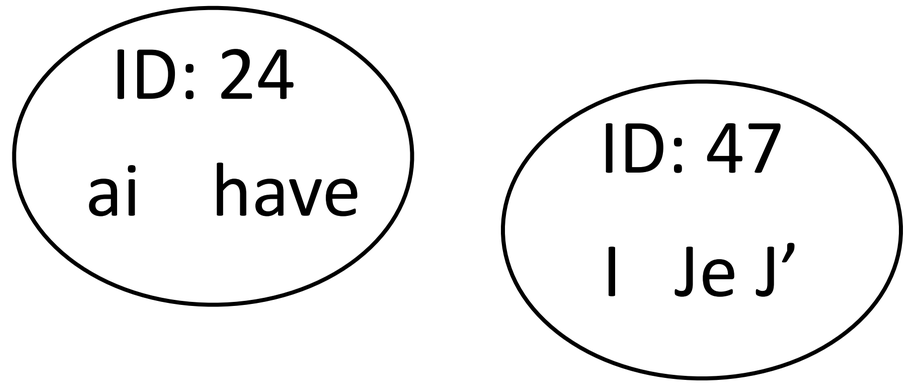
Multilingual Embeddings

- Input: corpora in many languages. Output: embeddings where similar words *in different languages* have similar embeddings

I have an apple
47 24 18 427

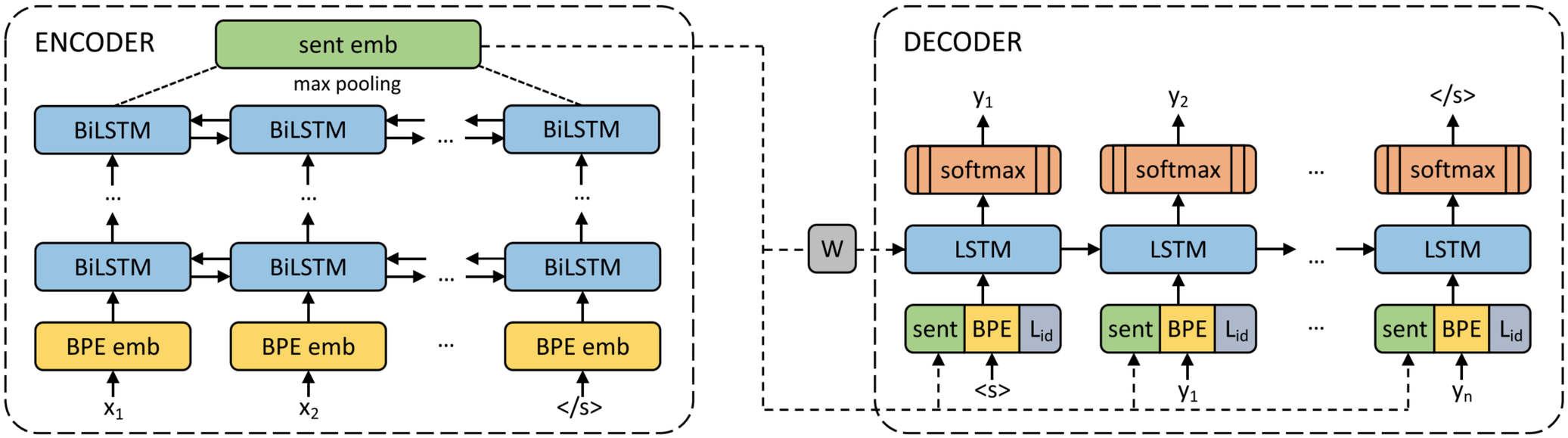
J’ ai des oranges
47 24 89 1981



- multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train “monolingual” embeddings over all these corpora
- Works okay but not all that well

Ammar et al. (2016)

Multilingual Sentence Embeddings



- Form BPE vocabulary over all corpora (50k merges); will include characters from every script

- Take a bunch of bitexts and train an MT model between a bunch of language pairs with shared parameters, use W as sentence embeddings

Artetxe et al. (2019)

Multilingual Sentence Embeddings

		EN	EN → XX												
			fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Zero-Shot Transfer, one NLI system for all languages:																
Conneau et al.	X-BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
(2018b)	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2
BERT uncased*	Transformer	81.4	–	74.3	70.5	–	–	–	–	62.1	–	–	63.8	–	–	58.3
Proposed method	BiLSTM	73.9	**71.9**	72.9	72.6	**72.8**	**74.2**	**72.1**	**69.7**	**71.4**	**72.0**	**69.2**	71.4	**65.5**	**62.2**	61.0

- Train a system for NLI (entailment/neutral/contradiction of a sentence pair) on English and evaluate on other languages

Artetxe et al. (2019)

Multilingual BERT

- Take top 104 Wikipedias, train BERT on all of them simultaneously
- What does this look like?

Beethoven may have proposed unsuccessfully to Therese Malfatti, the supposed dedicatee of "Für Elise"; his status as a commoner may again have interfered with those plans.

当人们在马尔法蒂身后发现这部小曲的手稿时，便误认为上面写的是“Für Elise”（即《给爱丽丝》）[51]。

Кита́й (официально — Кита́йская Наро́дная Респу́блика, сокращённо — КНР; кит. трад. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэньминь Гунхэго) — государство в Восточной Аз

Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

- Can transfer BERT directly across languages with some success
- …but this evaluation is on languages that all share an alphabet

Pires et al. (2019)

Multilingual BERT: Results

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- Urdu (Arabic script) => Hindi (Devanagari). Transfers well despite different alphabets!
- Japanese => English: different script and very different syntax

Pires et al. (2019)