Code-Switching as a Cross-Lingual Training Signal: an Example with Unsupervised Bilingual Embedding

Félix Gaschi[2,3], Ilias El Baamarani[1], Barbara Gendron[1], Parisa Rastin[2], Yannick Toussaint[2]

[1]École des Mines de Nancy, [2]LORIA, [3]SAS Posos

3rd Multilingual Representation Learning Workshop @ EMNLP 2023

Code-Switching (CS)

Code-switching (CS): words from **different languages** are found in the **same sentence**.

Language contamination (LC): **whole sentences** from other languages in a monolingual corpus.

Code-Switching (CS)

Code-switching (CS): words from **different languages** are found in the **same sentence**.

Language contamination (LC): **whole sentences** from other languages in a monolingual corpus.

Code-Switching (CS)

Code-switching (CS): words from **different languages** are found in the **same sentence**.

Language contamination (LC): **whole sentences** from other languages in a monolingual corpus.

(+) Always present in (supposedly) monolingual corpora [Blevins and Zettlemoyer, 2022]

Code-Switching (CS)

Code-switching (CS): words from **different languages** are found in the **same sentence**.

Language contamination (LC): **whole sentences** from other languages in a monolingual corpus.

- (+) Always present in (supposedly) monolingual corpora [Blevins and Zettlemoyer, 2022]
- (+) Support for cross-lingual generalization abilities in monolingual embeddings

Code-Switching (CS)

Code-switching (CS): words from **different languages** are found in the **same sentence**.

Language contamination (LC): **whole sentences** from other languages in a monolingual corpus.

(+) Always present in (supposedly) monolingual corpora [Blevins and Zettlemoyer, 2022]

(+) Support for cross-lingual generalization abilities in monolingual embeddings

(−) No context sharing because no token/sentence pairs

CS in monolingual corpora

Exemple 1： 1999年歐洲歌唱大賽(eurovision song contest 1999) 為歐洲歌唱大賽之第44屆比賽

Exemple 2： as a result , ” li ” (禮) , meaning ” ritual ” or ” etiquette , ”governed the conduct of the nobles , whilst ” xing ” (刑) , the rules of punishment

Exemple 3： 是一款由鬼游(ghost town games) 公司, team 17 行的烹模游. 玩家通多人合作或多角控制, 控制多游角色挑各种房里的机

Figure 1: Examples of code-switching

Cross-lingual embeddings

Trade-off required ressources / alignment robustness:

- Supervised methods: BDI requires parallel corpora [Mikolov et al., 2013]
- Fully unsupervised embedding alignment [Conneau et al., 2017] often lacks of robustness [Søgaard et al., 2018]
- A weak supervision signal brings more stability [Søgaard et al., 2018]

Cross-lingual embeddings

Trade-off required ressources / alignment robustness:

- Supervised methods: BDI requires parallel corpora [Mikolov et al., 2013]
- Fully unsupervised embedding alignment [Conneau et al., 2017] often lacks of robustness [Søgaard et al., 2018]
- A weak supervision signal brings more stability [Søgaard et al., 2018]

$\longrightarrow$ CS would act as a weak supervision signal bilingual embedding learning.

CS-enhanced cross-lingual embedding learning

CS-powered cross-lingual mappings: **tokens randomly replaced by their translation**.

- Alleviate the amount of parallel data [Krishnan et al., 2021]
- Improve PLMs' cross-lingual generalization [Qin et al., 2020, Yang et al., 2020]

$\longrightarrow$ Artificially-generated CS

CS-enhanced cross-lingual embedding learning

CS-powered cross-lingual mappings: **tokens randomly replaced by their translation**.

- Alleviate the amount of parallel data [Krishnan et al., 2021]
- Improve PLMs' cross-lingual generalization [Qin et al., 2020, Yang et al., 2020]

$\longrightarrow$ Artificially-generated CS

Ours:

(1) Identifies CS situations in monolingual corpora

(2) Uses **natural CS** to learn a bilingual embedding using orthogonal mapping across languages

Code-switching identfication

No supervision: don't want to rely on a full bilingual dictionnary!

$\longrightarrow$ Use **regular expressions** to detect CS situations.

Code-switching identfication

No supervision: don't want to rely on a full bilingual dictionnary!
⟶ Use **regular expressions** to detect CS situations.

Consequences:

- Scope reduction: languages written in different scripts (en-zh, en-ru, en-ar)
- Code-switching detection → script-switching detection?

Code-switching identfication

No supervision: don't want to rely on a full bilingual dictionnary!
$\longrightarrow$ Use **regular expressions** to detect CS situations.

Consequences:

- Scope reduction: languages written in different scripts (en-zh, en-ru, en-ar)
- Code-switching detection $\rightarrow$ script-switching detection?

Code-switching pair: pair of words from different scripts found **in the same context window**.

Training procedure

Based on monolingual skip-gram loss [Mikolov et al., 2013]:

$$L = -\frac{1}{|C|} \sum_{w_i \in C} \sum_{w_j \in \mathcal{N}(w_i)} \log P(w_j | w_i) \tag{1}$$

Training procedure

Based on monolingual skip-gram loss [Mikolov et al., 2013]:

$$L = -\frac{1}{|C|} \sum_{w_i \in C} \sum_{w_j \in \mathcal{N}(w_i)} \log P(w_j | w_i) \tag{1}$$

Replace the original negative sampling:

$$\log P(w_j | w_i) = \log \sigma(\tilde{x}_j^\top x_i) + \sum_{w_k \sim P_V}^{n} \log \sigma(-\tilde{x}_k^\top x_i) \tag{2}$$

Training procedure

Based on monolingual skip-gram loss [Mikolov et al., 2013]:

$$L = -\frac{1}{|C|} \sum_{w_i \in C} \sum_{w_j \in \mathcal{N}(w_i)} \log P(w_j | w_i) \tag{1}$$

Replace the original negative sampling:

$$\log P(w_j | w_i) = \log \sigma(\tilde{x}_j^\top x_i) + \sum_{w_k \sim P_V}^{n} \log \sigma(-\tilde{x}_k^\top x_i) \tag{2}$$

By computation on projected tokens in a CS pair (w_i^{src}, w_j^{tgt}):

$$\log P(w_j^{\text{tgt}} | w_i^{\text{src}}) = \log \sigma(\tilde{x}_j^{\text{tgt}\,\top} W x_i^{\text{src}}) + \sum_{w_k^{\text{tgt}} \sim \mathcal{U}_{V_{\text{tgt}}}}^{n} \log \sigma(-\tilde{x}_k^{\text{tgt}\,\top} W x_i^{\text{src}}) \tag{3}$$

Experimental setup

CS pairs extraction:

- Data is tokenized Wikipedia dumps
- FastText [Bojanowski et al., 2016] monolingual embeddings for 200,000 most frequent words
- Context window of width 5

Experimental setup

CS pairs extraction:

- Data is tokenized Wikipedia dumps
- FastText [Bojanowski et al., 2016] monolingual embeddings for 200,000 most frequent words
- Context window of width 5

Training pipeline:

- Modified skip-gram loss for **model initialisation**
- Add **orthogonalization steps** so W preserves distances
- **Refinement step** using `VecMap` self-learning procedure [Artetxe et al., 2018]

Code-switching in monolingual corpora

- Check for script-switching with a dictionnary
- Differentiate CS and token contamination

pair	number
en-ar	7,848,024
en-ru	50,182,802
en-zh	23,097,625

Figure: Counts of CS pairs

lang	tokens	token contamination			code-switching		
		coverage (%)	count	count digits	coverage (%)	count	count digits
ar	229M	44.9	1,043,396	6,511,347	38.0	486,764	6,360,450
ru	685M	55.1	5,237,773	26,063,394	50.7	4,158,232	25,637,900
zh	319M	47.6	1,720,247	3,220,332	39.4	1,174,912	3,117,309

Figure: Presence of English words in non-English monolingual corpora. An example is considered code-switched if it is in the vicinity of a non-English word.

Results of `CoSwitchMap`

method	en-ar	en-ru	en-zh
Methods with other self-learning procedures			
WP	$10.7_{\pm 9.9}$	$36.9_{\pm 1.4}$	$0.6_{\pm 0.8}$
MUSE	$30.9_{\pm 3.3}$	$41.7_{\pm 2.9}$	$0.0_{\pm 3.3}$
Different initializations for the same self-learning			
VecMap	$36.4_{\pm 1.8}$	$\mathbf{49.1}_{\pm 0.4}$	$0.0_{\pm 0.0}$
w/ MUSE init.	$37.4_{\pm 2.6}$	$48.3_{\pm 0.4}$	$0.0_{\pm 0.1}$
w/ WP init.	$38.6_{\pm 0.7}$	$45.8_{\pm 2.8}$	$0.1_{\pm 0.0}$
w/ identical init.	$\underline{39.8}_{\pm 0.3}$	$\underline{48.9}_{\pm 0.2}$	$36.8_{\pm 0.8}$
`CoSwitchMap` (ours)	$\mathbf{39.9}_{\pm 0.1}$	$\underline{49.0}_{\pm 0.3}$	$\mathbf{37.9}_{\pm 0.9}$
supervised	*43.0*	*52.7*	*43.3*

Figure: Comparison of `CoSwitchMap` with other unsupervised mapping-based methods. Top-1 accuracy of a nearest neighbout search with CSLS criterion for BLI.

Conclusion

Contributions:

- `CoSwitchMap` outperforms existing unsupervised mapping-based methods
- Natural CS constitutes a cross-lingual training signal for multilingual static embeddings
- Leverage naturally-occuring code-switching

Conclusion

Contributions:

- `CoSwitchMap` outperforms existing unsupervised mapping-based methods
- Natural CS constitutes a cross-lingual training signal for multilingual static embeddings
- Leverage naturally-occuring code-switching

Limitations:

- Only works with languages written in different scripts.
- Most of the CS situations are litteral translations of words in the vicinity.
- Demonstrates the utility of code-switching but still need for a more generalized method.

References I

- Min Xiao and Yuhong Guo (2014). *Distributed word representation learning for cross-lingual dependency parsing*. CoNLL 2014.
- Gouws and Søgaard (2015). *Simple task-specific bilingual word embeddings*. EMNLP 2018.
- Mikolov et al. (2013). *Exploiting similarities among languages for machine translation*. CoRR, abs/1309.4168.
- Conneau et al. (2017). *Word translation withour parallel data*. ICLR 2018.
- Søgaard et al. (2018). *On the limitations of unsupervised bilingual dictionary induction*. ACL 2018.
- Qin et al. (2020). *Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP*. IJCAI 2020.

References II

- Yang et al. (2020). *Alternating language modeling for cross-lingual pre-training*. AAAI 2020.
- Artetxe et al. (2018). *A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings*. ACL 2018.