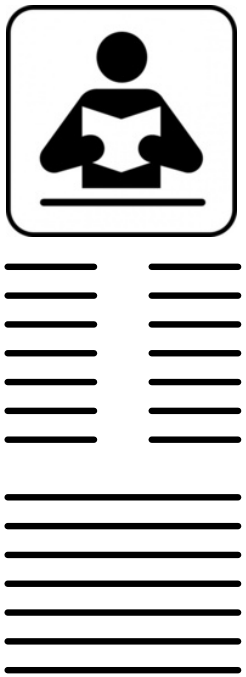
Corpus Acquisition from the Internet

Philipp Koehn
partially based on slides from Christian Buck

8 November 2022

Big Data

For many language pairs, lots of text available.



Text you read in your lifetime — 300 million words

Translated text available — billions of words

English text available — trillions of words

Mining the Web

- Largest source for text: the World Wide Web

Common Crawl

  - publicly available crawl of the web
  - hosted by Amazon Web Services, but can be downloaded
  - regularly updated (semi-annual)
  - 2-4 billion web pages per crawl

- Currently filling up hard drives in our lab

Monolingual Data

- Starting point: 35TB of text
- Processing pipeline [Buck et al., 2014]
  - language detection
  - deduplication
  - normalization of Unicode characters
  - sentence splitting
- Obtained corpora

Language	Lines (B)	Tokens (B)	Bytes	BLEU (WMT)
English	59.13	975.63	5.14 TB	-
German	3.87	51.93	317.46 GB	+0.5
Spanish	3.50	62.21	337.16 GB	-
French	3.04	49.31	273.96 GB	+0.6
Russian	1.79	21.41	220.62 GB	+1.2
Czech	0.47	5.79	34.67 GB	+0.6

Parallel Data

- Basic processing pipeline [Smith et al., 2013]
  - find parallel web pages (based on URL only)
  - align document by HTML structure
  - sentence splitting and tokenization
  - sentence alignment
  - filtering (remove boilerplate)

- Obtained corpora

	French	German	Spanish	Russian	Japanese	Chinese
Segments	10.2M	7.50M	5.67M	3.58M	1.70M	1.42M
Foreign Tokens	128M	79.9M	71.5M	34.7M	9.91M	8.14M
English Tokens	118M	87.5M	67.6M	36.7M	19.1M	14.8M
	Bengali	Farsi	Telugu	Somali	Kannada	Pashto
Segments	59.9K	44.2K	50.6K	52.6K	34.5K	28.0K
Foreign Tokens	573K	477K	336K	318K	305K	208K
English Tokens	537K	459K	358K	325K	297K	218K

- Much more work needed!

Data Cleaning and Subsampling

- Not all data useful – some may be harmful
- Removing data based on
  - domain relevance
  - alignment quality
  - redundancy
  - bad language (orthography, non-words)
  - machine translated or poorly translated
- Removing bad data always reduces training time
- Removing bad data sometimes helps quality
- Clean data approach (only using high quality data) helps in limited domains

corpus crawling

Finding Monolingual Text

- Simple Idea
  1. Download many websites
  2. Extract text from HTML
  3. Guess language of text
  4. Add to corpus
  5. Profit
- Turns out all these steps are quite involved

bilingual corpus crawling

Mining Bilingual Text

- Bilingual text = same text in different languages
- Usually: one side translation of the other
- Full page or interface/content only
- Potentially translation on same page
  e.g., Twitter, Facebook posts

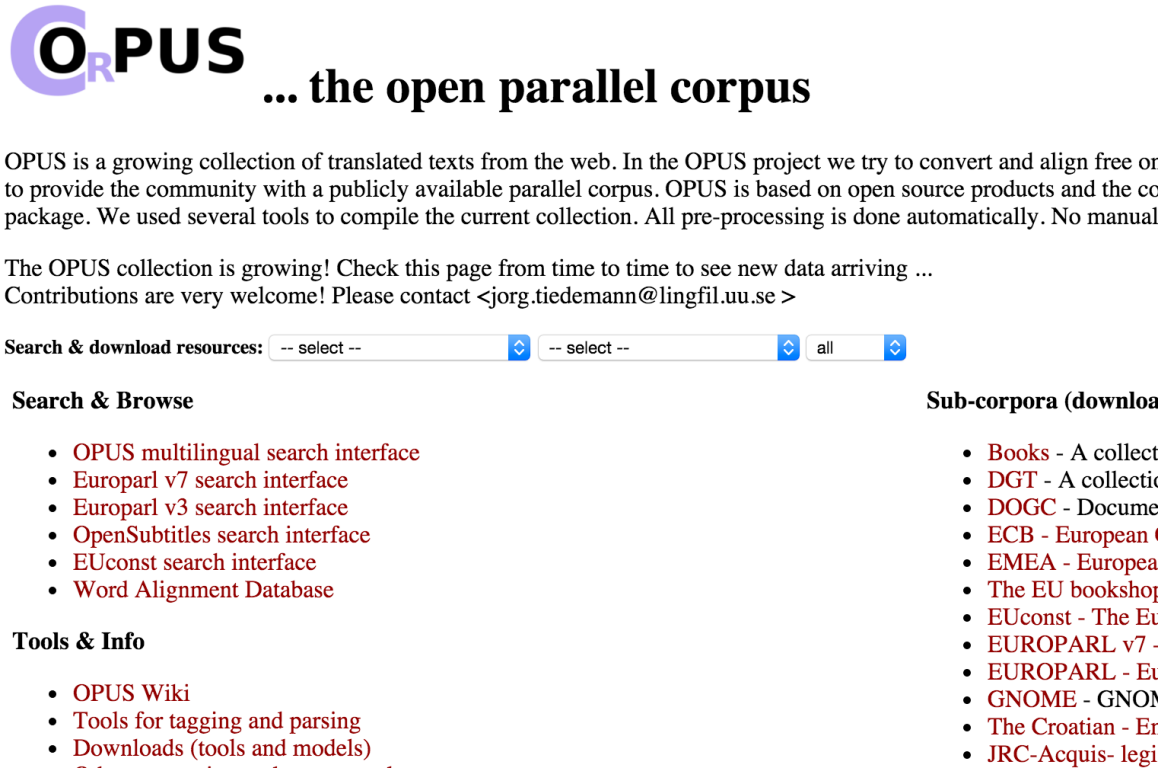
Pipeline

1. Identify web sites worth crawling
2. Crawl web site
3. Language detection — as before
4. Extract text from HTML — as before
5. Align documents
6. Align sentences
7. Clean corpus

identify web sites

Targeted Crawling

- A few web sites with a lot of parallel text, e.g.,
  - European Union, e.g., proceedings of the European Parliament
  - Canadian Hansards
  - United Nations
  - Project Syndicate
  - TED Talks
  - Movie / TV show subtitles
  - Global Voices
- Hand-written tools
  - crawling
  - text extraction
  - document alignment
- Few days effort per site



Broad Crawling

- Identify many web sites to crawl
  - has the phrase This page in English or variants
  - has link to language flag
  - known to have content in multiple languages (from CommonCrawl)
- Follow links
  - up to n links deep into site
  - up to n links in total
  - only follow links to web pages, not images, etc.
- Avoid crawling sites too deeply that do not have parallel text?
  (requires quick feedback from downstream processing)

document alignment

Document Alignment

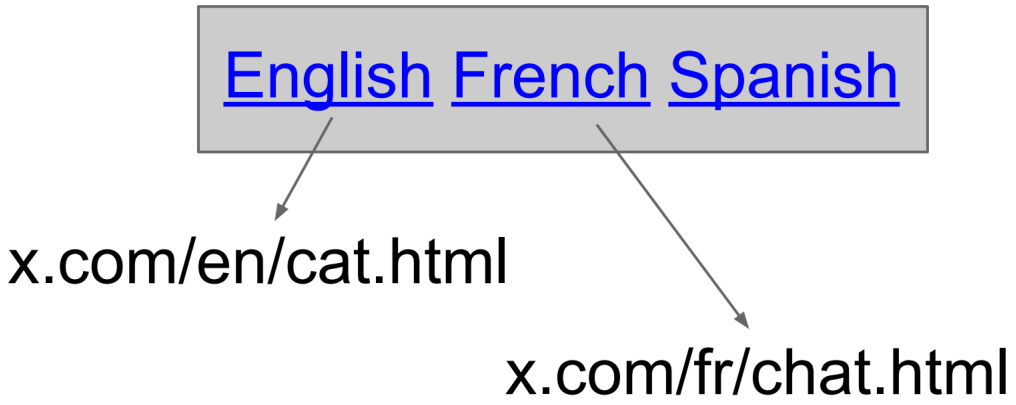
- Early Work: STRAND (Resnik 1998, 1999)
  (Structural Translation Recognition, Acquiring Natural Data)

- Pipeline

  1. candidate generation
  2. candidate ranking
  3. filtering
  4. optional: sentence alignment
  5. evaluation

Link Structure

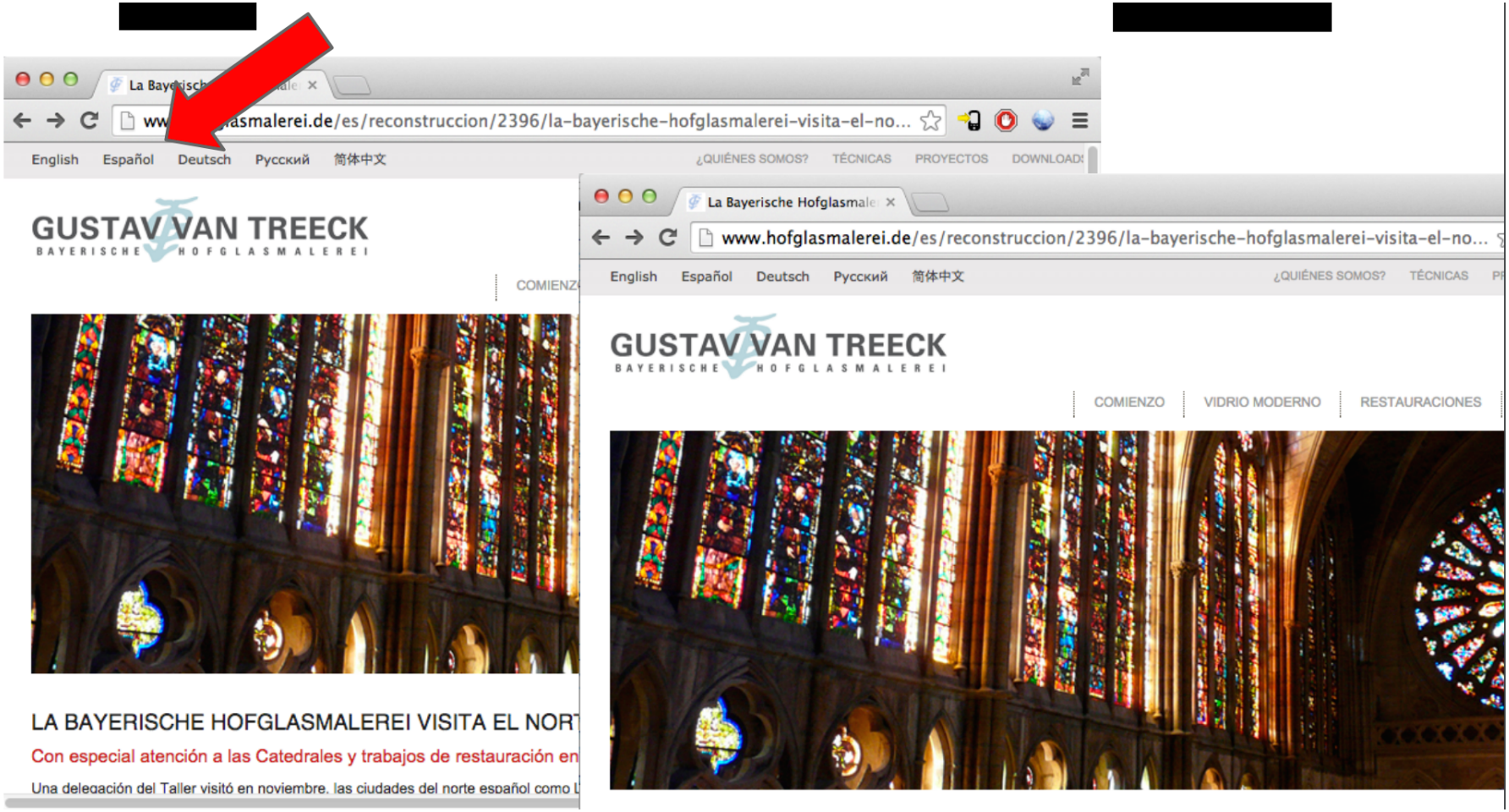
- Parent page: a page that links to different language versions



Parent Page Example

Sibling Page

- A page that links to its translation in another language



URL Matching

- Often URLs differ only slightly, often indicating language

xyz.com/en/	xyz.com/fr/
xyz.com/bla.htm	xyz.com/bla.htm?lang=FR
xyz.com/the_cat	xyz.fr/le_chat

Finding URL Patterns

- URLs with pattern =en

Count	Pattern
545875	lang=en
140420	lng=en
126434	LANG=en
110639	hl=en
99065	language=en
81471	tlng=en
56968	l=en
47504	locale=en
33656	langue=en
33503	lang=eng
19421	uil=English
15170	ln=en
14242	Language=EN
13948	lang=EN
12108	language=english
11997	lang=engcro
11646	store=en

Finding URL Patterns

- URLs with pattern `lang.*=.*`

Count	Pattern
13948	`lang=EN`
13456	`language=ca`
13098	`switchlang=1`
12960	`language=zh`
12890	`lang=Spanish`
12471	`lang=th`
12266	`langBox=US`
12108	`language=english`
12003	`lang=cz`
11997	`lang=engcro`
11635	`lang=sl`
11578	`lang=d`
11474	`lang=lv`
11376	`lang=NL`
11349	`lang=croeng`
11244	`lang=English`

Document Length

- Extract texts and compare lengths (Smith 2001)

$$\text{Length}(E) \approx C * \text{Length}(F)$$

learned,
language-specific parameter

- Document or sentence level

Document Object Model

- Translated web pages often retain similar structure

```
<html>                          <html>
  <body>                          <body>
  <h1>
  Where is the cat?
  </h1>
  The cat sat on                    El gato se sentó
  the mat.                          en la alfombra.
  </body>                         </body>
</html>                         </html>
```

- This includes links to the same images, etc.

Linearized Structure


```
[Start:html]
[Start:body]
[Start:h1]
[Chunk:17bytes]
[End:h1]
[Chunk:23bytes]
[End:body]
[End:html]
```

```
[Start:html]
[Start:body]
[Chunk:32bytes]
[End:body]
[End:html]
```

Levenshtein Alignment

[Start:html]	Keep
[Start:body]	Keep
[Start:h1]	Delete
[Chunk:17bytes]	Delete
[End:h1]	Delete
[Chunk:23bytes]	23 Bytes -> 32 Bytes
[End:body]	Keep
[End:html]	Keep

Content Similarity

- Simple things
  - same numbers or names in documents
  - often quite effective
- Use of lexicon
  - treat documents as bag of words
  - consider how many words in `EN` document have translations in `FR` document
- A bit more complex
  - semantic representations of documents content
  - bag of word vectors
  - neural network embeddings
- Major challenge: do this fast for $n \times m$ document pairs

sentence alignment

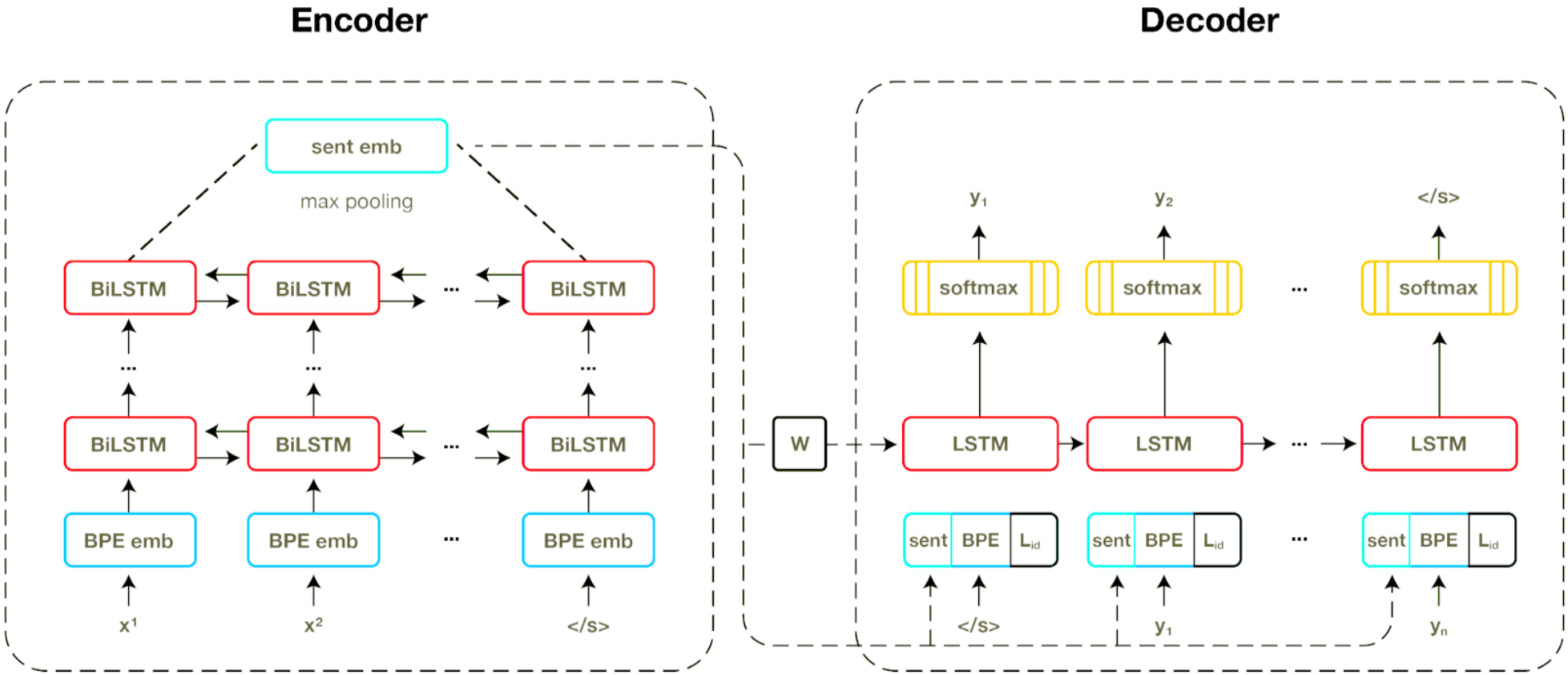
Sentence Alignment

- Much early work in 1990s, e.g., Gale and Church (1991)
  - find sequence of 1-1, 1-2, 0-1, etc., sentence alignment groups
  - good element in sequence = similar number of words
  - dynamic programming search for best sequence
- Featurized alignments
  - with dictionary (Hunalign)
  - with induced dictionary (Gargantua)
  - consider tags such as <P>
- Sensitive to noise — often large parts of page not translated

Sentence Pair Similarity

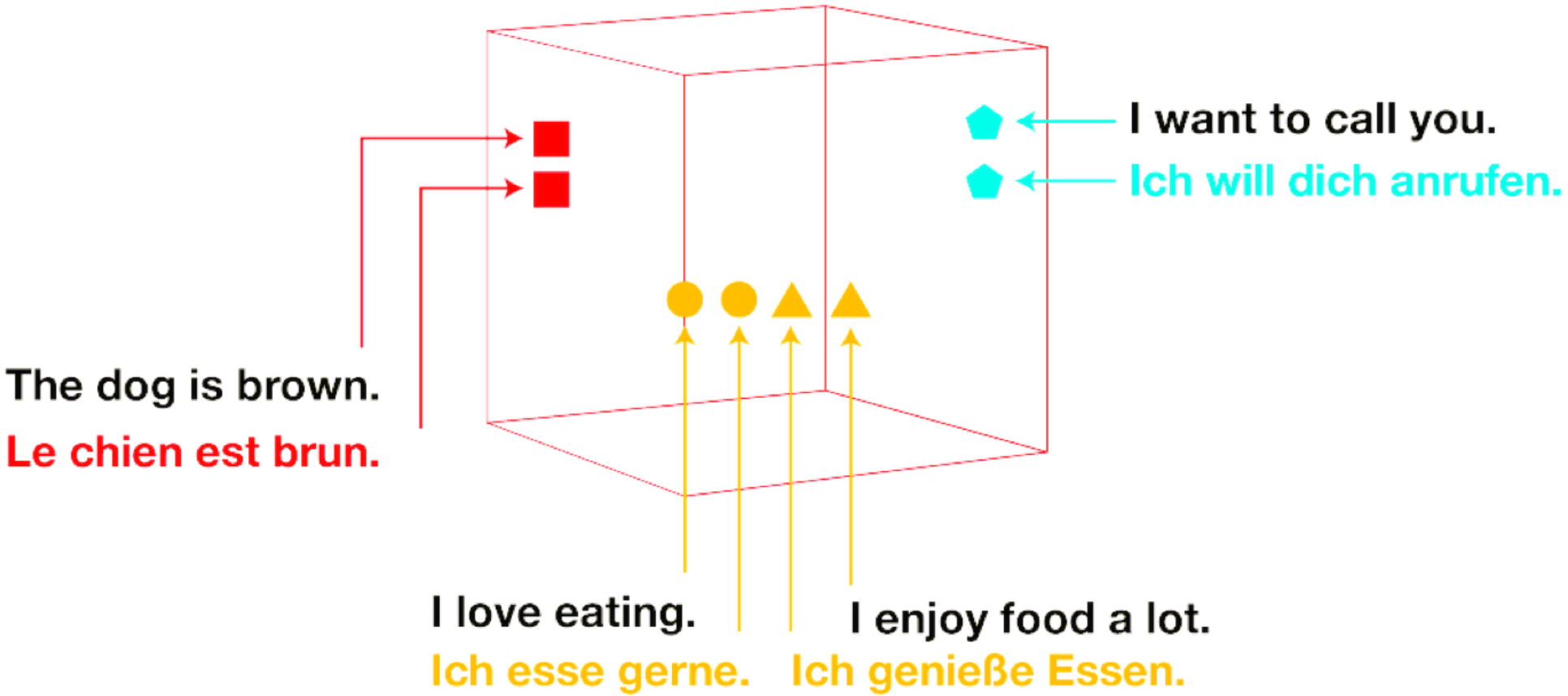
- Core Problem: both sentences must have same meaning
- Translate foreign sentence into English
  measure similarity with metrics like BLEU
- Words in one sentence have translation in the other
- Cross-lingual sentence embeddings

Sentence Embeddings

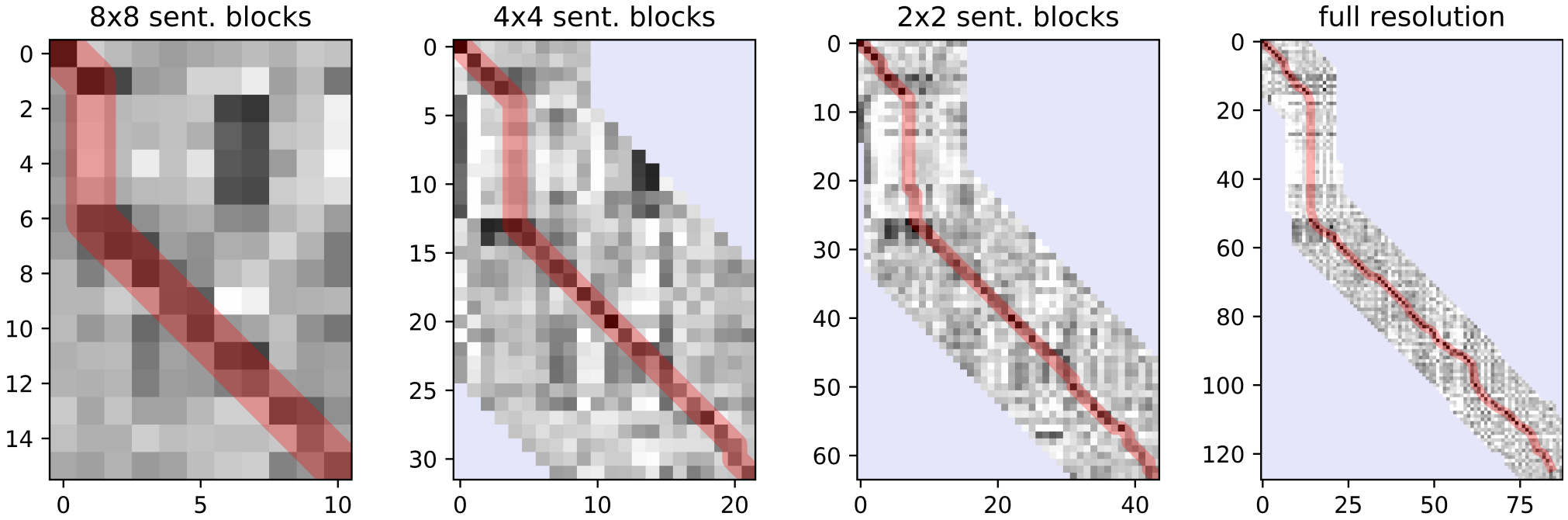


- LASER: Neural machine translation model with bottleneck feature

Sentence Embeddings



Vecalign



- Uses LASER sentence embeddings
- Linear time coarse-to-fine algorithm

sentence pair filtering

Filtering Bad Data

- Mismatched sentence pairs from errors in pipeline
- Non-literal translation
  e.g. news stories are notoriously non-literal
- Bad translations
- Machine translation
  - much of the parallel text on the Internet generated by Google Translate
  - detection hard — looks like very clean parallel data
  - maybe too clean (little reordering, very literal)
  - watermarking machine translation (Venugopal et al., 2011)
- How clean should it be?
  - trade-off between precision and recall unclear

Methods

- Dual cross-entropy
  - view sentence pair as input/output
  - score with neural machine translation model in both directions
  - scores should be low and similar
- LASER embeddings
- Feature-based approaches
  - matching numbers, named entities
  - language model probabilities
  - lexical translation probabilities
- Classifier
  - positive example: sentence pair from clean corpus
  - negative example: corrupted example (misalignment, words changed, ...)