Subword Mapping and Anchoring across Languages

Giorgos Vernikos & Andrei Popescu-Belis

EMNLP 2021

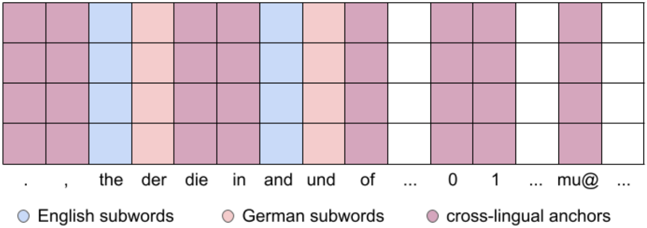
Shared Subword Vocabularies

Multilingual systems rely on shared subword vocabularies learned on concatenation of monolingual data from various languages.

Shared Subword Vocabularies

Multilingual systems rely on shared subword vocabularies learned on concatenation of monolingual data from various languages.

Subwords that appear in several languages (e.g shared words, punctuation, digits) function as **anchors** between languages that lead to improved performance (Conneau and Lample 2019).



Limitations of shared vocabularies

False positives

Identical subwords with different meanings,
e.g. *die* is a definite article in German and a verb in English

Limitations of shared vocabularies

False positives

Identical subwords with different meanings,
e.g. *die* is a definite article in German and a verb in English

False negatives

Different subwords with similar meanings,
e.g. *and* in English is the same as *und* in German

Subword Mapping and Anchoring across Languages (SMALA)

Create cross-lingual vocabularies that are parameter-efficient and exploit the similarity of concepts between different languages. Address the problem of false positives and false negatives by employing subword similarity to create cross-lingual anchors.

1. Subword Mapping
2. Anchoring of Similar Subwords

$$\text{the} \Leftrightarrow \text{der}$$
$$\text{in} \Leftrightarrow \text{in}$$
$$\text{and} \Leftrightarrow \text{und}$$
$$\text{is} \Leftrightarrow \text{ist}$$

Examples of alignments produced by SMALA

Subword Mapping and Anchoring across Languages (SMALA)

1. Subword Mapping

Learn separate subword vocabularies in each language.

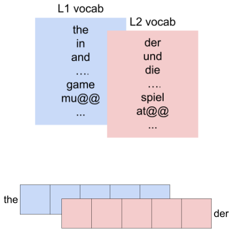
Subword Mapping and Anchoring across Languages (SMALA)

1. Subword Mapping

Learn separate subword vocabularies in each language.

⇓

Obtain subword representations using a distributional method, FastText (Bojanowski et al. 2017).



Subword Mapping and Anchoring across Languages (SMALA)

1. Subword Mapping

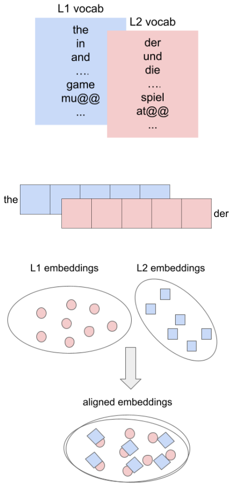
Learn separate subword vocabularies in each language.

⇓

Obtain subword representations using a distributional method, FastText (Bojanowski et al. 2017).

⇓

Align the subword representations using unsupervised alignment approach, VecMap (Artetxe et al. 2018).



2 Anchoring of Similar Subwords

Compute a similarity matrix from the aligned subwords.

	.	,	der	die	und	at@@	...
the	0.66	0.62	0.88	0.82	0.65	...	
,	0.73	0.86	0.65	0.64	0.77	...	
.	0.88	0.68	0.66	0.63	0.72	...	
of	0.62	0.65	0.78	0.67	0.68	...	
in	0.68	0.62	0.70	0.65	0.67	...	
mu@@	...	...	...	...	...	...	
...							

❷ Anchoring of Similar Subwords

Compute a similarity matrix from the aligned subwords.

$\Downarrow$

Extract subword alignments between two subwords $w_i^{L_1}$ and $w_j^{L_2}$ if and only if $w_j^{L_2}$ is the most similar subword to $w_i^{L_1}$ in $\mathcal{L}_2$ and vice versa (Jalili Sabet et al. 2020).

	.	,	der	die	und	at@@	...
the	0.66	0.62	0.88	0.82	0.65	...	
,	0.73	0.86	0.65	0.64	0.77	...	
.	0.88	0.68	0.66	0.63	0.72	...	
of	0.62	0.65	0.78	0.67	0.68	...	
in	0.68	0.62	0.70	0.65	0.67	...	
mu@@	...	...	...	...	...	...	
...							

② Anchoring of Similar Subwords

Compute a similarity matrix from the aligned subwords.

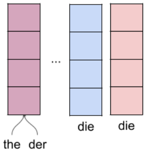
$\Downarrow$

Extract subword alignments between two subwords $w_i^{L_1}$ and $w_j^{L_2}$ if and only if $w_j^{L_2}$ is the most similar subword to $w_i^{L_1}$ in $\mathcal{L}_2$ and vice versa (Jalili Sabet et al. 2020).

$\Downarrow$

Tie the parameters (embeddings) of aligned subwords $\rightarrow$ cross-lingual anchors based on similarity.

	.	,	der	die	und	at@@	...
the	0.66	0.62	0.88	0.82	0.65	...	
,	0.73	0.86	0.65	0.64	0.77	...	
.	0.88	0.68	0.66	0.63	0.72	...	
of	0.62	0.65	0.78	0.67	0.68	...	
in	0.68	0.62	0.70	0.65	0.67	...	
mu@@	...	...	...	...	...	...	
...							



Experiments with XNLI

XNLI $\rightarrow$ determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise":

Language Model (LM) Transfer with SMALA:

- Start from a pretrained monolingual ($\mathcal{L}_1$) LM
- Add new embedding matrix for $\mathcal{L}_2$ and create cross-lingual anchors based on SMALA
- Further train model on Masked Language Modelling in $\mathcal{L}_1$&$\mathcal{L}_2$
- Fine-tune model on XNLI using data in $\mathcal{L}_1$
- Zero-shot inference on $\mathcal{L}_2$

Comparison to other methods:

- Parameter sharing
  - based on surface form: JOINT
  - based on similarity: OURS

Comparison to other methods:

- Parameter sharing
  - based on surface form: JOINT
  - based on similarity: OURS
- Initialization-based approaches
  - without any sharing: RAMEN (Tran 2020)
  - with sharing: OURS+ALIGN

Comparison to other methods:

- Parameter sharing
  - based on surface form: JOINT
  - based on similarity: OURS
- Initialization-based approaches
  - without any sharing: RAMEN (Tran 2020)
  - with sharing: OURS+ALIGN
- Multilingual Language Models
  - mBERT (Devlin et al. 2019)

Experiments with XNLI: Results

Method	Es	De	El	Ru	Ar
JOINT	70.0	64.4	61.2	56.2	45.8
OURS	74.2	70.6	70.0	65.4	62.3

Zero-shot classification scores on XNLI test set (Accuracy).

- sharing based on similarity $>$ sharing based on surface form

Experiments with XNLI: Results

Method	Es	De	El	Ru	Ar
JOINT	70.0	64.4	61.2	56.2	45.8
OURS	74.2	70.6	70.0	65.4	62.3
OURS+ALIGN	**76.5**	**72.8**	**72.9**	**70.2**	**67.0**
RAMEN	**76.5**	72.5	72.5	68.6	66.1

Zero-shot classification scores on XNLI test set (Accuracy).

- sharing based on similarity > sharing based on surface form
- better anchoring leads to more parameter-efficient vocabularies without sacrificing performance

Experiments with XNLI: Results

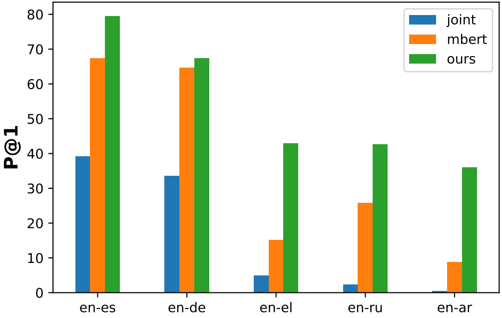
Method	Es	De	El	Ru	Ar
JOINT	70.0	64.4	61.2	56.2	45.8
OURS	74.2	70.6	70.0	65.4	62.3
OURS+ALIGN	**76.5**	**72.8**	**72.9**	**70.2**	**67.0**
RAMEN	**76.5**	72.5	72.5	68.6	66.1
mBERT	74.9	71.3	66.6	68.7	64.7

Zero-shot classification scores on XNLI test set (Accuracy).

- sharing based on similarity $>$ sharing based on surface form
- better anchoring leads to more parameter-efficient vocabularies without sacrificing performance
- competitive alternative for languages that are poorly modeled or not covered at all by multilingual LMs

Experiments with BLI: Results

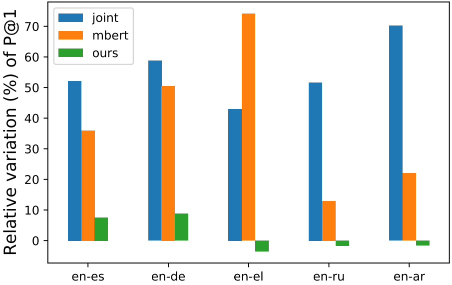
We compare the quality of representations created using SMALA vs. joint tokenization for **B**ilingual **L**exicon **I**nduction



- SMALA significantly **outperforms** JOINT and mBERT especially in more distant languages.

Experiments with BLI: Results on non-identical pairs

We remove test pairs with the same surface form
(e.g. (epic,epic) as a test pair for en-es)



- performance of JOINT and mBERT **deteriorates**, contrary to SMALA
- representations for the non-shared subwords are poorly aligned for JOINT and mBERT

Experiments with MT: Results

Languages	En-Ru		En-De		En-Ro		En-Ar	
Data	25M		5.85M		612k		239k	
	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$
JOINT	30.0	26.1	32.1	27.1	30.9	23.2	29.0	11.8
OURS	30.2	**26.6**	32.1	27.0	30.8	23.3	28.8	12.2

BLEU scores of baseline and our system for machine translation.

- comparable results to the baseline across languages and dataset sizes
- slight increase in distant language pairs (En-Ru and En-Ar)
- false positives/ negatives are less important due to strong cross-lingual signal (parallel data)

Experiments with MT: Ablation of FPs and FNs

Languages	En-Ru		En-De		En-Ro		En-Ar	
	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$
Sentences	49	2225	1674	2216	1249	1295	141	866
JOINT	39.2	27.6	33.1	27.0	31.6	24.6	37.8	16.2
OURS	42.2	28.0	33.0	27.0	32.0	24.8	40.4	16.6
Δ	+3.0	+0.4	-0.1	0.0	+0.4	+0.2	+2.6	+0.3

BLEU scores for sentences where 50% of tokens are false positives and / or false negatives.

- when number of false positives/ negatives increases our approach outperforms JOINT

Summary

SMALA: a novel approach to construct shared subword vocabularies.

- Improved performance in cases where there is no cross-lingual signal, such as XNLI.
- Viable alternative in cases with cross-lingual supervision, such as MT & improved performance in presence of multiple false positives/ negatives.

Future Work

We aim to:

- apply SMALA in settings of varying cross-lingual supervision where anchors play an important role, such as unsupervised machine translation,
- explore the quality / quantity trade-off of cross-lingual anchors and
- extend our approach to more than two languages.

Thank you for your attention!

References I

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018). "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 789–798. DOI: 10.18653/v1/P18-1073. URL: https://www.aclweb.org/anthology/P18-1073.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: Transactions of the Association for Computational Linguistics 5, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://www.aclweb.org/anthology/Q17-1010.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual Language Model Pretraining". In: Advances in Neural Information Processing Systems. Ed. by H. Wallach et al. Vol. 32. URL: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Jalili Sabet, Masoud et al. (2020). "SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings". In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1627–1643. DOI: 10.18653/v1/2020.findings-emnlp.147. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.147.

Tran, Ke (2020). From English To Foreign Languages: Transferring Pre-trained Language Models. arXiv: 2002.07306 [cs.CL].

Method	Es	De	El	Ru	Ar
JOINT	26%	25%	11%	9%	10%
OURS	44%	37%	33%	31%	30%

Percentage of cross-lingual anchors for each method (shared subwords).

- OURS is more parameter-efficient than JOINT especially for distant languages

Method	Data	Es	De	El	Ru	Ar
JOINT	mono	70.0 $\pm$ 0.2	64.4 $\pm$ 0.8	61.2 $\pm$ 0.9	56.2 $\pm$ 1.1	45.8 $\pm$ 0.4
OURS	mono	**74.2** $\pm$ 0.4	**70.6** $\pm$ 0.1	**70.0** $\pm$ 0.7	**65.4** $\pm$ 0.9	**62.3** $\pm$ 0.4
OURS+ALIGN	mono	76.5 $\pm$ 0.4	72.8 $\pm$ 0.5	72.9 $\pm$ 0.5	70.2 $\pm$ 0.6	67.0 $\pm$ 0.4
OURS+ALIGN	para	**77.1** $\pm$ 0.8	**74.1** $\pm$ 0.5	**75.1** $\pm$ 0.7	**71.9** $\pm$ 0.4	**67.8** $\pm$ 0.8
RAMEN	mono	76.5 $\pm$ 0.6	72.5 $\pm$ 0.8	72.5 $\pm$ 0.8	68.6 $\pm$ 0.7	66.1 $\pm$ 0.8
RAMEN	para	**77.3** $\pm$ 0.6	**74.1** $\pm$ 0.9	74.5 $\pm$ 0.6	**71.6** $\pm$ 0.8	**68.6** $\pm$ 0.6
mBERT	mono	74.9 $\pm$ 0.4	71.3 $\pm$ 0.6	66.6 $\pm$ 1.2	68.7 $\pm$ 1.1	64.7 $\pm$ 0.6

Zero-shot classification scores on XNLI test set (Accuracy): mean and standard deviation over 5 runs, when either monolingual or parallel corpora were used for alignment (or token matching for JOINT).

- use of parallel data improves results across the board

Method	Es	De	El	Ru	Ar
JOINT	70.0	64.4	61.2	56.2	45.8
−FP	68.5	61.7	62.6	53.6	44.8
−FN	74.3	70.0	70.2	65.8	63.1
OURS (−FP−FN)	74.2	70.6	70.0	65.4	62.3

Effect of removing false positives or false negatives in XNLI (accuracy).

- false negatives impact performance more than false positives