Bitext

Multilinguisme — lecture n°1

Guillaume Wisniewski

guillaume.wisniewski@u-paris.fr

Université de Paris & LLF

October 8th 2025

Definition

Bitext

- originally: translation studies [Harris, 1988]
- documents ⊕ translation in another language

Now...

- pair of texts...
- ... with some translational equivalence
- two kinds of bitexts:
  1. parallel texts = document + its translation
     ↪ *Candide* and its translation in English
  2. comparable corpora = documents in different languages about the same domain but necessarily translation of each other
     ↪ all the tweets about FIFA World Cup
- note: in translation studies parallel text = comparable corpora

Example (1)

The Rosetta Stone

Example (2)

Source for Translations (1)

Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)

Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)
- corpus size to train a state-of-the-art system:

	# parallel sentences
fra-eng	1.1B
fra-spa	742M
fra-eus	2.1M
fra-nep	684K

source: `http://opus.nlpl.eu/`

Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)
- corpus size to train a state-of-the-art system:

	# parallel sentences
fra-eng	1.1B
fra-spa	742M
fra-eus	2.1M
fra-nep	684K

source: `http://opus.nlpl.eu/`

↪ manual translation is not an option

5/1

Source for Translations (2)

Source for Translations (2)

1. international organization (ONU, European Parliament, Hansard, ...)
   ↪ most of their documents are freely available & translated into several languages
2. software manuals (in particular: free software)
3. classical books
4. subtitles

6/1

Source for comparable data

Source for comparable data

Twitter
↪ using hashtag to identify tweets about the same topic

Wikipedia
↪ using interlanguage links

7/1

Setup to build parallel corpora

1. data collection: identify source of documents & fetch them
2. pre-processing: extract text from documents (OCR, remove formatting or XML tags)
3. document alignment: identify translation(s) of a given document
4. sentence alignment: align paragraph and sentence for each parallel document pair

⇒ a lot of hand-crafted rules and heuristics except the last step

More Formally

The problem

- $\mathcal{D}_s = \{d_{s_1}, ..., d_{s_m}\}$ (source language)
- $\mathcal{D}_t = \{d_{t_1}, ..., d_{t_n}\}$ (target language)

↪ find (d_{t_i}, d_{t_j}) so that d_{t_i} is a translation of d_{t_j}

Challenges in Multilingual Data

- Partial or approximate translations
- Typologically distant languages
- Lack of reliable metadata
- Noisy or poorly segmented text (OCR, web crawl, etc.)

Documents Alignment: difficulties

What to do with:

00_70221.isv	Swedish (sv)
03_70821.pl	Polish (pl)
03_70221.isf	Swedish and Finish (sf)
03_70421.ien.ny_utg_970627	English (en), new edition ("ny utgåva")
02_60422.idenyutg.960626	German (de), new edition
08_80505.ity	German (ty = tyska)
08_80505.iho	Dutch (ho = holländska)
01_60123.pnl	Dutch (nl)
12B_help_sd_sve.rtf	Swedish (sve=svenska)
12B_help_sp_eng.rtf	English (eng)
om10aen.01	English (en)

parallel Scania corpus (sang1996ParallelScania)

Documents Alignment: difficulties

What to do with:

00_70221.isv	Swedish (sv)
03_70821.pl	Polish (pl)
03_70221.isf	Swedish and Finish (sf)
03_70421.ien.ny_utg_970627	English (en), new edition ("ny utgåva")
02_60422.idenyutg.960626	German (de), new edition
08_80505.ity	German (ty = tyska)
08_80505.iho	Dutch (ho = holländska)
01_60123.pnl	Dutch (nl)
12B_help_sd_sve.rtf	Swedish (sve=svenska)
12B_help_sp_eng.rtf	English (eng)
om10aen.01	English (en)

parallel Scania corpus (sang1996ParallelScania)

↪ basic assumption: similar names ⇒ translations
↪ in a perfect world: same filename ⊕ ISO-639 language code

Document Alignment: methods

Name mapping

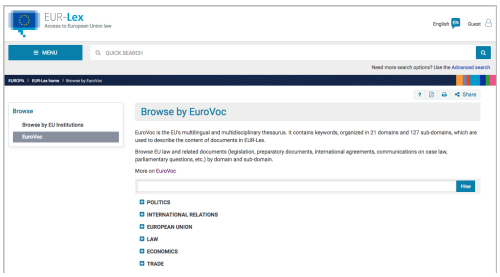
- handcrafted rules based on filenames
- tedious ⊕ source of errors

Content-Based similarity [Patry and Langlais, 2005]

- documents represented by a set of basic features (numbers, named entities, punctuations) ⊕ cosine similarity
- alignment score do we find the 'translation' of source words in the target documents?
- structure similarity
- looking for *hapax*
- using multilingual thesaurus (e.g. EuroVoc)

⇒ classifier

EuroVoc



Mining the Web (1): URL & Metadata

- Heuristic approach:
  - Identify URL patterns (e.g., `/en/` vs `/fr/`)
  - Detect cross-language links (tags `<link rel="alternate" hreflang=...>`)

  ⇒ high precision, low cost but (very) low recall
- Internal metadata:
  - Similar titles, comparable lengths, identical dates
  - Detection via content signatures (hash, fingerprint)

Mining the Web (2): Lexical and Multilingual Representations

1. Lexical representations
   - Automatic translation of titles or excerpts using dictionaries or MT systems
   - Similarity computation (cosine, Jaccard, etc.) between translated texts
   - ↪ translate French title into English and compare it with English title in the corpus
2. Multilingual vector representations
   - Use of multilingual embedding models (e.g. LASER, LaBSE, mUSE, NLLB-Embed)
   - For each document, compute an average vector of its sentence embeddings
   - Retrieve the nearest neighbour in the multilingual vector space

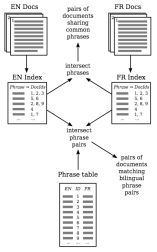
   ⇒ Standard current method (used in CCMatrix, CCAligned, etc.)
3. Sentence–document level similarity
   - Compare the distributions of sentence-level similarities
   - If many sentences show strong correspondences → documents are parallel
   - Used in modern bitext mining systems such as **schwenk-etal-2021-ccmatrix**

A solved task?

- WMT'16 shared task on document alignment **(buck-koehn-2016-findings)**
- 13 participants (mainly academic)
- evaluation: % of the aligned pages in the test set that are found (recall)
- results:

	Recall (%)
baseline (url matching)	59.8%
best system	95.0%
average (all non-bogus systems)	85.3%

The best system (gomes-pereira-lopes-2016-first)



- use a phrase table
  - ↪ by-product of the training of a phrase-based MT system
  - ↪ translation of *n*-grams
- intuition: parallel documents share more bilingual phrase pairs than non-parallel documents
- bootstrap / chicken-egg problem: only possible for resource-rich languages

(more) Modern Methods

Schwenk et al. (2021): Large-Scale Parallel Data Mining

Goal: Build massive multilingual parallel corpora directly from the web (Common Crawl).

Main Contributions:

- Introduction of two large-scale resources:
  - **CCAligned**: document-level alignment using web structure and metadata
  - **CCMatrix**: sentence-level alignment using multilingual embeddings
- Unified framework based on **LASER** sentence representations
- Coverage: 100+ languages and billions of aligned sentence pairs

Pipeline Overview:

1. Extract multilingual documents from Common Crawl
2. Apply language identification and text cleaning
3. Compute sentence embeddings using LASER
4. Match nearest neighbours in embedding space for alignment

Schwenk et al. (2021): Methodology and Impact

Alignment Methodology:

- Measure cosine similarity between LASER embeddings of sentences
- Filter out low-quality pairs using margin-based scoring
- Optional bilingual rescoring for increased precision

Advantages:

- Language-agnostic: same model for all languages
- Scalable to billions of sentence pairs
- High precision despite noisy web data

Impact:

- Provided high-quality training data for multilingual NMT
- Used in projects such as NLLB and OPUS
- Established LASER-based alignment as a **standard method** for bitext mining

Reference: Schwenk et al. (2021), *CCMatrix: Mining Billions of High-Quality Parallel Sentences*

LASER: Language-Agnostic SEntence Representations

(artetxe-schwenk-2019-massively)

What is it?

- Multilingual sentence encoder based on a `BiLSTM` architecture
- Trained on parallel data to produce aligned embeddings across languages
- Supports 93 languages 28 scripts, 30 language families with a shared vector space

↪ semantically similar sentences share nearby representations in a joint multilingual space

How can we use it?

- Mining parallel sentences (e.g., `CCMatrix`, `CCAligned`)
- Cross-lingual retrieval and classification
- Document alignment and multilingual clustering

FAISS: Facebook AI Similarity Search

What is FAISS?

- An open-source library developed by Facebook AI Research (FAIR)
- Designed for **efficient similarity search** and **clustering** of dense vectors
- Commonly used for large-scale tasks such as:
  - Nearest neighbour retrieval
  - Embedding-based search and alignment
  - Clustering in multilingual or multimodal spaces

Core idea:

- Represent items (sentences, images, etc.) as high-dimensional vectors
- Find the most similar vectors according to a distance metric (e.g., cosine or L2)
- Perform searches efficiently even with millions or billions of vectors

FAISS: Technical Principles and Strengths

Key Technical Features:

- **Index structures:**
  - Flat (exact) index: brute-force similarity search
  - IVF (Inverted File Index): partitions vectors into clusters for faster retrieval
  - HNSW and PQ (Product Quantisation): memory-efficient approximate search
- **GPU acceleration:** supports CUDA for large-scale parallel processing
- **Scalability:** handles billions of vectors efficiently

Advantages:

- High performance for both exact and approximate nearest neighbour search
- Flexible integration with deep learning frameworks (PyTorch, NumPy)
- Open-source and optimised for large-scale embedding spaces

Use in bitext mining:

- Matches sentence embeddings across languages (e.g., LASER, LaBSE)
- Core component of pipelines such as CCMatrix and CCAligned

Vector Size and Efficiency in FAISS-based Systems

Dimensionality of Representations:

- Typically between **256 and 1024 dimensions**
- Example models:
  - LASER: 1024-dim BiLSTM embeddings
  - LaBSE / mUSE: 768-dim Transformer embeddings

Impact of Dimensionality:

- Higher dimensions ⇒ richer semantic representation
- But also higher **memory footprint** and **search cost**
- FAISS mitigates this with:
  - **Quantisation** (e.g., Product Quantisation, OPQ)
  - **Compressed indices** for billion-scale data

Typical trade-off: precision vs. efficiency in large-scale similarity search.

Some tools & corpora for bitexts mining

Bitextor

Open-source pipeline to harvest parallel data from multilingual websites and (optionally) align segments.

Typical inputs/steps

Crawl/download HTML → language ID → document pairing (URL/matrix + content features) → sentence segmentation → segment alignment → post-processing.

▷ `https://github.com/bitextor/bitextor`

▷ **Licence:** GPL-3.0

Bicleaner (and Bicleaner AI)

A supervised (and neural, in `Bicleaner AI`) classifier to score and filter sentence pairs by translation likelihood.

Features

Length ratios; character/number patterns; LM-based fluency; lexical/probabilistic dictionaries; optional "hard rules".

Use in pipeline

Takes raw bitext (e.g., from `Bitextor` or `LASER`/`LaBSE` mining) and outputs quality-filtered parallel data at a chosen threshold.

▷ `https://github.com/bitextor/bicleaner`

▷ **Licence:** GPL-3.0

ParaCrawl Pipeline

A large-scale, end-to-end workflow for acquiring, aligning, and cleaning web parallel data for many European language pairs.

Typical pipeline: Crawling/Common Crawl → Bitextor (document pairing) → Bifixer (normalise/deduplicate) → Bicleaner (quality filtering) → packaging.

▷ `https://paracrawl.eu`

▷ **Corpus licence:** CC0 (public domain); software components include GPL-3.0 projects (e.g., Bitextor/Bicleaner).

CCAligned and CCMatrix Corpora

(schwenk-etal-2021-ccmatrix)

Large-scale multilingual parallel datasets mined automatically from the *Common Crawl*. Sentences are aligned using multilingual sentence embeddings (`LASER` / `LaBSE`) and similarity search.

↪ **Languages:** 100+

↪ **Size:** over 4.5 billion parallel sentence pairs

↪ **Domain:** general web (cross-domain)

↪ **Licence:** CC-BY 4.0

↪ **Access:** `https://opus.nlpl.eu/CCMatrix.php`

2 construction strategies:

- `CCAligned`: document matching, then sentence filtering (with LASER)

↪ more constrained / higher quality

- `CCMatrix`: direct sentence matching

↪ larger corpus

ParaCrawl Corpus

(banon-etal-2020-paracrawl)

A European initiative to collect, align, and publicly release large web-scale parallel corpora. Built using the Bitextor–Bifixer–Bicleaner pipeline to ensure document alignment and quality filtering.

↪ **Languages:** 40+ European and associated languages

↪ **Size:** 100 million – 1 billion sentence pairs (depending on pair)

↪ **Domain:** general web content

↪ **Licence:** CC0 (public domain)

↪ **Access:** `https://paracrawl.eu` / `https://opus.nlpl.eu/ParaCrawl.php`