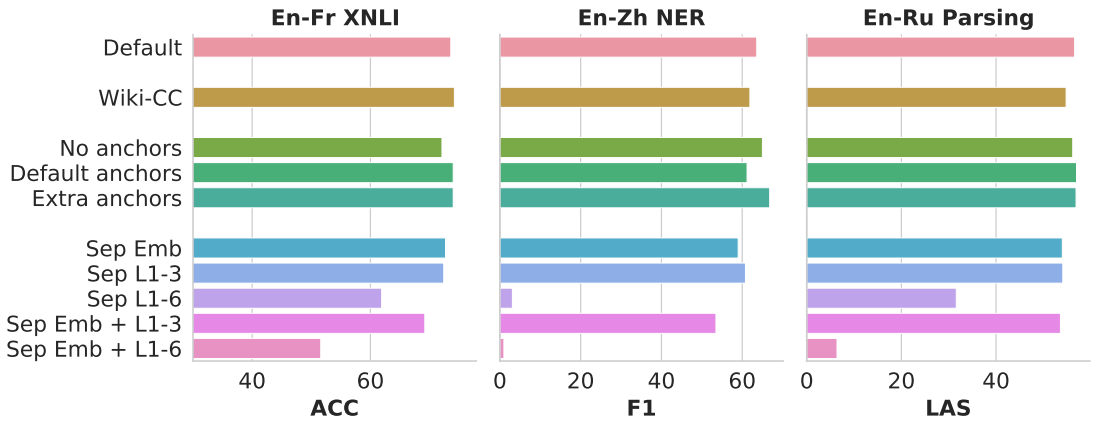

Figure 3: Cross-lingual transfer of bilingual MLM on three tasks and language pairs under different settings. Others tasks and languages pairs follows similar trend. See Tab. 1 for full results.

Model	Domain	BPE Merges	Anchors Pts	Share Param.	Softmax	XNLI (Acc) fr	ru	zh	Δ	NER (F1) fr	ru	zh	Δ	Parsing (LAS) fr	ru	zh	Δ
Default	Wiki-Wiki	80k	all	all	shared	73.6	68.7	68.3	0.0	79.8	60.9	63.6	0.0	73.2	56.6	28.8	0.0
Domain Similarity (§4.1)																	
Wiki-CC	Wiki-CC	-	-	-	-	74.2	65.8	66.5	-1.4	74.0	49.6	61.9	-6.2	71.3	54.8	25.2	-2.5
Anchor Points (§4.2)																	
No anchors	-	40k/40k	0	-	-	72.1	67.5	67.7	-1.1	74.0	57.9	65.0	-2.4	72.3	56.2	27.4	-0.9
Default anchors	-	40k/40k	-	-	-	74.0	68.1	68.9	+0.1	76.8	56.3	61.2	-3.3	73.0	57.0	28.3	-0.1
Extra anchors	-	-	extra	-	-	74.0	69.8	72.1	+1.8	76.1	59.7	66.8	-0.5	73.3	56.9	29.2	+0.3
Parameter Sharing (§4.3)																	
Sep Emb	-	40k/40k	0*	Sep Emb	lang-specific	72.7	63.6	60.8	-4.5	75.5	57.5	59.0	-4.1	71.7	54.0	27.5	-1.8
Sep L1-3	-	40k/40k	-	Sep L1-3	-	72.4	65.0	63.1	-3.4	74.0	53.3	60.8	-5.3	69.7	54.1	26.4	-2.8
Sep L1-6	-	40k/40k	-	Sep L1-6	-	61.9	43.6	37.4	-22.6	61.2	23.7	3.1	-38.7	61.7	31.6	12.0	-17.8
Sep Emb + L1-3	-	40k/40k	0*	Sep Emb + L1-3	lang-specific	69.2	61.7	56.4	-7.8	73.8	46.8	53.5	-10.0	68.2	53.6	23.9	-4.3
Sep Emb + L1-6	-	40k/40k	0*	Sep Emb + L1-6	lang-specific	51.6	35.8	34.4	-29.6	56.5	5.4	1.0	-47.1	50.9	6.4	1.5	-33.3

Table 1: Dissecting bilingual MLM based on zero-shot cross-lingual transfer performance. - denote the same as the first row (**Default**). Δ denote the difference of average task performance between a model and **Default**.

has not been carefully measured.

We present a controlled study of the impact of anchor points on cross-lingual transfer performance by varying the amount of shared subword vocabulary across languages. Instead of using a single joint BPE with 80k merges, we use language-specific BPE with 40k merges for each language. We then build vocabulary by taking the union of the vocabulary of two languages and train a bilingual MLM (**Default anchors**). To remove anchor points, we add a language prefix to each word in the vocabulary before taking the union. Bilingual MLM (**No anchors**) trained with such data has no shared vocabulary across languages. However, it still has a single softmax prediction layer shared across languages and tied with input embeddings.

As Wu and Dredze (2019) suggest there may also be correlation between cross-lingual performance and anchor points, we additionally increase anchor points by using a bilingual dictionary to create code switch data for training bilingual MLM (**Extra anchors**). For two languages, ℓ_1 and ℓ_2, with bilingual dictionary entries d_{ℓ_1,ℓ_2}, we add anchors to the training data as follows. For each training word w_{ℓ_1} in the bilingual dictionary, we either leave it as is (70% of the time) or randomly replace it with one of the possible translations from the dictionary (30% of the time). We change at most 15% of the words in a batch and sample word translations from PanLex (Kamholz et al., 2014) bilingual dictionaries, weighted according to their translation quality [1].

The second group of Tab. 1 shows cross-lingual transfer performance under the three anchor point conditions. Anchor points have a clear effect on performance and more anchor points help, especially in the less closely related language pairs (e.g. English-Chinese has a larger effect than English-French with over 3 points improvement on NER and XNLI). However, surprisingly, effective transfer is still possible with no anchor points. Com-

[1] Although we only consider pairs of languages, this procedure naturally scales to multiple languages, which could produce larger gains in future work.