Bilingual Embeddings

Table 1: Dataset details

Dataset	Ru Vocab	En Vocab	Ru Words	En Words	Ru-En words
NC	51K	28K	5.5M	5.7M	5.5M
NC + UNPar trunc.	190K	100K	120M	129M	5.5M
NC + Wiki trunc.	300K	150K	100M	120M	5.5M

Note: Limited parallel from NC + Added monolingual from UNPar -min-count for vocabulary = 5 or Wiki -min-count = 10 (all lowercased)

Table 2: Evaluation Details: (a) and (b)

a) 12 en words (lowercase) which are present in “nc.en” and their set of google translations
b) 29 en words (lowercase) which are not present in “nc.en” but are present in “nc-unpar.en” and “nc-wiki.en”

Evaluation Words (a)	TopK recall if nearest K words include any one of the translations
shop	set([u'магазин' , u'цех' , u'мастерская' , u'лавка' , u'предприятие' , u'заведение' , u'занятие' , u'профессия' , u'учреждение' , u'цеховой'])
you	set([u'вас' , u'вам' , u'вы' , u'вами' , u'тебя' , u'тебе' , u'ты' , u'тобой'])
bring	set([u'приносить' , u'приводить' , u'доводить' , u'нести' , u'привозить' , u'доставлять' , u'вызывать' , u'возбуждать' , u'заносить' , u'завезти' , u'вывезти' , u'подносить' , u'заводить' , u'причинять' , u'заставлять' , u'пригонять' , u'влечь за собой' , u'убеждать' , u'натащить'])
eat	set([u'есть' , u'кушать' , u'съедать' , u'поедать' , u'закусить' , u'поглощать' , u'разъедать' , u'грызть' , u'разрушать' , u'лизать' , u'мучить' , u'терзать'])
legal	set([u'правовой' , u'юридический' , u'законный' , u'легальный' , u'узаконенный'])
food	set([u'еда' , u'питание' , u'пища' , u'продовольствие' , u'корм' , u'провизия' , u'съестные припасы'])
house	set([u'дом' , u'жилище' , u'театр' , u'здание' , u'палата' , u'гостиница' , u'семья' , u'рубка' , u'хозяйство' , u'бордель' , u'род' , u'торговый дом' , u'династия' , u'монастырь' , u'постоялый двор' , u'работный дом' , u'торговая фирма' , u'представление' , u'пансион при школе' , u'колледж университета' , u'лондонская биржа' , u'религиозное братство' , u'зрители' , u'публика' , u'жить' , u'приютить' , u'вмещать' , u'расквартировывать' , u'убирать' , u'поселить' , u'загонять' , u'предоставлять жилище' , u'обеспечивать жильем' , u'помещать' , u'вмещаться' , u'помещаться'])
home	set([u'дома' , u'домой' , u'в цель' , u'туго' , u'в точку' , u'крепко' , u'до отказа' , u'до конца' , u'дом' , u'жилище' , u'родина' , u'семья' , u'приют' , u'домашний очаг' , u'колыбель' , u'кров' , u'пансионат' , u'финиш' , u'метрополия' , u'семейный круг' , u'место распространения' , u'домашний' , u'родной' , u'внутренний' , u'отечественный' , u'семейный' , u'относящийся к метрополии' , u'сыгранный на своем поле' ,

	u'возвращаться домой' , u'жить' , u'направлять домой' , u'доходить' , u'посылать домой' , u'предоставлять жилье'])
beautiful	set([u'прекрасный' , u'красивый' , u'превосходный' , u'прекрасное' , u'красотка' , u'красивые люди'])
smart	set([u'умный' , u'элегантный' , u'сообразительный' , u'нарядный' , u'ловкий' , u'остроумный' , u'сильный' , u'быстрый' , u'находчивый' , u'модный' , u'щеголеватый' , u'проворный' , u'резкий' , u'энергичный' , u'острый' , u'продувной' , u'суровый' , u'прыткий' , u'изящно' , u'щеголевато' , u'саднить' , u'вызывать жгучую боль' , u'причинять боль' , u'жечь' , u'болеть' , u'испытывать жгучую боль' , u'страдать' , u'жгучая боль' , u'печаль'])
intelligent	set([u'мный' , u'разумный' , u'смышленый' , u'понимающий' , u'понятливый'])
environment	set([u'среда' , u'окружающая среда' , u'окружение' , u'окружающая обстановка' , u'контекст' , u'состояние'])

Evaluation Words (b)	Count wiki	Count unpar	TopK recall if nearest K words include any one of the translations
album	23251	14	set([u'альбом'])
islander	11380	243	set([u'островитянин'])
creek	8039	14	set([u'ручей', u'бухта', u'залив', u'бухточка', u'приток'])
kansas	5646	16	set([u'канзас'])
piano	3658	14	set([u'пианино', u'фортепьяно', u'рояль', u'пиано', u'фортепьянный'])
expedition	4315	132	set([u'экспедиция', u'быстрота', u'поспешность'])
comics	3842	19	set([u'комиксы'])
tours	2121	714	set([u'турне', u'путешествие', u'тур', u'экскурсия', u'гастроли', u'поездка', u'прогулка', u'вояж', u'объезд'])
plantation	1873	152	set([u'плантация', u'насаждение', u'колония', u'внедрение', u'колонизация'])
purple	1858	10	set([u'пурпурный', u'фиолетовый', u'пурпурный', u'багровый', u'пышный', u'порфироносный', u'царский', u'пурпур', u'багроветь', u'сан кардинала', u'багрянец', u'порфира'])
painter	1856	19	set([u"художник",u"живописец",u"маляр",u"носовой фалинь"])
paragraph	101967	502	set([u'пункт', u'параграф', u'абзац'])
honour	12728	2587	set([u'честь', u'слава', u'почет', u'почести', u'почтение', u'уважение', u'награды', u'честность', u'благородство', u'ордена', u'добродетель', u'почтить', u'чтить', u'соблюдать', u'выполнять', u'почитать', u'уважать', u'оплатить', u'чествовать', u'удостаивать', u'акцептировать'])
offences	8143	448	set([u'преступлений', u'преступление', u'правонарушение', u'нарушение', u'оскорбление', u'обида', u'нападение', u'проступок', u'наступление'])
carriage	2510	859	set([u'перевозка', u'транспорт', u'каретка', u'вагон', u'экипаж', u'коляска', u'тележка', u'салазки', u'осанка', u'суппорт', u'шасси', u'лафет', u'рама', u'посадка', u'вагонетка', u'проведение', u'переноска', u'станок', u'выполнение', u'косоур', u'осуществление'])

chairperson	3818	121	set([u'председатель'])
modality	909	107	set([u'модальность'])
reimbursement	5028	71	set([u'возмещение', u'компенсация', u'оплата', u'компенсирование'])
biennial	2693	157	set([u'двухгодичный', u'двухлетний'])
orally	4812	322	set([u'устно', u'ртом'])
receipt	3252	236	set([u'получение', u'квитанция', u'приход', u'рецепт', u'приходный', u'расписываться', u'схлынуть'])
vacancy	2990	2389	set([u'вакансия', u'пустота', u'пропуск', u'пробел', u'бессмысленность', u'бездеятельность', u'безучастность', u'рассеянность'])
neighbour	501	351	set([u"сосед",u"соседка",u"соседний",u"ближний",u"смежный",u"местный",u"граничить",u"дружить"])
botany	19	369	set([u"ботаника"])
hut	30	508	set([u"хижина",u"хата",u"барак",u"лачуга",u"хибарка",u"хибара")]
seep	10	53	set([u"просачиваться",u"проникать",u"протекать",u"распространяться",u"просачивание"])
goat	18	665	set([u"козел", u"коза", u"осел", u"козел отпущения", u"мелкий скот", u"дурень", u"остолоп"])
cupboard	13	40	set([u"шкаф",u"буфет",u"чулан",u"стенной шкаф"])
scarf	6	102	set([u"шарф",u"кашне",u"галстук",u"скос",u"косой срез или кромка",u"соединение замком",u"сращивать",u"резать вкось",u"скашивать",u"отесывать края",u"отесывать углы",u"делать пазы",u"делать выемки",u"делать продольный разрез",u"соединять замком"])

Table 3: Results

Model	Dataset	Top-10 Recall	Words	Top-30 Recall	Words
Crosslingual-cca					
NC alignments	NC	0.083	you	0.083	you
NC alignments	NC + UNPar trunc.	0.25	food, environment, you	0.25	food, environment, you
		0.13	offences, honour, comics, biennial	0.17	receipt, offences, honour, comics, biennial
	NC + Wiki trunc.	0.083	eat	0.17	eat, food
		0.034	comics	0.07	offences, comics
Google dict	NC +	0.167	beautiful, you	0.333	beautiful, food, environment, you

	UNPar trunc.	0.2	offences, honour, neighbour, comics, carriage, biennial	0.2	offences, honour, neighbour, comics, carriage, biennial
	NC + Wiki trunc.	0.167	shop, you	0.167	shop, you
		0.1	offences,paragraph,comics	0.17	offences,purple,paragraph,reimbursement,comics
Original-Bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 0.001 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC	0.5	food, house, legal, environment, home, you	0.583	shop,food, house, legal, environment, home, you
Bi-weight=4	NC	0.583	beautiful, food, house, legal, environment, home, you	0.583	beautiful, food, house, legal, environment, home, you
Modified-Bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC	0.417	house, legal, environment, home, you	0.5	food, house, legal, environment, home, you
Bi-weight=4	NC	0.5	food, house, legal, environment, home, you	0.5	food, house, legal, environment, home, you
Bi-weight=1	NC + UNPar trunc.	0.167	home, you	0.167	home, you
		0.07	offences,neighbour	0.1	offences,neighbour,painter
	NC + Wiki trunc.	0.25	shop, food, you	0.25	shop, food, you
		0.03	seep	0.1	seep, offences, album
Bi-weight=4	NC + UNPar trunc.	0.25	food, home, you	0.33	shop, food, home, you
		0.034	offences	0.07	offences,honour
	NC + Wiki trunc.	0.083	you	0.083	you, shop
		0.07	offences, paragraph	0.1	offences, paragraph, honour
Bi-weight=10	NC + UNPar trunc.	0.25	food, home, you	0.33	food,house,home,you
		0.1	offences,kansas,biennial	0.1	offences,kansas,biennial
	NC + Wiki trunc.	0.167	shop, you	0.417	shop,food,house,home,you
		0.07	offenses, paragraph	0.17	offences, paragraph, seep, honour, comics

Note: NC alignments are got from doing fast_align on NC set and then choosing pairs of words based on most count in alignments. Google dict are the top 20K words translated from google available from MultiCCA paper.

Table 4: Results (Balanced)

Balancing-Modified bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC + UNPar trunc.	0.5	shop,food,house,legal,home, you	0.583	shop,food,house,legal,environment, home,you
		0.07	paragraph,kansas	0.17	offences,album,paragraph,kansas, neighbour
	NC + Wiki trunc.	0.5	Shop,house,legal, environment, home, you	0.58	shop,food,house,legal, environment,home,you
		0.24	receipt,offences,plantation, paragraph,piano,honour, neighbour	0.38	receipt,offences,purple,plantation, paragraph,piano,honour,kansas, neighbour,painter, reimbursement
Bi-weight=4	NC + UNPar trunc.	0.33	shop,food,home,you	0.5	shop,food,house,environment, home, you
		0.07	paragraph,kansas	0.1	offences,paragraph,kansas
	NC + Wiki trunc.	0.167	home, you	0.67	shop,beautiful,food,house,legal,environment,home,you,
		0.24	receipt,offences,purple,paragraph,honour,painter,orally	0.31	receipt,offences,purple,paragraph,honour,kansas,painter,reimbursement,orally
Bi-weight=10	NC + UNPar trunc.	0.17	home,you	0.33	shop,food,home,you
		0.07	paragraph,kansas	0.17	offences,paragraph,kansas,neighbour,orally
	NC + Wiki trunc.	0.25	house,home,you	0.33	food,house,home,you
		0.2	purple,paragraph,honour,kansas,painter,orally	0.28	offences,purple,paragraph,honour,kansas,painter,reimbursement,orally

Sanity Check Comparison:

Original-Bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 0.001 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC	0.5	food, house, legal, environment, home, you	0.583	shop,food, house, legal, environment, home, you
Bi-weight=4	NC	0.583	beautiful, food, house, legal, environment, home, you	0.583	beautiful, food, house, legal, environment, home, you

Modified-Bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC	0.417	house, legal, environment, home, you	0.5	food, house, legal, environment, home, you
Bi-weight=4	NC	0.5	food, house, legal, environment, home, you	0.5	food, house, legal, environment, home, you

Balancing Comparison:

Modified-Bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC + UNPar trunc.	0.167	home, you	0.167	home, you
		0.07	offences,neighbour	0.1	offences,neighbour,painter
	NC + Wiki trunc.	0.25	shop, food, you	0.25	shop, food, you
		0.03	seep	0.1	seep, offences, album

Balancing	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC + UNPar trunc.	0.5	shop,food,house,legal,home, you	0.583	shop,food,house,legal,environment, home,you
		0.07	paragraph,kansas	0.17	offences,album,paragraph,kansas, neighbour
	NC + Wiki trunc.	0.5	Shop,house,legal, environment, home, you	0.58	shop,food,house,legal, environment,home,you
		0.24	receipt,offences,plantation, paragraph,piano,honour, neighbour	0.38	receipt,offences,purple,plantation, paragraph,piano,honour,kansas, neighbour,painter, reimbursement

Comparison with CCA:

Crosslingual-cca					
NC alignments	NC	0.083	you	0.083	you
NC alignments	NC + UNPar trunc.	0.25	food, environment, you	0.25	food, environment, you
		0.13	offences, honour, comics, biennial	0.17	receipt, offences, honour, comics, biennial
	NC + Wiki trunc.	0.083	eat	0.17	eat, food
		0.034	comics	0.07	offences, comics

Balancing-Modified bivec	Dim = 40 Iters = 10 Neg = 10		Samp = 1e-3 Alpha = 0.025 Threads = 20		
Bi-weight=1	NC	0.417	house, legal, environment, home, you	0.5	food, house, legal, environment, home, you
Bi-weight=1	NC + UNPar trunc.	0.5	shop,food,house,legal,home, you	0.583	shop,food,house,legal,environment, home,you
		0.07	paragraph,kansas	0.17	offences,album,paragraph,kansas, neighbour
	NC + Wiki trunc.	0.5	Shop,house,legal, environment, home, you	0.58	shop,food,house,legal, environment,home,you
		0.24	receipt,offences,plantation, paragraph,piano,honour, neighbour	0.38	receipt,offences,purple,plantation, paragraph,piano,honour,kansas, neighbour,painter, reimbursement