Code-Switched Named Entity Recognition with Embedding Attention

Changhan Wang, Kyunghyun Cho and Douwe Kiela

July 19th, 2018

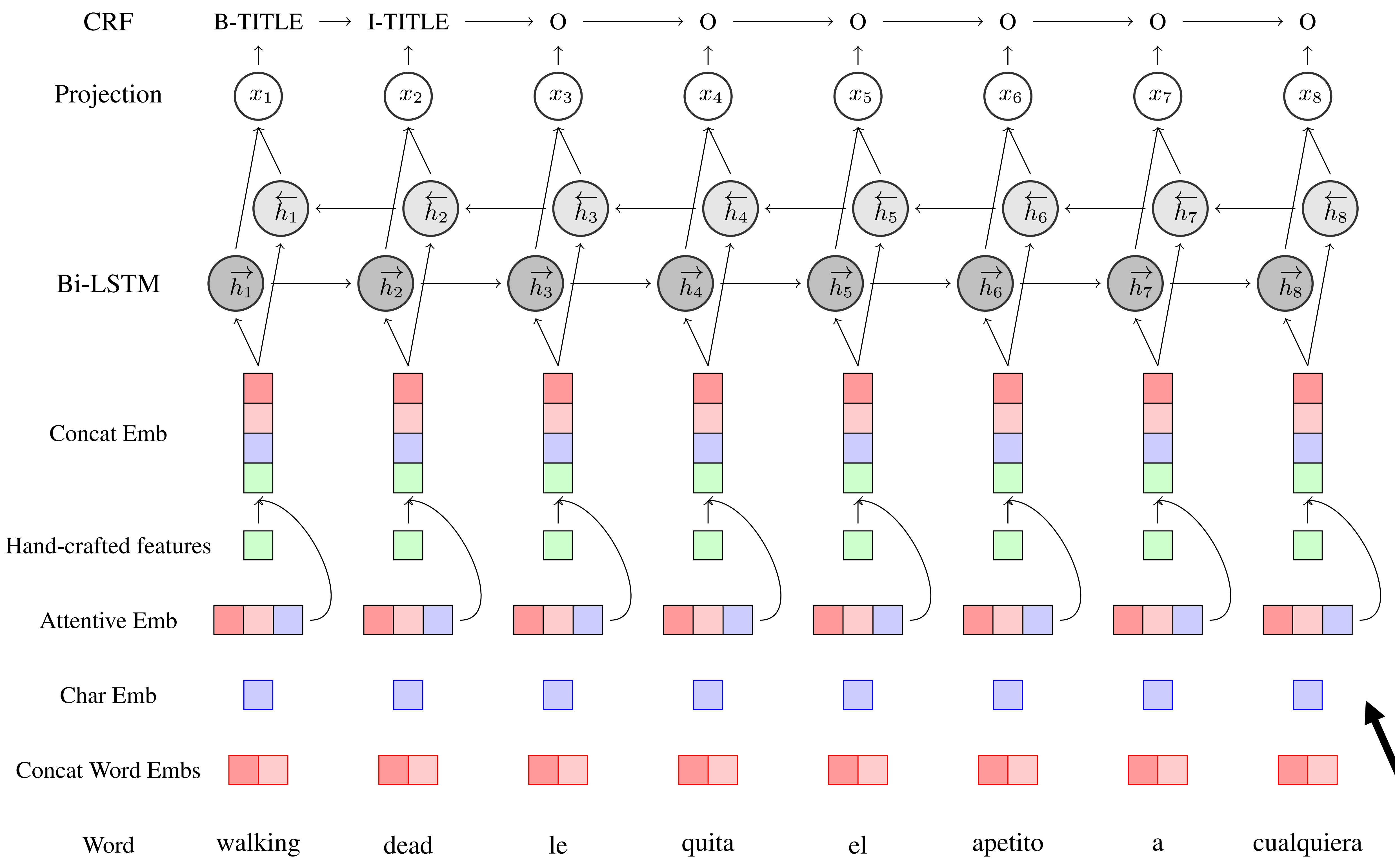


facebook
Artificial Intelligence

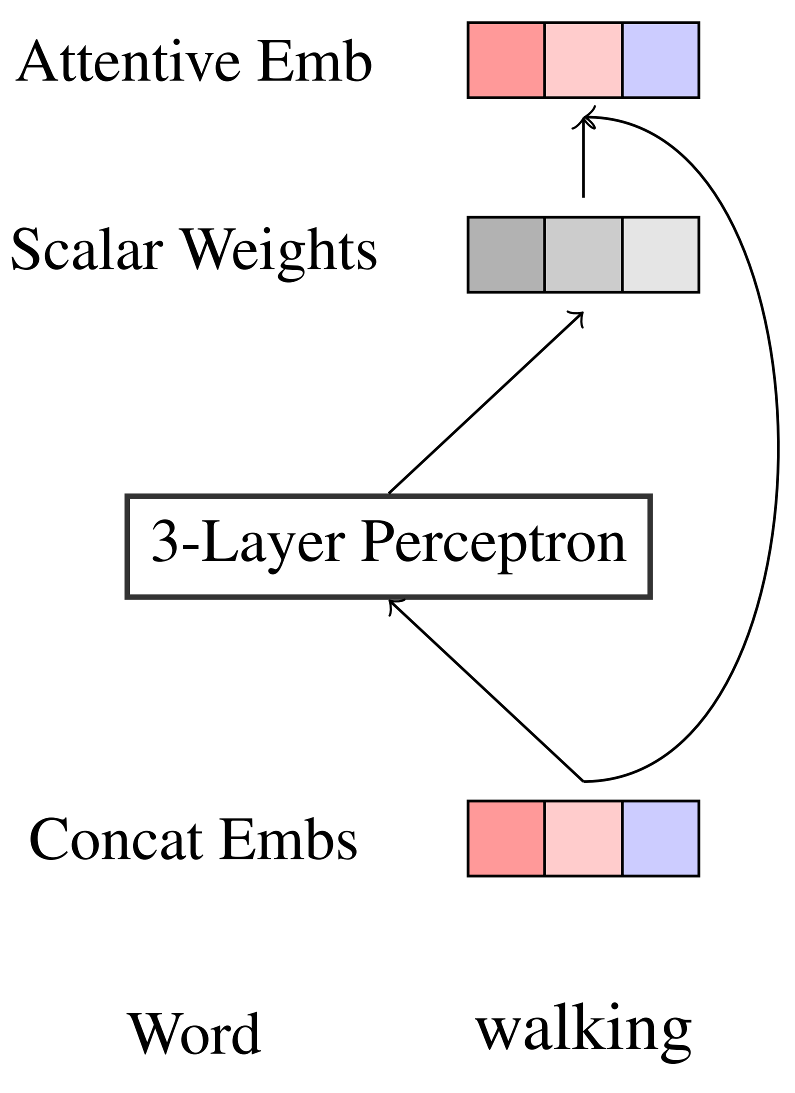
Highlights

- 1st place on MSA-EGY
- 3rd place on ENG-SPA
- Same architecture with slightly different pre-processing
- With minimal hand-crafted features, without gazetteers

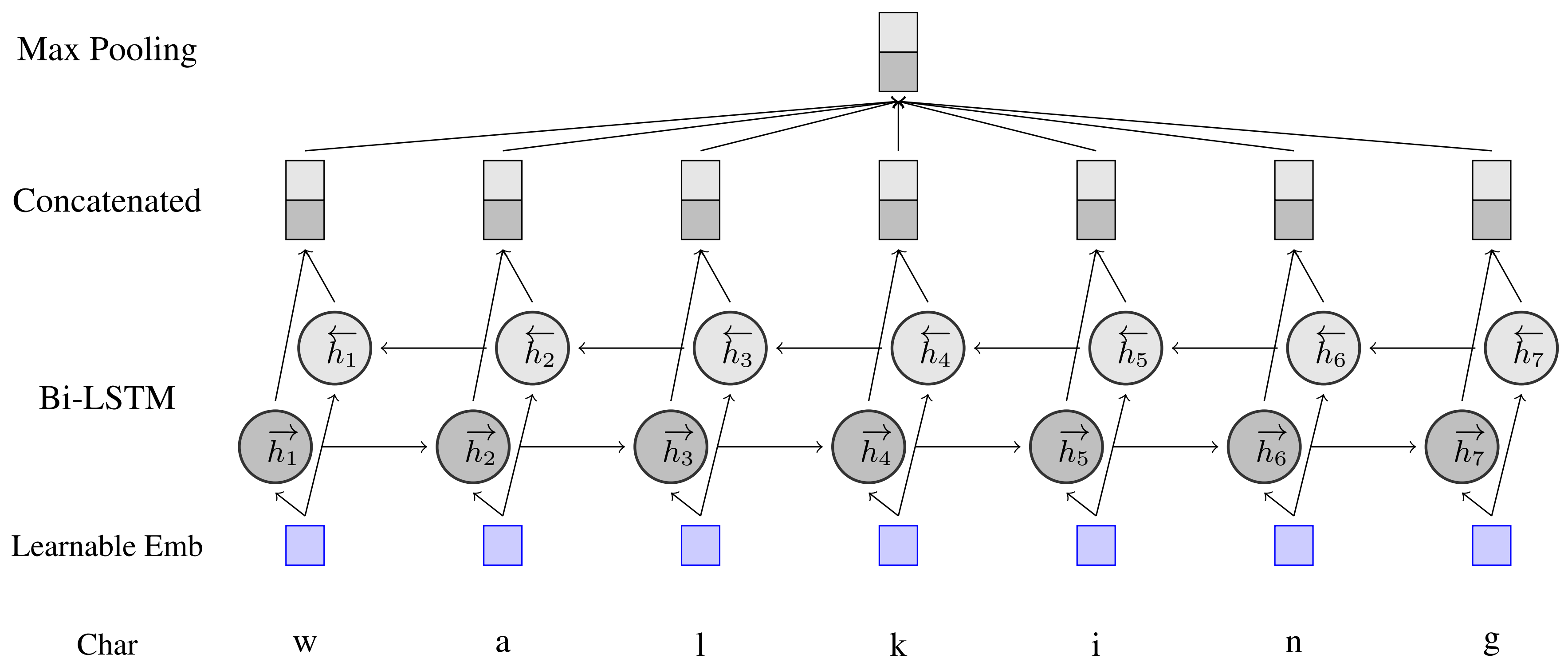
System Overview



Attentive Embedding

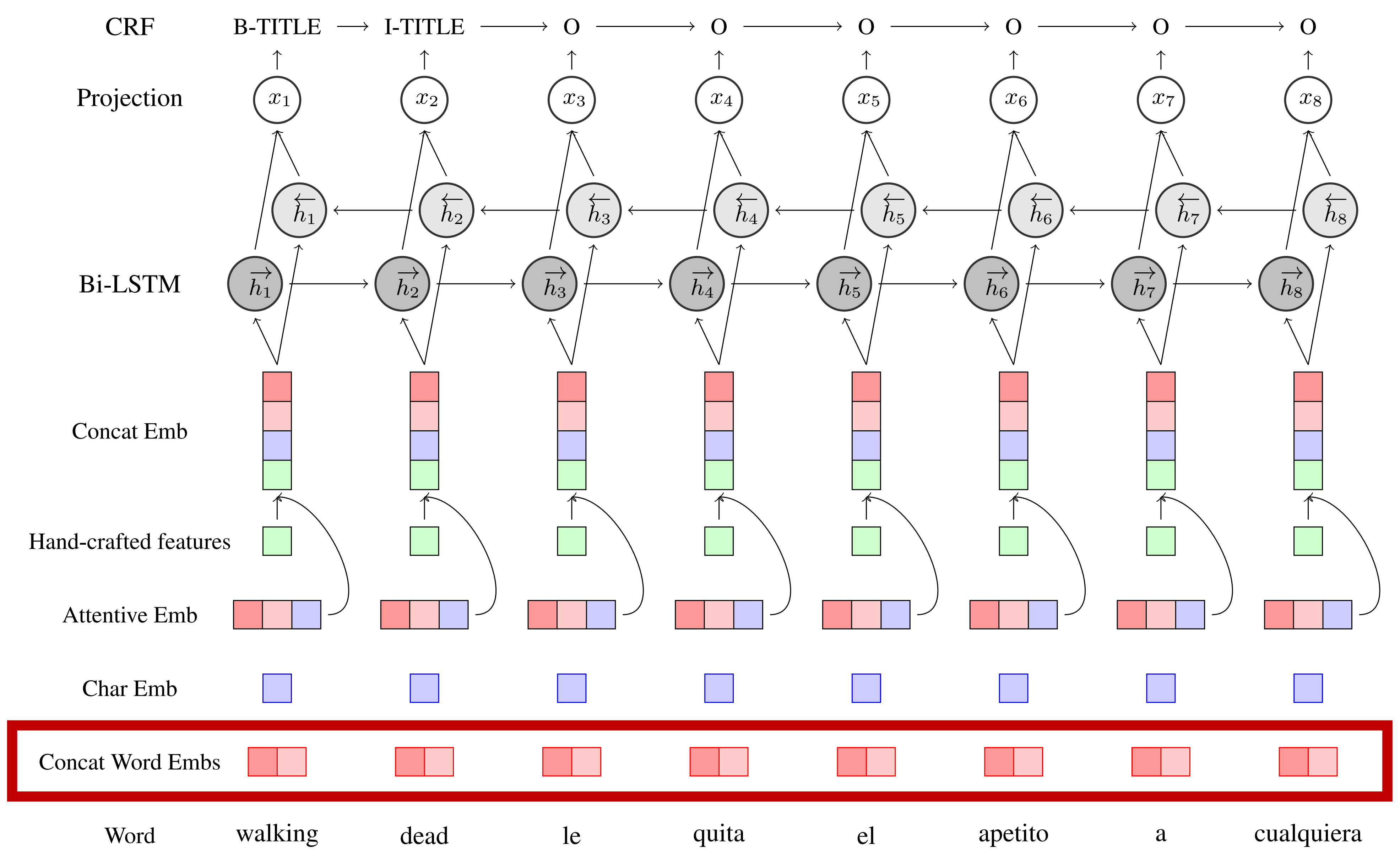


Character Embedding



Pre-trained Word Embedding

- fastText monolingual word vectors trained on Common Crawl + Wikipedia data
- Huge training dataset, better generalization
- Kept fixed during training
- Initialized OOV words with normal dist. with zero mean and 0.1 variance



fastText train set

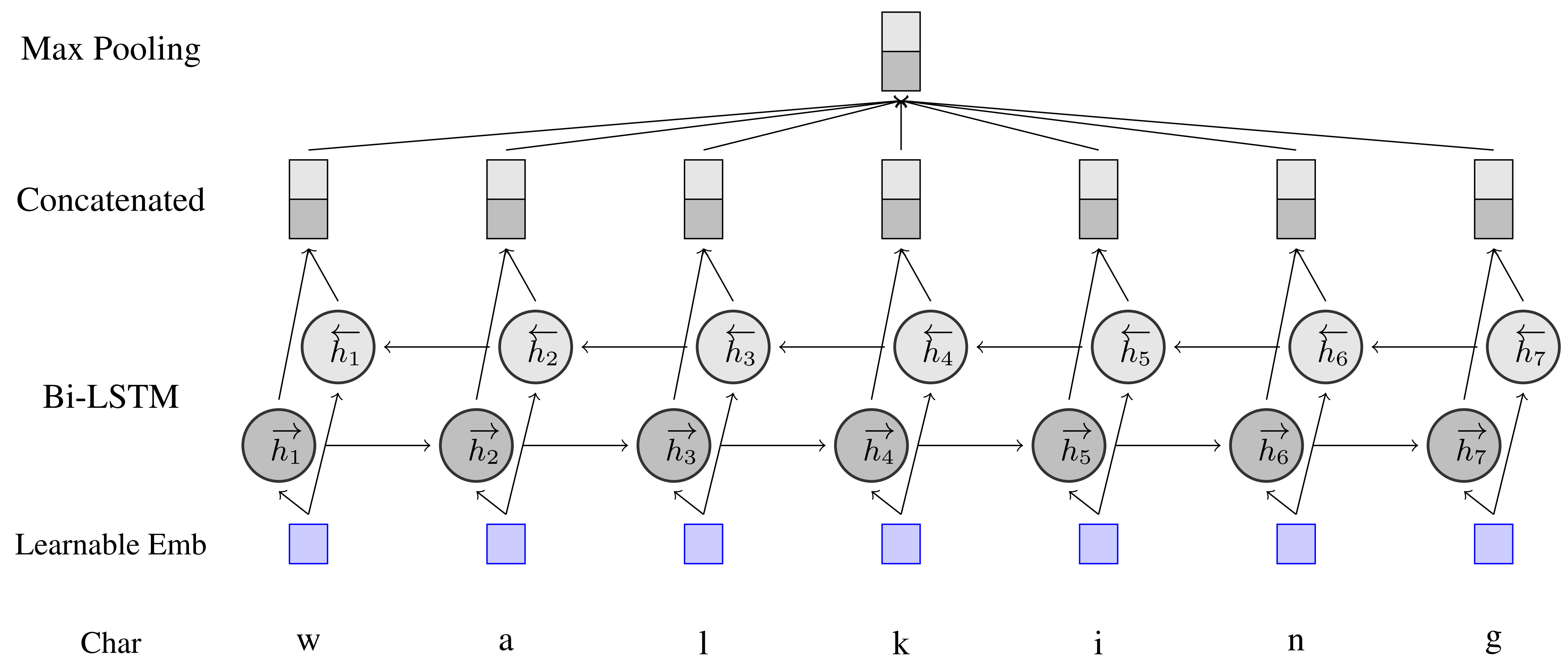
	Tokens	Vocab	Coverage
ENG	600B	200M	64.94%
SPA	72B	200M	82.14%
MSA	25B	200M	88.74%
EGY	129M	361k	64.30%

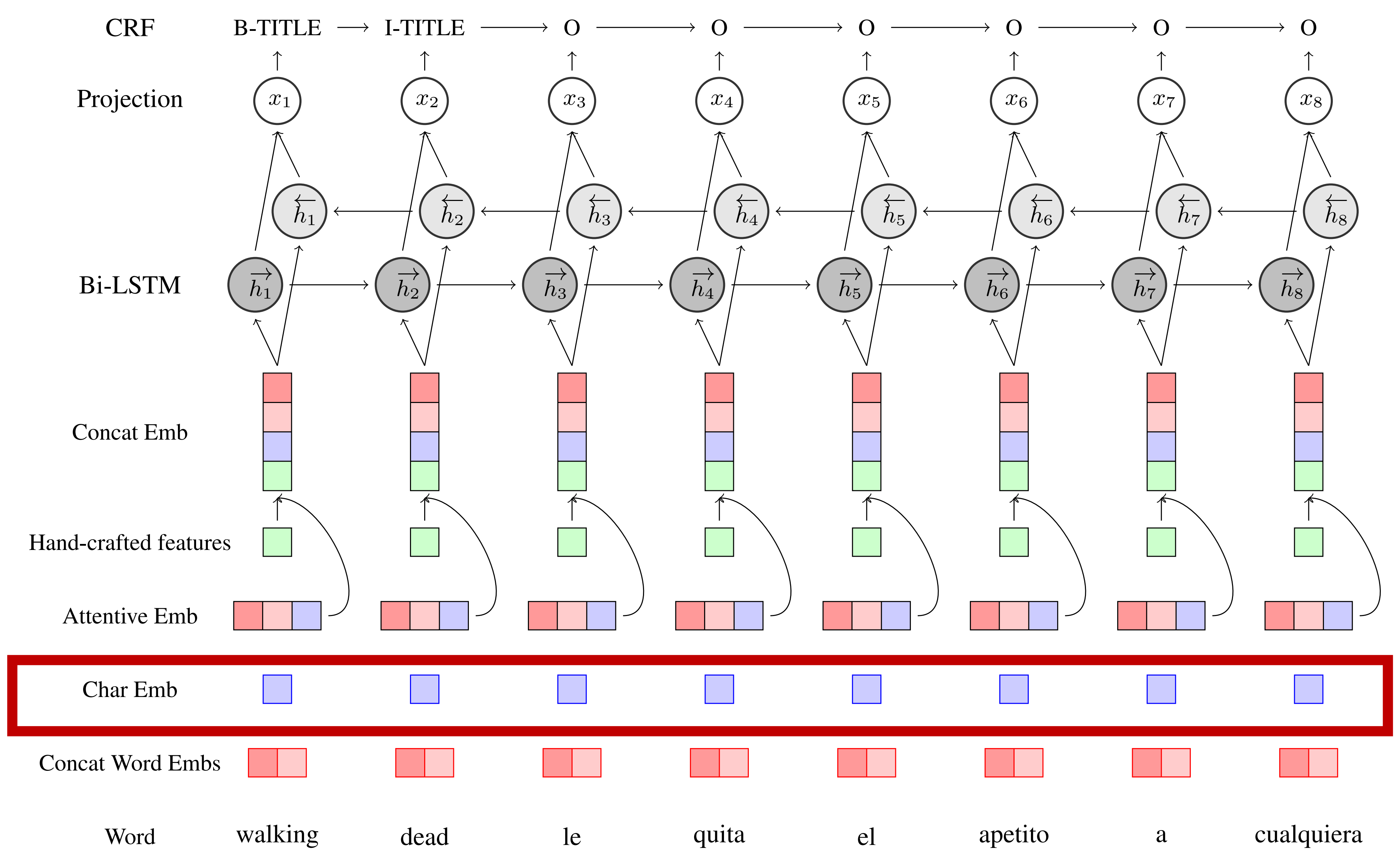
Code switching train set

	Tokens
ENG-SPA	616k
MSA-EGY	204k

Character Embedding

- Imperfect pre-processing and word matching, < 90% pre-trained word vector coverage
- To capture morphological similarities especially for OOV words (e.g. awesome & awesomeee)
- Updated during training, complementary to generic fixed word embedding



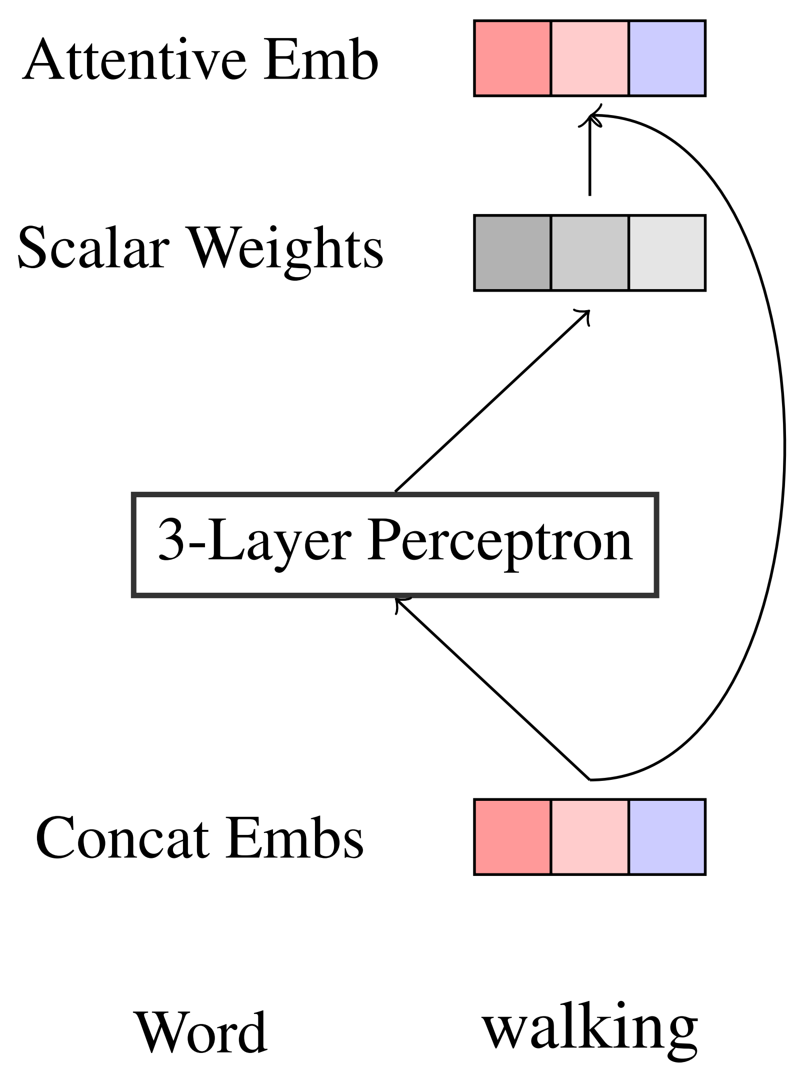


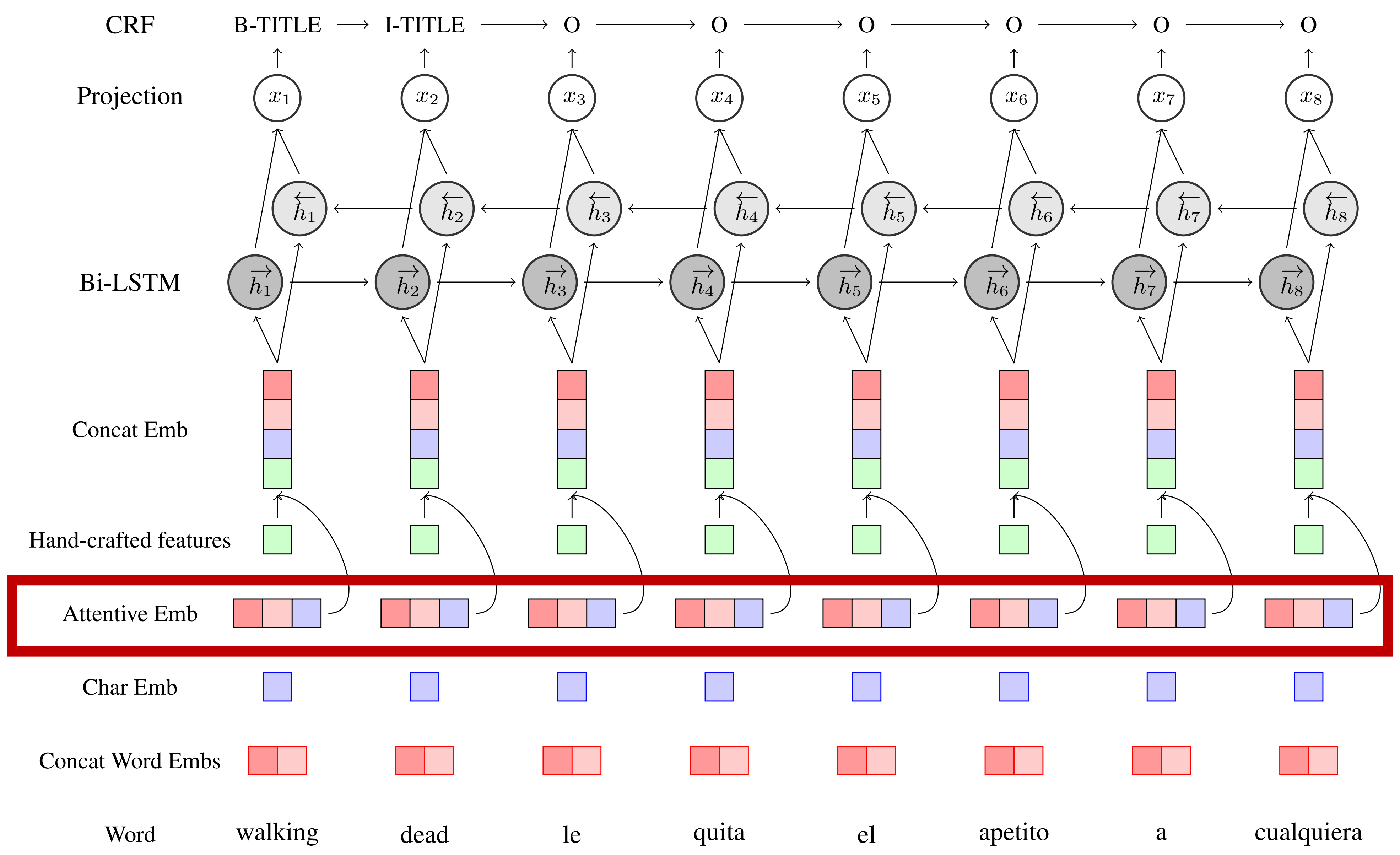
Attentive Embedding

- Context-Attentive Embeddings for Improved Sentence Representations (arXiv:1804.07983)

- Learnt scalar weights for each word/char embedding of each token
- Scaled L2-normalized embeddings with the weights

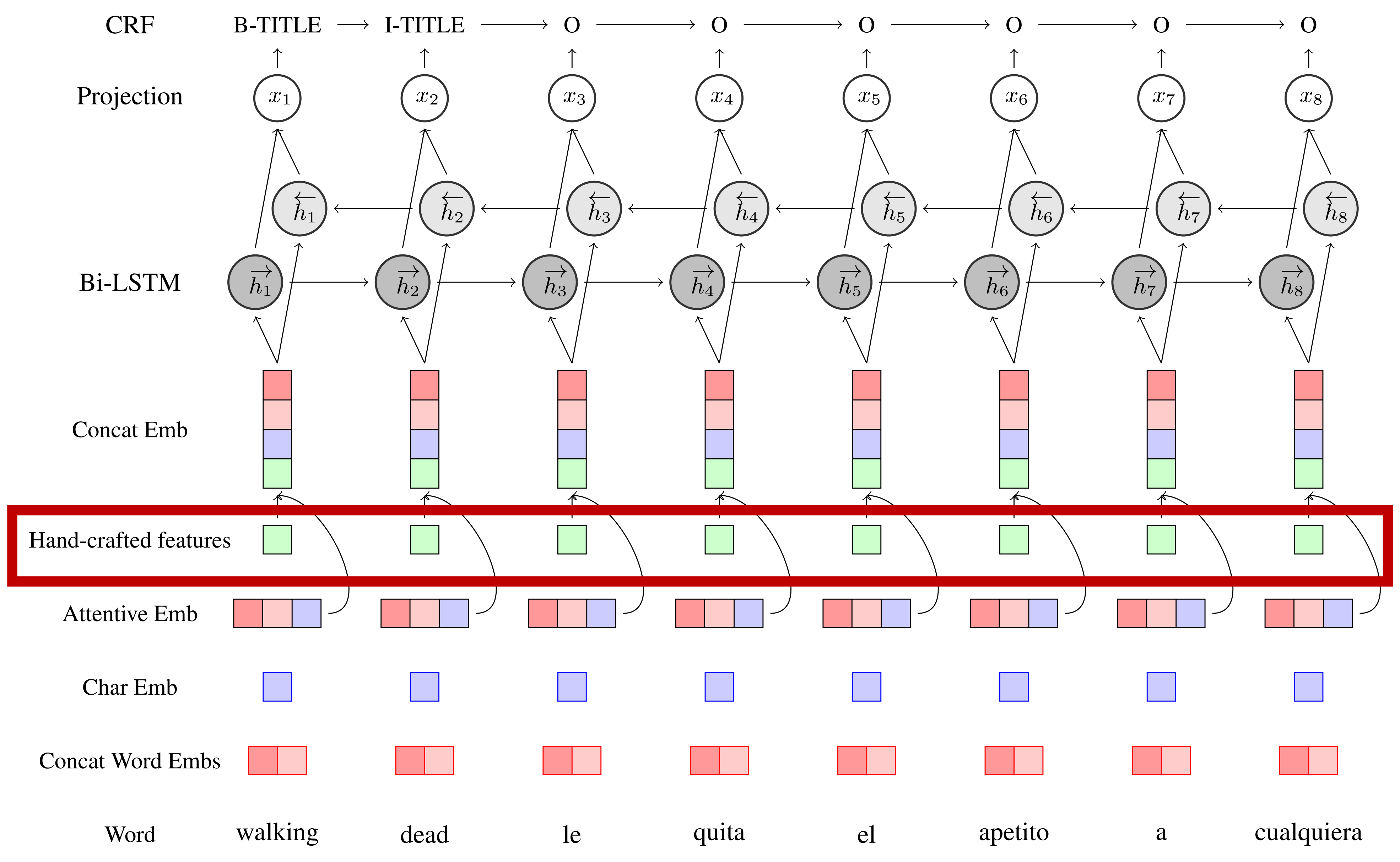
- Higher activations for dominating embeddings for each token (e.g. ENG vectors for ENG words, SPA vectors for SPA words, char embedding for OOV words)





Hand-Crafted Features

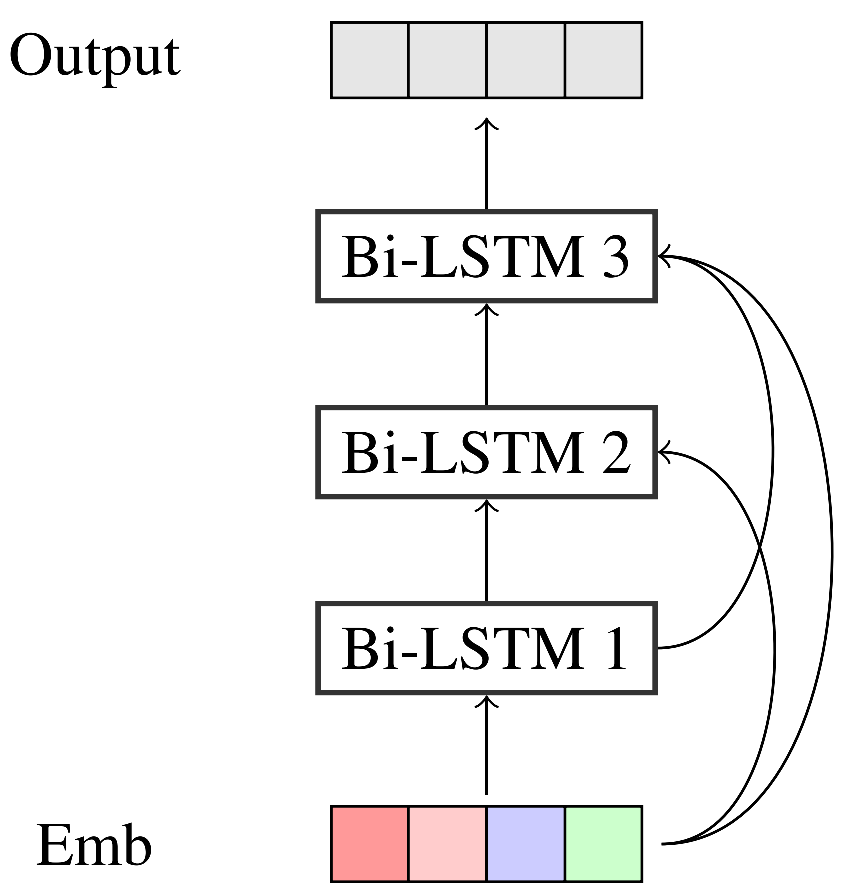
- Capitalization features: all uppercase, initial uppercase, all lowercase
- Trained embedding for the 3 types
- Improved performance on the Person, Location and Organization categories
- Possible to add other hand-crafted features

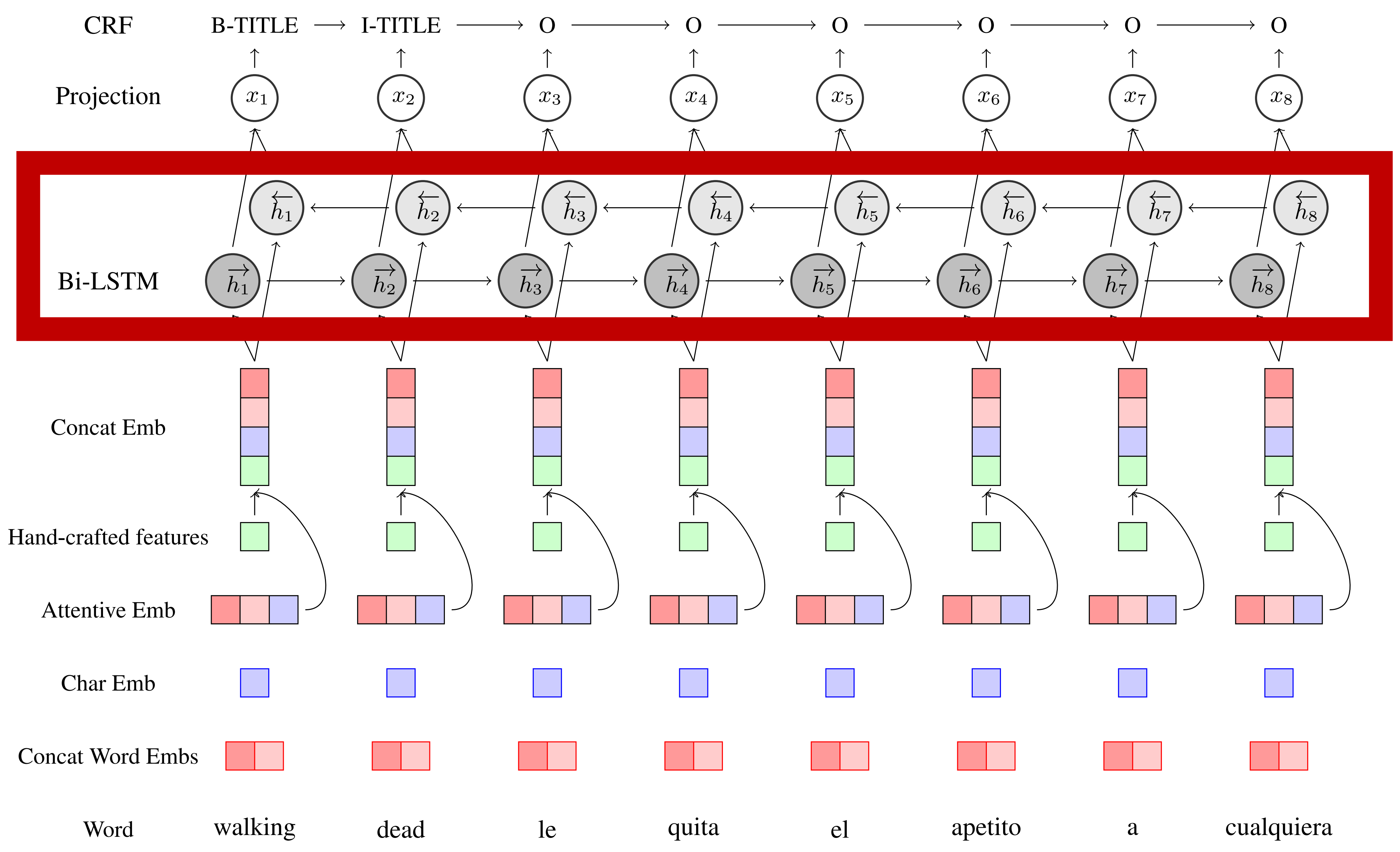


(Shortcut Stacked) Bi-LSTM

- Bi-LSTM to capture sequential dependencies and contextualized features

- Shortcut stacked version to get deep representations (shortcut-stacked sentence encoders for multi-domain inference, Nie and Bansal, 2017)





Conditional Random Field

- Instead of labeling individually (with Softmax for example)
- Taking transitional probabilities/constraints into account as well

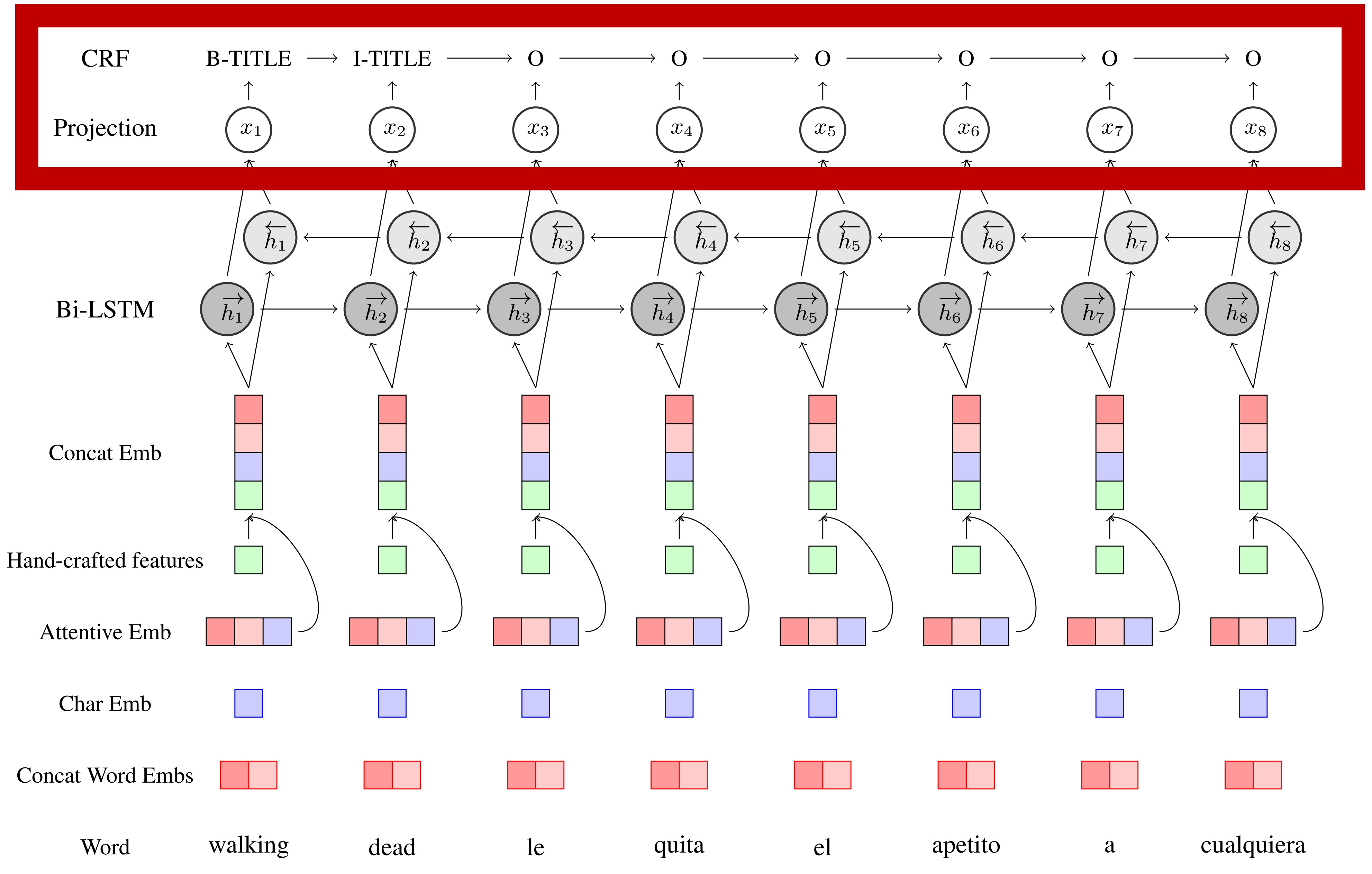
e.g.

- I-* is always after B-*
- B-TITLE to I-LOC is invalid

$$p(\mathbf{s}|\mathbf{x};\mathbf{w}) = \frac{\exp(\mathbf{w}\cdot\Phi(\mathbf{x},\mathbf{s}))}{\sum_{\mathbf{s}'}\exp(\mathbf{w}\cdot\Phi(\mathbf{x},\mathbf{s}'))} = \frac{\exp(\sum_j \mathbf{w}\cdot\phi_j(\mathbf{x},j,s_{j-1},s_j))}{\sum_{\mathbf{s}'}\exp(\sum_j \mathbf{w}\cdot\phi_j(\mathbf{x},j,s'_{j-1},s'_j))} = \frac{\prod_j \exp(\psi_j(\mathbf{w},\mathbf{x},j,s_{j-1},s_j))}{\sum_{\mathbf{s}'}\prod_j \exp(\psi_j(\mathbf{w},\mathbf{x},j,s'_{j-1},s'_j))}.$$

$$\psi_j(\mathbf{w},\mathbf{x},j,p,q) = \mathbf{W}_{[q,:]}^{\intercal}\mathbf{x}_j + \mathbf{B}_{[p,q]}$$

Unary + Transition



Pre-Processing

User generated data is noisy (special tags, typos, misspellings, etc.):

- " @PattyB_14 : Este weekend es largo ! A celebrar mi bday allllllll weeeekend looooooong " #PARY wuutt wutt

Replacement rules

- URLs to <url>
- User tags (starting with "@") to <user>
- Hashtags (starting with \# but not followed by a number) to <hash_tag>
- Punctuation tokens to <punct>
- Integer and real numbers to <num>
- [num]:[num] to <time>
- [num]-[num] to <range>
- Unicode emojis tokens to <emoji>

- MSA-EGY data has very few Unicode/ASCII emojis and tokenization is imperfect (~70% to ~71.6% test F1 after removing heading and trailing punctuations)

Results

- 1st place on MSA-EGY, 3rd place on ENG-SPA

Model	Dev F1	Test F1
Baseline	68.17	60.28
Ours	67.74	62.39

Table 1: Results for ENG-SPA.

(Small dev set)

Model	Dev F1	Test F1
Baseline	79.55	70.08
Ours	81.41	71.62

Table 2: Results for MSA-EGY.

- Good at Person and Location categories
- The Title category is difficult
- The Other category is extremely difficult for ENG-SPA (only 3 examples for MSA-EGY)

	Precision	Recall	Entity F1	# Train
EVENT	56.25	20.00	29.51	232
GROUP	69.77	30.93	42.86	718
LOC	70.75	69.23	69.98	811
ORG	62.50	27.23	37.93	2810
OTHER	14.29	1.71	3.08	324
PER	76.52	68.15	72.09	4701
PROD	63.76	47.53	54.46	1369
TIME	51.58	37.09	43.24	577
TITLE	49.14	25.79	33.83	824
Overall	70.62	55.88	62.39	12366

Table 1: ENG-SPA test performance breakdown.

	Precision	Recall	Entity F1	# Train
EVENT	78.18	61.43	68.80	535
GROUP	69.77	76.92	73.17	1799
LOC	76.19	67.84	71.78	3275
ORG	66.14	67.20	66.67	1504
OTHER	100.00	100.00	100.00	116
PER	77.29	69.53	73.21	5705
PROD	76.47	78.79	77.61	538
TIME	64.29	72.00	67.92	466
TITLE	31.58	60.00	41.38	896
Overall	73.95	69.42	71.62	14834

Table 2: MSA-EGY test performance breakdown.

Results

- 1st place on MSA-EGY, 3rd place on ENG-SPA

Model	Dev F1	Test F1
Baseline	68.17	60.28
Ours	67.74	62.39

Table 1: Results for ENG-SPA.

(Small dev set)

Model	Dev F1	Test F1
Baseline	79.55	70.08
Ours	81.41	71.62

Table 2: Results for MSA-EGY.

- Good at Person and Location categories
- The Title category is difficult
- The Other category is extremely difficult for ENG-SPA (only 3 examples for MSA-EGY)

	Precision	Recall	Entity F1	# Train
EVENT	56.25	20.00	29.51	232
GROUP	69.77	30.93	42.86	718
LOC	70.75	69.23	69.98	811
ORG	62.50	27.23	37.93	2810
OTHER	14.29	1.71	3.08	324
PER	76.52	68.15	72.09	4701
PROD	63.76	47.53	54.46	1369
TIME	51.58	37.09	43.24	577
TITLE	49.14	25.79	33.83	824
Overall	70.62	55.88	62.39	12366

Table 1: ENG-SPA test performance breakdown.

	Precision	Recall	Entity F1	# Train
EVENT	78.18	61.43	68.80	535
GROUP	69.77	76.92	73.17	1799
LOC	76.19	67.84	71.78	3275
ORG	66.14	67.20	66.67	1504
OTHER	100.00	100.00	100.00	116
PER	77.29	69.53	73.21	5705
PROD	76.47	78.79	77.61	538
TIME	64.29	72.00	67.92	466
TITLE	31.58	60.00	41.38	896
Overall	73.95	69.42	71.62	14834

Table 2: MSA-EGY test performance breakdown.

Results

- 1st place on MSA-EGY, 3rd place on ENG-SPA

Model	Dev F1	Test F1
Baseline	68.17	60.28
Ours	67.74	62.39

Table 1: Results for ENG-SPA.

(Small dev set)

Model	Dev F1	Test F1
Baseline	79.55	70.08
Ours	81.41	71.62

Table 2: Results for MSA-EGY.

- Good at Person and Location categories
- The Title category is difficult
- The Other category is extremely difficult for ENG-SPA (only 3 examples for MSA-EGY)

	Precision	Recall	Entity F1	# Train
EVENT	56.25	20.00	29.51	232
GROUP	69.77	30.93	42.86	718
LOC	70.75	69.23	69.98	811
ORG	62.50	27.23	37.93	2810
OTHER	14.29	1.71	3.08	324
PER	76.52	68.15	72.09	4701
PROD	63.76	47.53	54.46	1369
TIME	51.58	37.09	43.24	577
TITLE	49.14	25.79	33.83	824
Overall	70.62	55.88	62.39	12366

Table 1: ENG-SPA test performance breakdown.

	Precision	Recall	Entity F1	# Train
EVENT	78.18	61.43	68.80	535
GROUP	69.77	76.92	73.17	1799
LOC	76.19	67.84	71.78	3275
ORG	66.14	67.20	66.67	1504
OTHER	100.00	100.00	100.00	116
PER	77.29	69.53	73.21	5705
PROD	76.47	78.79	77.61	538
TIME	64.29	72.00	67.92	466
TITLE	31.58	60.00	41.38	896
Overall	73.95	69.42	71.62	14834

Table 2: MSA-EGY test performance breakdown.

Other findings

- Pre-processing is important for user generated data: cleaning noisy data and clustering special tokens led to significant changes in pre-trained word vector coverage and final performance
- Small dev set (ENG-SPA train/dev/test: 50.8k/832/15.6k) brings difficulties to hyper-parameter tuning
- Attention mechanism for combining different embeddings is an open question: more analysis and evaluation to better understand and improve it

Summary & Discussion

- A system based on mainstream recurrent neural network models and techniques

Possible improvements

- Better pre-processing (higher word vector coverage)
- Use contextualized attention (taking token context into account)
- Make use of other linguistic features and gazetteers
- Model ensemble

Thank you!

facebook
Artificial Intelligence

Questions?