Neural Cross-Lingual Named Entity Recognition with Minimal Resources

Jiateng Xie, Zhilin Yang, Graham Neubig,
Noah A. Smith, and Jaime Carbonell

Conference on Empirical Methods in Natural Language Processing
Oct.31 - Nov. 4, 2018

Presented by Vivian Lin

Named Entity Recognition

Input: sentence

Output: labels for each token (location, organization, person, none, etc)

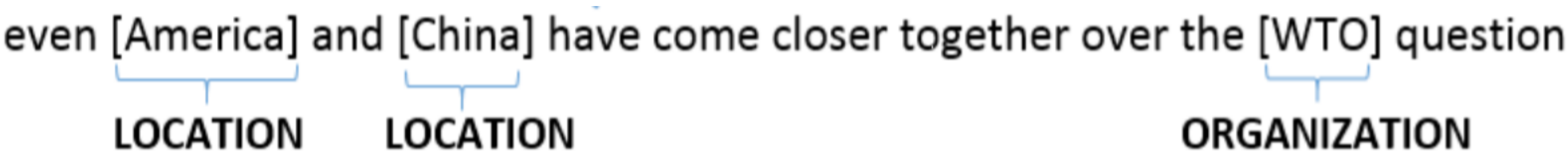


Figure: Abdel-Hady, et al. 2014. Unsupervised active learning of CRF model for cross-lingual named entity recognition. In *Proceedings of the 6th IAPR TC 3 International Workshop (ANNPR 2014).*

Cross-Lingual Named Entity Recognition

Goal: Construct NER training examples for the low-resource language using existing NER examples in a high-resource language

- "Source": high-resource language
- "Target": low-resource language

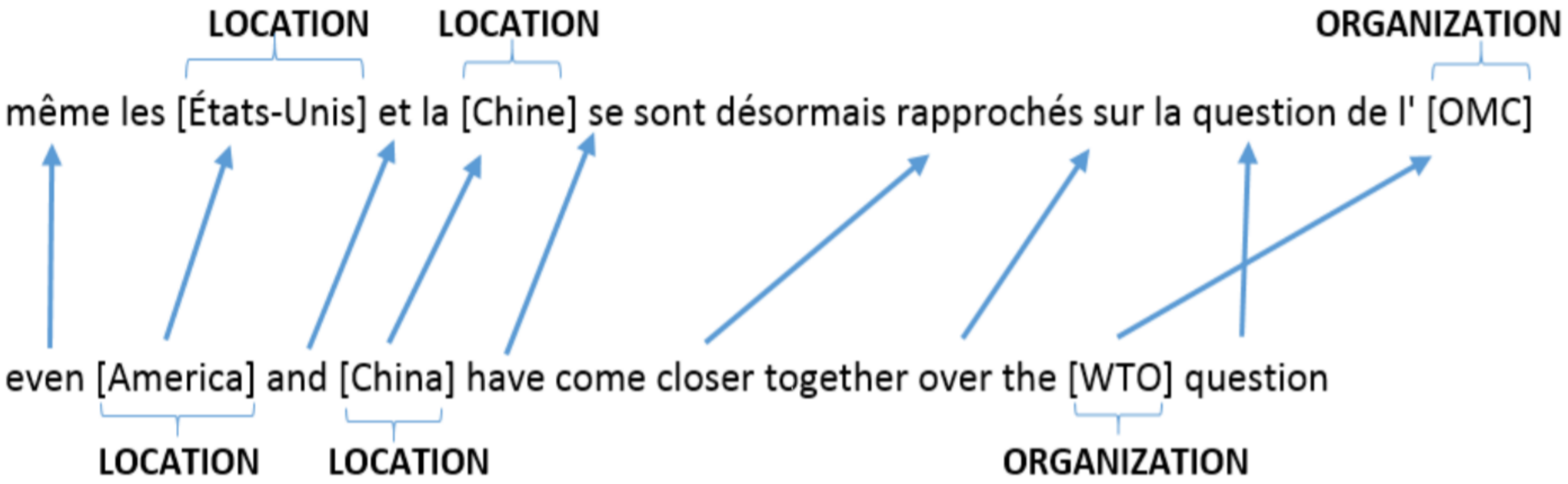


Figure: Abdel-Hady, et al. 2014. Unsupervised active learning of CRF model for cross-lingual named entity recognition. In *Proceedings of the 6th IAPR TC 3 International Workshop (ANNPR 2014).*

Cross-Lingual Named Entity Recognition

Challenge 1: Performing lexical mappings can be difficult for low-resource languages

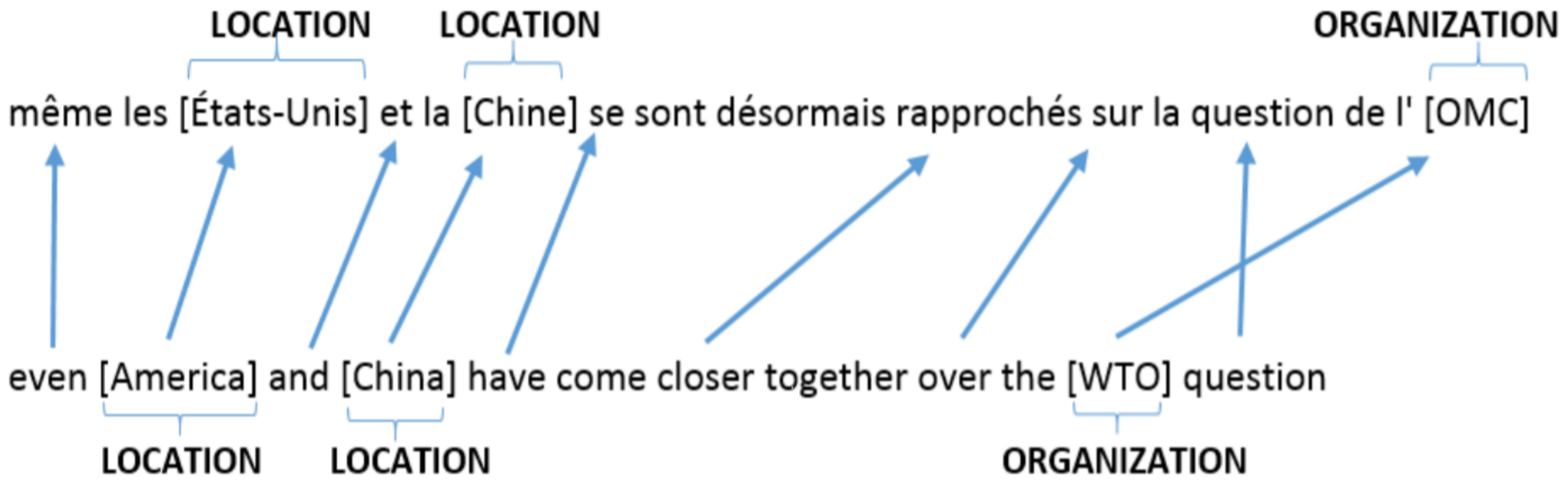


Figure: Abdel-Hady, et al. 2014. Unsupervised active learning of CRF model for cross-lingual named entity recognition. In *Proceedings of the 6th IAPR TC 3 International Workshop (ANNPR 2014).*

Cross-Lingual Named Entity Recognition

Challenge 2: Languages have different word orderings for the same sentence

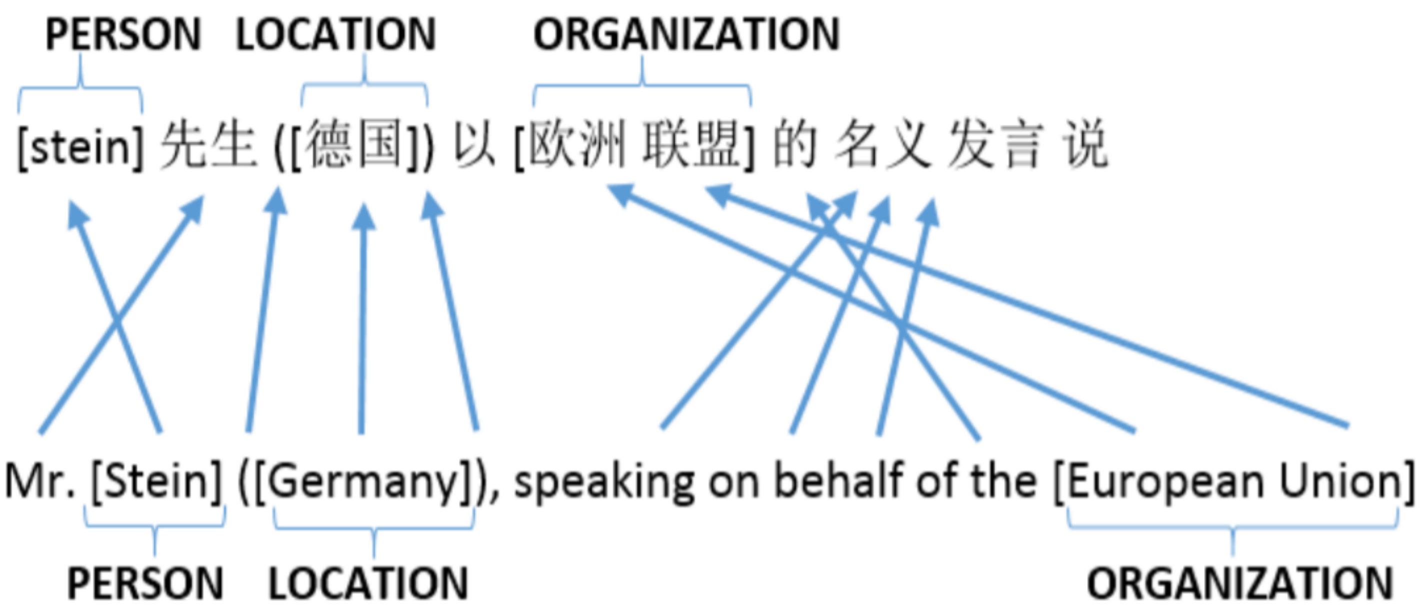
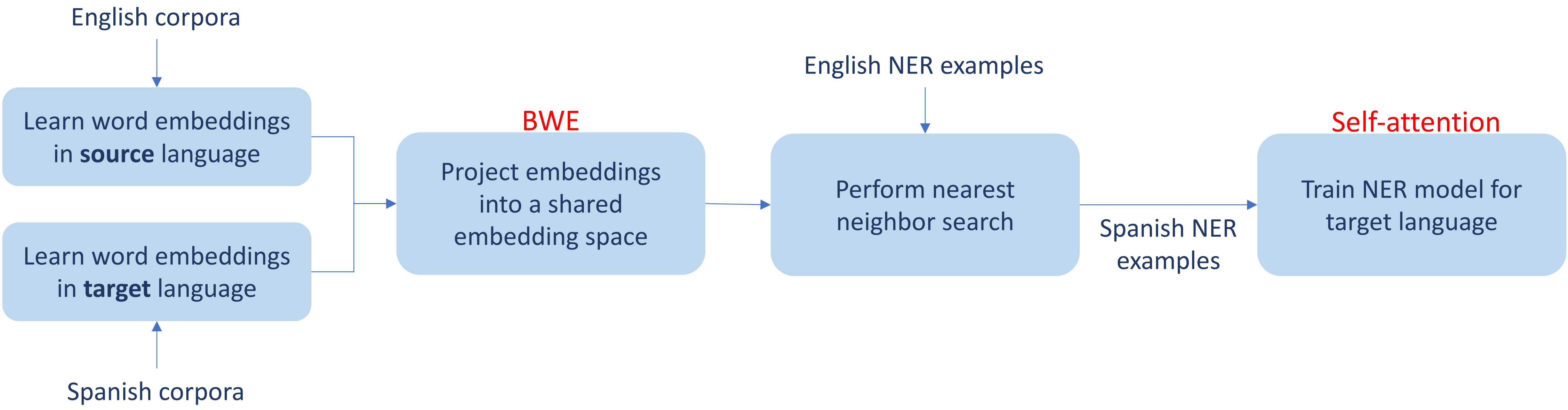


Figure: Abdel-Hady, et al. 2014. Unsupervised active learning of CRF model for cross-lingual named entity recognition. In *Proceedings of the 6th IAPR TC 3 International Workshop (ANNPR 2014).*

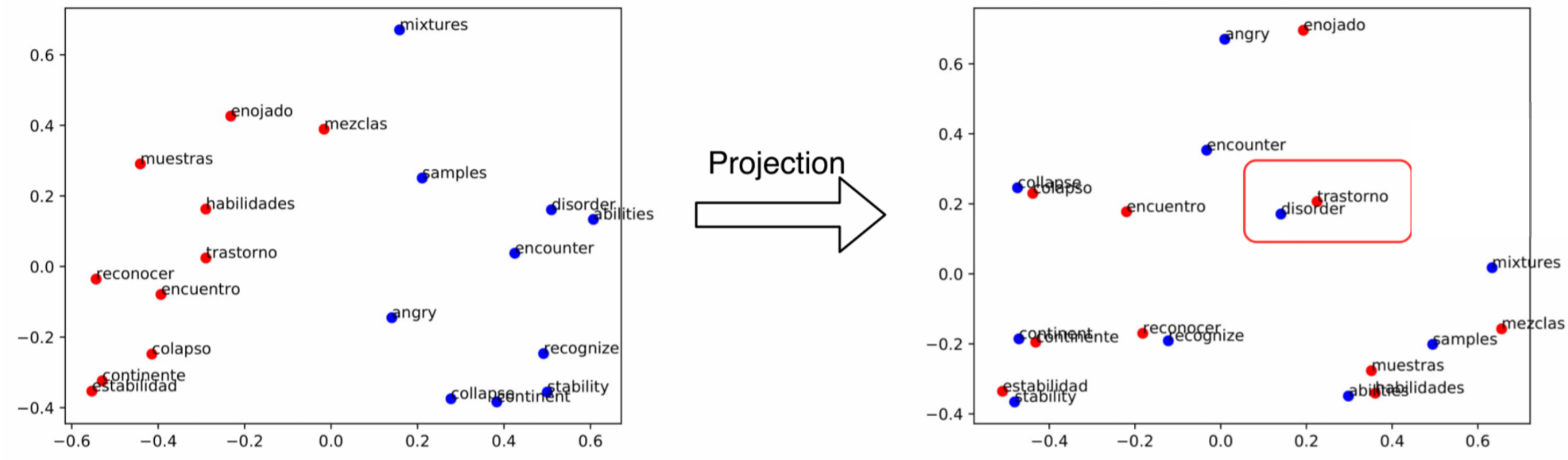
Approach

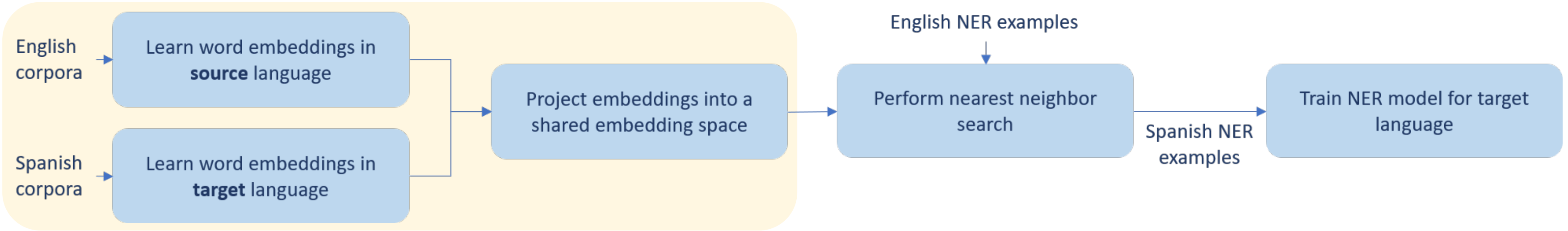
- Challenge 1: Performing lexical mappings can be difficult for low-resource languages
  - Solution: bilingual word embeddings (BWE)
  - Benefit: Doesn’t require a large number of parallel resources
- Challenge 2: Languages have different word orderings for the same sentence
  - Solution: self-attention
  - Benefit: self-attention is order-invariant
- Resources (limited to imitate resources available for low-resource languages)
  - Labeled NER examples in the source language
  - Monolingual corpora in the source and target languages
  - A small dictionary
- Demonstrated with translation from English to Spanish, German, Dutch, and Uyghur

Pipeline


English corpora
Learn word embeddings in **source** language
Learn word embeddings in **target** language
Spanish corpora
BWE
Project embeddings into a shared embedding space
English NER examples
Perform nearest neighbor search
Spanish NER examples
Self-attention
Train NER model for target language

Pipeline: Bilingual Word Embeddings





Pipeline: Bilingual Word Embeddings

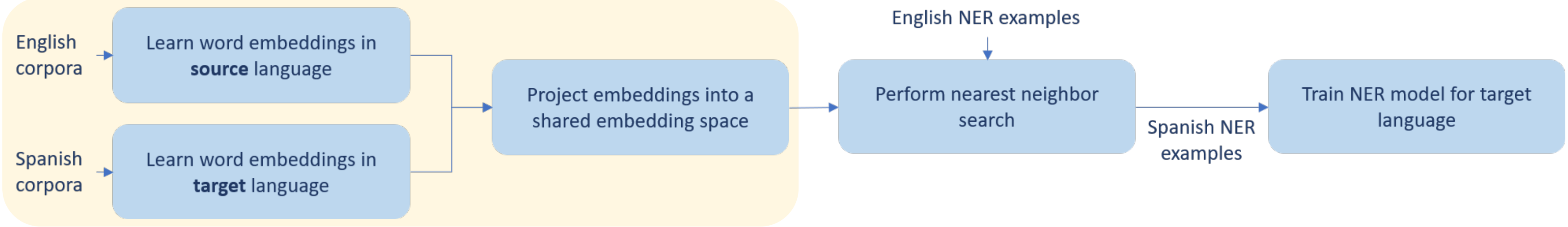
Dictionary

$$\{x_i, y_i\}_{i=1}^{D}$$

Objective function

$$\min_{W} \sum_{i=1}^{d} ||Wx_i - y_i||^2 \text{ s.t. } WW^\top = I$$

parameter matrix



Pipeline: Bilingual Word Embeddings

Dictionary

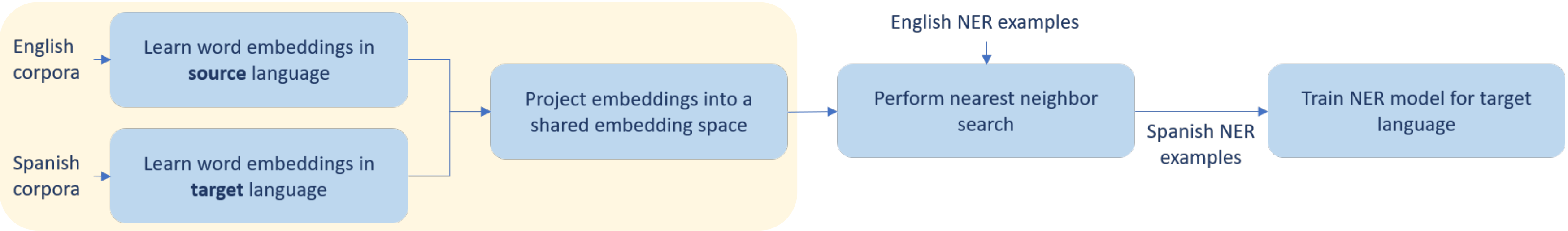
$$\{x_i, y_i\}_{i=1}^{D}$$

Equivalent Objective function

$$\min_{W} \sum_{i=1}^{d} \|Wx_i - y_i\|^2 \ \text{ s.t. } \ WW^\top = I \quad \longleftrightarrow \quad \max_{W} \mathrm{Tr}(X_D W Y_D^\top) \ \text{ s.t. } \ WW^\top = I$$

parameter matrix

embedding matrices



Pipeline: Bilingual Word Embeddings

Dictionary

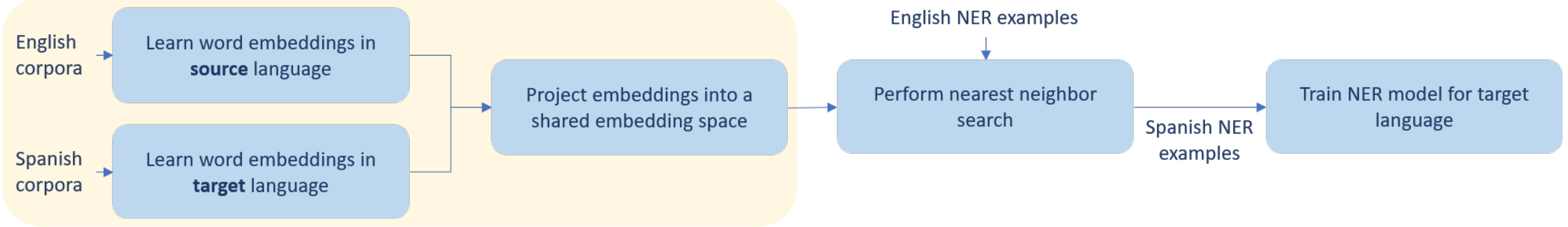
$$\{x_i, y_i\}_{i=1}^{D}$$

Equivalent Objective function

$$\min_W \sum_{i=1}^{d} \|Wx_i - y_i\|^2 \text{ s.t. } WW^\top = I \longleftrightarrow \max_W \text{Tr}(X_D W Y_D^\top) \text{ s.t. } WW^\top = I$$

Solution

$W = UV^\top$, where U and V are given by the SVD: $Y_D^\top X_D = U \sum V^\top$

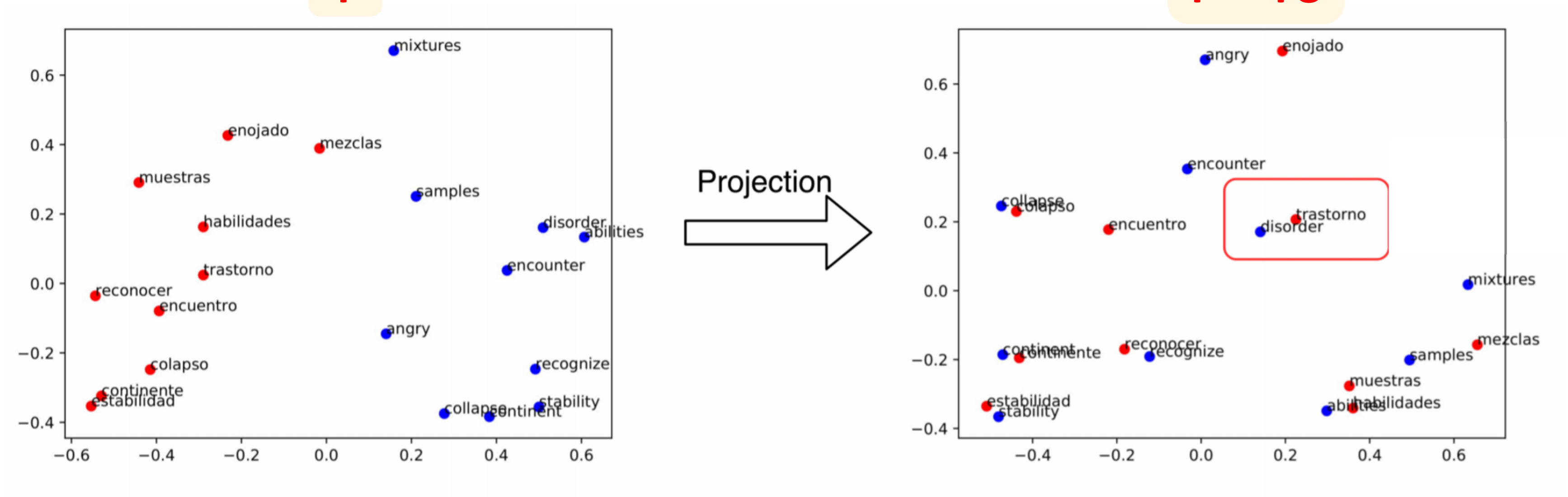


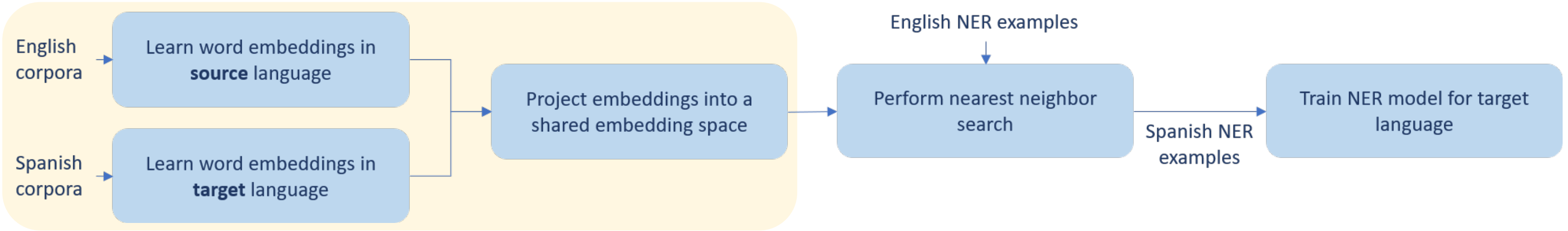
Pipeline: Bilingual Word Embeddings

X
Y

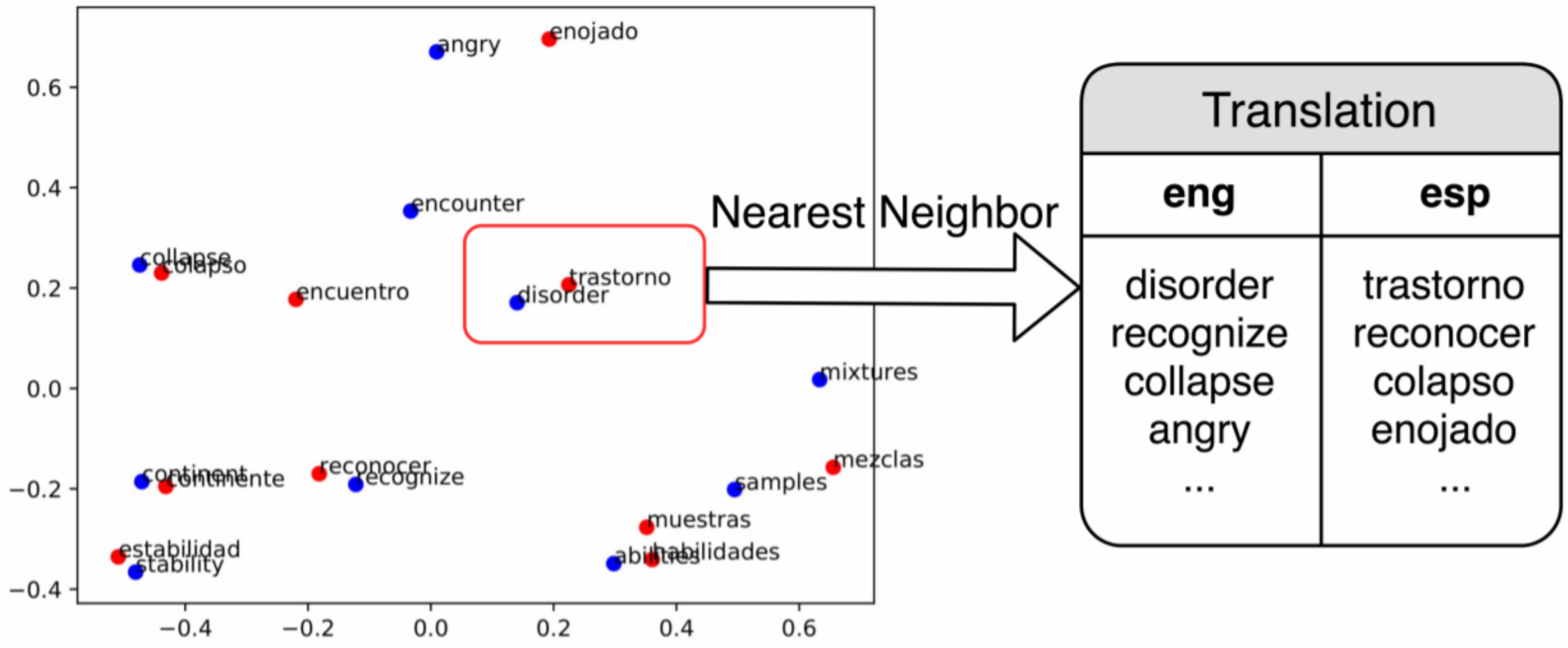
$$S = YWX^T = YU(XV)^T$$

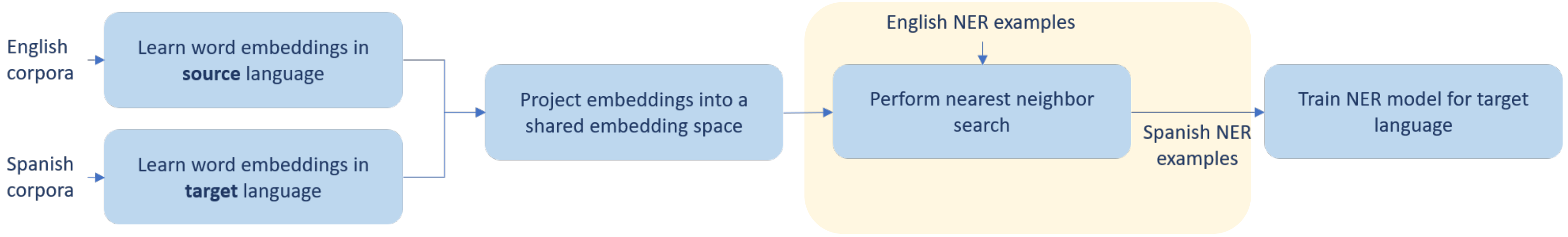
X' = XV
Y' = YU





Pipeline: Translation





Pipeline: NER Model

Model Input: Sentences in the low-resource language
Model output: NER labels for input sentences
Training data: NER examples translated to the low-resource

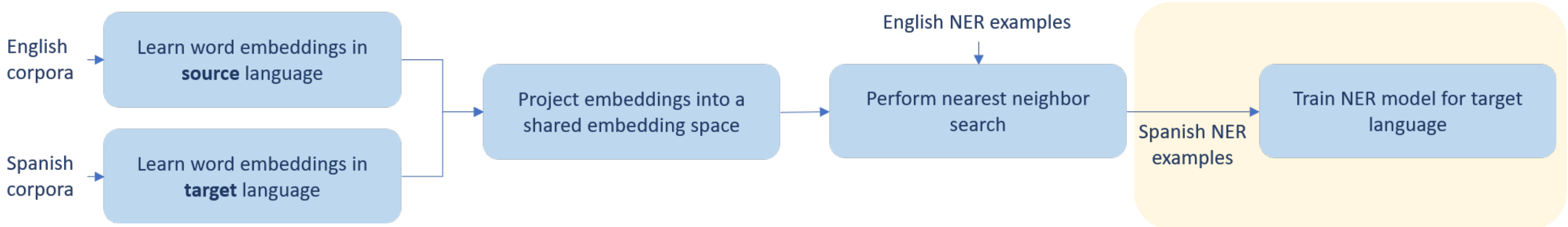
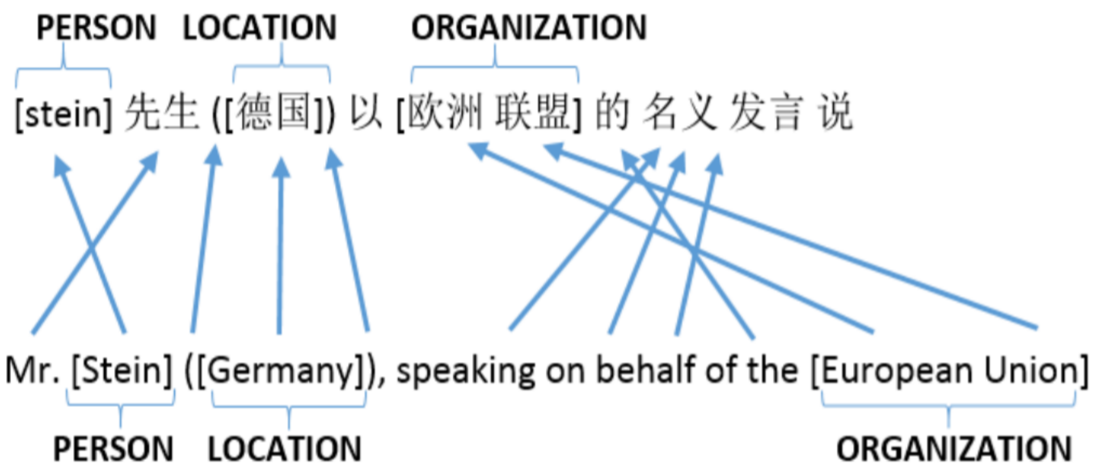


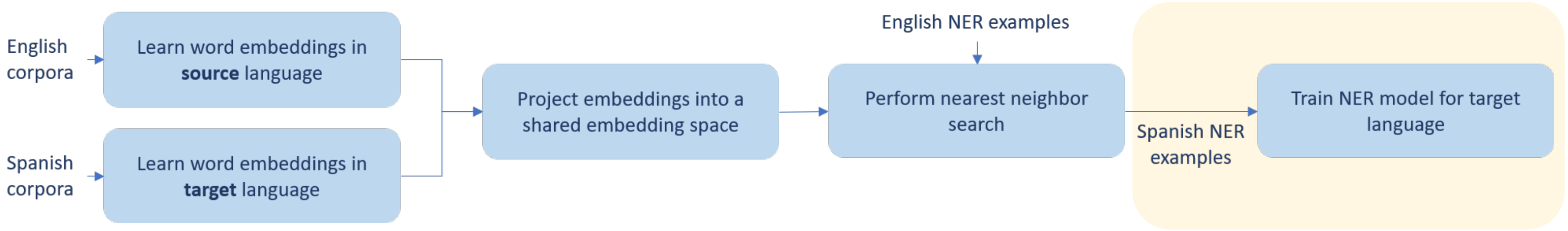
Figure: Abdel-Hady, et al. 2014. Unsupervised active learning of CRF model for cross-lingual named entity recognition. In *Proceedings of the 6th IAPR TC 3 International Workshop (ANNPR 2014).*

Pipeline: NER Model

Challenge: Word to word translation doesn't account for word orderings

- Translated NER training data uses "corrupted" sentences (words are wrongly ordered)

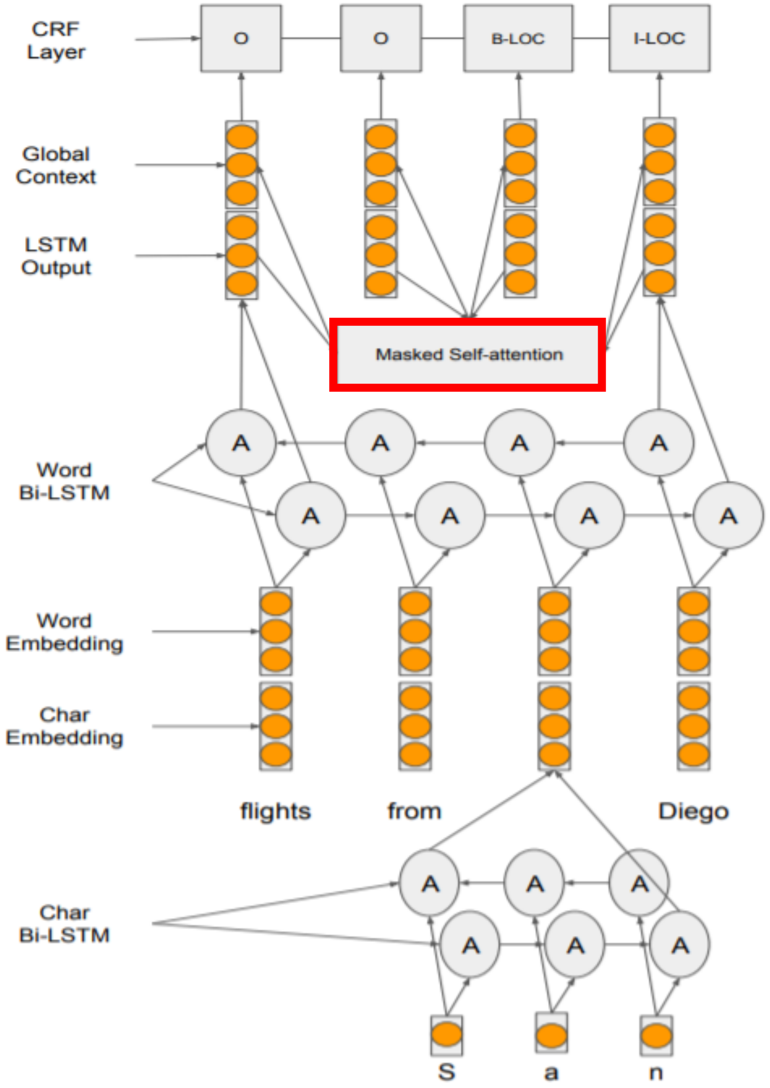


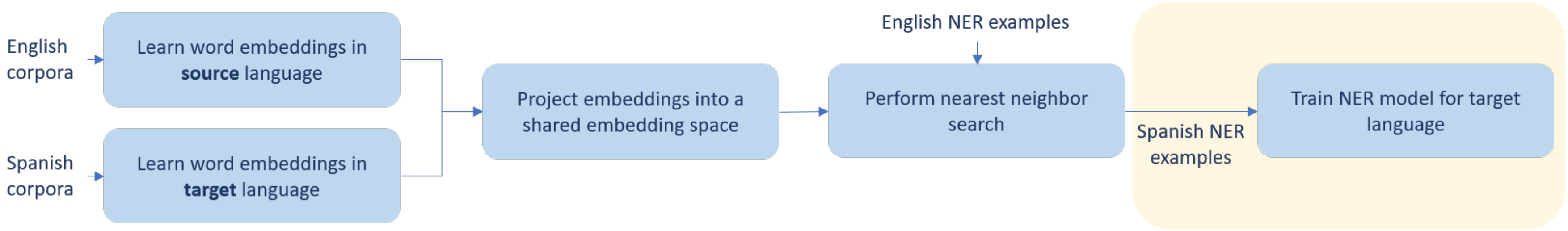


Pipeline: NER Model

Solution: Self-attention layer

- Each word is associated with a context feature vector, produced using all of the words in a sentence
  - i.e. feature vectors are order-invariant





Experiments: Proof of Concept

- Benchmark datasets: CoNLL 2002 and 2003 datasets for NER
  - English, German, Dutch, Spanish
- Source language English tested with target languages German, Dutch, and Spanish
- Vocabulary size: 100,000
- Considered three dictionaries, obtained in different ways

Experiments: Proof of Concept

Baseline (bilingual dictionary)

	Model	Spanish	Dutch	German	Extra Resources
*	Täckström et al. (2012)	59.30	58.40	40.40	parallel corpus
*	Nothman et al. (2013)	61.0	64.00	55.80	Wikipedia
*	Tsai et al. (2016)	60.55	61.60	48.10	Wikipedia
*	Ni et al. (2017)	65.10	65.40	58.50	Wikipedia, parallel corpus, 5K dict.
*†	Mayhew et al. (2017)	65.95	66.50	**59.11**	Wikipedia, 1M dict.
*	Mayhew et al. (2017) (only Eng. data)	51.82	53.94	50.96	1M dict.
	Our methods:				
	BWET (id.c.)	71.14 ± 0.60	70.24 ± 1.18	57.03 ± 0.25	–
	BWET (id.c.) + self-att.	$\mathbf{72.37} \pm 0.65$	70.40 ± 1.16	$\mathbf{57.76} \pm 0.12$	–
	BWET (adv.)	70.54 ± 0.85	70.13 ± 1.04	55.71 ± 0.47	–
	BWET (adv.) + self-att.	71.03 ± 0.44	$\mathbf{71.25} \pm 0.79$	56.90 ± 0.76	–
	BWET	71.33 ± 1.26	69.39 ± 0.53	56.95 ± 1.20	10K dict.
	BWET + self-att.	71.67 ± 0.86	70.90 ± 1.09	57.43 ± 0.95	10K dict.
*	BWET on data from Mayhew et al. (2017)	66.53 ± 1.12	69.24 ± 0.66	55.39 ± 0.98	1M dict.
*	BWET + self-att. on data from Mayhew et al. (2017)	66.90 ± 0.65	69.31 ± 0.49	55.98 ± 0.65	1M dict.
*	Our supervised results	86.26 ± 0.40	86.40 ± 0.17	78.16 ± 0.45	annotated corpus

* used additional resources

Experiments: Proof of Concept

	Model	Spanish	Dutch	German	Extra Resources
*	Täckström et al. (2012)	59.30	58.40	40.40	parallel corpus
*	Nothman et al. (2013)	61.0	64.00	55.80	Wikipedia
*	Tsai et al. (2016)	60.55	61.60	48.10	Wikipedia
*	Ni et al. (2017)	65.10	65.40	58.50	Wikipedia, parallel corpus, 5K dict.
*†	Mayhew et al. (2017)	65.95	66.50	**59.11**	Wikipedia, 1M dict.
*	Mayhew et al. (2017) (only Eng. data)	51.82	53.94	50.96	1M dict.
	Our methods:				
	BWET (id.c.)	71.14 ± 0.60	70.24 ± 1.18	57.03 ± 0.25	–
	BWET (id.c.) + self-att.	$\mathbf{72.37} \pm 0.65$	70.40 ± 1.16	$\mathbf{57.76} \pm 0.12$	–
	BWET (adv.)	70.54 ± 0.85	70.13 ± 1.04	55.71 ± 0.47	–
	BWET (adv.) + self-att.	71.03 ± 0.44	$\mathbf{71.25} \pm 0.79$	56.90 ± 0.76	–
	BWET	71.33 ± 1.26	69.39 ± 0.53	56.95 ± 1.20	10K dict.
	BWET + self-att.	71.67 ± 0.86	70.90 ± 1.09	57.43 ± 0.95	10K dict.
*	BWET on data from Mayhew et al. (2017)	66.53 ± 1.12	69.24 ± 0.66	55.39 ± 0.98	1M dict.
*	BWET + self-att. on data from Mayhew et al. (2017)	66.90 ± 0.65	69.31 ± 0.49	55.98 ± 0.65	1M dict.
*	Our supervised results	86.26 ± 0.40	86.40 ± 0.17	78.16 ± 0.45	annotated corpus

The model performs worst on German text

- German capitalization patterns are different than those of English
- The model is overfitting to English capitalization patterns

* used additional resources

Experiments: Uyghur

	Model	Uyghur Unsequestered Set	Extra Resources
*†	Mayhew et al. (2017)	**51.32**	Wikipedia, 100K dict.
*	Mayhew et al. (2017) (only Eng. data)	27.20	Wikipedia, 100K dict.
	BWET	25.73 ± 0.89	5K dict.
	BWET + self-att.	26.38 ± 0.34	5K dict.
*	BWET on data from Mayhew et al. (2017)	30.20 ± 0.98	Wikipedia, 100K dict.
*	BWET + self-att. on data from Mayhew et al. (2017)	30.68 ± 0.45	Wikipedia, 100K dict.
*	Combined (see text)	31.61 ± 0.46	Wikipedia, 100K dict., 5K dict.
*	Combined + self-att.	$\mathbf{32.09} \pm 0.61$	Wikipedia, 100K dict., 5K dict.

Competitive performance, despite lesser resources

* used additional resources

Contributions and Remaining Work

- Addresses the low-resource language problem in supervision
  - Translates NER training data from a high-resource language to a low-resource language
  - Adds a self-attention layer to an existing model architecture, accounting for word mis-orderings after translation
  - Even with less supervision, the proposed approach performs competitively to the state-of-the-art
- Continuing challenges
  - Language-specific patterns (capitalization, characters used)
    - Differing capitalization patterns across languages make cross-lingual NER more difficult
    - [WILL BE EDITED FURTHER] If language A uses different characters than language B, this limits the ways in which the seed dictionary can be produced (i.e. it is more difficult to obtain the resources necessary to perform BWE).
      - Uyghur is written in Arabic script, but English is written in the Latin alphabet
  - Lacks theoretical guarantees for translation
  - Requires no NER training labels (unsupervised) for the target language, but does require a small dictionary (resources) for source-target word translation