(50m)	**21.2**	**(50m)**	21.0	(50m)
Hungarian	16.5	(30m)	**16.8**	**(30m)**	16.6	(15m)
Estonian	**21.8**	**(20m)**	21.6	(20m)	21.4	(20m)
Maltese	33.5	(5m)	**34.1**	**(7m)**	30.3	(2m)

- Best results with Vecalign, except for Estonian

global sentence alignment

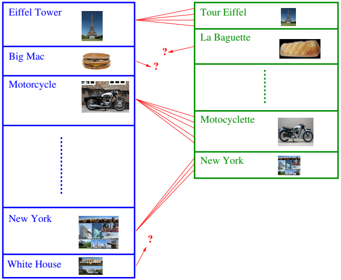
Bitext Mining in Wikipedia

Some statistics

- 300 different languages
- Huge differences in size:
  - 1M+ articles: 15 languages
    (major European languages, ru, vi, ja, zh)
  - 100k+ articles: 47 languages
  - 10k+ articles: 81 languages
  - long tail . . .
- English by far the biggest (5.8M articles, 208M sentences)
- Cebuano has many articles produced by a bot
  (5.4M articles, 67M sentences)

Bitext Mining in Wikipedia

Local mining



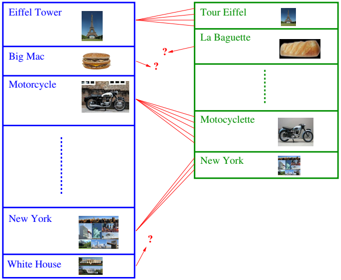
- Only articles with link

+ Seems logical
+ Very fast

– Ignored articles
– Many simple sentences

Bitext Mining in Wikipedia

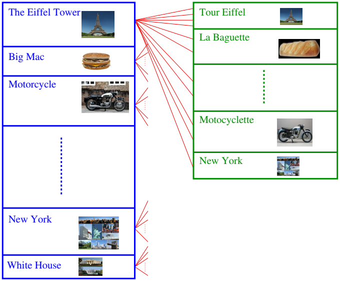
Local mining



- Only articles with link

+ Seems logical
+ Very fast
– Ignored articles
– Many simple sentences

Global mining



- Always consider all sent.

– Increased complexity
± Lower recall ?
+ Generic: any corpus

Bitext Mining in Wikipedia

Global mining

- Compare **all** sentences of two Wikipedia
- - Computationally more challenging: 134M×51M distances
- + Ability to handle two languages even though there are only few articles in common
- + Margin criterion:
  excludes short sentences which differ in NE only
- - Potentially increased risk of misalignment and a lower recall

We chose global mining for this study (more generic)

Bitext Mining in Wikipedia

Processing pipeline

- Sentence splitting (very difficult for Thai)
- Deduplication
- Language identification with fasttext

→ $\approx 600M$ sentences for > 180 languages
(each with more than 50k sentences)

Bitext Mining in Wikipedia

Processing pipeline

- Sentence splitting (very difficult for Thai)
- Deduplication
- Language identification with fasttext

→ $\approx 600M$ sentences for > 180 languages (each with more than 50k sentences)

Complexity issues

- English/German Wikipedia:
  - 134M × 51M sentences
  - $513 + 204$GB memory to store LASER embeddings
  - 6.8×10^{15} distance calculations

$\Rightarrow$ Optimization and compression are needed !!

Efficient Mining with FAISS

FAISS library

- Library for efficient similarity search and clustering of dense vectors
- `https://github.com/facebookresearch/faiss`
- Mainly used for indexing images
  but can operate on any arbitrary vectors

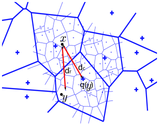
⇒ Used here for efficient large-scale bitext mining

- Can be scaled to search in billions of sentences

Efficient Mining with FAISS

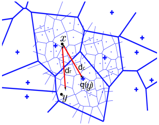
FAISS index types

- Define *N* Voronoi cells
- Quantizers:
  - PCA, not enough compression
  - Product: `OPQ64,IVF32768,PQ64`, 55x compression
  - Scalar: `PCAR128,IVF32768,SQ8`, 28x compression



Efficient Mining with FAISS

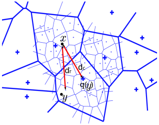
FAISS index types



- Define N Voronoi cells
- Quantizers:
  - PCA, not enough compression
  - Product: `OPQ64,IVF32768,PQ64`, 55x compression
  - Scalar: `PCAR128,IVF32768,SQ8`, 28x compression
- English FAISS index: 9.2GB

Efficient Mining with FAISS

FAISS index types



- Define *N* Voronoi cells
- Quantizers:
  - PCA, not enough compression
  - Product: `OPQ64,IVF32768,PQ64`, 55x compression
  - Scalar: `PCAR128,IVF32768,SQ8`, 28x compression
- English FAISS index: 9.2GB
- English/German mining: 3h30 on 8 GPUS

Efficient Mining with FAISS

Overall complexity

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
    - a couple of hours, run on parallel on standard server

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
  - a couple of hours, run on parallel on standard server
- Sentence embeddings with LASER (>7M sents/h)
  - total of 100h, can be run in parallel on cluster

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
  - a couple of hours, run on parallel on standard server
- Sentence embeddings with LASER ($>$7M sents/h)
  - total of 100h, can be run in parallel on cluster
- Train and create FAISS index for each language (CPU)
  - English $\approx$ 4h, total 21h

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
    - a couple of hours, run on parallel on standard server
- Sentence embeddings with LASER (>7M sents/h)
    - total of 100h, can be run in parallel on cluster
- Train and create FAISS index for each language (CPU)
    - English ≈ 4h, total 21h
- Mine bitext for each language pair
    - total of ≈ 1000h

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
    - a couple of hours, run on parallel on standard server
- Sentence embeddings with LASER (>7M sents/h)
    - total of 100h, can be run in parallel on cluster
- Train and create FAISS index for each language (CPU)
    - English ≈ 4h, total 21h
- Mine bitext for each language pair
    - total of ≈ 1000h

⇒ **Total of 43 days on one GPU**

Efficient Mining with FAISS

Overall complexity

- Deduplication and LID
  - a couple of hours, run on parallel on standard server
- Sentence embeddings with LASER (>7M sents/h)
  - total of 100h, can be run in parallel on cluster
- Train and create FAISS index for each language (CPU)
  - English ≈ 4h, total 21h
- Mine bitext for each language pair
  - total of ≈ 1000h

⇒ **Total of 43 days on one GPU**

or much less on many GPUs . . .

Threshold Optimization

Precision/recall trade-off

- Margin-based mining has only one parameter:
  the margin between the closest and the average distance
    - large margin: high precision, low recall
    - small margin: lower precision, higher recall
- We have no gold-alignments to optimize this parameter

Threshold Optimization

Precision/recall trade-off

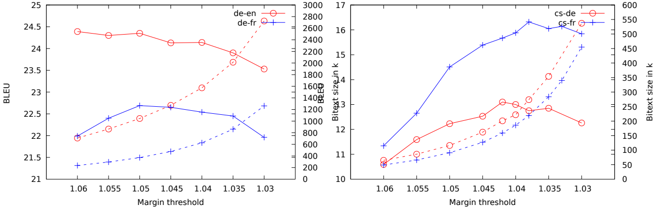
- Margin-based mining has only one parameter: the margin between the closest and the average distance
  - large margin: high precision, low recall
  - small margin: lower precision, higher recall
- We have no gold-alignments to optimize this parameter

Task oriented threshold optimization

- Mine bitexts for thresholds in range [1.01–1.06]
- Train NMT systems for increasing amounts of data
- Evaluate each one and keep best one

Threshold Optimization on Europarl

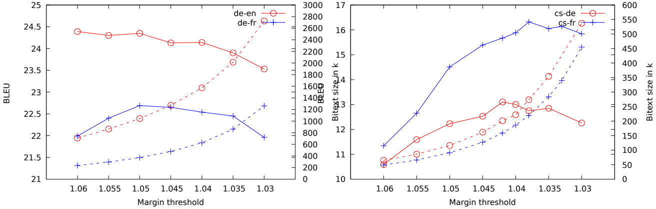
BLEU score for NMT trained on Wikipedia only



Precision/recall trade-off

Threshold Optimization on Europarl

BLEU score for NMT trained on Wikipedia only

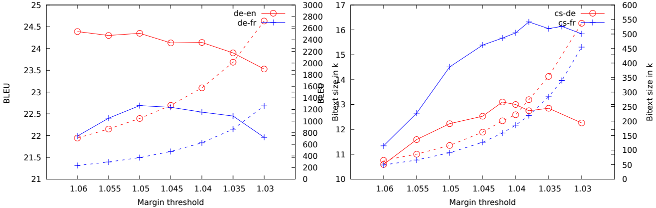


Precision/recall trade-off

- Threshold on margin of 1.04 best for most conditions

Threshold Optimization on Europarl

BLEU score for NMT trained on Wikipedia only



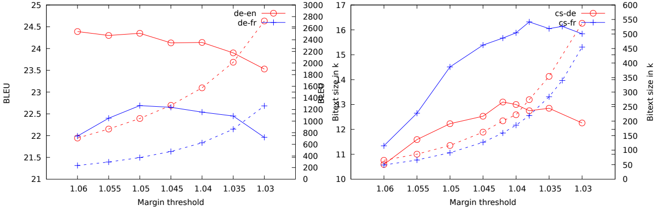
Precision/recall trade-off

- Threshold on margin of 1.04 best for most conditions

Bitexts	de-en	de-fr	cs-de	cs-fr
Mined Wikipedia	1.0M 24.4	372k 22.7	201k 13.1	219k 16.3
Europarl	1.0M 21.2	370k 21.1	200k 12.6	220k 19.2

Threshold Optimization on Europarl

BLEU score for NMT trained on Wikipedia only

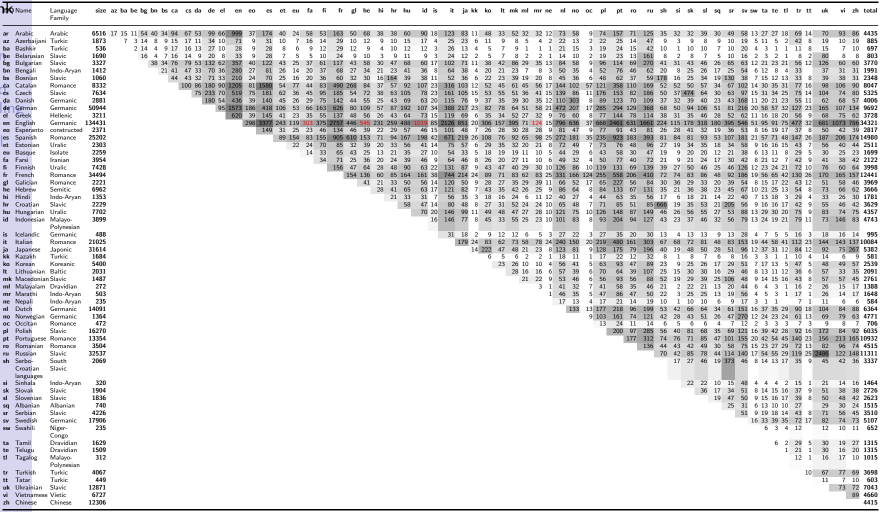


Precision/recall trade-off

- Threshold on margin of 1.04 best for most conditions
- WikiMatrix bitexts outperform Europarl

Bitexts	de-en	de-fr	cs-de	cs-fr
Mined Wikipedia	1.0M 24.4	372k 22.7	201k 13.1	219k 16.3
Europarl	1.0M 21.2	370k 21.1	200k 12.6	220k 19.2

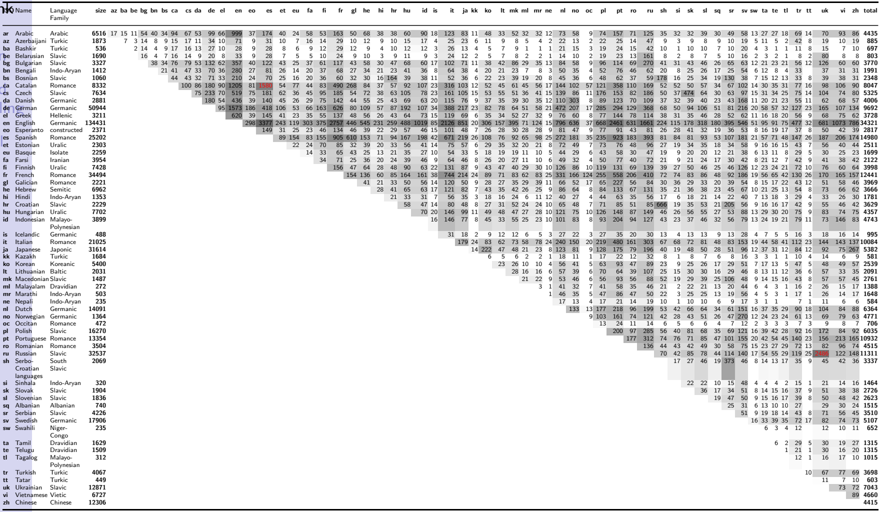
WikiMatrix: 85 Languages, 1620Pairs



ISO	Name	Language Family	size	az	ba	be	bg	bn	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	id	is	it	ja	kk	ko	lt	mk	ml	mr	ne	nl	no	oc	pl	pt	ro	ru	sh	si	sk	sl	sq	sr	sv	sw	ta	te	tl	tr	tt	uk	vi	zh	total
ar	Arabic	Arabic	6516	17	15	11	54	40	34	94	67	53	99	66	999	37	174	40	24	58	53	163	50	68	38	38	60	90	18	123	83	11	48	33	52	32	32	12	73	58	9	74	157	71	125	35	32	32	39	30	49	58	13	27	27	18	69	14	70	93	86	4435
az	Azerbaijani	Turkic	1873		7	3	14	8	9	15	17	11	34	10	71	9	31	10	7	16	14	29	10	12	9	10	15	17	4	25	23	6	11	9	8	5	4	2	22	13	2	22	25	14	47	9	3	9	8	9	10	19	5	11	5	2	42	8	19	10	19	885
ba	Bashkir	Turkic	536			2	14	4	9	17	16	13	27	10	28	9	28	8	6	9	12	29	12	9	4	10	12	12	3	26	13	4	5	7	9	1	1	1	21	15	3	19	24	15	42	10	1	10	10	7	10	20	4	3	1	1	11	8	15	7	10	697
be	Belarusian	Slavic	1690				16	4	7	16	14	9	20	8	33	9	28	7	5	5	10	24	9	10	3	9	11	9	3	24	12	2	5	7	8	2	2	1	14	10	2	19	23	13	161	8	2	8	7	5	10	16	2	3	2	1	8	2	80	8	8	803
bg	Bulgarian	Slavic	3327					38	34	76	79	53	132	62	357	40	122	43	25	37	61	117	43	58	30	47	68	60	17	102	71	11	38	42	86	29	35	13	84	58	9	96	114	69	270	41	31	43	46	26	65	63	12	21	23	21	56	12	126	60	60	3770
bn	Bengali	Indo-Aryan	1412						21	41	47	33	70	36	280	27	81	26	14	20	37	68	27	34	21	23	41	36	8	64	38	4	20	21	23	7	8	3	50	35	4	52	76	46	62	20	8	25	26	17	25	54	6	12	8	4	33		37	31	31	1991
bs	Bosnian	Slavic	1060							44	43	32	71	33	210	24	70	25	16	20	36	60	32	30	16	164	39	38	11	52	36	6	22	23	39	19	20	8	45	36	6	48	62	37	59	178	16	25	34	19	130	38	7	15	12	13	33	8	39	38	31	2348
ca	Catalan	Romance	8332								100	86	180	90	1205	81	1580	54	77	44	83	490	268	84	37	57	92	107	23	316	103	12	52	45	61	45	56	17	144	102	57	121	358	110	169	52	52	50	57	34	67	102	14	30	35	31	77	16	98	106	90	8047
cs	Czech	Slavic	7634									75	233	70	519	75	181	62	36	45	95	185	54	72	38	63	105	78	23	161	105	15	53	55	51	36	41	15	139	86	11	176	153	82	186	50	37	474	64	30	63	97	15	31	34	25	75	14	104	74	80	5325
da	Danish	Germanic	2881										180	54	436	39	140	45	26	29	75	142	44	55	25	43	69	63	20	115	76	9	37	35	39	30	35	12	110	303	8	89	123	70	109	37	32	39	40	23	43	168	11	20	21	23	55	11	62	68	57	4006
de	German	Germanic	50944											95	1573	186	418	106	53	66	163	626	80	109	57	87	192	107	34	388	217	23	82	78	64	51	58	21	472	207	17	285	294	129	368	68	50	94	106	51	81	216	20	58	57	32	127	23	165	107	134	9692
el	Greek	Hellenic	3211												620	39	145	41	23	35	55	137	48	56	26	43	64	73	15	119	69	6	35	34	52	27	32	9	76	60	8	77	144	78	114	38	31	35	46	28	52	62	11	16	18	20	56	9	68	75	62	3728
en	English	Germanic	134431													298	3377	243	119	303	375	2757	446	545	231	259	488	1019	85	2126	851	20	306	157	395	71	124	15	796	636	37	668	2461	631	1661	224	115	178	318	180	395	546	51	95	91	75	477	32	681	1073	786	34321
eo	Esperanto	constructed	2371														149	31	25	23	46	134	46	39	22	29	57	46	15	101	48	7	26	28	30	28	28	9	81	47	9	77	91	43	81	26	28	41	32	19	36	53	8	16	19	17	37	8	50	42	39	2817
es	Spanish	Romance	25202															89	154	83	155	905	610	153	71	94	167	198	42	671	219	26	108	76	92	65	98	25	272	181	35	235	923	183	393	81	84	81	93	53	107	181	21	57	71	48	147	26	187	206	174	14980
et	Estonian	Uralic	2303																22	24	70	85	32	39	20	33	56	41	14	75	57	6	29	35	32	20	21	8	72	49	7	73	76	48	96	27	19	34	35	18	34	58	9	16	16	15	43	7	56	40	44	2511
eu	Basque	Isolate	2259																	14	33	65	43	25	13	21	35	27	10	54	33	5	18	19	19	11	10	5	44	29	6	43	58	30	47	19	9	20	20	12	21	38	6	13	11	8	29	5	30	25	23	1699
fa	Farsi	Iranian	3954																		34	71	25	36	20	24	39	46	9	64	46	8	26	20	27	11	10	6	49	32	4	50	77	40	72	21	9	21	24	17	30	42	8	21	12	7	42	9	41	38	42	2122
fi	Finnish	Uralic	7428																			156	47	64	28	48	90	63	22	131	87	9	43	47	40	29	30	10	126	86	10	119	131	69	139	39	27	50	46	25	46	126	12	23	24	21	72	10	76	60	64	3998
fr	French	Romance	34494																				154	136	60	85	164	161	38	744	214	24	89	71	83	62	83	25	331	166	124	255	558	206	410	72	74	83	86	48	92	186	19	56	65	42	130	26	170	165	157	12441
gl	Galician	Romance	2221																					41	21	33	50	56	14	120	50	9	28	27	35	29	39	11	66	52	17	65	227	56	84	30	36	29	33	20	39	54	8	15	17	22	43	12	51	58	46	3969
he	Hebrew	Semitic	6962																						28	41	65	63	17	121	82	7	43	35	42	26	25	9	86	64	8	84	133	67	131	35	21	36	38	23	45	67	10	21	25	13	54	8	73	66	62	3666
hi	Hindi	Indo-Aryan	1353																							21	33	31	7	56	35	3	18	16	24	6	11	12	40	27	4	44	63	35	56	17	6	18	21	14	22	40	7	13	18	3	29	4	33	26	30	1781
hr	Croatian	Slavic	2229																								58	47	14	80	48	8	27	31	52	24	24	10	65	48	7	71	85	51	85	666	19	35	53	21	205	56	9	16	16	17	42	9	55	46	42	3629
hu	Hungarian	Uralic	7702																									70	20	146	99	11	49	48	47	27	28	10	121	75	10	126	148	87	149	46	26	56	55	27	53	88	13	29	30	20	75	9	83	74	75	4357
id	Indonesian	Malayo-Polynesian	3899																										16	146	77	8	45	33	55	25	23	10	101	83	8	93	204	94	127	43	23	37	46	32	56	79	13	24	19	21	79	11	73	146	83	4743
is	Icelandic	Germanic	488																											31	18	2	9	12	12	6	5	3	27	22	3	27	35	20	30	13	4	13	13	9	13	28	4	7	5	5	16	3	18	16	14	995
it	Italian	Romance	21025																												179	24	83	62	73	58	78	24	240	150	20	219	480	161	303	67	68	72	81	48	83	153	19	44	58	41	112	23	144	143	137	10084
ja	Japanese	Japonic	31614																													14	222	47	48	21	23	8	123	81	9	128	175	79	196	40	19	48	50	28	51	96	12	37	31	12	84	12	92	75	267	5382
kk	Kazakh	Turkic	1684																														6	5	6	2	2	1	18	11	1	17	22	12	32	8	1	8	7	6	8	16	3	3	1	1	10	4	14	6	9	581
ko	Korean	Koreanic	5400																															23	26	10	10	4	56	41	5	63	93	47	89	23	9	25	26	17	29	51	7	17	13	5	47	5	48	49	57	2539
lt	Lithuanian	Baltic	2031																																28	16	16	6	57	39	6	70	64	39	107	25	15	30	30	16	29	46	8	13	11	12	36	6	57	33	35	2091
mk	Macedonian	Slavic	1487																																	21	22	9	53	46	6	56	93	56	88	52	19	29	39	25	106	48	9	14	15	16	43	8	57	57	45	2761
ml	Malayalam	Dravidian	272																																		3	1	41	32	7	41	58	35	46	21	2	22	21	13	20	44	6	4	3	1	16	2	26	15	17	1388
mr	Marathi	Indo-Aryan	503																																			1	46	35	5	47	86	47	50	22	3	25	25	13	19	56	4	5	3	1	17	1	26	14	17	1648
ne	Nepali	Indo-Aryan	235																																				17	13	4	17	21	14	19	10	1	10	10	6	9	17	3	1	1		7	1	11	6	6	584
nl	Dutch	Germanic	14091																																					133	13	177	218	96	199	53	42	66	64	34	61	151	16	37	35	29	90	18	104	84	88	6364
no	Norwegian	Germanic	1364																																						9	103	161	74	121	42	28	43	51	26	47	270	12	24	23	24	61	13	69	79	63	4771
oc	Occitan	Romance	472																																							13	24	11	14	6	5	6	6	4	7	12	2	3	3	3	7	3	9	8	7	706
pl	Polish	Slavic	16270																																								200	97	285	56	40	81	68	35	69	121	16	39	42	28	92	16	172	84	92	6035
pt	Portuguese	Romance	13354																																									177	312	74	76	71	85	47	101	155	20	42	54	45	140	23	156	213	165	10932
ro	Romanian	Romance	3504																																										136	44	43	42	49	30	58	75	15	23	27	29	72	13	82	96	74	4515
ru	Russian	Slavic	32537																																											70	42	85	78	44	114	140	17	54	55	29	119	25	2486	122	148	11311
sh	Serbo-Croatian languages	South Slavic	2069																																												17	27	46	19	373	46	8	14	13	17	35	9	45	42	36	3337
si	Sinhala	Indo-Aryan	320																																													22	22	10	15	48	4	4	4	2	15	1	21	14	16	1464
sk	Slovak	Slavic	1904																																														36	17	34	51	8	14	15	16	37	9	51	38	38	2726
sl	Slovenian	Slavic	1836																																															19	47	50	9	15	16	17	39	8	50	48	42	2623
sq	Albanian	Albanian	740																																																25	31	6	13	10	10	27		29	30	24	1515
sr	Serbian	Slavic	4226																																																	51	9	19	18	14	43	8	71	56	45	3510
sv	Swedish	Germanic	17906																																																		16	33	39	35	72	17	82	74	73	5107
sw	Swahili	Niger-Congo	235																																																			6	3	4	12		12	10	11	652
ta	Tamil	Dravidian	1629																																																				6	2	29	5	30	19	27	1315
te	Telugu	Dravidian	1509																																																					1	21	1	30	16	20	1315
tl	Tagalog	Malayo-Polynesian	312																																																						12	1	16	17	10	1015
tr	Turkish	Turkic	4067																																																							10	67	77	69	3698
tt	Tatar	Turkic	449																																																								11	7	10	603
uk	Ukrainian	Slavic	12871																																																									73	72	7043
vi	Vietnamese	Vietic	6727																																																										89	4660
zh	Chinese	Chinese	12306																																																											4415

- With English: Indonesian 1M, Hebrew 545k, Farsi 303k or Marathi 124k

WikiMatrix: 85 Languages, 1620Pairs



ISO	Name	Language Family	size	az	ba	be	bg	bn	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	id	is	it	ja	kk	ko	lt	mk	ml	mr	ne	nl	no	oc	pl	pt	ro	ru	sh	si	sk	sl	sq	sr	sv	sw	ta	te	tl	tr	tt	uk	vi	zh	total
ar	Arabic	Arabic	6516	17	15	11	54	40	34	94	67	53	99	66	999	37	174	40	24	58	53	163	50	68	38	38	60	90	18	123	83	11	48	33	52	32	32	12	73	58	9	74	157	71	125	35	32	32	39	30	49	58	13	27	27	18	69	14	70	93	86	4435
az	Azerbaijani	Turkic	1873		7	3	14	8	9	15	17	11	34	10	71	9	31	10	7	16	14	29	10	12	9	10	15	17	4	25	23	6	11	9	8	5	4	2	22	13	2	22	25	14	47	9	3	9	8	9	10	19	5	11	5	2	42	8	19	10	19	885
ba	Bashkir	Turkic	536			2	14	4	9	17	16	13	27	10	28	9	28	8	6	9	12	29	12	9	4	10	12	12	3	26	13	4	5	7	9	1	1	1	21	15	3	19	24	15	42	10	1	10	10	7	10	20	4	3	1	1	11	8	15	7	10	697
be	Belarusian	Slavic	1690				16	4	7	16	14	9	20	8	33	9	28	7	5	5	10	24	9	10	3	9	11	9	3	24	12	2	5	7	8	2	2	1	14	10	2	19	23	13	161	8	2	8	7	5	10	16	2	3	2	1	8	2	80	8	8	803
bg	Bulgarian	Slavic	3327					38	34	76	79	53	132	62	357	40	122	43	25	37	61	117	43	58	30	47	68	60	17	102	71	11	38	42	86	29	35	13	84	58	9	96	114	69	270	41	31	43	46	26	65	63	12	21	23	21	56	12	126	60	60	3770
bn	Bengali	Indo-Aryan	1412						21	41	47	33	70	36	280	27	81	26	14	20	37	68	27	34	21	23	41	36	8	64	38	4	20	21	23	7	8	3	50	35	4	52	76	46	62	20	8	25	26	17	25	54	6	12	8	4	33		37	31	31	1991
bs	Bosnian	Slavic	1060							44	43	32	71	33	210	24	70	25	16	20	36	60	32	30	16	164	39	38	11	52	36	6	22	23	39	19	20	8	45	36	6	48	62	37	59	178	16	25	34	19	130	38	7	15	12	13	33	8	39	38	31	2348
ca	Catalan	Romance	8332								100	86	180	90	1205	81	1580	54	77	44	83	490	268	84	37	57	92	107	23	316	103	12	52	45	61	45	56	17	144	102	57	121	358	110	169	52	52	50	57	34	67	102	14	30	35	31	77	16	98	106	90	8047
cs	Czech	Slavic	7634									75	233	70	519	75	181	62	36	45	95	185	54	72	38	63	105	78	23	161	105	15	53	55	51	36	41	15	139	86	11	176	153	82	186	50	37	474	64	30	63	97	15	31	34	25	75	14	104	74	80	5325
da	Danish	Germanic	2881										180	54	436	39	140	45	26	29	75	142	44	55	25	43	69	63	20	115	76	9	37	35	39	30	35	12	110	303	8	89	123	70	109	37	32	39	40	23	43	168	11	20	21	23	55	11	62	68	57	4006
de	German	Germanic	50944											95	1573	186	418	106	53	66	163	626	80	109	57	87	192	107	34	388	217	23	82	78	64	51	58	21	472	207	17	285	294	129	368	68	50	94	106	51	81	216	20	58	57	32	127	23	165	107	134	9692
el	Greek	Hellenic	3211												620	39	145	41	23	35	55	137	48	56	26	43	64	73	15	119	69	6	35	34	52	27	32	9	76	60	8	77	144	78	114	38	31	35	46	28	52	62	11	16	18	20	56	9	68	75	62	3728
en	English	Germanic	134431													298	3377	243	119	303	375	2757	446	545	231	259	488	1019	85	2126	851	20	306	157	395	71	124	15	796	636	37	668	2461	631	1661	224	115	178	318	180	395	546	51	95	91	75	477	32	681	1073	786	34321
eo	Esperanto	constructed	2371														149	31	25	23	46	134	46	39	22	29	57	46	15	101	48	7	26	28	30	28	28	9	81	47	9	77	91	43	81	26	28	41	32	19	36	53	8	16	19	17	37	8	50	42	39	2817
es	Spanish	Romance	25202															89	154	83	155	905	610	153	71	94	167	198	42	671	219	26	108	76	92	65	98	25	272	181	35	235	923	183	393	81	84	81	93	53	107	181	21	57	71	48	147	26	187	206	174	14980
et	Estonian	Uralic	2303																22	24	70	85	32	39	20	33	56	41	14	75	57	6	29	35	32	20	21	8	72	49	7	73	76	48	96	27	19	34	35	18	34	58	9	16	16	15	43	7	56	40	44	2511
eu	Basque	Isolate	2259																	14	33	65	43	25	13	21	35	27	10	54	33	5	18	19	19	11	10	5	44	29	6	43	58	30	47	19	9	20	20	12	21	38	6	13	11	8	29	5	30	25	23	1699
fa	Farsi	Iranian	3954																		34	71	25	36	20	24	39	46	9	64	46	8	26	20	27	11	10	6	49	32	4	50	77	40	72	21	9	21	24	17	30	42	8	21	12	7	42	9	41	38	42	2122
fi	Finnish	Uralic	7428																			156	47	64	28	48	90	63	22	131	87	9	43	47	40	29	30	10	126	86	10	119	131	69	139	39	27	50	46	25	46	126	12	23	24	21	72	10	76	60	64	3998
fr	French	Romance	34494																				154	136	60	85	164	161	38	744	214	24	89	71	83	62	83	25	331	166	124	255	558	206	410	72	74	83	86	48	92	186	19	56	65	42	130	26	170	165	157	12441
gl	Galician	Romance	2221																					41	21	33	50	56	14	120	50	9	28	27	35	29	39	11	66	52	17	65	227	56	84	30	36	29	33	20	39	54	8	15	17	22	43	12	51	58	46	3969
he	Hebrew	Semitic	6962																						28	41	65	63	17	121	82	7	43	35	42	26	25	9	86	64	8	84	133	67	131	35	21	36	38	23	45	67	10	21	25	13	54	8	73	66	62	3666
hi	Hindi	Indo-Aryan	1353																							21	33	31	7	56	35	3	18	16	24	6	11	12	40	27	4	44	63	35	56	17	6	18	21	14	22	40	7	13	18	3	29	4	33	26	30	1781
hr	Croatian	Slavic	2229																								58	47	14	80	48	8	27	31	52	24	24	10	65	48	7	71	85	51	85	666	19	35	53	21	205	56	9	16	16	17	42	9	55	46	42	3629
hu	Hungarian	Uralic	7702																									70	20	146	99	11	49	48	47	27	28	10	121	75	10	126	148	87	149	46	26	56	55	27	53	88	13	29	30	20	75	9	83	74	75	4357
id	Indonesian	Malayo-Polynesian	3899																										16	146	77	8	45	33	55	25	23	10	101	83	8	93	204	94	127	43	23	37	46	32	56	79	13	24	19	21	79	11	73	146	83	4743
is	Icelandic	Germanic	488																											31	18	2	9	12	12	6	5	3	27	22	3	27	35	20	30	13	4	13	13	9	13	28	4	7	5	5	16	3	18	16	14	995
it	Italian	Romance	21025																												179	24	83	62	73	58	78	24	240	150	20	219	480	161	303	67	68	72	81	48	83	153	19	44	58	41	112	23	144	143	137	10084
ja	Japanese	Japonic	31614																													14	222	47	48	21	23	8	123	81	9	128	175	79	196	40	19	48	50	28	51	96	12	37	31	12	84	12	92	75	267	5382
kk	Kazakh	Turkic	1684																														6	5	6	2	2	1	18	11	1	17	22	12	32	8	1	8	7	6	8	16	3	3	1	1	10	4	14	6	9	581
ko	Korean	Koreanic	5400																															23	26	10	10	4	56	41	5	63	93	47	89	23	9	25	26	17	29	51	7	17	13	5	47	5	48	49	57	2539
lt	Lithuanian	Baltic	2031																																28	16	16	6	57	39	6	70	64	39	107	25	15	30	30	16	29	46	8	13	11	12	36	6	57	33	35	2091
mk	Macedonian	Slavic	1487																																	21	22	9	53	46	6	56	93	56	88	52	19	29	39	25	106	48	9	14	15	16	43	8	57	57	45	2761
ml	Malayalam	Dravidian	272																																		3	1	41	32	7	41	58	35	46	21	2	22	21	13	20	44	6	4	3	1	16	2	26	15	17	1388
mr	Marathi	Indo-Aryan	503																																			1	46	35	5	47	86	47	50	22	3	25	25	13	19	56	4	5	3	1	17	1	26	14	17	1648
ne	Nepali	Indo-Aryan	235																																				17	13	4	17	21	14	19	10	1	10	10	6	9	17	3	1	1		7	1	11	6	6	584
nl	Dutch	Germanic	14091																																					133	13	177	218	96	199	53	42	66	64	34	61	151	16	37	35	29	90	18	104	84	88	6364
no	Norwegian	Germanic	1364																																						9	103	161	74	121	42	28	43	51	26	47	270	12	24	23	24	61	13	69	79	63	4771
oc	Occitan	Romance	472																																							13	24	11	14	6	5	6	6	4	7	12	2	3	3	3	7	3	9	8	7	706
pl	Polish	Slavic	16270																																								200	97	285	56	40	81	68	35	69	121	16	39	42	28	92	16	172	84	92	6035
pt	Portuguese	Romance	13354																																									177	312	74	76	71	85	47	101	155	20	42	54	45	140	23	156	213	165	10932
ro	Romanian	Romance	3504																																										136	44	43	42	49	30	58	75	15	23	27	29	72	13	82	96	74	4515
ru	Russian	Slavic	32537																																											70	42	85	78	44	114	140	17	54	55	29	119	25	2486	122	148	11311
sh	Serbo-Croatian languages	South Slavic	2069																																												17	27	46	19	373	46	8	14	13	17	35	9	45	42	36	3337
si	Sinhala	Indo-Aryan	320																																													22	22	10	15	48	4	4	4	2	15	1	21	14	16	1464
sk	Slovak	Slavic	1904																																														36	17	34	51	8	14	15	16	37	9	51	38	38	2726
sl	Slovenian	Slavic	1836																																															19	47	50	9	15	16	17	39	8	50	48	42	2623
sq	Albanian	Albanian	740																																																25	31	6	13	10	10	27		29	30	24	1515
sr	Serbian	Slavic	4226																																																	51	9	19	18	14	43	8	71	56	45	3510
sv	Swedish	Germanic	17906																																																		16	33	39	35	72	17	82	74	73	5107
sw	Swahili	Niger-Congo	235																																																			6	3	4	12		12	10	11	652
ta	Tamil	Dravidian	1629																																																				6	2	29	5	30	19	27	1315
te	Telugu	Dravidian	1509																																																					1	21	1	30	16	20	1315
tl	Tagalog	Malayo-Polynesian	312																																																						12	1	16	17	10	1015
tr	Turkish	Turkic	4067																																																								10	67	77	69	3698
tt	Tatar	Turkic	449																																																									11	7	10	603
uk	Ukrainian	Slavic	12871																																																										73	72	7043
vi	Vietnamese	Vietic	6727																																																											89	4660
zh	Chinese	Chinese	12306																																																												4415

- Russian/Ukrainian 2.5M, Catalan/Spanish 1.6M

WikiMatrix: 85 Languages, 1620Pairs

ISO	Name	Language Family	size	az	ba	be	bg	bn	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	id	is	it	ja	kk	ko	lt	mk	ml	mr	ne	nl	no	oc	pl	pt	ro	ru	sh	si	sk	sl	sq	sr	sv	sw	ta	te	tl	tr	tt	uk	vi	zh	total
ar	Arabic	Arabic	6516	17	15	11	54	40	34	94	67	53	99	66	999	37	174	40	24	58	53	163	50	68	38	38	60	90	18	123	83	11	48	33	52	32	32	12	73	58	9	74	157	71	125	35	32	32	39	30	49	58	13	27	27	18	69	14	70	93	86	4435
az	Azerbaijani	Turkic	1873		7	3	14	8	9	15	17	11	34	10	71	9	31	10	7	16	14	29	10	12	9	10	15	17	4	25	23	6	11	9	8	5	4	2	22	13	2	22	25	14	47	9	3	9	8	9	10	19	5	11	5	2	42	8	19	10	19	885
ba	Bashkir	Turkic	536			2	14	4	9	17	16	13	27	10	28	9	28	8	6	9	12	29	12	9	4	10	12	12	3	26	13	4	5	7	9	1	1	1	21	15	3	19	24	15	42	10	1	10	10	7	10	20	4	3	1	1	11	8	15	7	10	697
be	Belarusian	Slavic	1690				16	4	7	16	14	9	20	8	33	9	28	7	5	5	10	24	9	10	3	9	11	9	3	24	12	2	5	7	8	2	2	1	14	10	2	19	23	13	161	8	2	8	7	5	10	16	2	3	2	1	8	2	80	8	8	803
bg	Bulgarian	Slavic	3327					38	34	76	79	53	132	62	357	40	122	43	25	37	61	117	43	58	30	47	68	60	17	102	71	11	38	42	86	29	35	13	84	58	9	96	114	69	270	41	31	43	46	26	65	63	12	21	23	21	56	12	126	60	60	3770
bn	Bengali	Indo-Aryan	1412						21	41	47	33	70	36	280	27	81	26	14	20	37	68	27	34	21	23	41	36	8	64	38	4	20	21	23	7	8	3	50	35	4	52	76	46	62	20	8	25	26	17	25	54	6	12	8	4	33		37	31	31	1991
bs	Bosnian	Slavic	1060							44	43	32	71	33	210	24	70	25	16	20	36	60	32	30	16	164	39	38	11	52	36	6	22	23	39	19	20	8	45	36	6	48	62	37	59	178	16	25	34	19	130	38	7	15	12	13	33	8	39	38	31	2348
ca	Catalan	Romance	8332								100	86	180	90	1205	81	1580	54	77	44	83	490	268	84	37	57	92	107	23	316	103	12	52	45	61	45	56	17	144	102	57	121	358	110	169	52	52	50	57	34	67	102	14	30	35	31	77	16	98	106	90	8047
cs	Czech	Slavic	7634									75	233	70	519	75	181	62	36	45	95	185	54	72	38	63	105	78	23	161	105	15	53	55	51	36	41	15	139	86	11	176	153	82	186	50	37	474	64	30	63	97	15	31	34	25	75	14	104	74	80	5325
da	Danish	Germanic	2881										180	54	436	39	140	45	26	29	75	142	44	55	25	43	69	63	20	115	76	9	37	35	39	30	35	12	110	303	8	89	123	70	109	37	32	39	40	23	43	168	11	20	21	23	55	11	62	68	57	4006
de	German	Germanic	50944											95	1573	186	418	106	53	66	163	626	80	109	57	87	192	107	34	388	217	23	82	78	64	51	58	21	472	207	17	285	294	129	368	68	50	94	106	51	81	216	20	58	57	32	127	23	165	107	134	9692
el	Greek	Hellenic	3211												620	39	145	41	23	35	55	137	48	56	26	43	64	73	15	119	69	6	35	34	52	27	32	9	76	60	8	77	144	78	114	38	31	35	46	28	52	62	11	16	18	20	56	9	68	75	62	3728
en	English	Germanic	134431													298	3377	243	119	303	375	2757	446	545	231	259	488	1019	85	2126	851	20	306	157	395	71	124	15	796	636	37	668	2461	631	1661	224	115	178	318	180	395	546	51	95	91	75	477	32	681	1073	786	34321
eo	Esperanto	constructed	2371														149	31	25	23	46	134	46	39	22	29	57	46	15	101	48	7	26	28	30	28	28	9	81	47	9	77	91	43	81	26	28	41	32	19	36	53	8	16	19	17	37	8	50	42	39	2817
es	Spanish	Romance	25202															89	154	83	155	905	610	153	71	94	167	198	42	671	219	26	108	76	92	65	98	25	272	181	35	235	923	183	393	81	84	81	93	53	107	181	21	57	71	48	147	26	187	206	174	14980
et	Estonian	Uralic	2303																22	24	70	85	32	39	20	33	56	41	14	75	57	6	29	35	32	20	21	8	72	49	7	73	76	48	96	27	19	34	35	18	34	58	9	16	16	15	43	7	56	40	44	2511
eu	Basque	Isolate	2259																	14	33	65	43	25	13	21	35	27	10	54	33	5	18	19	19	11	10	5	44	29	6	43	58	30	47	19	9	20	20	12	21	38	6	13	11	8	29	5	30	25	23	1699
fa	Farsi	Iranian	3954																		34	71	25	36	20	24	39	46	9	64	46	8	26	20	27	11	10	6	49	32	4	50	77	40	72	21	9	21	24	17	30	42	8	21	12	7	42	9	41	38	42	2122
fi	Finnish	Uralic	7428																			156	47	64	28	48	90	63	22	131	87	9	43	47	40	29	30	10	126	86	10	119	131	69	139	39	27	50	46	25	46	126	12	23	24	21	72	10	76	60	64	3998
fr	French	Romance	34494																				154	136	60	85	164	161	38	744	214	24	89	71	83	62	83	25	331	166	124	255	558	206	410	72	74	83	86	48	92	186	19	56	65	42	130	26	170	165	157	12441
gl	Galician	Romance	2221																					41	21	33	50	56	14	120	50	9	28	27	35	29	39	11	66	52	17	65	227	56	84	30	36	29	33	20	39	54	8	15	17	22	43	12	51	58	46	3969
he	Hebrew	Semitic	6962																						28	41	65	63	17	121	82	7	43	35	42	26	25	9	86	64	8	84	133	67	131	35	21	36	38	23	45	67	10	21	25	13	54	8	73	66	62	3666
hi	Hindi	Indo-Aryan	1353																							21	33	31	7	56	35	3	18	16	24	6	11	12	40	27	4	44	63	35	56	17	6	18	21	14	22	40	7	13	18	3	29	4	33	26	30	1781
hr	Croatian	Slavic	2229																								58	47	14	80	48	8	27	31	52	24	24	10	65	48	7	71	85	51	85	666	19	35	53	21	205	56	9	16	16	17	42	9	55	46	42	3629
hu	Hungarian	Uralic	7702																									70	20	146	99	11	49	48	47	27	28	10	121	75	10	126	148	87	149	46	26	56	55	27	53	88	13	29	30	20	75	9	83	74	75	4357
id	Indonesian	Malayo-Polynesian	3899																										16	146	77	8	45	33	55	25	23	10	101	83	8	93	204	94	127	43	23	37	46	32	56	79	13	24	19	21	79	11	73	146	83	4743
is	Icelandic	Germanic	488																											31	18	2	9	12	12	6	5	3	27	22	3	27	35	20	30	13	4	13	13	9	13	28	4	7	5	5	16	3	18	16	14	995
it	Italian	Romance	21025																												179	24	83	62	73	58	78	24	240	150	20	219	480	161	303	67	68	72	81	48	83	153	19	44	58	41	112	23	144	143	137	10084
ja	Japanese	Japonic	31614																													14	222	47	48	21	23	8	123	81	9	128	175	79	196	40	19	48	50	28	51	96	12	37	31	12	84	12	92	75	267	5382
kk	Kazakh	Turkic	1684																														6	5	6	2	2	1	18	11	1	17	22	12	32	8	1	8	7	6	8	16	3	3	1	1	10	4	14	6	9	581
ko	Korean	Koreanic	5400																															23	26	10	10	4	56	41	5	63	93	47	89	23	9	25	26	17	29	51	7	17	13	5	47	5	48	49	57	2539
lt	Lithuanian	Baltic	2031																																28	16	16	6	57	39	6	70	64	39	107	25	15	30	30	16	29	46	8	13	11	12	36	6	57	33	35	2091
mk	Macedonian	Slavic	1487																																	21	22	9	53	46	6	56	93	56	88	52	19	29	39	25	106	48	9	14	15	16	43	8	57	57	45	2761
ml	Malayalam	Dravidian	272																																		3	1	41	32	7	41	58	35	46	21	2	22	21	13	20	44	6	4	3	1	16	2	26	15	17	1388
mr	Marathi	Indo-Aryan	503																																			1	46	35	5	47	86	47	50	22	3	25	25	13	19	56	4	5	3	1	17	1	26	14	17	1648
ne	Nepali	Indo-Aryan	235																																				17	13	4	17	21	14	19	10	1	10	10	6	9	17	3	1	1		7	1	11	6	6	584
nl	Dutch	Germanic	14091																																					133	13	177	218	96	199	53	42	66	64	34	61	151	16	37	35	29	90	18	104	84	88	6364
no	Norwegian	Germanic	1364																																						9	103	161	74	121	42	28	43	51	26	47	270	12	24	23	24	61	13	69	79	63	4771
oc	Occitan	Romance	472																																							13	24	11	14	6	5	6	6	4	7	12	2	3	3	3	7	3	9	8	7	706
pl	Polish	Slavic	16270																																								200	97	285	56	40	81	68	35	69	121	16	39	42	28	92	16	172	84	92	6035
pt	Portuguese	Romance	13354																																									177	312	74	76	71	85	47	101	155	20	42	54	45	140	23	156	213	165	10932
ro	Romanian	Romance	3504																																										136	44	43	42	49	30	58	75	15	23	27	29	72	13	82	96	74	4515
ru	Russian	Slavic	32537																																											70	42	85	78	44	114	140	17	54	55	29	119	25	2486	122	148	11311
sh	Serbo-Croatian languages	South Slavic	2069																																												17	27	46	19	373	46	8	14	13	17	35	9	45	42	36	3337
si	Sinhala	Indo-Aryan	320																																													22	22	10	15	48	4	4	4	2	15	1	21	14	16	1464
sk	Slovak	Slavic	1904																																														36	17	34	51	8	14	15	16	37	9	51	38	38	2726
sl	Slovenian	Slavic	1836																																															19	47	50	9	15	16	17	39	8	50	48	42	2623
sq	Albanian	Albanian	740																																																25	31	6	13	10	10	27		29	30	24	1515
sr	Serbian	Slavic	4226																																																	51	9	19	18	14	43	8	71	56	45	3510
sv	Swedish	Germanic	17906																																																		16	33	39	35	72	17	82	74	73	5107
sw	Swahili	Niger-Congo	235																																																			6	3	4	12		12	10	11	652
ta	Tamil	Dravidian	1629																																																				6	2	29	5	30	19	27	1315
te	Telugu	Dravidian	1509																																																					1	21	1	30	16	20	1315
tl	Tagalog	Malayo-Polynesian	312																																																						12	1	16	17	10	1015
tr	Turkish	Turkic	4067																																																							10	67	77	69	3698
tt	Tatar	Turkic	449																																																								11	7	10	603
uk	Ukrainian	Slavic	12871																																																									73	72	7043
vi	Vietnamese	Vietic	6727																																																										89	4660
zh	Chinese	Chinese	12306																																																											4415

- Between Romance languages fr, es, it and pt 480k–923k

WikiMatrix: 85 Languages, 1620Pairs

ISO	Name	Language Family	size	az	ba	be	bg	bn	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	id	is	it	ja	kk	ko	lt	mk	ml	mr	ne	nl	no	oc	pl	pt	ro	ru	sh	si	sk	sl	sq	sr	sv	sw	ta	te	tl	tr	tt	uk	vi	zh	total
ar	Arabic	Arabic	6516	17	15	11	54	40	34	94	67	53	99	66	999	37	174	40	24	58	53	163	50	68	38	38	60	90	18	123	83	11	48	33	52	32	32	12	73	58	9	74	157	71	125	35	32	32	39	30	49	58	13	27	27	18	69	14	70	93	86	4435
az	Azerbaijani	Turkic	1873		7	3	14	8	9	15	17	11	34	10	71	9	31	10	7	16	14	29	10	12	9	10	15	17	4	25	23	6	11	9	8	5	4	2	22	13	2	22	25	14	47	9	3	9	8	9	10	19	5	11	5	2	42	8	19	10	19	885
ba	Bashkir	Turkic	536			2	14	4	9	17	16	13	27	10	28	9	28	8	6	9	12	29	12	9	4	10	12	12	3	26	13	4	5	7	9	1	1	1	21	15	3	19	24	15	42	10	1	10	10	7	10	20	4	3	1	1	11	8	15	7	10	697
be	Belarusian	Slavic	1690				16	4	7	16	14	9	20	8	33	9	28	7	5	5	10	24	9	10	3	9	11	9	3	24	12	2	5	7	8	2	2	1	14	10	2	19	23	13	161	8	2	8	7	5	10	16	2	3	2	1	8	2	80	8	8	803
bg	Bulgarian	Slavic	3327					38	34	76	79	53	132	62	357	40	122	43	25	37	61	117	43	58	30	47	68	60	17	102	71	11	38	42	86	29	35	13	84	58	9	96	114	69	270	41	31	43	46	26	65	63	12	21	23	21	56	12	126	60	60	3770
bn	Bengali	Indo-Aryan	1412						21	41	47	33	70	36	280	27	81	26	14	20	37	68	27	34	21	23	41	36	8	64	38	4	20	21	23	7	8	3	50	35	4	52	76	46	62	20	8	25	26	17	25	54	6	12	8	4	33		37	31	31	1991
bs	Bosnian	Slavic	1060							44	43	32	71	33	210	24	70	25	16	20	36	60	32	30	16	164	39	38	11	52	36	6	22	23	39	19	20	8	45	36	6	48	62	37	59	178	16	25	34	19	130	38	7	15	12	13	33	8	39	38	31	2348
ca	Catalan	Romance	8332								100	86	180	90	1205	81	1580	54	77	44	83	490	268	84	37	57	92	107	23	316	103	12	52	45	61	45	56	17	144	102	57	121	358	110	169	52	52	50	57	34	67	102	14	30	35	31	77	16	98	106	90	8047
cs	Czech	Slavic	7634									75	233	70	519	75	181	62	36	45	95	185	54	72	38	63	105	78	23	161	105	15	53	55	51	36	41	15	139	86	11	176	153	82	186	50	37	474	64	30	63	97	15	31	34	25	75	14	104	74	80	5325
da	Danish	Germanic	2881										180	54	436	39	140	45	26	29	75	142	44	55	25	43	69	63	20	115	76	9	37	35	39	30	35	12	110	303	8	89	123	70	109	37	32	39	40	23	43	168	11	20	21	23	55	11	62	68	57	4006
de	German	Germanic	50944											95	1573	186	418	106	53	66	163	626	80	109	57	87	192	107	34	388	217	23	82	78	64	51	58	21	472	207	17	285	294	129	368	68	50	94	106	51	81	216	20	58	57	32	127	23	165	107	134	9692
el	Greek	Hellenic	3211												620	39	145	41	23	35	55	137	48	56	26	43	64	73	15	119	69	6	35	34	52	27	32	9	76	60	8	77	144	78	114	38	31	35	46	28	52	62	11	16	18	20	56	9	68	75	62	3728
en	English	Germanic	134431													298	3377	243	119	303	375	2757	446	545	231	259	488	1019	85	2126	851	20	306	157	395	71	124	15	796	636	37	668	2461	631	1661	224	115	178	318	180	395	546	51	95	91	75	477	32	681	1073	786	34321
eo	Esperanto	constructed	2371														149	31	25	23	46	134	46	39	22	29	57	46	15	101	48	7	26	28	30	28	28	9	81	47	9	77	91	43	81	26	28	41	32	19	36	53	8	16	19	17	37	8	50	42	39	2817
es	Spanish	Romance	25202															89	154	83	155	905	610	153	71	94	167	198	42	671	219	26	108	76	92	65	98	25	272	181	35	235	923	183	393	81	84	81	93	53	107	181	21	57	71	48	147	26	187	206	174	14980
et	Estonian	Uralic	2303																22	24	70	85	32	39	20	33	56	41	14	75	57	6	29	35	32	20	21	8	72	49	7	73	76	48	96	27	19	34	35	18	34	58	9	16	16	15	43	7	56	40	44	2511
eu	Basque	Isolate	2259																	14	33	65	43	25	13	21	35	27	10	54	33	5	18	19	19	11	10	5	44	29	6	43	58	30	47	19	9	20	20	12	21	38	6	13	11	8	29	5	30	25	23	1699
fa	Farsi	Iranian	3954																		34	71	25	36	20	24	39	46	9	64	46	8	26	20	27	11	10	6	49	32	4	50	77	40	72	21	9	21	24	17	30	42	8	21	12	7	42	9	41	38	42	2122
fi	Finnish	Uralic	7428																			156	47	64	28	48	90	63	22	131	87	9	43	47	40	29	30	10	126	86	10	119	131	69	139	39	27	50	46	25	46	126	12	23	24	21	72	10	76	60	64	3998
fr	French	Romance	34494																				154	136	60	85	164	161	38	744	214	24	89	71	83	62	83	25	331	166	124	255	558	206	410	72	74	83	86	48	92	186	19	56	65	42	130	26	170	165	157	12441
gl	Galician	Romance	2221																					41	21	33	50	56	14	120	50	9	28	27	35	29	39	11	66	52	17	65	227	56	84	30	36	29	33	20	39	54	8	15	17	22	43	12	51	58	46	3969
he	Hebrew	Semitic	6962																						28	41	65	63	17	121	82	7	43	35	42	26	25	9	86	64	8	84	133	67	131	35	21	36	38	23	45	67	10	21	25	13	54	8	73	66	62	3666
hi	Hindi	Indo-Aryan	1353																							21	33	31	7	56	35	3	18	16	24	6	11	12	40	27	4	44	63	35	56	17	6	18	21	14	22	40	7	13	18	3	29	4	33	26	30	1781
hr	Croatian	Slavic	2229																								58	47	14	80	48	8	27	31	52	24	24	10	65	48	7	71	85	51	85	666	19	35	53	21	205	56	9	16	16	17	42	9	55	46	42	3629
hu	Hungarian	Uralic	7702																									70	20	146	99	11	49	48	47	27	28	10	121	75	10	126	148	87	149	46	26	56	55	27	53	88	13	29	30	20	75	9	83	74	75	4357
id	Indonesian	Malayo-Polynesian	3899																										16	146	77	8	45	33	55	25	23	10	101	83	8	93	204	94	127	43	23	37	46	32	56	79	13	24	19	21	79	11	73	146	83	4743
is	Icelandic	Germanic	488																											31	18	2	9	12	12	6	5	3	27	22	3	27	35	20	30	13	4	13	13	9	13	28	4	7	5	5	16	3	18	16	14	995
it	Italian	Romance	21025																												179	24	83	62	73	58	78	24	240	150	20	219	480	161	303	67	68	72	81	48	83	153	19	44	58	41	112	23	144	143	137	10084
ja	Japanese	Japonic	31614																													14	222	47	48	21	23	8	123	81	9	128	175	79	196	40	19	48	50	28	51	96	12	37	31	12	84	12	92	75	267	5382
kk	Kazakh	Turkic	1684																														6	5	6	2	2	1	18	11	1	17	22	12	32	8	1	8	7	6	8	16	3	3	1	1	10	4	14	6	9	581
ko	Korean	Koreanic	5400																															23	26	10	10	4	56	41	5	63	93	47	89	23	9	25	26	17	29	51	7	17	13	5	47	5	48	49	57	2539
lt	Lithuanian	Baltic	2031																																28	16	16	6	57	39	6	70	64	39	107	25	15	30	30	16	29	46	8	13	11	12	36	6	57	33	35	2091
mk	Macedonian	Slavic	1487																																	21	22	9	53	46	6	56	93	56	88	52	19	29	39	25	106	48	9	14	15	16	43	8	57	57	45	2761
ml	Malayalam	Dravidian	272																																		3	1	41	32	7	41	58	35	46	21	2	22	21	13	20	44	6	4	3	1	16	2	26	15	17	1388
mr	Marathi	Indo-Aryan	503																																			1	46	35	5	47	86	47	50	22	3	25	25	13	19	56	4	5	3	1	17	1	26	14	17	1648
ne	Nepali	Indo-Aryan	235																																				17	13	4	17	21	14	19	10	1	10	10	6	9	17	3	1	1		7	1	11	6	6	584
nl	Dutch	Germanic	14091																																					133	13	177	218	96	199	53	42	66	64	34	61	151	16	37	35	29	90	18	104	84	88	6364
no	Norwegian	Germanic	1364																																						9	103	161	74	121	42	28	43	51	26	47	270	12	24	23	24	61	13	69	79	63	4771
oc	Occitan	Romance	472																																							13	24	11	14	6	5	6	6	4	7	12	2	3	3	3	7	3	9	8	7	706
pl	Polish	Slavic	16270																																								200	97	285	56	40	81	68	35	69	121	16	39	42	28	92	16	172	84	92	6035
pt	Portuguese	Romance	13354																																									177	312	74	76	71	85	47	101	155	20	42	54	45	140	23	156	213	165	10932
ro	Romanian	Romance	3504																																										136	44	43	42	49	30	58	75	15	23	27	29	72	13	82	96	74	4515
ru	Russian	Slavic	32537																																											70	42	85	78	44	114	140	17	54	55	29	119	25	2486	122	148	11311
sh	Serbo-Croatian languages	South Slavic	2069																																												17	27	46	19	373	46	8	14	13	17	35	9	45	42	36	3337
si	Sinhala	Indo-Aryan	320																																													22	22	10	15	48	4	4	4	2	15	1	21	14	16	1464
sk	Slovak	Slavic	1904																																														36	17	34	51	8	14	15	16	37	9	51	38	38	2726
sl	Slovenian	Slavic	1836																																															19	47	50	9	15	16	17	39	8	50	48	42	2623
sq	Albanian	Albanian	740																																																25	31	6	13	10	10	27		29	30	24	1515
sr	Serbian	Slavic	4226																																																	51	9	19	18	14	43	8	71	56	45	3510
sv	Swedish	Germanic	17906																																																		16	33	39	35	72	17	82	74	73	5107
sw	Swahili	Niger-Congo	235																																																			6	3	4	12		12	10	11	652
ta	Tamil	Dravidian	1629																																																				6	2	29	5	30	19	27	1315
te	Telugu	Dravidian	1509																																																					1	21	1	30	16	20	1315
tl	Tagalog	Malayo-Polynesian	312																																																						12	1	16	17	10	1015
tr	Turkish	Turkic	4067																																																							10	67	77	69	3698
tt	Tatar	Turkic	449																																																								11	7	10	603
uk	Ukrainian	Slavic	12871																																																									73	72	7043
vi	Vietnamese	Vietic	6727																																																										89	4660
zh	Chinese	Chinese	12306																																																											4415

- Japanese/Korean 222k, Japanese/Russian 196k, Indonesian/Vietnamese 146k, Hebrew/fr,es,it,ru 120–150k

Large-Scale Bitext Mining

Scaling up !

Large-Scale Bitext Mining

Scaling up !

- Can we apply the same global mining approach to a much bigger corpus ?

Large-Scale Bitext Mining

Scaling up !

- Can we apply the same global mining approach to a much bigger corpus ?
- 10 snapshot of curated common crawl corpus (Wenzek et al, arxiv'19)
- 36 billion unique sentences (50× bigger than Wikipedia)

Large-Scale Bitext Mining

Scaling up !

- Can we apply the same global mining approach to a much bigger corpus ?
- 10 snapshot of curated common crawl corpus (Wenzek et al, arxiv'19)
- 36 billion unique sentences (50× bigger than Wikipedia)

⇒ Substantial computational and storage challenges

  - Mining Russian against Japanese: 3 × 2.9 billion sentences
  - $\approx 8.7 \cdot 10^{18}$ distances (6 months on 8 GPUs)
  - optimized and highly parallelized processing

Mining in the Whole Internet

CCMatrix

- 36 billion sentences collected on the Internet in 39 languages

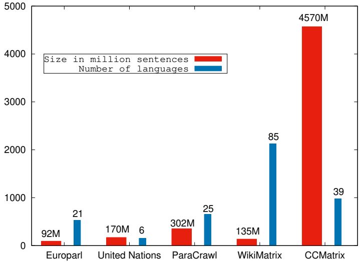
⇒ More than 4.5 billion parallel sentences in 39 languages

Mining in the Whole Internet

CCMatrix

- 36 billion sentences collected on the Internet in 39 languages

⇒ More than 4.5 billion parallel sentences in 39 languages



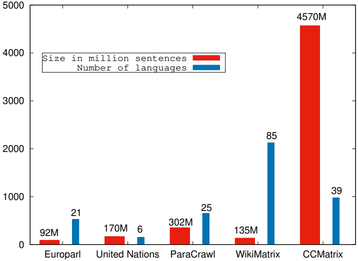
⇒ By far the largest collection of high quality mined bitexts

Mining in the Whole Internet

CCMatrix

- 36 billion sentences collected on the Internet in 39 languages

⇒ More than 4.5 billion parallel sentences in 39 languages

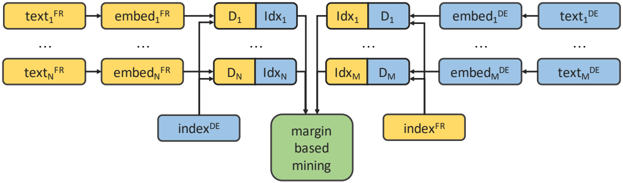


⇒ By far the largest collection of high quality mined bitexts

- Expected to cover many topics: politics, sports, tourism, daily life, . . .

Complexity Optimization

Example: mining French/English



- Split monolingual texts into many parts
- Calculate forward and backward distances in parallel

⇒ Extract bitexts when all distances are available

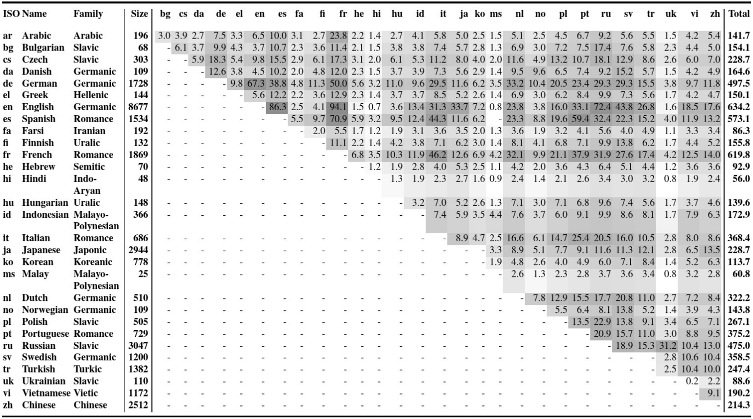
CCMatrix

ISO	Name	Family	Size	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	hu	id	it	ja	ko	ms	nl	no	pl	pt	ru	sv	tr	uk	vi	zh	Total
ar	Arabic	Arabic	196	3.0	3.9	2.7	7.5	3.3	6.5	10.0	3.1	2.7	23.8	2.2	1.4	2.7	4.1	5.8	5.0	2.5	1.5	5.1	2.5	4.5	6.7	9.2	5.6	5.5	1.5	4.2	5.4	141.7
bg	Bulgarian	Slavic	68	-	6.1	3.7	9.9	4.3	3.7	10.7	2.3	3.6	11.4	2.1	1.5	3.8	3.8	7.4	5.7	2.8	1.3	6.9	3.0	7.2	7.5	17.4	7.6	5.8	2.3	4.4	5.0	154.1
cs	Czech	Slavic	303	-	-	5.9	18.3	5.4	9.8	15.5	2.9	6.1	17.3	3.1	2.0	6.1	5.3	11.2	8.0	4.0	2.0	11.6	4.9	13.2	10.7	18.1	12.9	8.6	2.6	6.0	7.0	228.7
da	Danish	Germanic	109	-	-	-	12.6	3.8	4.5	10.2	2.0	4.8	12.0	2.3	1.5	3.7	3.9	7.3	5.6	2.9	1.4	9.5	9.6	6.5	7.4	9.2	15.2	5.7	1.5	4.2	4.9	164.6
de	German	Germanic	1728	-	-	-	-	9.8	67.3	38.8	4.8	11.3	50.0	5.6	3.2	11.0	9.6	29.5	11.6	6.2	3.5	33.2	10.4	20.5	23.4	29.3	29.3	15.5	3.8	9.7	11.8	497.5
el	Greek	Hellenic	144	-	-	-	-	-	5.6	12.2	2.2	3.6	12.9	2.3	1.4	3.7	3.7	8.5	5.2	2.6	1.4	6.9	3.0	6.2	8.4	9.9	7.3	5.6	1.7	4.2	4.7	150.1
en	English	Germanic	8677	-	-	-	-	-	-	86.3	2.5	4.1	94.1	1.5	0.7	3.6	13.4	31.3	33.7	7.2	0.8	23.8	3.8	16.0	33.1	72.4	43.8	26.8	1.6	18.5	17.6	634.2
es	Spanish	Romance	1534	-	-	-	-	-	-	-	5.5	9.7	70.9	5.9	3.2	9.5	12.4	44.3	11.6	6.2	-	23.3	8.8	19.6	59.4	32.4	22.3	15.2	4.0	11.9	13.2	573.1
fa	Farsi	Iranian	192	-	-	-	-	-	-	-	-	2.0	5.5	1.7	1.2	1.9	3.1	3.6	3.5	2.0	1.3	3.6	1.9	3.2	4.1	5.6	4.0	4.9	1.1	3.3	3.4	86.3
fi	Finnish	Uralic	132	-	-	-	-	-	-	-	-	-	11.1	2.2	1.4	4.2	3.8	7.1	6.2	3.0	1.4	8.1	4.1	6.8	7.1	9.9	13.8	6.2	1.7	4.4	5.2	155.8
fr	French	Romance	1869	-	-	-	-	-	-	-	-	-	-	6.8	3.5	10.3	11.9	46.2	12.6	6.9	4.2	32.1	9.9	21.1	37.9	31.9	27.6	17.4	4.2	12.5	14.0	619.8
he	Hebrew	Semitic	70	-	-	-	-	-	-	-	-	-	-	-	1.2	1.9	2.8	4.0	5.3	2.5	1.1	4.2	2.0	3.6	4.3	6.4	5.1	4.4	1.2	3.6	3.6	92.9
hi	Hindi	Indo-Aryan	48	-	-	-	-	-	-	-	-	-	-	-	-	1.3	1.9	2.3	2.7	1.6	0.9	2.4	1.4	2.1	2.6	3.4	3.0	3.2	0.8	1.9	2.4	56.0
hu	Hungarian	Uralic	148	-	-	-	-	-	-	-	-	-	-	-	-	-	3.2	7.0	5.2	2.6	1.3	7.1	3.0	7.1	6.8	9.6	7.4	5.6	1.7	3.7	4.6	139.6
id	Indonesian	Malayo-Polynesian	366	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.4	5.9	3.5	4.4	7.6	3.7	6.0	9.1	9.9	8.6	8.1	1.7	7.9	6.3	172.9
it	Italian	Romance	686	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.9	4.7	2.5	16.6	6.1	14.7	25.4	20.5	16.0	10.5	2.8	8.0	8.6	368.4
ja	Japanese	Japonic	2944	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.3	8.9	5.1	7.7	9.1	11.6	11.3	12.1	2.8	6.5	13.5	228.7
ko	Korean	Koreanic	778	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.9	4.8	2.6	4.0	4.9	6.0	7.1	8.4	1.4	5.2	6.3	113.7
ms	Malay	Malayo-Polynesian	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.6	1.3	2.3	2.8	3.7	3.6	3.4	0.8	3.2	2.8	60.8
nl	Dutch	Germanic	510	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.8	12.9	15.5	17.7	20.8	11.0	2.7	7.2	8.4	322.2
no	Norwegian	Germanic	109	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.5	6.4	8.1	13.8	5.2	1.4	3.9	4.3	143.8
pl	Polish	Slavic	505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13.5	22.9	13.8	9.1	3.4	6.5	7.1	267.1
pt	Portuguese	Romance	729	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.9	15.7	11.0	3.0	8.8	9.5	375.2
ru	Russian	Slavic	3047	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18.9	15.3	31.2	10.4	13.0	475.0
sv	Swedish	Germanic	1200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.8	10.6	10.4	358.5
tr	Turkish	Turkic	1382	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.5	10.4	10.0	247.4
uk	Ukrainian	Slavic	110	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2	2.2	88.6
vi	Vietnamese	Vietic	1172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9.1	190.2
zh	Chinese	Chinese	2512	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	214.3

Table 1: CCMatrix: size of mined sentences **(in millions)** for each language pair.

- French/Spanish: 71M

CCMatrix



ISO	Name	Family	Size	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	hu	id	it	ja	ko	ms	nl	no	pl	pt	ru	sv	tr	uk	vi	zh	Total
ar	Arabic	Arabic	196	3.0	3.9	2.7	7.5	3.3	6.5	10.0	3.1	2.7	23.8	2.2	1.4	2.7	4.1	5.8	5.0	2.5	1.5	5.1	2.5	4.5	6.7	9.2	5.6	5.5	1.5	4.2	5.4	141.7
bg	Bulgarian	Slavic	68	-	6.1	3.7	9.9	4.3	3.7	10.7	2.3	3.6	11.4	2.1	1.5	3.8	3.8	7.4	5.7	2.8	1.3	6.9	3.0	7.2	7.5	17.4	7.6	5.8	2.3	4.4	5.0	154.1
cs	Czech	Slavic	303	-	-	5.9	18.3	5.4	9.8	15.5	2.9	6.1	17.3	3.1	2.0	6.1	5.3	11.2	8.0	4.0	2.0	11.6	4.9	13.2	10.7	18.1	12.9	8.6	2.6	6.0	7.0	228.7
da	Danish	Germanic	109	-	-	-	12.6	3.8	4.5	10.2	2.0	4.8	12.0	2.3	1.5	3.7	3.9	7.3	5.6	2.9	1.4	9.5	9.6	6.5	7.4	9.2	15.2	5.7	1.5	4.2	4.9	164.6
de	German	Germanic	1728	-	-	-	-	9.8	67.3	38.8	4.8	11.3	50.0	5.6	3.2	11.0	9.6	29.5	11.6	6.2	3.5	33.2	10.4	20.5	23.4	29.3	29.3	15.5	3.8	9.7	11.8	497.5
el	Greek	Hellenic	144	-	-	-	-	-	5.6	12.2	2.2	3.6	12.9	2.3	1.4	3.7	3.7	8.5	5.2	2.6	1.4	6.9	3.0	6.2	8.4	9.9	7.3	5.6	1.7	4.2	4.7	150.1
en	English	Germanic	8677	-	-	-	-	-	-	86.3	2.5	4.1	94.1	1.5	0.7	3.6	13.4	31.3	33.7	7.2	0.8	23.8	3.8	16.0	33.1	72.4	43.8	26.8	1.6	18.5	17.6	634.2
es	Spanish	Romance	1534	-	-	-	-	-	-	-	5.5	9.7	70.9	5.9	3.2	9.5	12.4	44.3	11.6	6.2	-	23.3	8.8	19.6	59.4	32.4	22.3	15.2	4.0	11.9	13.2	573.1
fa	Farsi	Iranian	192	-	-	-	-	-	-	-	-	2.0	5.5	1.7	1.2	1.9	3.1	3.6	3.5	2.0	1.3	3.6	1.9	3.2	4.1	5.6	4.0	4.9	1.1	3.3	3.4	86.3
fi	Finnish	Uralic	132	-	-	-	-	-	-	-	-	-	11.1	2.2	1.4	4.2	3.8	7.1	6.2	3.0	1.4	8.1	4.1	6.8	7.1	9.9	13.8	6.2	1.7	4.4	5.2	155.8
fr	French	Romance	1869	-	-	-	-	-	-	-	-	-	-	6.8	3.5	10.3	11.9	46.2	12.6	6.9	4.2	32.1	9.9	21.1	37.9	31.9	27.6	17.4	4.2	12.5	14.0	619.8
he	Hebrew	Semitic	70	-	-	-	-	-	-	-	-	-	-	-	1.2	1.9	2.8	4.0	5.3	2.5	1.1	4.2	2.0	3.6	4.3	6.4	5.1	4.4	1.2	3.6	3.6	92.9
hi	Hindi	Indo-Aryan	48	-	-	-	-	-	-	-	-	-	-	-	-	1.3	1.9	2.3	2.7	1.6	0.9	2.4	1.4	2.1	2.6	3.4	3.0	3.2	0.8	1.9	2.4	56.0
hu	Hungarian	Uralic	148	-	-	-	-	-	-	-	-	-	-	-	-	-	3.2	7.0	5.2	2.6	1.3	7.1	3.0	7.1	6.8	9.6	7.4	5.6	1.7	3.7	4.6	139.6
id	Indonesian	Malayo-Polynesian	366	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.4	5.9	3.5	4.4	7.6	3.7	6.0	9.1	9.9	8.6	8.1	1.7	7.9	6.3	172.9
it	Italian	Romance	686	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.9	4.7	2.5	16.6	6.1	14.7	25.4	20.5	16.0	10.5	2.8	8.0	8.6	368.4
ja	Japanese	Japonic	2944	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.3	8.9	5.1	7.7	9.1	11.6	11.3	12.1	2.8	6.5	13.5	228.7
ko	Korean	Koreanic	778	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.9	4.8	2.6	4.0	4.9	6.0	7.1	8.4	1.4	5.2	6.3	113.7
ms	Malay	Malayo-Polynesian	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.6	1.3	2.3	2.8	3.7	3.6	3.4	0.8	3.2	2.8	60.8
nl	Dutch	Germanic	510	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.8	12.9	15.5	17.7	20.8	11.0	2.7	7.2	8.4	322.2
no	Norwegian	Germanic	109	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.5	6.4	8.1	13.8	5.2	1.4	3.9	4.3	143.8
pl	Polish	Slavic	505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13.5	22.9	13.8	9.1	3.4	6.5	7.1	267.1
pt	Portuguese	Romance	729	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.9	15.7	11.0	3.0	8.8	9.5	375.2
ru	Russian	Slavic	3047	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18.9	15.3	31.2	10.4	13.0	475.0
sv	Swedish	Germanic	1200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.8	10.6	10.4	358.5
tr	Turkish	Turkic	1382	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.5	10.4	10.0	247.4
uk	Ukrainian	Slavic	110	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2	2.2	88.6
vi	Vietnamese	Vietic	1172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9.1	190.2
zh	Chinese	Chinese	2512	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	214.3

Table 1: CCMatrix: size of mined sentences **(in millions)** for each language pair.

- Norwegian/Swedish: 14M

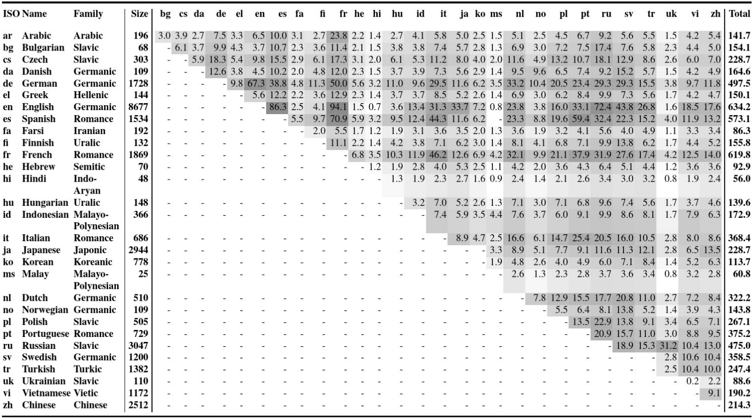
CCMatrix

ISO	Name	Family	Size	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	hu	id	it	ja	ko	ms	nl	no	pl	pt	ru	sv	tr	uk	vi	zh	Total
ar	Arabic	Arabic	196	3.0	3.9	2.7	7.5	3.3	6.5	10.0	3.1	2.7	23.8	2.2	1.4	2.7	4.1	5.8	5.0	2.5	1.5	5.1	2.5	4.5	6.7	9.2	5.6	5.5	1.5	4.2	5.4	141.7
bg	Bulgarian	Slavic	68	-	6.1	3.7	9.9	4.3	3.7	10.7	2.3	3.6	11.4	2.1	1.5	3.8	3.8	7.4	5.7	2.8	1.3	6.9	3.0	7.2	7.5	17.4	7.6	5.8	2.3	4.4	5.0	154.1
cs	Czech	Slavic	303	-	-	5.9	18.3	5.4	9.8	15.5	2.9	6.1	17.3	3.1	2.0	6.1	5.3	11.2	8.0	4.0	2.0	11.6	4.9	13.2	10.7	18.1	12.9	8.6	2.6	6.0	7.0	228.7
da	Danish	Germanic	109	-	-	-	12.6	3.8	4.5	10.2	2.0	4.8	12.0	2.3	1.5	3.7	3.9	7.3	5.6	2.9	1.4	9.5	9.6	6.5	7.4	9.2	15.2	5.7	1.5	4.2	4.9	164.6
de	German	Germanic	1728	-	-	-	-	9.8	67.3	38.8	4.8	11.3	50.0	5.6	3.2	11.0	9.6	29.5	11.6	6.2	3.5	33.2	10.4	20.5	23.4	29.3	29.3	15.5	3.8	9.7	11.8	497.5
el	Greek	Hellenic	144	-	-	-	-	-	5.6	12.2	2.2	3.6	12.9	2.3	1.4	3.7	3.7	8.5	5.2	2.6	1.4	6.9	3.0	6.2	8.4	9.9	7.3	5.6	1.7	4.2	4.7	150.1
en	English	Germanic	8677	-	-	-	-	-	-	86.3	2.5	4.1	94.1	1.5	0.7	3.6	13.4	31.3	33.7	7.2	0.8	23.8	3.8	16.0	33.1	72.4	43.8	26.8	1.6	18.5	17.6	634.2
es	Spanish	Romance	1534	-	-	-	-	-	-	-	5.5	9.7	70.9	5.9	3.2	9.5	12.4	44.3	11.6	6.2	-	23.3	8.8	19.6	59.4	32.4	22.3	15.2	4.0	11.9	13.2	573.1
fa	Farsi	Iranian	192	-	-	-	-	-	-	-	-	2.0	5.5	1.7	1.2	1.9	3.1	3.6	3.5	2.0	1.3	3.6	1.9	3.2	4.1	5.6	4.0	4.9	1.1	3.3	3.4	86.3
fi	Finnish	Uralic	132	-	-	-	-	-	-	-	-	-	11.1	2.2	1.4	4.2	3.8	7.1	6.2	3.0	1.4	8.1	4.1	6.8	7.1	9.9	13.8	6.2	1.7	4.4	5.2	155.8
fr	French	Romance	1869	-	-	-	-	-	-	-	-	-	-	6.8	3.5	10.3	11.9	46.2	12.6	6.9	4.2	32.1	9.9	21.1	37.9	31.9	27.6	17.4	4.2	12.5	14.0	619.8
he	Hebrew	Semitic	70	-	-	-	-	-	-	-	-	-	-	-	1.2	1.9	2.8	4.0	5.3	2.5	1.1	4.2	2.0	3.6	4.3	6.4	5.1	4.4	1.2	3.6	3.6	92.9
hi	Hindi	Indo-Aryan	48	-	-	-	-	-	-	-	-	-	-	-	-	1.3	1.9	2.3	2.7	1.6	0.9	2.4	1.4	2.1	2.6	3.4	3.0	3.2	0.8	1.9	2.4	56.0
hu	Hungarian	Uralic	148	-	-	-	-	-	-	-	-	-	-	-	-	-	3.2	7.0	5.2	2.6	1.3	7.1	3.0	7.1	6.8	9.6	7.4	5.6	1.7	3.7	4.6	139.6
id	Indonesian	Malayo-Polynesian	366	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.4	5.9	3.5	4.4	7.6	3.7	6.0	9.1	9.9	8.6	8.1	1.7	7.9	6.3	172.9
it	Italian	Romance	686	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.9	4.7	2.5	16.6	6.1	14.7	25.4	20.5	16.0	10.5	2.8	8.0	8.6	368.4
ja	Japanese	Japonic	2944	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.3	8.9	5.1	7.7	9.1	11.6	11.3	12.1	2.8	6.5	13.5	228.7
ko	Korean	Koreanic	778	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.9	4.8	2.6	4.0	4.9	6.0	7.1	8.4	1.4	5.2	6.3	113.7
ms	Malay	Malayo-Polynesian	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.6	1.3	2.3	2.8	3.7	3.6	3.4	0.8	3.2	2.8	60.8
nl	Dutch	Germanic	510	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.8	12.9	15.5	17.7	20.8	11.0	2.7	7.2	8.4	322.2
no	Norwegian	Germanic	109	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.5	6.4	8.1	13.8	5.2	1.4	3.9	4.3	143.8
pl	Polish	Slavic	505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13.5	22.9	13.8	9.1	3.4	6.5	7.1	267.1
pt	Portuguese	Romance	729	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.9	15.7	11.0	3.0	8.8	9.5	375.2
ru	Russian	Slavic	3047	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18.9	15.3	31.2	10.4	13.0	475.0
sv	Swedish	Germanic	1200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.8	10.6	10.4	358.5
tr	Turkish	Turkic	1382	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.5	10.4	10.0	247.4
uk	Ukrainian	Slavic	110	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2	2.2	88.6
vi	Vietnamese	Vietic	1172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9.1	190.2
zh	Chinese	Chinese	2512	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	214.3

Table 1: CCMatrix: size of mined sentences **(in millions)** for each language pair.

- Chinese/Japanese: 13.5M

CCMatrix



ISO	Name	Family	Size	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	hu	id	it	ja	ko	ms	nl	no	pl	pt	ru	sv	tr	uk	vi	zh	Total
ar	Arabic	Arabic	196	3.0	3.9	2.7	7.5	3.3	6.5	10.0	3.1	2.7	23.8	2.2	1.4	2.7	4.1	5.8	5.0	2.5	1.5	5.1	2.5	4.5	6.7	9.2	5.6	5.5	1.5	4.2	5.4	141.7
bg	Bulgarian	Slavic	68	-	6.1	3.7	9.9	4.3	3.7	10.7	2.3	3.6	11.4	2.1	1.5	3.8	3.8	7.4	5.7	2.8	1.3	6.9	3.0	7.2	7.5	17.4	7.6	5.8	2.3	4.4	5.0	154.1
cs	Czech	Slavic	303	-	-	5.9	18.3	5.4	9.8	15.5	2.9	6.1	17.3	3.1	2.0	6.1	5.3	11.2	8.0	4.0	2.0	11.6	4.9	13.2	10.7	18.1	12.9	8.6	2.6	6.0	7.0	228.7
da	Danish	Germanic	109	-	-	-	12.6	3.8	4.5	10.2	2.0	4.8	12.0	2.3	1.5	3.7	3.9	7.3	5.6	2.9	1.4	9.5	9.6	6.5	7.4	9.2	15.2	5.7	1.5	4.2	4.9	164.6
de	German	Germanic	1728	-	-	-	-	9.8	67.3	38.8	4.8	11.3	50.0	5.6	3.2	11.0	9.6	29.5	11.6	6.2	3.5	33.2	10.4	20.5	23.4	29.3	29.3	15.5	3.8	9.7	11.8	497.5
el	Greek	Hellenic	144	-	-	-	-	-	5.6	12.2	2.2	3.6	12.9	2.3	1.4	3.7	3.7	8.5	5.2	2.6	1.4	6.9	3.0	6.2	8.4	9.9	7.3	5.6	1.7	4.2	4.7	150.1
en	English	Germanic	8677	-	-	-	-	-	-	86.3	2.5	4.1	94.1	1.5	0.7	3.6	13.4	31.3	33.7	7.2	0.8	23.8	3.8	16.0	33.1	72.4	43.8	26.8	1.6	18.5	17.6	634.2
es	Spanish	Romance	1534	-	-	-	-	-	-	-	5.5	9.7	70.9	5.9	3.2	9.5	12.4	44.3	11.6	6.2	-	23.3	8.8	19.6	59.4	32.4	22.3	15.2	4.0	11.9	13.2	573.1
fa	Farsi	Iranian	192	-	-	-	-	-	-	-	-	2.0	5.5	1.7	1.2	1.9	3.1	3.6	3.5	2.0	1.3	3.6	1.9	3.2	4.1	5.6	4.0	4.9	1.1	3.3	3.4	86.3
fi	Finnish	Uralic	132	-	-	-	-	-	-	-	-	-	11.1	2.2	1.4	4.2	3.8	7.1	6.2	3.0	1.4	8.1	4.1	6.8	7.1	9.9	13.8	6.2	1.7	4.4	5.2	155.8
fr	French	Romance	1869	-	-	-	-	-	-	-	-	-	-	6.8	3.5	10.3	11.9	46.2	12.6	6.9	4.2	32.1	9.9	21.1	37.9	31.9	27.6	17.4	4.2	12.5	14.0	619.8
he	Hebrew	Semitic	70	-	-	-	-	-	-	-	-	-	-	-	1.2	1.9	2.8	4.0	5.3	2.5	1.1	4.2	2.0	3.6	4.3	6.4	5.1	4.4	1.2	3.6	3.6	92.9
hi	Hindi	Indo-Aryan	48	-	-	-	-	-	-	-	-	-	-	-	-	1.3	1.9	2.3	2.7	1.6	0.9	2.4	1.4	2.1	2.6	3.4	3.0	3.2	0.8	1.9	2.4	56.0
hu	Hungarian	Uralic	148	-	-	-	-	-	-	-	-	-	-	-	-	-	3.2	7.0	5.2	2.6	1.3	7.1	3.0	7.1	6.8	9.6	7.4	5.6	1.7	3.7	4.6	139.6
id	Indonesian	Malayo-Polynesian	366	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.4	5.9	3.5	4.4	7.6	3.7	6.0	9.1	9.9	8.6	8.1	1.7	7.9	6.3	172.9
it	Italian	Romance	686	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.9	4.7	2.5	16.6	6.1	14.7	25.4	20.5	16.0	10.5	2.8	8.0	8.6	368.4
ja	Japanese	Japonic	2944	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.3	8.9	5.1	7.7	9.1	11.6	11.3	12.1	2.8	6.5	13.5	228.7
ko	Korean	Koreanic	778	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.9	4.8	2.6	4.0	4.9	6.0	7.1	8.4	1.4	5.2	6.3	113.7
ms	Malay	Malayo-Polynesian	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.6	1.3	2.3	2.8	3.7	3.6	3.4	0.8	3.2	2.8	60.8
nl	Dutch	Germanic	510	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.8	12.9	15.5	17.7	20.8	11.0	2.7	7.2	8.4	322.2
no	Norwegian	Germanic	109	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.5	6.4	8.1	13.8	5.2	1.4	3.9	4.3	143.8
pl	Polish	Slavic	505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13.5	22.9	13.8	9.1	3.4	6.5	7.1	267.1
pt	Portuguese	Romance	729	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.9	15.7	11.0	3.0	8.8	9.5	375.2
ru	Russian	Slavic	3047	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18.9	15.3	31.2	10.4	13.0	475.0
sv	Swedish	Germanic	1200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.8	10.6	10.4	358.5
tr	Turkish	Turkic	1382	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.5	10.4	10.0	247.4
uk	Ukrainian	Slavic	110	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2	2.2	88.6
vi	Vietnamese	Vietic	1172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9.1	190.2
zh	Chinese	Chinese	2512	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	214.3

Table 1: CCMatrix: size of mined sentences **(in millions)** for each language pair.

- Hindi with Chinese, Japanese, Korean, Indonesian: ≈2M

Example Parallel Sentences

Japanese	国内レベルで進捗状況を監視するためには、良質かつアクセス可能な適時のデータ収集や、地域的なフォローアップと検証が必要となります *(Monitoring progress at the national level requires quality, accessible and timely data collection and regional follow-up and verification.)*
Russian	Для мероприятий по отслеживанию прогресса на национальном уровне необходимо обеспечить сбор качественных, доступных и актуальных данных, а также проведение последующей деятельности и обзора на региональном уровне. *(For activities to track progress at the national level, it is necessary to ensure the collection of quality, accessible and relevant data, as well as follow-up and review at the regional level.)*

Example Parallel Sentence

Malay	Tahun ketiga pengajian biasanya dibelanjakan ke luar negara di institusi rakan kongsi di Timur Tengah atau Afrika Utara. *(The third year of study is usually spent abroad at partner institutions in the Middle East or North Africa.)*
Chinese	研究的第三年通常是在中东或北非伙伴机构在国外度过。 *(The third year of the study is usually spent abroad in partner institutions in the Middle East or North Africa.)*

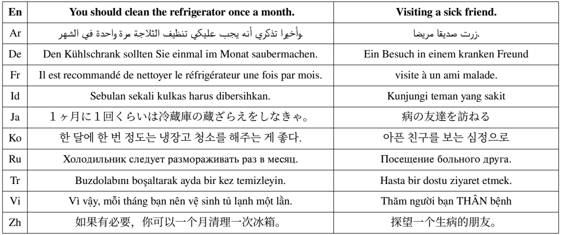
Example Parallel Sentence

Arabic	ذهب الحراس إلى الأبواب البلوط الكبيرة وهيلين وجاك نزل إلى اسطبلات لجعل لأجل اثنين من الخيول. *(The guards went to the large oak doors, Helen and Jack came down to stables to make for two horses.)*
Hebrew	השומרים חזרו אל דלתות עץ האלון הגדולות והלנה וג'ק ירדו לאורוות להכין שני סוסים. *(The guards returned to the large oak doors and Helena and Jack went down to the stables to make two horses.)*

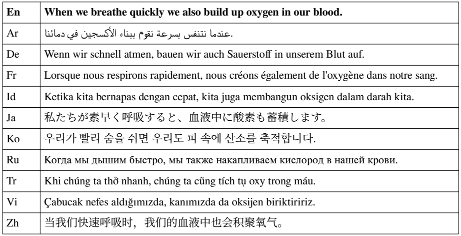
Example Parallel Sentence

Finish	Tarkista aina väliin, että olet oikealla tiellä. *(Always check in between that you are on the right track.)*
Tamil	எப்போதும் உறுதிப்படுத்திக் கொள்ளுங்கள் நீங்கள் சரியான வழியில் செல்வீர்கள். *(Always make sure you go the right way.)*

11-way Parallel Sentences

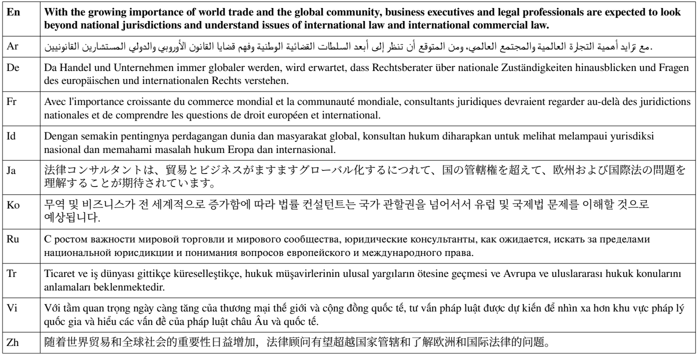


En	You should clean the refrigerator once a month.	Visiting a sick friend.
Ar	وأخيرا تذكري أنه يجب عليكي تنظيف الثلاجة مرة واحدة في الشهر.	زرت صديقا مريضا.
De	Den Kühlschrank sollten Sie einmal im Monat saubermachen.	Ein Besuch in einem kranken Freund
Fr	Il est recommandé de nettoyer le réfrigérateur une fois par mois.	visite à un ami malade.
Id	Sebulan sekali kulkas harus dibersihkan.	Kunjungi teman yang sakit
Ja	１ヶ月に１回くらいは冷蔵庫の蔵ざらえをしなきゃ。	病の友達を訪ねる
Ko	한 달에 한 번 정도는 냉장고 청소를 해주는 게 좋다.	아픈 친구를 보는 심정으로
Ru	Холодильник следует размораживать раз в месяц.	Посещение больного друга.
Tr	Buzdolabını boşaltarak ayda bir kez temizleyin.	Hasta bir dostu ziyaret etmek.
Vi	Vì vậy, mỗi tháng bạn nên vệ sinh tủ lạnh một lần.	Thăm người bạn THÂN bệnh
Zh	如果有必要，你可以一个月清理一次冰箱。	探望一个生病的朋友。



En	When we breathe quickly we also build up oxygen in our blood.
Ar	عندما نتنفس بسرعة نقوم ببناء الأكسجين في دمائنا.
De	Wenn wir schnell atmen, bauen wir auch Sauerstoff in unserem Blut auf.
Fr	Lorsque nous respirons rapidement, nous créons également de l'oxygène dans notre sang.
Id	Ketika kita bernapas dengan cepat, kita juga membangun oksigen dalam darah kita.
Ja	私たちが素早く呼吸すると、血液中に酸素も蓄積します。
Ko	우리가 빨리 숨을 쉬면 우리도 피 속에 산소를 축적합니다.
Ru	Когда мы дышим быстро, мы также накапливаем кислород в нашей крови.
Tr	Khi chúng ta thở nhanh, chúng ta cũng tích tụ oxy trong máu.
Vi	Çabucak nefes aldığımızda, kanımızda da oksijen biriktiririz.
Zh	当我们快速呼吸时，我们的血液中也会积聚氧气。

11-way Parallel Sentences



En	**With the growing importance of world trade and the global community, business executives and legal professionals are expected to look beyond national jurisdictions and understand issues of international law and international commercial law.**
Ar	مع تزايد أهمية التجارة العالمية والمجتمع العالمي، ومن المتوقع أن تنظر إلى أبعد السلطات القضائية الوطنية وفهم قضايا القانون الأوروبي والدولي المستشارين القانونيين.
De	Da Handel und Unternehmen immer globaler werden, wird erwartet, dass Rechtsberater über nationale Zuständigkeiten hinausblicken und Fragen des europäischen und internationalen Rechts verstehen.
Fr	Avec l'importance croissante du commerce mondial et la communauté mondiale, consultants juridiques devraient regarder au-delà des juridictions nationales et de comprendre les questions de droit européen et international.
Id	Dengan semakin pentingnya perdagangan dunia dan masyarakat global, konsultan hukum diharapkan untuk melihat melampaui yurisdiksi nasional dan memahami masalah hukum Eropa dan internasional.
Ja	法律コンサルタントは、貿易とビジネスがますますグローバル化するにつれて、国の管轄権を超えて、欧州および国際法の問題を理解することが期待されています。
Ko	무역 및 비즈니스가 전 세계적으로 증가함에 따라 법률 컨설턴트는 국가 관할권을 넘어서서 유럽 및 국제법 문제를 이해할 것으로 예상됩니다.
Ru	С ростом важности мировой торговли и мирового сообщества, юридические консультанты, как ожидается, искать за пределами национальной юрисдикции и понимания вопросов европейского и международного права.
Tr	Ticaret ve iş dünyası gittikçe küreselleştikçe, hukuk müşavirlerinin ulusal yargıların ötesine geçmesi ve Avrupa ve uluslararası hukuk konularını anlamaları beklenmektedir.
Vi	Với tầm quan trọng ngày càng tăng của thương mại thế giới và cộng đồng quốc tế, tư vấn pháp luật được dự kiến để nhìn xa hơn khu vực pháp lý quốc gia và hiểu các vấn đề của pháp luật châu Âu và quốc tế.
Zh	随着世界贸易和全球社会的重要性日益增加，法律顾问有望超越国家管辖和了解欧洲和国际法律的问题。

Evaluating CCMatrix

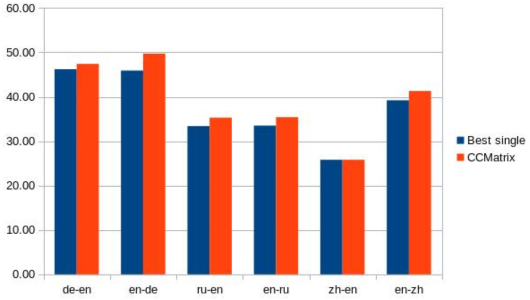
WMT'19

- De-facto standard for NMT progress, strong competition
- Train NMT systems **on mined data only**, no human bitexts

Evaluating CCMatrix

WMT'19

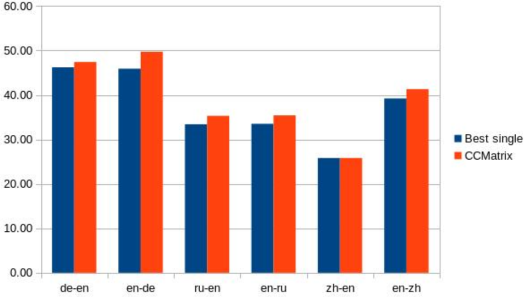
- De-facto standard for NMT progress, strong competition
- Train NMT systems **on mined data only**, no human bitexts
- Newstest 2018:



Evaluating CCMatrix

WMT'19

- De-facto standard for NMT progress, strong competition
- Train NMT systems **on mined data only**, no human bitexts
- Newstest 2018:

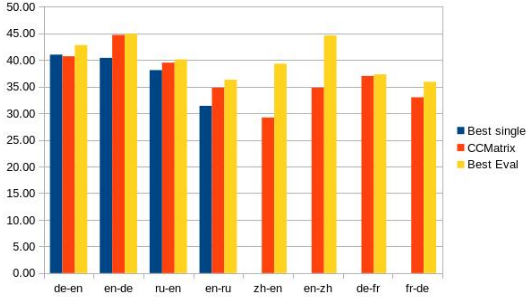


- We outperform all best single systems, +3.8 BLEU en-de

Evaluating CCMatrix

WMT'19

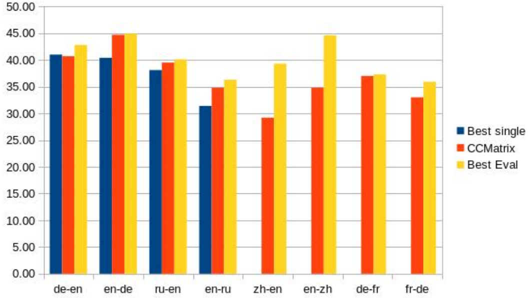
- De-facto standard for NMT progress, strong competition
- Train NMT systems **on mined data only**, no human bitexts
- Newstest 2019:



Evaluating CCMatrix

WMT'19

- De-facto standard for NMT progress, strong competition
- Train NMT systems **on mined data only**, no human bitexts
- Newstest 2019:



- en-de/de-fr: on-pair with eval system (BT, sys.comb)

BLEU Scores on TED Test Sets

	ar	bg	cs	da	de	el	en	es	fa	fi	fr	he	hi	id	it	ja	ko	ms	nl	no	pl	pt	ru	sv	tr	uk	vi	zh
ar	-	16.6	11.5	14.9	15.5	17.5	28.7	22.4	8.7	7.3	19.6	10.9	12.6	17.5	18.6	8.5	2.8	10.2	15.8	16.2	10.2	20.7	13.8	18.8	7.6	7.1	19.0	9.9
bg	9.5	-	19.3	24.9	22.9	24.4	36.3	27.8	9.8	11.9	26.1	13.7	14.4	22.0	23.3	9.8	3.4	12.9	21.8	22.2	15.0	26.1	19.3	24.1	9.9	13.9	21.8	10.5
cs	7.0	21.9	-	21.8	21.0	20.0	29.2	24.0	7.9	13.8	24.1	10.8	13.9	18.8	20.5	9.9	3.5	8.5	21.3	22.6	16.1	22.5	18.1	22.6	9.2	11.4	20.2	10.3
da	6.5	25.3	18.1	-	26.8	23.5	44.1	28.9	8.4	14.3	27.6	12.6	15.8	22.2	24.3	10.3	3.5	14.7	28.2	32.1	16.3	27.0	18.9	33.4	10.3	9.9	19.9	9.3
de	9.2	24.0	19.8	30.4	-	22.3	35.8	28.1	9.7	13.1	27.6	13.7	18.4	22.5	24.3	10.5	4.5	11.7	26.8	20.5	15.5	26.9	18.3	26.6	11.4	12.3	22.6	11.3
el	9.8	24.4	17.2	25.0	21.3	-	35.8	28.4	9.3	11.8	26.7	13.8	16.0	21.5	23.8	9.7	3.2	13.4	22.1	24.1	13.9	26.4	17.5	24.1	10.1	10.3	22.2	10.9
en	16.6	35.7	24.5	42.2	32.6	34.6	-	42.4	15.7	17.6	36.6	24.8	25.2	33.4	34.1	12.3	5.8	24.8	33.3	43.2	18.4	41.2	21.9	38.2	16.1	19.2	29.5	15.1
es	11.6	26.4	19.4	28.7	25.1	26.2	41.1	-	10.9	14.4	30.7	15.7	17.9	24.2	29.9	10.9	4.5	-	26.2	25.0	15.7	32.8	19.4	26.9	11.4	13.9	24.7	12.3
fa	7.0	14.7	10.7	14.4	14.2	14.6	25.1	18.2	-	5.7	17.2	8.0	9.1	16.9	15.2	7.6	2.3	9.3	14.2	11.8	8.5	17.5	12.3	15.6	8.1	6.4	17.7	8.8
fi	4.7	14.2	13.2	17.7	14.6	13.8	21.6	18.1	4.0	-	16.1	7.8	11.9	13.3	14.1	9.5	2.8	2.2	16.1	13.7	10.4	14.8	11.6	16.6	7.2	6.3	16.1	8.0
fr	10.4	25.9	19.7	29.5	25.4	26.6	40.2	32.6	10.6	13.8	-	15.6	18.8	25.5	29.6	10.8	4.6	13.7	26.7	24.8	16.3	31.2	20.0	27.2	12.0	13.4	24.6	10.3
he	9.6	19.0	13.8	19.8	18.2	18.9	33.4	24.0	7.7	9.7	23.2	-	12.8	18.8	20.2	7.9	3.1	8.4	18.3	17.2	11.7	22.7	15.9	21.1	8.1	7.5	18.8	9.2
hi	4.1	10.3	7.9	11.8	14.2	11.4	24.3	16.5	3.6	6.5	17.2	6.9	-	13.4	12.9	6.5	1.9	6.7	12.9	9.6	7.1	15.4	12.4	13.9	6.8	3.8	15.3	6.5
id	8.3	19.9	14.3	21.4	19.6	19.1	31.9	24.7	9.9	9.8	23.7	11.7	16.4	-	20.2	10.1	4.6	19.0	20.8	21.2	12.8	23.5	16.0	21.2	10.2	9.6	23.7	11.5
it	11.0	24.2	18.4	26.4	24.1	25.4	36.9	32.5	10.2	13.0	30.4	14.1	16.8	23.1	-	10.7	3.8	13.2	24.8	25.3	15.4	30.8	18.7	26.9	11.2	12.1	23.7	11.4
ja	3.8	7.4	5.9	8.2	7.8	7.7	11.8	11.6	4.3	4.6	10.7	4.0	9.1	9.6	9.2	-	3.1	5.5	8.4	7.4	6.1	9.9	7.3	8.1	4.6	3.3	12.0	6.9
ko	4.2	8.6	7.4	9.8	10.0	8.8	15.4	13.7	5.1	5.6	13.2	5.2	10.8	12.5	10.5	11.0	-	6.7	10.2	9.9	6.7	12.1	8.8	11.3	6.1	3.8	14.3	8.1
ms	7.4	11.9	8.8	14.2	13.3	13.6	29.1	-	9.0	5.2	16.9	6.4	13.0	20.9	18.0	8.6	2.5	-	14.0	14.8	8.8	19.4	12.9	20.1	7.8	4.8	23.8	10.1
nl	8.7	21.8	18.0	28.3	25.6	21.9	35.8	28.3	9.3	12.8	27.9	13.4	16.9	22.7	23.7	10.4	3.7	12.2	-	18.1	15.5	26.6	17.2	25.4	10.6	9.8	22.0	10.7
no	9.7	23.3	19.8	34.0	21.9	24.2	45.2	27.0	5.3	12.4	25.6	12.7	13.5	22.9	26.8	10.8	3.5	16.8	19.1	-	13.1	27.6	18.8	32.5	9.7	11.0	16.4	10.7
pl	6.5	17.3	16.2	19.5	16.8	15.9	22.3	20.0	6.5	10.2	20.0	9.3	12.9	16.3	17.4	9.4	3.0	9.8	17.5	14.4	-	18.8	15.4	17.4	7.6	10.6	17.7	8.5
pt	11.5	26.6	19.3	29.1	25.0	27.3	43.0	35.7	11.0	13.0	31.9	16.0	18.5	25.9	30.5	10.6	4.2	15.1	26.3	24.3	16.2	-	19.7	26.6	11.5	13.2	25.1	10.0
ru	8.0	19.6	15.9	18.8	18.1	18.3	24.2	21.8	8.6	10.6	22.1	11.2	16.0	17.8	19.0	10.1	3.5	11.7	18.2	19.3	14.1	20.7	-	19.7	8.4	21.1	19.3	10.3
sv	11.1	24.4	20.2	34.3	26.3	24.4	40.8	28.8	9.4	14.9	27.7	15.4	18.7	24.3	26.3	10.8	4.6	15.6	25.7	30.3	14.7	28.3	18.8	-	12.0	11.2	24.9	12.1
tr	7.5	16.4	12.8	16.2	16.4	16.6	25.0	20.3	8.7	9.6	19.9	9.2	15.8	17.5	16.9	9.6	4.1	10.8	16.6	14.7	10.8	18.8	13.3	18.2	-	7.3	19.2	9.9
uk	5.0	17.2	12.0	13.8	14.7	13.2	23.1	18.6	5.4	8.0	17.7	6.9	8.7	12.8	15.3	7.0	1.7	6.2	13.1	12.9	12.3	17.2	23.0	13.6	5.6	-	14.5	7.3
vi	8.0	16.8	12.9	17.1	16.5	17.0	25.8	21.7	8.7	9.3	21.0	9.9	15.7	21.3	18.3	9.6	4.3	16.5	17.6	16.1	11.0	20.6	14.1	19.0	9.4	9.1	-	10.8
zh	6.3	11.8	9.3	11.2	12.2	12.3	18.3	16.0	6.9	7.4	15.2	7.6	12.3	14.8	13.4	9.6	3.5	9.7	12.8	12.6	8.4	14.0	11.2	13.8	6.8	6.1	18.1	-

- Same NMT system for all language pairs (despite huge difference in bitext size)
- Best: BLEU 45.2 for Norwegian/English

CCMatrix: What's Next ?

Scaling even further

- Scaling to 32 crawls, 100 languages

$\Rightarrow$ $\approx$ 10 billion bitexts

- Further improvements on WMT evaluation

CCMatrix: What's Next ?

Scaling even further

- Scaling to 32 crawls, 100 languages

⇒ ≈ 10 billion bitexts

- Further improvements on WMT evaluation

Sharing our results

- Looking for means to share these CCMatrix bitexts

parallel sentence filtering

Filtering for What?

- We have intuitive notions of useful training data
  - source and target match in meaning
  - both are well-formed text

Filtering for What?

- We have intuitive notions of useful training data
  - source and target match in meaning
  - both are well-formed text
- But: the right question is: does it help to build a better MT system
- We do not know how to answer that

Types of Noise

- Misaligned sentences
- Disfluent language (from MT, bad translations)
- Wrong language data (e.g., French in German–English corpus)
- Untranslated sentences
- Short segments (e.g., dictionaries)
- Mismatched domain

Mismatched Sentences

- Artificial created by randomly shuffling sentence order
- Added to existing parallel corpus in different amounts

5%		10%		20%		50%		100%	
	24.0		24.0		23.9	26.1	23.9	25.3	23.4
	-0.0		-0.0		-0.1	-1.1	-0.1	-1.9	-0.6

- Bigger impact on NMT (green, left) than SMT (blue, right)

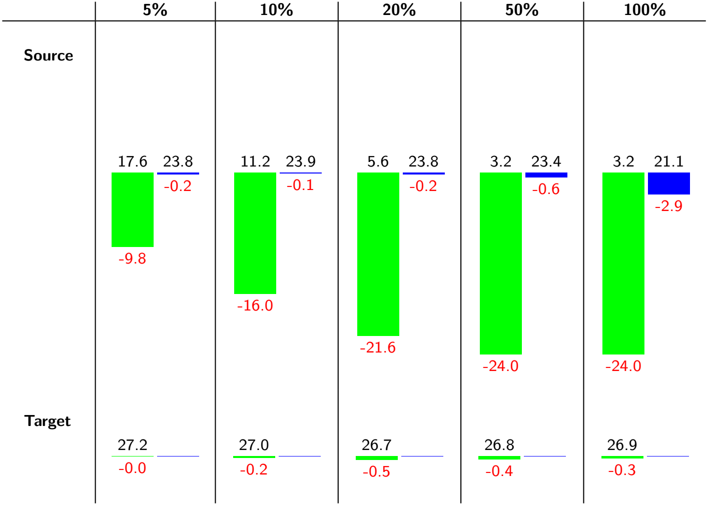
Misordered Words

- Artificial created by randomly shuffling words in each sentence

	5%		10%		20%		50%		100%	
Source		24.0 -0.0		23.6 -0.4		23.9 -0.1	26.6 -0.6	23.6 -0.4	25.5 -1.7	23.7 -0.3
Target		24.0 -0.0		24.0 -0.0		23.4 -0.6	26.7 -0.5	23.2 -0.8	26.1 -1.1	22.9 -1.1

- Similar impact on NMT than SMT, worse for source reshuffle

Untranslated Sentences



Copy Noise

- Harmfulness of copy noise also discovered by Ott, Auli, Granger, Ranzato (Facebook FAIR)
  - noticed link to beam search decoding
  - proposed remedies at inference time
- Motivated overlap penalty as feature in data filtering

Zipporah

Zipporah: A Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora

Zipporah: Motivation & Objective

Motivation

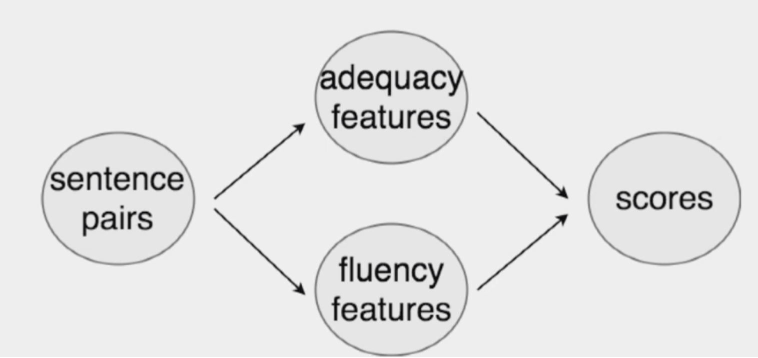
- Often have large pool of noisy parallel data
- Need to perform fast data-selection to select a higher-quality subset of this data

Objective

- Design a function to rank the sentence pairs
- Select the best sentences under some size constraint

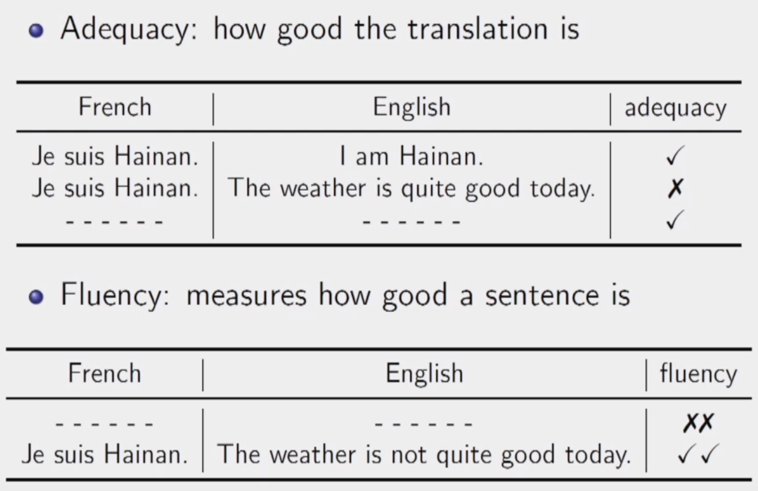
Zipporah: Features

Features



Zipporah: Features

Features



- Adequacy: how good the translation is

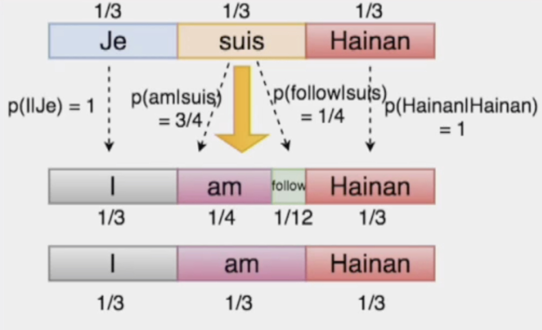
French	English	adequacy
Je suis Hainan.	I am Hainan.	✓
Je suis Hainan.	The weather is quite good today.	✗
- - - - - -	- - - - - -	✓

- Fluency: measures how good a sentence is

French	English	fluency
- - - - - -	- - - - - -	✗✗
Je suis Hainan.	The weather is not quite good today.	✓✓

Zipporah: Adequacy Features

- Dictionary
  - P(I—Je) = 1
  - P(am—suis) = 3/4
  - P(follow—suis) = 1/4



Zipporah: Adequacy Features

$$XEnt(p,q) = \sum_i p(i) log \frac{1}{q(j)} \tag{10}$$

- A(en—fr) = 1/3 * log 3 + 1/3 * log 4 + 1/3 log 3 = 1.1945
- Also compute A(fr—en), given e2f dictionary
- For each sentence pair, define A(fr—en) + A(en—fr) as the adequacy feature
- Small when the translation is good (and literal)

Zipporah: Fluency Features

$$F(s) = -\frac{log(p_{LM}(s))}{lengh(s)} \tag{11}$$

- Tarn ngram LMs for both languages
- For each sentence, we compute the $F(s)$
- For each sentence pair, define $F(en) + F(fr)$ as the fluency feature
- Small when the sentence pair is fluent

Zipporah: Scoring Function

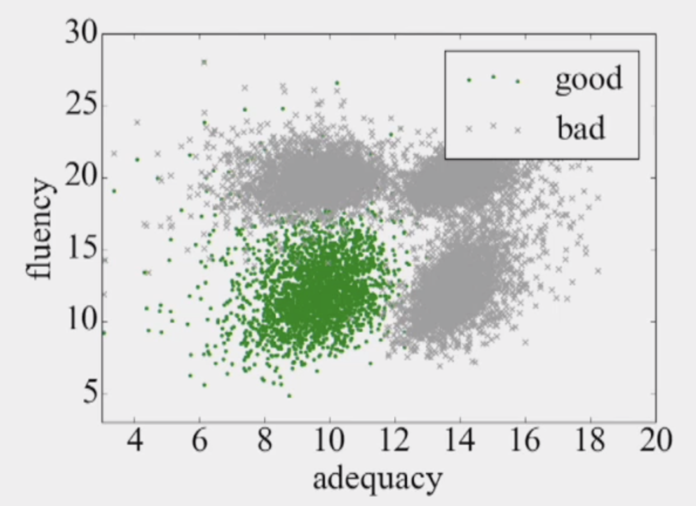
- Goal: train a classifier to distiinguish between good and bad data
- Have good data (true parallel sentences)
- Need bad data. Preferably one that covers all types f bad data in the feature space
- Auto generate bad data from good data

Zipporah: Generating Bad Data

Starting from a good dev corpus

- Shuffly individual words within sentences (bad fluency)
- shuffle sentences (bad adequacy)
- Shuffle both (bad both)

Zipporah: Bad Data vs Good Data



Zipporah: Logistic Regression Classifier

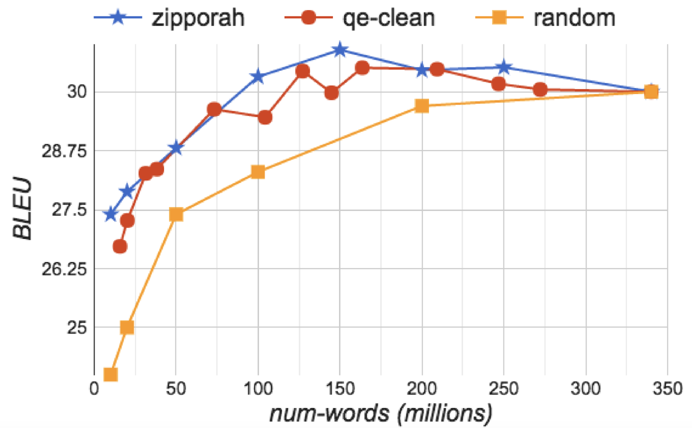
- Task: To separate the good parallel data from (synthetic) bad parallel data
- Method: Logistic regression classifier with polynomials of features.
- Use the trained weights to compute a signed-distance to the decision boundary as score

Zipporah: Baselines

- Random selection
- QE Clean: Uses LM scores and word-alignment scores to perform data selection

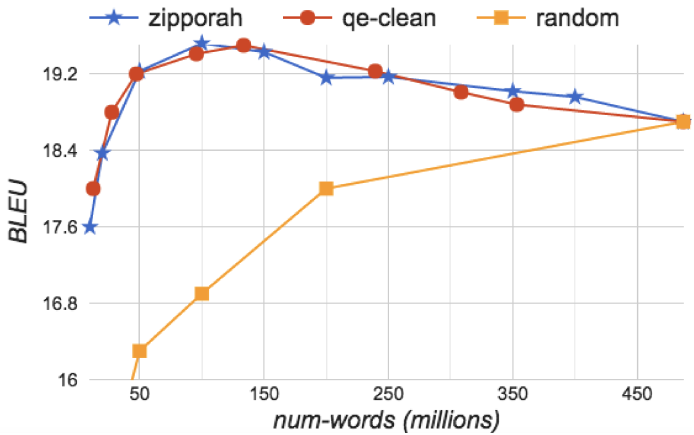
Zipporah: Results

French-English: Ted Talks Dataset



Zipporah: Results

German-English: Newstest 11 Datasett



WMT Shared Task on Sentence Pair Filtering

WMT Shared Task on Sentence Pair Filtering

WMT Shared Task on Sentence Pair Filtering

- Shared Task in 2018: High Resouce
  - German–English
  - 1 billion words of noisy parallel data
  - 100+ million words of clean parallel data
- Shared Task in 2019: Low Resource
  - Sinhala–English and Nepali–English
  - 50-60 million words of noisy parallel data
  - 3-4 million words of *relatively* clean parallel data

Task Definition

- Given
  - very noise web crawled corpus
  - sentence-aligned
  - 50-60 billion English words

Task Definition

- Given
  - very noise web crawled corpus
  - sentence-aligned
  - 50-60 billion English words
- Submission: sentence-level quality score for each sentence pair

Task Definition

- Given
  - very noise web crawled corpus
  - sentence-aligned
  - 50-60 billion English words
- Submission: sentence-level quality score for each sentence pair
- Evaluation
  - subselection of training corpus based on quality threshold
    - 1 million English words
    - 5 million English words
  - machine translation performance on undisclosed test sets
    - statistical machine translation (Moses)
    - neural machine translation (fairseq)

Provided Resources

- Noisy parallel corpus
    - English sentence
    - foreign sentence
    - Hunalign score
- Training data

Provided Resources

- Noisy parallel corpus
  - English sentence
  - foreign sentence
  - Hunalign score
- Training data
- Development pack
  - script to subsample corpora
  - Moses configuration file to build and test SMT system
  - Fairseq scripts to build and test NMT system
  - Development and test sets: Wikipedia translations

Clean Parallel Corpora

Nepali	Sentence Pairs	English Words
Bible (two translations)	61,645	1,507,905
Global Voices	2,892	75,197
Penn Tree Bank	4,199	88,758
GNOME/KDE/Ubuntu	494,994	2,018,631
Nepali Dictionary	9,916	25,058

Sinhala	Sentence Pairs	English Words
Open Subtitles	601,164	3,594,769
GNOME/KDE/Ubuntu	45,617	150,513

Development and Test Sets

- Evaluation on translations of Wikipedia content

  Two New Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English, Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, Marc'Aurelio Ranzato, *arXiv:1902.01382*

	Nepali		**Sinhala**	
	Sentence Pairs	**English Words**	**Sentence Pairs**	**English Words**
dev	2,559	46,274	2,898	53,479
dev test	2,835	51,458	2,766	50,985
test	2,924	54,062	2,905	52,851

Participants

Acronym	Participant and System Description Citation
AFRL	Air Force Research Lab, USA
DiDi	DiDi, USA
Facebook	Facebook, USA
Helsinki	University of Helsinki, Finland
IITP	Indian Institute of Technology Patna, India
Webinterpret	WebInterpret Inc., USA
NRC	National Research Council, Canada
Stockholm	Stockholm University, Sweden
SUNY Buffalo	State University of New York, USA
Sciling	Sciling S.L., Spain
TALP-UPC	TALP, Universitat Politècnica de Catalunya, Spain

methods

AFRL

AFRL

- Uses coverage metric and quality metric.
- Coverage metric discourages addition of sentence pairs that have vocab already included in selected set
- Quality metric based on comparing machine translation of foreign sentences with English sent using METEOR MT metric

DiDi

DiDi

- Dual cross-entropy based on monolingual language models to find pairs where each sentence has similar probability
- Cynical data selection that prefers to select representative subset
- Length-ratio and using character-set based language identification

Facebook

Facebook

- Ensemble
- Matching of cross-lingual sentence embeddings feature
- Dual cross entropy based on neural translation model scores
- Open source Ziporah and Bicleaner

NRC

NRC

- Filtering rules based on lang ID, length ratio, mismatched numbers, near duplicates
- Cross-lingual semantic evaluation metric (Yisi-2) that uses:
  - cross-lingual word embeddings
  - transformer model language model pretrained based on XLM
  - optimized to distinguish between clean parallel data and synthetic noisy parallel data
- Reranking to increase coverage

Sciling

Sciling

- Build translation models on clean data
- translate non-english to English in noisy data
- Similarity between machine translation and given English sentence
- Filtering rules for sent length, source-target overlap, and lang identification

Stockholm

Stockholm

- Filtering (excessive numbers, too few words, sentence length, too long, etc)
- Mono-lingual word embeddings with FastText
- learn projection between emebdding spaces based on word alignment from parallel data
- Cosine similarity between English word to best matching projection of the word

TALP-UPC

TALP-UPC

- Monolingual word embeddings with FastText
- Unsupervised word ealignment
- Word mover's distance between sentences
- Filtering rules (sent length, lang identification, num mismatches)

Helsinki

TALP-UPC

- Clean the clean parallel data using filter rules (sent length, sents with long words, XML, HTML, tags, wrong script)
- Obtain word alignments from this clean data
- Noisy parallel data is scored using word alignments
- Filtered with language models, lang identifiication, ratio of chars in correct script, punctuation, number matching, length mismatch

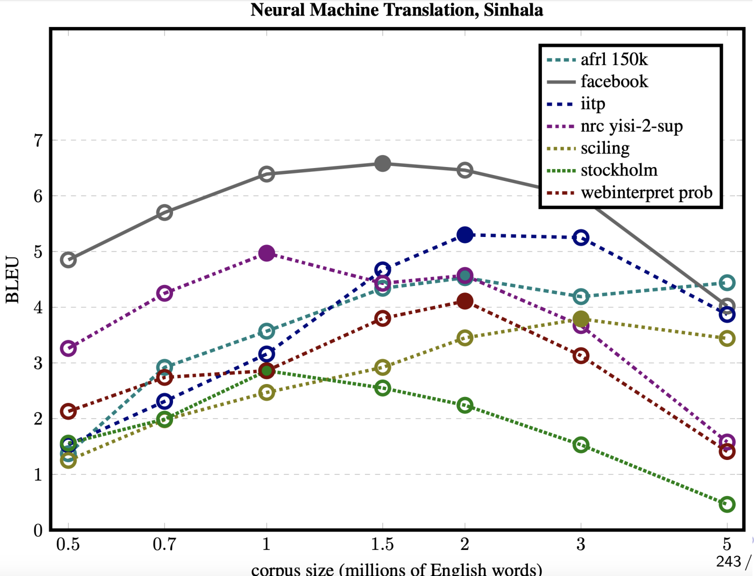
Webinterpret

Webinterpret

- Filtering rules based on language identification and sent length
- Coverage ranking incrementally adds sentence pairs to increase vocan and ngram coverage
- Adequacy ranking considers IBM Model 1 word translation scores

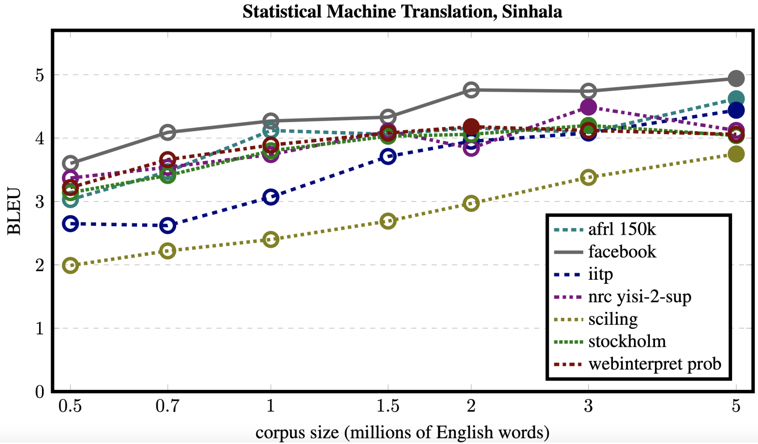
Different Subset Sizes

Neural Machine Translation, Sinhala



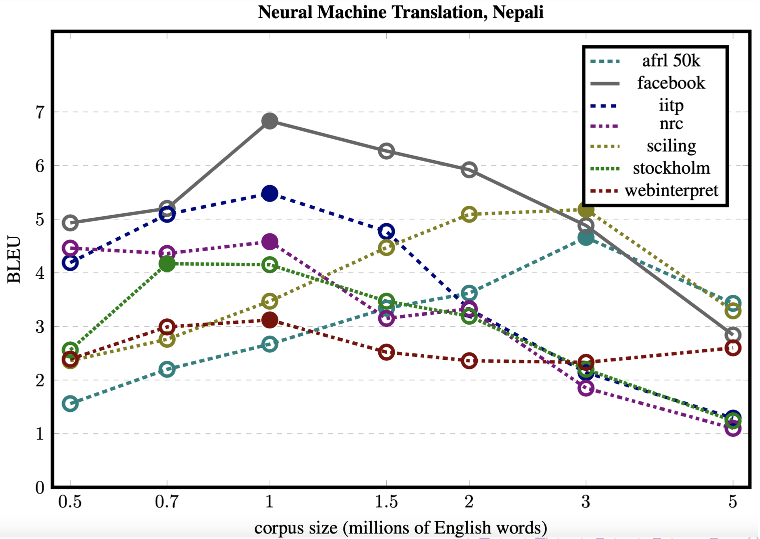
Different Subset Sizes

Statistical Machine Translation, Sinhala



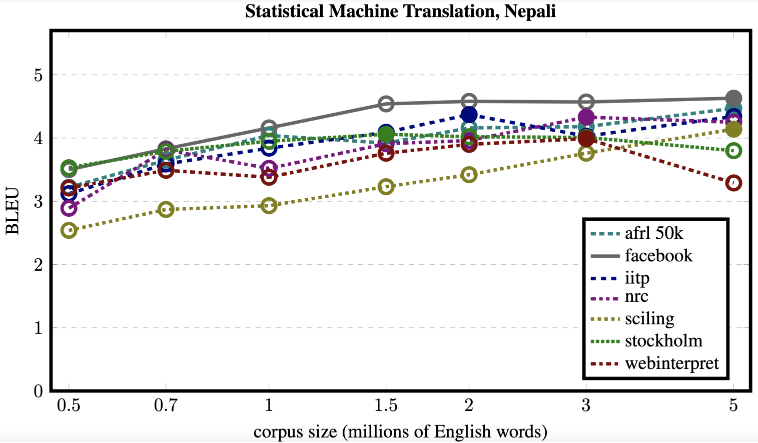
Different Subset Sizes

Neural Machine Translation, Nepali



Different Subset Sizes

Statistical Machine Translation, Nepali



Things Learned

Commonalities Learned from Submissions

Pre-Filtering Rules

- Discard some data based on deterministic filtering rules
    - too short or too long
    - too many non-words
    - average token length is too short or too long
    - mismatched lengths
    - names, numbers, email addresses, URLs do not match between both sides
    - too similar, indicating simple copying
    - language identification

Embeddings

- Cross-lingual sentence embeddings
  - central to best performing system
  - LASER (Artexte and Schwenk, 2018)
- Word embeddings
  - monolingual spaces, mapped unsupervised or using dictionaries
  - bilingually trained word embeddings

Use of Machine Translation Models

- Quality scores on translations
  - translate foreign into English
  - score with METEOR, BLEU, Levenshtein distance
- Cross-entropy filtering
  - force-translate foreign into given English
  - consider translation model score

Scoring Functions

- N-gram or neural language models on clean data
- Language models trained on the provided raw data as contrast
- Neural translation models
- Bag-of-words lexical translation probabilities
- Off-the-shelf tools: Zipporah, Bicleaner

Learning Weights for Scoring Functions

- Large number of scoring functions → averaging scores inadequate
- Learning weights to optimize MT quality computationally intractable
- Solution: train classifier to distinguish between good and bad sentence pairs
    - good sentence pairs from clean corpus
    - bad sentence pairs from provided data, or synthetic noise

Low-Resource Corpus Filtering using Multilingual Sentence Embeddings

Approach

- Leverage LASER multilingual embeddings as a tool to measure parallel sentence quality
- Margin-based scoring function to score sentence pairs

Scoring Function

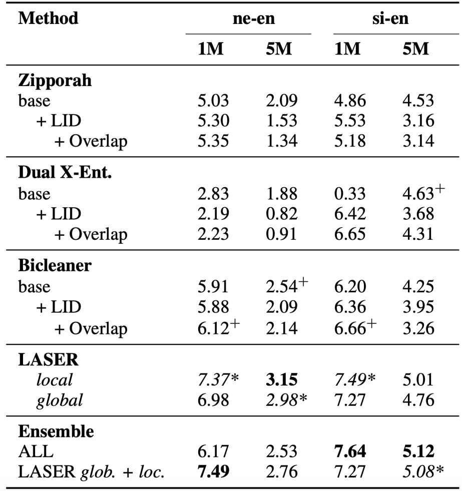
$$\frac{2k\ cos(x,y)}{\sum_{y'\in NN_k(x)} cos(x,y') + \sum_{x'\in NN_k(y)} cos(x'y))}$$

where

- $NN_k(x)$ denotes the k nearest neighbors of x in the other language and analogously for $NN_k(y)$.
- pool of sentences are deduplicated
- Global: pool of neighbors can be from global (all = clean + noisy data)
- Local: pool of neighbors can be from local (only from noisy data)

Results

Dev Test Results

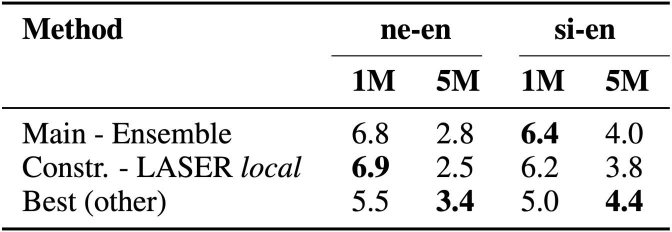


Method	**ne-en**		**si-en**	
	1M	**5M**	**1M**	**5M**
Zipporah				
base	5.03	2.09	4.86	4.53
+ LID	5.30	1.53	5.53	3.16
+ Overlap	5.35	1.34	5.18	3.14
Dual X-Ent.				
base	2.83	1.88	0.33	4.63^{+}
+ LID	2.19	0.82	6.42	3.68
+ Overlap	2.23	0.91	6.65	4.31
Bicleaner				
base	5.91	2.54^{+}	6.20	4.25
+ LID	5.88	2.09	6.36	3.95
+ Overlap	6.12^{+}	2.14	6.66^{+}	3.26
LASER				
local	*7.37**	**3.15**	*7.49**	5.01
global	6.98	*2.98**	7.27	4.76
Ensemble				
ALL	6.17	2.53	**7.64**	**5.12**
LASER *glob.* + *loc.*	**7.49**	2.76	7.27	*5.08**

Bold=best scores, Italics*= runner up

Results

Test Results



Method	**ne-en**		**si-en**	
	1M	**5M**	**1M**	**5M**
Main - Ensemble	6.8	2.8	**6.4**	4.0
Constr. - LASER *local*	**6.9**	2.5	6.2	3.8
Best (other)	5.5	**3.4**	5.0	**4.4**

Welalin
Terima Kasih
Kiitos
Salamat
Grazie
Dankon
Multumesc
Merci
Asante
Spasibo
Thank
Matondo
Mator Nuwun
Nirringrazzjak
Ua Tsaug Rau Koj
You
Dank Je
Raibh Maith Agat
Mochchakkeram
Obrigado
Arigato
Kia Ora
Cam on ban